From Predictive Algorithms to Automatic Generation of Anomalies*

Sendhil Mullainathan Ashesh Rambachan†

January 22, 2024

Abstract

Machine learning algorithms can find predictive signals that researchers fail to notice; yet they are notoriously hard-to-interpret. How can we extract theoretical insights from these black boxes? History provides a clue. Facing a similar problem – how to extract theoretical insights from their intuitions – researchers often turned to "anomalies:" carefully constructed examples that highlight flaws in an existing theory and spur the development of new ones. Canonical examples include the Allais paradox and the Kahneman-Tversky choice experiments for expected utility theory. We suggest anomalies can similarly be used to extract theoretical insights from black-box predictive algorithms. We develop procedures to automatically generate anomalies for an existing theory when given a predictive algorithm. We cast anomaly generation as an adversarial game between a theory and a falsifier; as such, our procedures produce adversarial examples on which the black-box algorithm predicts we would likely observe violations of our existing theory if we were to collect data As an illustration, we generate anomalies for expected utility theory using simulated lottery choice data consistent with cumulative prospect theory. Our procedures recover known anomalies and discover new ones implied by the probability weighting function. In incentivized experiments, subjects violate expected utility theory on our algorithmically generated anomalies at similar rates to the Allais paradox and Common ratio effect.

*We thank Peter G. Chang for exceptional research assistance. We also thank audiences at Harvard, Georgetown, Stanford, UCLA, UCL, Bristol, Warwick, Yale, and the NBER Summer Institute Digital Economics and AI session, Nikhil Agarwal, Rohan Alur, Raf Batista, Alex Imas, Roshni Sahoo, Suproteem Sarkar, Josh Schwartzstein, Cassidy Shubatt, Richard Thaler, Keyon Vafa, and especially our discussant Colin F. Camerer for helpful comments. We are grateful to the Center for Applied Artificial Intelligence at the Booth School of Business for generous funding. All errors are our own.

†Mullainathan: University of Chicago and NBER (Sendhil.Mullainathan@uchicago.edu). Rambachan: Massachusetts Institute of Technology (asheshr@mit.edu).

1 Introduction

How do we improve economic theories? There is of course no single answer, but one common pattern stands out across many fields. Consider the celebrated "Allais paradox." Allais (1953) felt expected utility theory did not match his intuition about how people actually make risky choices. To highlight that inconsistency, he crafted two hypothetical menus of lotteries (see Table 1); his intuition suggested that people's choices on this pair would be inconsistent with expected utility theory. When data confirmed Allais' intuition (e.g., Slovic and Tversky, 1974; Kahneman and Tversky, 1979), it led to a fundamental reappraisal of expected utility theory. More examples like the Allais paradox were constructed, eventually pointing the way to a new theory: cumulative prospect theory (Tversky and Kahneman, 1992).[1,2]

(a) Menu A

Lottery 0	\$1 million		
	100%		
Lottery 1	\$1 million	\$0	\$5 million
	89%	1%	10%

(b) Menu B

Lottery 0	\$0	\$1 million
	89%	11%
Lottery 1	\$0	\$5 million
	90%	10%

Table 1: Menus of lotteries in the Allais paradox (Allais, 1953).

Notes: We highlight in green the hypothetical choices on these two menus. Allais (1953) originally denominated the payoffs in French Francs, and we reproduce the version of the Allais paradox used in Slovic and Tversky (1974).

Despite its importance, the Allais paradox does not have a natural place in our empirical toolkit. It is not a measurement breakthrough in collecting new kinds of data on risky choice. It is not an econometric breakthrough such as novel test statistic for whether expected utility theory is misspecified.[3] Rather it is a pair of menus, whose brilliance lay in its precise construction to highlight where expected utility theory might fail and thereby to reveal how we might improve it. The Allais paradox is a specific instance of what we will call an "anomaly" in this paper.

[1]For example, Allais (1953); Kahneman and Tversky (1979) produced the Certainty effect or Common ratio effect, Slovic and Lichtenstein (1983); Tversky and Thaler (1990) produced anomalies to highlight framing effects and preference reversals, and finally Kahneman and Tversky (1984); Tversky and Kahneman (1991) produced anomalies to highlight loss aversion.

[2]Blavatskyy, Ortmann and Panchenko (2022) conducted a meta-analysis of 81 experiments in 29 papers that test variations of the Allais paradox, finding that its empirical strength depends on features of the experimental design, such as whether the choices are incentivized, etc.

[3]Constructing test statistics and hypothesis tests for model misspecification is a celebrated and foundational literature in econometrics and economic theory. See, for example, Sargan (1958); Afriat (1967, 1973); Hansen (1982); Varian (1982); Conlisk (1989); Choi et al. (2014); Bugni, Canay and Shi (2015); Kitamura and Stoye (2018); Polisson, Quah and Renou (2020); Dembo et al. (2021) among many others.

Anomalies are neither anachronistic nor idiosyncratic. They remain relevant: even after cumulative prospect theory, they continue to play a key role in advancing theories of risky choice.[4] They are used across many fields: decision-making under risk is not exceptional. They have played a crucial role in the development of asset pricing, game theory, intertemporal choice, and many other fields.[5] This then is one common pattern of how economic theories are improved over time. Researchers construct anomalies that highlight inconsistencies between existing theories and their intuitions; researchers invest great effort in robustly testing anomalies; and if they hold, new theories are proposed to resolve them.

We are interested in anomalies because we believe they offer a familiar solution to a novel problem: how can we use machine learning algorithms to improve economic theories? Machine learning algorithms could play a crucial role in economic theory because they can uncover novel predictive signals that existing theories do not model and researchers may not notice.[6] But there is a challenging obstacle. Predictive algorithms are notoriously black-boxes. Even when they predict better than existing theories, it is hard to know what they have discovered. Their discoveries are buried in the opacity of an intricate set of parameters.

In this paper, we develop procedures to algorithmically construct anomalies from predictive algorithms. Our procedures output anomalies for an existing theory that researchers can examine, like Allais. However, researchers produce anomalies themselves by constrasting their intuition with an existing theory and then crafting examples that highlight the contrast. Our procedures instead contrast the theory with a black-box predictive algorithm, automatically searching for minimal examples on which the theory cannot explain the black box's predictions. The resulting generated anomalies are then a natural place to collect data and look for possible inconsistencies between our existing theory and nature.

Building such procedures and analyzing their properties requires that we first develop a framework for anomaly generation that simultaneously models theories, machine learning

[4]Recent examples include salience theory (Bordalo, Gennaioli and Shleifer, 2012, 2022), betweenness preferences and certainty independence (Cerreia-Vioglio, Dillenberger and Ortoleva, 2015, 2020), simplicity preferences (Oprea, 2022; Puri, 2022), and cognitive uncertainty (Enke and Graeber, 2023; Enke and Shubatt, 2023).

[5]For example, Richard H. Thaler's series of articles entitled "Anomalies" in *The Journal of Economic Perspectives* highlighted anomalies in asset pricing (e.g., Lamont and Thaler, 2003), game theory (e.g., Camerer and Thaler, 1995), international finance (e.g., Froot and Thaler, 1990), public finance (Hines and Thaler, 1995), decision-making under uncertainty (e.g., Kahneman, Knetsch and Thaler, 1991), intertemporal choice (Loewenstein and Thaler, 1989), and auction theory (Thaler, 1988). See also Loewenstein and Prelec (1992) for further discussion of anomalies for intertemporal choice.

[6]Mullainathan and Spiess (2017); Athey (2017); Camerer (2019) provide broad overviews on the role of machine learning in economics. See Peysakhovich and Naecker (2017); Peterson et al. (2021) for applications in choice under risk and uncertainty, Hartford, Wright and Leyton-Brown (2016); Wright and Leyton-Brown (2017); Fudenberg and Liang (2019); Hirasawa, Kandori and Matsushita (2022) in strategic behavior in normal-form games, and Gu, Kelly and Xiu (2018); Kelly and Xiu (2023) in asset pricing.

algorithms, and anomalies. The first, somewhat surprising, challenge in building such a framework is to model theories. Of course, any individual theory is already formal, but different theories are formalized in quite different ways. Expected utility theory is a collection of axioms that restrict preference relations over lotteries, Nash equilibrium is an equilibrium condition on choices in normal-form games, and the capital asset pricing model is a model of homogeneous investors optimizing in a frictionless marketplace. We require a single framework that simultaneously captures the essence of any economic theory, despite this diversity. To tackle this challenge, we note that all theories share a common functionality: they posit some restrictive underlying structure that enables them to derive novel implications. For example, in classical choice theory, any choice from any single menu is allowed. But if a person has chosen item a over item b in one menu and item b over item c in another, it implies that they will choose a over c on any menu containing both.

To capture this common functionality, we focus on settings summarized by some input features (x) and some modeled outcome (y^*); for example, x can be a menu of lotteries and y^* can be a choice probability. In this context, we model theories as *black box* mappings, which take examples (collections of (x, y^*)-pairs) and return correspondences. These correspondences summarize the logical implications of the theory: for any given x, the returned correspondence specifies what y^* are allowed. Expected utility theory, for example, derives implications about an individual's choice behavior in new menus based on the choices they have made on other menus. In this framework, we define anomalies as *minimal* collections of examples that are incompatible with the theory.

We introduce four assumptions on such theory mappings and establish two results that serve as the basis of our anomaly generation procedures. First, we provide a representation result: a theory can be equivalently represented as an *allowable function class*, which summarizes all mappings between the features and the theory's modeled outcome that are consistent with its underlying structure (whatever that may be). Second, we show that anomalies always exist; because theories are not vacuous in our framework, there exist collections of examples they cannot explain. These two results allow us operationalize any theory for purposes of anomaly generation. A theory can be analyzed as if it searches for allowable functions that fit the examples they are given, and an anomaly is a collection of examples precisely constructed to foil this search.

To generate anomalies given an estimated prediction function, we observe that searching for anomalies can be viewed as an adversarial game between a falsifier and the theory. The falsifier proposes collections of features and the estimated prediction function evaluated on those features, and the theory attempts to explain them by fitting an allowable function. The falsifier's payoff is increasing in the theory's average loss on the proposed collection, and

the theory's payoff is decreasing in its average loss. An anomaly arises if the falsifier finds a collection of examples that the theory cannot fit. In other words, anomalies are *adversarial* collections of examples that induce a positive loss for the theory in such a game.

Our first anomaly generation procedure directly optimizes the falsifier's adversarial problem as a max-min optimization program over a theory's allowable functions. We analyze the statistical properties of a feasible implementation of the falsifier's max-min program, establishing finite sample bounds on how well it approximates its population analog. Practically optimizing this max-min program may be challenging – the falsifier's maximization over collections of features will typically be non-concave, and so standard optimization techniques may not apply (e.g., Rockafellar, 1970; Freund and Schapire, 1996). We leverage recent results in adversarial learning and non-convex/concave min-max optimization to develop a gradient descent ascent procedure and analyze its properties (Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020). The resulting gradient descent ascent procedure generates anomalies by iteratively updating the falsifier's proposed collection of examples to maximize the average loss of the theory's best-responding allowable function.

There often, however, exists additional structure in theories that this procedure does not exploit. Theories often behave as if they have a lower-dimensional representation of the input; that is, there exists some pair of feature values that all allowable functions assign the same modeled outcome value and it is as if the theory collapses these features together. Some anomalies, like the Allais paradox, illustrate what this representation misses: they reveal a dimension that is relevant for the theory's modeled outcome but which the theory's lower dimensional representation fails to capture. Our second procedure is an example morphing procedure to generate these *representational* anomalies for a theory. Given an initial feature value, the example morphing procedure searches for nearby feature values across which the theory's allowable functions do not vary but across which the estimated prediction function varies.

Both of our anomaly generation procedures take two inputs: a theory (represented as a set of allowable functions) and a black-box predictive algorithm. How do we construct the black-box algorithm? The current human practice of anomaly generation suggests two possibilities. One way researchers construct anomalies is by contrasting an existing theory with their intuitions about the world, as Allais likely did. By way of analogy, we could begin with a large collected dataset and fit a supervised learning algorithm to these data. The resulting predictive model would serve the analogy of human "intuition" for our procedures – an empirically derived object that captures statistical patterns. Alternatively, researchers construct anomalies to contrast the existing theory with an alternative theory. For example, Kahneman-Tversky may have produced their anomalies by contrasting expected utility

theory with cumulative prospect theory. Analogously, we could simulate data from an alternative theory and fit a predictive algorithm to that simulated data. As such, we can use our procedure to construct anomalies implied either by empirical data or alternative theories.

As an illustration, we apply our anomaly generation procedures to the second case: we algorithmically generate logical anomalies for expected utility theory that are implied by cumulative prospect theory, as Kahneman and Tversky may have done. While interesting in its own right, illustrating our procedure in this way serves an important purpose. Were we to apply our anomaly generation procedures on collected lottery choice data, we would have no sense of what anomalies we should expect to find. By contrast, since cumulative prospect theory has been well-studied by theorists for decades, we can compare our algorithmically generated anomalies against known anomalies for expected utility theory constructed by researchers, such as those produced in Allais (1953), Kahneman and Tversky (1979), and many others. We now have a floor: do our anomaly reproduce known anomalies for expected utility theory implied by cumulative prospect theory?

Our procedures in fact reach this floor, recovering known anomalies for expected utility theory over the specified search space. They also go further, uncovering novel anomalies that – to our knowledge – are nowhere in the existing literature. Even in this well-trodden domain, our anomaly generation procedures more exhaustively contrast an existing theory against the implications of an alternative theory. Though these anomalies are novel, they have an intuitive interpretation. One of cumulative prospect theory's core insights is that our perceptions of probabilities exhibit diminishing sensitivity: a shift in probability from 1% to 10% looms larger in our minds than a shift from 41% to 50% (Tversky and Kahneman, 1992). This is captured by the well-known "s-shaped" probability weighting function. Importantly, diminishing sensitivity in our perceptions of probabilities means that there are ways to manipulate lottery probabilities that would affect our choices, even though expected utility theory predicts our choices should be unchanged. Indeed, well-known logical anomalies like the Allais paradox are exactly crafted to illustrate such a choice reversal. Our algorithmically generated anomalies fall into several distinct categories that illustrate new ways to produce choice reversals across pairs of lottery menus implied by the probability weighting function that are inconsistent with expected utility theory.

Automatically producing such anomalies is where our procedures end and is the primary contribution of our work. Yet having generated these novel anomalies, one cannot help but wonder: do they actually hold in real data? While robustly answering that question is obviously beyond the scope of the present paper, we can take a first step. We recruit participants on Prolific to make incentivized choices on a set of our algorithmically generated anomalies. We design our survey to mirror recent work testing known anomalies like the Allais paradox

and the Common ratio effect (e.g., Harless and Camerer, 1994; Blavatskyy, Ortmann and Panchenko, 2022; Blavatskyy, Panchenko and Ortmann, 2022; Jain and Nielsen, 2023; McGranaghan et al., Forthcoming), providing a benchmark for how "significant" these novel anomalies are. On our algorithmically generated anomalies, participants exhibit behavior that is inconsistent with expected utility theory at rates similar to the well-established anomalies of behavioral economics. These preliminary findings suggest these algorithmically generated anomalies merit the kind of rigorous experimental scrutiny given to known anomalies generated by researchers.

Our ultimate goal of course is not to revisit choice under risk not generate novel anomalies for expected utility theory. Rather, this specific illustration demonstrates the broader potential for our anomaly generation procedures. Their success in generating novel anomalies in a well-trodden domain suggests they could be valuable in many other areas. Indeed, our algorithmic procedures are broadly applicable beyond choice under risk and can be used in any domain where there exists a formal theory and rich data that the theory seeks to explain. Our procedures exploit the fact that supervised machine learning algorithms often uncover novel empirical patterns, ones that our existing theories may not capture. Rather than leaving us with a black box predictive algorithm, however, our procedures return anomalies – small collections of examples that may help researchers evaluate and improve theories.

Our work sits in a rapidly growing literature that seeks to integrate machine learning into the scientific process across various fields. Carleo et al. (2019); Raghu and Schmidt (2020); Pion-Tonachini et al. (2021); Krenn et al. (2022); Wang et al. (2023) provide recent reviews on the use of machine learning across the physical sciences, such as biology, chemistry, mathematics, and physics. Substantial progress has already been made in exploring how machine learning interacts with economic theories. Recent work compares the out-of-sample predictive performance of black-box machine learning models against that of economic theories in choice under risk and strategic behavior in normal form games, measuring the "completeness" of economic theories (Fudenberg et al., 2022). Andrews et al. (2022) develops conformal inference procedures to measure the out-of-distribution predictive performance of economic theories. When a supervised machine learning model predicts some outcome of interest accurately out-of-sample, researchers often attempt to open the black-box prediction function and investigate particular hypotheses of researchers (Camerer, 2019). See, for example, Peysakhovich and Naecker (2017) and Peterson et al. (2021) for choice under risk, Wright and Leyton-Brown (2017); Hirasawa, Kandori and Matsushita (2022) for strategic behavior in normal-form games, Mullainathan and Obermeyer (2021) for medical decision-making, and Kleinberg et al. (2018); Rambachan (2022); Sunstein (2022) for judicial decision-making. By contrast, we use supervised machine learning algorithms

as stepping stones to automatically generate anomalies for an existing theory, rather than relying on researchers to directly inspect the black-box prediction function.

Fudenberg and Liang (2019) use supervised machine learning algorithms to predict on which normal-form games will observed play differ from alternative theories of strategic behavior. They use the resulting prediction function to generate new normal-form games where a particular theory will predict poorly. This intuitive procedure can be formally reinterpreted as a heuristic solution to our adversarial characterization of anomalies tailored to the models of strategic behavior they study. Ludwig and Mullainathan (2023) develop a morphing procedure for images based on generative adversarial networks in order to uncover implicit characteristics of defendant mug-shots that affect pretrial release decisions. Our procedures are general-purpose, enabling researchers to search for anomalies given any formal theory that places restrictions on the relationship between some features x and modeled outcome y^*.

2 Theories and the anomaly generation problem

Whatever their mathematical formalism, economic theories all share a common functionality: they model some underlying restrictive structure in order to derive novel implications about an economic domain. In this section, we develop an econometric framework that analyzes theories as black box mappings that return correspondences between some input features (x) and some modeled outcome (y), summarizing all implications drawn by the theory. We introduce four assumptions on these theory mappings so they behave as if they have some underlying structure (whatever that may be), thereby capturing this shared functionality of theories. We establish two results that serve as the foundation for our algorithmic procedures for anomaly generation.

2.1 Setting and theories

Let $x \in \mathcal{X}$ be some vector of features and $y^* \in \mathcal{Y}^*$ be some modeled outcome in an economic domain. Any pair $(x, y^*) \in \mathcal{X} \times \mathcal{Y}^*$ is an *example*, and $D := \{(x_1, y_1^*), \ldots, (x_n, y_n^*)\}$ is a finite collection of examples. We let $\mathcal{D}$ denote all collections of examples, $\mathcal{F}$ the collection of all mappings $f(\cdot)\colon \mathcal{X} \to \mathcal{Y}^*$, and $\mathcal{C}$ the collection of all correspondences $c(\cdot)\colon \mathcal{X} \rightrightarrows \mathcal{Y}^*$.

Rather than focusing on any particular mathematical model, we define a theory as a mapping that returns correspondences between the features and modeled outcome given examples.

Definition 1. A *theory* consists of the pair $(T(\cdot), \mathcal{M})$, where $T(\cdot)\colon \mathcal{D} \to \mathcal{C}$ is a mapping from examples to correspondences between the features and modeled outcome, and $\mathcal{M}$ is some finite set with elements $m \in \mathcal{M}$.

Given examples D, a theory $T(\cdot)$ returns a correspondence summarizing all implications it draws about the relationship between the features and modeled outcome. We write $T(\cdot; D) \in \mathcal{C}$ to be the theory's correspondence when applied to examples D, and $T(x; D) \subseteq \mathcal{Y}^*$ to be the theory's implications about the modeled outcome at feature x. All else about the economic domain is collapsed into the theory's *modeled contexts* $m \in \mathcal{M}$. The theory refines its underlying structure within a modeled context and does not extrapolate across modeled contexts. We take a theory's modeled contexts $\mathcal{M}$ as a primitive throughout the paper, and we primarily focus on the behavior of its correspondence $T(\cdot)$.

Definition 1 is necessarily abstract in order to capture the diversity of economic theories. To make it concrete, we next illustrate how three popular economic domains map into this framework. In Appendix C, we provide additional examples such as choice under risk over certainty equivalents or valuation tasks and multi-attribute discrete choice.

Example: choice under risk Consider individuals making choices from menus of two lotteries over $J > 1$ monetary payoffs.[7] The features are a complete description of the menu of lotteries $x = (z_0, p_0, z_1, p_1)$, where $z_0, z_1 \in \mathbb{R}^J$ are the payoffs and $p_0, p_1 \in \Delta^{J-1}$ are the probabilities associated with lottery 0 and lottery 1 respectively. The features may also, for example, include information about how each lottery is presented (e.g., presented as a two-stage lottery) or the ordering of lotteries in the menu. The modeled outcome is the choice probability $y^* \in [0, 1]$ for lottery 1, and the modeled contexts $m \in \mathcal{M}$ are each individual.

Given examples D, expected utility theory (with strict preferences) searches for utility functions $u(\cdot)$ that rationalize the lottery choice probabilities, meaning that $u(\cdot)$ satisfies $y^* = \arg\max_{k\in\{0,1\}} \sum_{j=1}^{J} p_k(j)u(z_k(j))$ for all $(x, y^*) \in D$. On any new menu of lotteries x, expected utility theory returns $T(x; D)$, where $y^* \in T(x; D)$ if and only if $y^* \in \arg\max_{k\in\{0,1\}} \sum_{j=1}^{J} p_k(j)u(z_k(j))$ for some utility function $u(\cdot)$ rationalizing D.

In our framework, incorporating noise yields an alternative theory $T(\cdot)$. For instance, expected utility theory with idiosyncratic errors allows the individual to mistakenly select the wrong lottery with some probability $\epsilon \in [0, 0.5]$. Given examples D, expected utility theory with idiosyncratic errors searches for utility functions $u(\cdot)$ and an error rate ϵ satisfying $y^* = (1-\epsilon)1\{\sum_{j=1}^{J} p_1(j)u(z_1(j)) \geq \sum_{j=1}^{J} p_0(j)u(z_0(j))\} + \epsilon 1\{\sum_{j=1}^{J} p_1(j)u(z_1(j)) < \sum_{j=1}^{J} p_0(j)u(z_0(j))\}$ for all $(x, y^*) \in D$. More complex models of noisy choices can of course be captured by our framework.[8] ▲

[7] Recent work such as Erev et al. (2010, 2017); Peysakhovich and Naecker (2017); Peterson et al. (2021) collect large experimental datasets of individuals making choices from menus of risky lotteries and fit black-box predictive algorithms to flexibly model people's risky choices.

[8] Harless and Camerer (1994) analyze expected utility theory with idiosyncratic errors. Ballinger and Wilcox (1997); Loomes (2005); Hey (2005) consider expected utility theory with i.i.d. additive utility noise, McGranaghan et al. (Forthcoming) consider a more general model of noisy expected utility theory, and Enke

Example: play in normal-form games Consider individuals playing $J \times J$ normal-form games.[9] Let $\{1, \ldots, J\}$ denote the actions available to the row and column players, and $\pi_{row}(j,k)$, $\pi_{col}(j,k)$ denote the payoff to the row player and column player respectively from action profile (j,k). The features are a complete description of the normal-form payoff matrix with $x = (\pi_{row}(1,1), \pi_{col}(1,1), \ldots, \pi_{row}(J,J), \pi_{col}(J,J))'$. The modeled outcome is the row player's strategy profile, which is a probability distribution over actions $y^* \in \Delta^{J-1}$. The modeled contexts $m \in \mathcal{M}$ are again each individual.

Given examples D, Nash equilibrium returns $T(x;D)$ satisfying $T(x;D) = \{y^*\}$ for all $(x,y^*) \in D$ and $y^* \in T(x;D)$ for any $x \notin D$ if and only if there exists some $y^*_{col} \in \Delta^{J-1}$ such that $\sum_{j=1}^{J}\sum_{k=1}^{J} y^*(j)y^*_{col}(k)\pi_{row}(j,k) \geq \sum_{j=1}^{J}\sum_{k=1}^{J} \tilde{y}^*(j)y^*_{col}(k)\pi_{row}(j,k)$ for all $\tilde{y}^* \in \Delta^{J-1}$ and $\sum_{j=1}^{J}\sum_{k=1}^{J} y^*(j)y^*_{col}(k)\pi_{col}(j,k) \geq \sum_{j=1}^{J}\sum_{k=1}^{J} y^*(j)\tilde{y}^*(k)\pi_{col}(j,k)$ for all $\tilde{y}^* \in \Delta^{J-1}$. Alternatively, for instance, "level-0" strategic behavior is a theory $T(\cdot)$ satisfying $T(x;D) = \{y^*\}$ for all $(x,y^*) \in D$ and $T(x;D) = \{(1/J, \ldots, 1/J)\}$ for $x \notin D$ if and only if $y^* = (1/J, \ldots, 1/J)$ for all $(x,y^*) \in D$. More sophisticated behavioral models of strategic behavior can also be captured in our framework. ▲

Example: asset pricing Consider the evolution of $J \geq 1$ risky asset returns over time. The features x enumerate the expected return for all assets, the full variance-covariance matrix of asset returns, and possibly higher-order moments of asset returns over a particular time period. The modeled outcome $y^* \in \mathbb{R}$ is the expected return of some asset j in the next period, and each modeled context $m \in \mathcal{M}$ is an asset. Given examples D, the capital asset pricing model (CAPM) provides a procedure for calculating the expected market return $\bar{y}_{market}$, the risk-free rate $\bar{y}_{risk\text{-}free}$, and the asset's covariance with the market return β. On any new period x, CAPM returns $T(x;D)$, where $y^* \in T(x;D)$ if and only if $y^* = \bar{y}_{risk\text{-}free} + \beta\,(\bar{y}_{market} - \bar{y}_{risk\text{-}free})$. ▲

2.2 Incompatible examples and logical anomalies

The examples D are incompatible with a theory $T(\cdot)$ if its underlying structure cannot accommodate the configuration of features and modeled outcomes. Otherwise, the examples D are compatible with theory $T(\cdot)$.

Definition 2. A collection of examples $D \in \mathcal{D}$ is

i. *compatible* with theory $T(\cdot)$ if $T(x;D) \neq \varnothing$ for all $x \in \mathcal{X}$.

and Shubatt (2023) consider expected utility theory with complexity-dependent noise.

[9] Recent work such as Wright and Leyton-Brown (2010); Hartford, Wright and Leyton-Brown (2016); Wright and Leyton-Brown (2017); Fudenberg and Liang (2019); Hirasawa, Kandori and Matsushita (2022) collect large experimental datasets of individuals selecting actions in normal-form games and fit black-box predictive algorithms to flexibly model people's chosen strategy profiles.

ii. *incompatible* with theory $T(\cdot)$ if $T(x; D) = \varnothing$ for all $x \in \mathcal{X}$.

It may be difficult for researchers to understand what drives the failure of the theory's underlying structure on any particular collection of examples. Researchers like Allais are therefore not simply interested in characterizing all possible collections that are incompatible with a theory; rather they construct minimally incompatible collections, which we refer to as logical anomalies.

Definition 3. A collection of examples $D \in \mathcal{D}$ is a *logical anomaly* for theory $T(\cdot)$ if D is incompatible with theory $T(\cdot)$ and $\widetilde{D}$ is compatible with theory $T(\cdot)$ for all $\widetilde{D} \subset D$.

A logical anomaly is a minimally incompatible collection of examples in the sense that $T(\cdot)$ is compatible with any of its subsets. To make this concrete, we illustrate how two famous logical anomalies for expected utility theory map into Definition 3.

Example: the Allais paradox Consider the Allais paradox for expected utility theory in Table 1. The Allais paradox is a pair of examples consisting of the menus of lotteries x_A, x_B and the associated modeled outcomes $y_A^* = 0, y_B^* = 1$. The independence axiom of expected utility theory implies that the choice on menu x_A must be the same as the choice on menu x_B; that is, for any $D \in \mathcal{D}$, $T(x_A; D) = T(x_B; D)$. The Allais paradox is therefore an incompatible collection of examples for expected utility theory. Since any single choice (x_A, y_A^*) or (x_B, y_B^*) is compatible with expected utility theory, the Allais paradox further satisfies Definition 3. ▲

Example: the Certainty effect Consider the Certainty effect for expected utility theory introduced by Kahneman and Tversky (1979) in Table 2. The Certainty effect is a pair

(a) Menu A

Lottery 0	\$4000	\$0
	80%	20%
Lottery 1	\$3000	
	100%	

(b) Menu B

Lottery 0	\$4000	\$0
	20%	80%
Lottery 1	\$3000	\$0
	\$25	75%

Table 2: Menus of lotteries in the Certainty effect (Kahneman and Tversky, 1979).

Notes: We highlight in green the conjectured choices on these two menus.

of examples again consisting of the menus of lotteries x_A, x_B and the associated modeled outcomes $y_A^* = 0, y_B^* = 1$. The independence axiom of expected utility theory implies that the choice on menu x_A must be the same as the choice on menu x_B, like the Allais paradox.

This pair of examples is therefore incompatible with expected utility theory, yet any single choice (x_A, y_A^*) or (x_B, y_B^*) alone is compatible with expected utility theory. The Certainty effect therefore satisfies Definition 3. ▲

We discuss logical anomalies for our other examples of economic theories in Appendix C. For Nash equilibrium, we provide a logical anomaly that consists of a single normal-form game on which a level-0 or level-1 thinker would select a strategy profile that places positive probability on dominated actions. For CAPM, any asset whose expected return does not satisfy the asset pricing equation is a logical anomaly (e.g., an asset's expected return may depend on higher moments or other asset characteristics).

Finally, whether a particular collection of examples is a logical anomaly depends on the researcher's exact specification of theory $T(\cdot)$. For instance, in choice under risk, any (x, y^*) with choice probability $y^* \in (0, 1)$ is a logical anomaly for expected utility theory without idiosyncratic errors (ignoring possible indifferences). This need not be a logical anomaly if we incorporate alternative models of noisy choices.

2.3 Representation result and existence of logical anomalies

We introduce four assumptions on the properties of theory's correspondence $T(\cdot)$. These assumptions place restrictions on $T(\cdot)$ such that it behaves as if it has some underlying structure, whatever that may be.

Assumption 1 (Compatibility)**.** $T(\cdot)$ is either compatible or incompatible with any $D \in \mathcal{D}$.

Assumption 2 (Consistency)**.** If $T(\cdot)$ is compatible with $D \in \mathcal{D}$, then $T(x; D) = \{y^*\}$ for all $(x, y^*) \in D$.

Assumption 3 (Refinement)**.** For any $D, D' \in \mathcal{D}$ with $D \subseteq D'$, $T(x; D') \subseteq T(x; D)$ for all $x \in \mathcal{X}$.

Assumption 4 (Non-trivial implications)**.** There exists $D \in \mathcal{D}$ and $x \notin D$ such that $T(x; D) \subset \mathcal{Y}^*$.

Assumption 1 states $T(\cdot)$ is either compatible or incompatible with any collection of examples. Assumption 2 states that whenever $T(\cdot)$ is compatible with a particular collection of examples, it is further consistent with all provided examples. Assumption 3 states that the theory refines its implications as more examples are provided. Finally, Assumption 4 states that there exists some collection of examples and an unseen feature value at which theory $T(\cdot)$ derives non-trivial implications.

Our previous examples of economic theories satisfy these assumptions. Consider first expected utility theory. First, expected utility theory satisfies Assumption 1 and Assumption 2. For any collection D of menus and choice probabilities, either (i) there exists no rationalizing utility function in which case expected utility theory is incompatible with D, or (ii) there exists a rationalizing utility function. Second, for any pair D, D' satisfying $D \subseteq D'$, the rationalizing utility functions for D' must be a subset of the rationalizing utility functions for D. This implies expected utility theory satisfies Assumption 3. Finally, consider any $(x, y^*) \in D$ with $x = (p_1, z_1, p_0, z_0)$ and $y^* \in \{0, 1\}$. The independence axiom implies the same choice would be made on all other menus $x' = (\alpha p_1 + (1-\alpha)\tilde{p}, \alpha z_1 + (1-\alpha)\tilde{z}, \alpha p_0 + (1-\alpha)\tilde{p}, \alpha z_0 + (1-\alpha)\tilde{z})$ for any lottery $(\tilde{p}, \tilde{z})$ and $\alpha \in [0, 1)$.[10] Expected utility theory therefore satisfies Assumption 4. Appendix C discusses our other examples.

For any theory $T(\cdot)$ satisfying Assumptions 1-4, we establish that there exist logical anomalies and such a theory can be equivalently represented by an allowable function class. To state this result, we say a mapping $f(\cdot) \in \mathcal{F}$ is *consistent* with $D \in \mathcal{D}$ if $f(x) = y^*$ for all $(x, y^*) \in D$. A collection D is *inconsistent* with function class $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ if there exists no $f(\cdot) \in \widetilde{\mathcal{F}}$ that is consistent with D.

Proposition 2.1.

i. Any theory $T(\cdot)$ satisfies Assumptions 1-4 if and only if there exists a function class $\mathcal{F}^T \subset \mathcal{F}$ that is inconsistent with some collection of examples and satisfies, for all $x \in \mathcal{X}$ and $D \in \mathcal{D}$,

$$T(x; D) = \left\{ f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ and } f(\cdot) \text{ is consistent with } D \right\}. \tag{1}$$

ii. There exists logical anomalies for any theory $T(\cdot)$ satisfying Assumptions 1-4.

We call $\mathcal{F}^T$ the *allowable function class* of theory $T(\cdot)$. The allowable function class $\mathcal{F}^T$ summarizes all mappings from features to the modeled outcome that are consistent with theory $T(\cdot)$'s underlying structure, however that may be mathematically modeled. As a result, theory $T(\cdot)$ can be analyzed as if it simply searches for any allowable functions $f(\cdot) \in \mathcal{F}^T$ that are consistent with the given examples $D \in \mathcal{D}$. Furthermore, the theory is not compatible with all possible collections of examples. In fact, there exist logical anomalies for any theory $T(\cdot)$ satisfying Assumptions 1-4. By establishing the existence of logical anomalies and placing theories into a tractable allowable function representation irrespective

[10] We write the compound lottery that yields lottery (p, z) with probability $\alpha \in [0, 1)$ and lottery (p', z') with probability $(1-\alpha)$ as $(\alpha p + (1-\alpha)p', \alpha z + (1-\alpha)z')$.

of its economic domain or mathematical structure, Proposition 2.1 serves as the launching point of our anomaly generation procedures.

We provide the complete proof in Appendix A, and we briefly sketch our proof strategy here. It is clear that the allowable function representation (1) satisfies Assumptions 1-3. To show it also satisfies Assumption 4, consider the smallest collection of examples $D_{min} \in \mathcal{D}$ that is inconsistent with $\mathcal{F}^T$. For any $(x, y^*) \in D_{min}$, Assumption 4 is satisfied for $D = D_{min} \setminus \{(x, y^*)\}$ and x. For this choice, $T(x; D) \subset \mathcal{Y}^*$ must be satisfied since otherwise $\mathcal{F}^T$ could not be inconsistent with D_{min}. This establishes necessity. To show sufficiency, we construct an allowable function representation $\mathcal{F}^T \subset \mathcal{F}$ for any theory $T(\cdot)$ satisfying Assumptions 1-4. To do so, we define $\mathcal{D}^{\neg T}$ as all incompatible collections of examples for $T(\cdot)$, which is non-empty by Assumption 4. We define $\mathcal{F}^{\neg T}$ to be all mappings that are consistent with any $D \in \mathcal{D}^{\neg T}$. We construct the allowable functions as $\mathcal{F}^T = \mathcal{F} \setminus \mathcal{F}^{\neg T}$, and the proof establishes that this construction satisfies Equation (1). This proves part (i). To show part (ii), we establish that there exists a smallest, incompatible collection of examples for theory $T(\cdot)$. This must be a logical anomaly by Definition 3.

Incompatible collections of examples and logical anomalies have a simple characterization in terms of a theory's allowable functions $\mathcal{F}^T$ and a loss function.

Proposition 2.2. *Suppose theory $T(\cdot)$ satisfies Assumptions 1-4, and consider any loss function $\ell \colon \mathcal{Y}^* \times \mathcal{Y}^* \to \mathbb{R}_+$ satisfying $\ell(y, y') = 0$ if and only if $y = y'$.*

i. The collection of examples $D \in \mathcal{D}$ is incompatible with $T(\cdot)$ if and only if

$$\min_{f(\cdot) \in \mathcal{F}^T} |D|^{-1} \sum_{(x, y^*) \in D} \ell\left(f(x), y^*\right) > 0. \tag{2}$$

ii. If there exist no incompatible collection with fewer than $n > 1$ examples, then any incompatible collection with n examples is also a logical anomaly.

This is an immediate consequence of Definitions 2-3 and Proposition 2.1. Searching for incompatible collections of examples is equivalent to searching for collections that induce a strictly positive loss for the theory's allowable functions. Furthermore, we can search for logical anomalies by iteratively searching for larger incompatible collections.

Importantly, this characterization of incompatible collections of examples (2) can be reinterpreted as an adversarial game between the theory (the min-player) and a falsifier. The falsifier proposes examples D to the theory, and the theory attempts to explain them by fitting its allowable functions. The theory's payoffs are decreasing in its average loss, and the falsifier wishes to search for examples that induce a positive loss for the theory's

best-responding allowable function. We build on this characterization of logical anomalies to develop our anomaly generation procedures.

Before continuing, our model of theories builds on a classic literature on measuring the predictive success and restrictiveness of economic theories, tracing back to Selten and Krischker (1983) and Selten (1991). Selten (1991) measures the predictive success of a theory as the comparison between the fraction of correct predictions it makes and the fraction of outcomes it deems possible.[11] Fudenberg, Gao and Liang (2020) measure the "restrictiveness" of economic theories, generalizing Selten's definition. Our existence result for logical anomalies establishes that any black-box theory satisfying our axiomatization must be restrictive in the sense that there exist some minimal collections of examples that it is incompatible with.

2.4 Observable data and empirical anomalies

To this point, we analyzed the behavior of theory $T(\cdot)$ on collections of examples $D \in \mathcal{D}$. Our goal is to ultimately contrast theory $T(\cdot)$ with nature in order to generate hypotheses about how it may be improved. Yet observable data may suffer from a variety of typical econometric problems, such as measurement error, endogeneity, or unobserved variables, that produce additional empirical variation that is not modeled by the theory. Bridging the world from theory to data therefore requires assumptions on how data map onto the theory's examples.

In this paper, we form this bridge by supposing each modeled context $m \in \mathcal{M}$ is associated with some joint distribution over $(X_i, Y_i) \sim P_m(\cdot)$, where $Y_i \in \mathcal{Y}$ is an observed outcome. We assume $P_m(X_i = x) > 0$ for all $x \in \mathcal{X}$. The observed outcome is statistically related to the theory's modeled outcome. The *empirical* modeled outcome of theory $T(\cdot)$ is

$$f_m^*(x) := \mathbb{E}_m \left[g(Y_i) \mid X_i = x \right] \tag{3}$$

for some researcher-specified function $g(\cdot)$, where $\mathbb{E}_m[\cdot]$ denotes the expectation under $P_m(\cdot)$. The empirical modeled outcome is some identified functional of each modeled context's underlying joint distribution. Indeed, researchers often first estimate choice probabilities from data on discrete choices, strategy profiles in normal-form games from data on actions, or expected returns from data on historical realized returns.

For the rest of the paper, our goal is to discover *empirical anomalies* for theory $T(\cdot)$ in modeled context m, if they exist. Given modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot)$,

[11] Harless and Camerer (1994) measure the predictive success of alternative theories for decision-making under risk and propose methods for aggregating evidence of predictive success across experiments. See also Beatty and Crawford (2011) for an application to consumer demand.

an empirical example is now any pair $(x, f_m^*(x))$. We search for collections of empirical examples $D = \{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$ that are logical anomalies for theory $T(\cdot)$.[12]

Before continuing, we note that this definition of the empirical modeled outcome is of course restrictive. It implies that any residual variation in the observed outcome Y_i given the observed features X_i within a modeled context is irrelevant for the underlying structure that the theory purports to model. Nonetheless, the researcher retains substantial flexibility to specify the function $g(\cdot)$ in order to capture whatever aspects of the conditional distribution $Y_i \mid X_i$ they deem relevant. We view this as a desirable attribute of our framework.

3 An adversarial algorithm for anomalies

In this section, we develop our first procedure to generate empirical anomalies when given access to a theory's allowable functions $\mathcal{F}^T$ and a black-box predictive algorithm.

Consider modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot)$. For $x_{1:n} := (x_1, \ldots, x_n)$, let

$$\mathcal{E}_m(x_{1:n}) := \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell(f(x_i), f_m^*(x_i)) \tag{4}$$

be theory $T(\cdot)$'s loss on the empirical examples $D = \{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$. Proposition 2.2 establishes the collection D is incompatible with $T(\cdot)$ if and only if $\mathcal{E}_m(x_{1:n}) > 0$. Furthermore, the collection is also an empirical anomaly in modeled context m if there exists no smaller collection of empirical examples that is incompatible with $T(\cdot)$. If we had oracle access to the true function $f_m^*(\cdot)$, we could therefore search for empirical anomalies by: first, searching for collections of empirical examples that are incompatible with $T(\cdot)$, or equivalently feature vectors $x_{1:n}$ satisfying $\mathcal{E}_m(x_{1:n}) > 0$; and second, iterating that search over successively larger collections.

For any collection of empirical examples with size $n \geq 1$, we can directly solve the falsifier's adversarial problem in the following optimization program

$$\max_{x_{1:n}} \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell(f(x_i), f_m^*(x_i)), \tag{5}$$

which searches for empirical examples that generate large positive loss for the theory's best-responding allowable function (if they exist). Our first procedure for generating empirical anomalies is an iterative search procedure based on this max-min program. For some maximal size $\overline{n} \geq 1$, we iterate over $n = 1, \ldots, \overline{n}$ and solve the adversarial game (5), letting n^*

[12] Our discussion in the main text focuses on searching for empirical anomalies in a single modeled context. In Appendix D, we extend our algorithmic procedures to search for empirical anomalies across multiple modeled contexts.

denote the smallest collection size for which the optimal value of the max-min program is strictly positive. Any feature vector $x_{1:n^*}$ with $\mathcal{E}_m(x_{1:n^*}) > 0$ is an empirical anomaly by Proposition 2.2. We can then search for other empirical anomalies by searching for other feature vectors in the set $\{x_{1:n^*} : \mathcal{E}_m(x_{1:n^*}) > 0\}$.

Of course, this iterative search procedure is not directly feasible. First, we do not observe the true function $f_m^*(\cdot)$, and it instead must be estimated from the observable data. Second, solving the max-min program may be quite difficult as both the inner minimization over the theory's allowable functions and the outer maximization over feature vectors may be intractable. We next tackle both of these challenges, constructing our first feasible search procedure for empirical anomalies.

3.1 Statistical analysis of plug-in max-min optimization

Recall the true function $f_m^*(\cdot)$ in modeled context m is given by $f_m^*(x) = \mathbb{E}_m[g(Y_i) \mid X_i = x]$ for some researcher-specified function $g(\cdot)$. Suppose we observe a random sample $(X_i, Y_i) \sim P_m(\cdot)$ i.i.d. for $i = 1, \ldots, N_m$ from modeled context m, and we construct an estimator $\widehat{f}_m^*(\cdot) \in \mathcal{F}$ for the true function. For example, this estimator may be constructed using any black box, supervised machine learning algorithm that predicts $g(Y_i)$ based on the features X_i such as deep neural networks, or classic nonparametric regression techniques (e.g., Chen, 2007).

We solve the falsifier's *plug-in* max-min program

$$\max_{x_{1:n}} \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}_m^*(x_i)\right). \tag{6}$$

In order to analyze the plug-in program's error for the infeasible program (5), we assume the researcher has access to approximate optimization routines that can solve the inner minimization and outer maximization problems up to some errors.

Assumption 5 (Approximate optimization)**.**

i. For any $x_{1:n}$ and $\widehat{f}_m^*(\cdot) \in \mathcal{F}$, the approximate inner minimization routine returns an allowable function $\widetilde{f}(\cdot; x_{1:n}) \in \mathcal{F}^T$ satisfying

$$n^{-1} \sum_{i=1}^{n} \ell\left(\widetilde{f}(x_i; x_{1:n}), \widehat{f}_m^*(x_i)\right) \leq \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}_m^*(x_i)\right) + \delta \tag{7}$$

for some $\delta \geq 0$.

ii. For any $f(\cdot; x_{1:n}) \in \mathcal{F}^T$ and $\widehat{f}_m^*(\cdot) \in \mathcal{F}$, the approximate outer maximization routine

returns $\widetilde{x}_{1:n}$ satisfying

$$n^{-1}\sum_{i=1}^{n}\ell\left(f(\widetilde{x}_i;\widetilde{x}_{1:n}),\widehat{f}_m^*(\widetilde{x}_i)\right)\geq\max_{x_{1:n}}n^{-1}\sum_{i=1}^{n}\ell\left(f(x_i,x_{1:n}),\widehat{f}_m^*(x_i)\right)-\nu \tag{8}$$

for some $\nu\geq 0$.

Our analysis provides a finite sample bound on the plug-in program's error that explicitly depends on the optimization errors introduced by the approximate optimization routines.

Define $\widetilde{f}^T(\cdot;x_{1:n})$ to be the allowable function returned when the approximate inner minimization routine solves $\min_{f(\cdot)\in\mathcal{F}^T}n^{-1}\sum_{i=1}^{n}\ell\left(f(x_i),\widehat{f}_m^*(x_i)\right)$ at any feature values $x_{1:n}$. Analogously define $\widetilde{x}_{1:n}$ to be the feature values returned when the approximate outer maximization routine solves $\max_{x_{1:n}}n^{-1}\sum_{i=1}^{n}\ell\left(\widetilde{f}^T(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)$. Define the optimal values of the plug-in and population programs

$$\widehat{\mathcal{E}}_m:=n^{-1}\sum_{i=1}^{n}\ell\left(\widetilde{f}^T(\widetilde{x}_i,\widetilde{x}_{1:n}),\widehat{f}_m^*(\widetilde{x}_i)\right)\text{ and }\mathcal{E}_m=\max_{x_{1:n}}\min_{f(\cdot)\in\mathcal{F}^T}n^{-1}\sum_{i=1}^{n}\ell\left(f(x_i),f_m^*(x_i)\right) \tag{9}$$

respectively.

Proposition 3.1. *Suppose the loss function $\ell(\cdot,\cdot)$ is differentiable with gradients bounded by some $K<\infty$ and convex in its second argument. Then, for any $n\geq 1$,*

$$\left\|\widehat{\mathcal{E}}_m-\mathcal{E}_m\right\|\leq(\delta+\nu)+3K\|\widehat{f}_m^*(\cdot)-f_m^*(\cdot)\|_\infty, \tag{10}$$

where $\|f_1(\cdot)-f_2(\cdot)\|_\infty=\sup_{x\in\mathcal{X}}|f_1(x)-f_2(x)|$ is the supremum norm between two functions $f_1(\cdot),f_2(\cdot)\in\mathcal{F}$.

The error of the plug-in max-min program for the infeasible max-min program is bounded by the optimization error introduced by the approximate optimization routines and the estimation error of $\widehat{f}_m^*(\cdot)$ for the true function $f_m^*(\cdot)$. The estimation error contributes to the bound through the worst-case (supremum norm) error of $\widehat{f}_m^*(\cdot)$ for $f_m^*(\cdot)$. Equivalently, if we could exactly optimize and set $\delta,\nu=0$, the rate at which the plug-in optimal value converges to the population optimal value is bounded by the rate at which $\widehat{f}_m^*(\cdot)$ converges uniformly to the true function $f_m^*(\cdot)$. While strong, it is unsurprising that this strong form of convergence is sufficient to control the plug-in's error as the max-min optimization program explores the mapping $x\to f_m^*(x)$ and possibly extrapolates in searching for incompatible collections of empirical examples.

Importantly, the finite sample bound in Proposition 3.1 is agnostic, applying to any choice of the researcher's estimator $\widehat{f}_m^*(\cdot)$. By introducing additional regularity conditions

and for particular choices of the researcher's estimator $\widehat{f}_m^*(\cdot)$, existing work provides high-probability bounds on the worst-case error $\|\widehat{f}_m^*(\cdot) - f_m^*(\cdot)\|_\infty$ in terms of the sample size N_m and other primitives of the problem, such as the dimensionality of the features x. For example, see Belloni et al. (2015); Chen and Christensen (2015); Cattaneo, Farrell and Feng (2020, among many others) for recent results on the supremum norm convergence for a large class of series based estimators for $f_m^*(\cdot)$, reproducing kernel Hilbert space methods (e.g., Yang, Bhattacharya and Pati, 2017; Fischer and Steinwart, 2020), and deep neural networks (e.g., Imaizumi, 2023). Proposition 3.1 can therefore be combined with these existing results to provide high-probability bounds on the error of the plug-in max-min program.

3.2 Gradient descent ascent optimization

While Proposition 3.1 analyzes its statistical properties, this still leaves open the question of how to practically solve the inner minimization and outer maximization of the plug-in max-min program. The falsifier's manipulation of the features induces both variation in the theory's chosen allowable function and the true function $f_m^*(\cdot)$, making the outer maximization program difficult.

To tackle this problem, we notice that the plug-in max-min program (6) has connections to a recent computer science literature on adversarial learning (e.g., Madry et al., 2017; Akhtar and Mian, 2018; Kolter and Madry, 2018). In adversarial learning, "data-poisoning attacks" are studied to understand the robustness of black-box predictive algorithms. Given an estimated neural network for image classification, for example, we search for small perturbations to particular pixel values that would lead the neural network to (humorously) classify a picture of a pig as an airliner or (more dangerously) fail to notice a stop sign in a self-driving car. The resulting perturbed images are referred to as "adversarial examples." The plug-in max-min program's search for logical anomalies can be reinterpreted as a type of data-poisoning attack on the theory's allowable functions $\mathcal{F}^T$. The falsifier searches for collections of empirical examples that simultaneously poison the performance of all allowable functions $f(\cdot) \in \mathcal{F}^T$. The resulting logical anomalies are adversarial examples for the theory.

This connection to adversarial learning is more than mere intellectual curiosity. We can exploit the connection between the plug-in max-min program and data-poisoning attacks in adversarial learning in order to develop a feasible gradient descent ascent (GDA) optimization routine. Recent results on non-convex/concave max-min optimization (e.g., Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020) in adversarial learning provide optimization guarantees on its performance.

We first simplify the inner minimization over the theory's allowable functions. We assume the theory's allowable functions can be flexibly parametrized, meaning $\mathcal{F}^T = \{f_\theta(\cdot) \colon \theta \in \Theta\}$

for some (possibly high-dimensional) parameter vector θ and compact parameter space Θ. In expected utility theory, for example, we may construct such a parameterization using a flexible sieve basis or a class of neural networks for the possible utility functions. The inner minimization over the theory's allowable functions then becomes

$$\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \ell \left(f_\theta(x_i), \widehat{f}^*_m(x_i) \right). \tag{11}$$

For particular parametrizations and loss functions, this may be convex and so it can be solved accurately using convex optimization methods. Otherwise, we can apply standard gradient descent procedures with random initializations since it is equivalent to an empirical risk minimization problem. We can therefore implement an approximate inner minimization routine using standard optimization methods, and so we maintain our high-level Assumption 5(i).

By contrast, the outer maximization over features remains difficult as varying the feature vector simultaneously induces variation in the estimated function $\widehat{f}^*_m(\cdot)$, the theory's allowable function $f_\theta(\cdot)$ and the theory's best-fitting parameter vector $\theta \in \Theta$. The outer maximization problem will therefore typically be non-concave. We can nonetheless use a gradient-based optimization procedure. As notation, let $\widehat{\mathcal{E}}_m(x_{1:n}, \theta) := n^{-1} \sum_{i=1}^{n} \ell \left(f_\theta(x_i), \widehat{f}^*_m(x_i) \right)$ and we assume $\widehat{\mathcal{E}}_m(x_{1:n}, \theta)$ is differentiable in $x_{1:n}$ for all $\theta \in \Theta$. For a collection of initial feature values $x^0_{1:n}$, maximum number of iterations $S > 0$, and some chosen step size sequence $\{\eta_s\}_{s=0}^{S} > 0$, we iterate over $s = 0, \ldots, S$ and calculate at each iteration

$$\theta^{s+1} = \arg\min_{\theta \in \Theta} \widehat{\mathcal{E}}_m(x^s_{1:n}; \theta) \tag{12}$$

$$x^{s+1}_{1:n} = x^s_{1:n} + \eta_s \nabla \widehat{\mathcal{E}}_m(x^s_{1:n}; \theta^{s+1}). \tag{13}$$

At each iteration s, we construct an approximate solution to the inner minimization problem θ^{s+1}, and we then take a gradient ascent step on the feature values plugging in θ^{s+1}. Algorithm 1 summarizes our practical implementation of the gradient descent ascent algorithm.

Recent results in non-convex/concave max-min optimization imply that such a gradient descent ascent algorithm converges to an approximate stationary point of the outer maximization problem (Jin, Netrapalli and Jordan, 2019), loosely meaning that $\nabla \widehat{\mathcal{E}}_m(x_{1:n}, \theta) \approx 0$ at the returned feature and parameter vectors. We state this result formally in Appendix E.

Algorithm 1: Feasible gradient descent ascent for empirical anomalies.

Input: Estimated prediction function $\widehat{f}_m^*(\cdot)$, collection size n, maximum iterations S, step size sequence $\{\eta_s\}_{s=0}^S$, initial feature vector $x_{1:n}^0$.

$s \leftarrow 0$;
while $s < S$ **do**
  $\theta^{s+1} \leftarrow \arg\min_{\theta \in \Theta} \widehat{\mathcal{E}}_m(x_{1:n}^s; \theta)$;
  $x_{1:n}^{s+1} \leftarrow x_{1:n}^s + \eta_s \nabla \widehat{\mathcal{E}}_m(x_{1:n}^t; \theta^{s+1})$;
  $s \leftarrow s + 1$;
return $\{(x_1^S, \widehat{f}_m^*(x_1^S)), \ldots, (x_n^S, \widehat{f}_m^*(x_n^S))\}$.

4 Representational anomalies and example morphing

Our adversarial algorithm exploits no structure about theory $T(\cdot)$ beyond its allowable functions. If a strengthened Assumption 4 is satisfied, then theory $T(\cdot)$ has a lower-dimensional representation of the features, meaning $T(\cdot)$ behaves as if it always pools together some distinct feature values. In this case, researchers may be interested in uncovering what we call "representational anomalies," which highlight ways in which the theory fails to capture some relevant dimension along which modeled outcomes vary. We propose an example morphing algorithm to generate such representational anomalies.

4.1 Representational equivalence and logical anomalies

To this point, we modeled theory $T(\cdot)$ as a mapping that draws implications about the relationship between the features and modeled outcomes from any collection of examples, placing no assumptions on how $T(\cdot)$ behaves across feature values. However, theories often draw the same implications at distinct feature values x, x', which we formalize in the following definition.

Definition 4. Features $x_1, x_2 \in \mathcal{X}$ are *representationally equivalent* under theory $T(\cdot)$ if $T(x_1; D) = T(x_2; D)$ for all $D \in \mathcal{D}$.

Proposition 4.1. *Suppose theory $T(\cdot)$ satisfies Assumptions 1-4. Features x_1, x_2 are representationally equivalent if and only if $f(x_1) = f(x_2)$ for all $f(\cdot) \in \mathcal{F}^T$.*

Two features are representationally equivalent if theory $T(\cdot)$ always behaves as if it derives the same implications at their values. This has a simple interpretation in terms of a theory's allowable functions — all allowable functions assign the same modeled outcome value to the two features.

To build intuition, a theory $T(\cdot)$ has representationally equivalent features whenever it ignores any particular feature in an economic domain. Consider choice under risk and sup-

pose we include as a feature whether or not a lottery in the menu is presented as a compound lottery. Expected utility theory is silent on whether this presentational choice would influence an individual's decision. Any pairs of menus x_1, x_2 whose constituent lotteries have the same final payoffs and probabilities over those final payoffs yet differ in their presentation are representationally equivalent under expected utility theory.

While ignoring a particular feature is sufficient for a theory $T(\cdot)$ to have representationally equivalent features, it is perhaps surprisingly not necessary. Consider again choice under risk and suppose we define the feature vector to only consist of the final payoffs and probabilities associated of the constituent lotteries. Expected utility theory of course does not ignore any of these features in modeling risky choices. Any utility function $u(\cdot)$ is associated with an allowable function $f(\cdot) \in \mathcal{F}^T$ given by $f(x_1) = \arg\max \left\{ \sum_{j=1}^{J} p_{0j} u(z_{0j}), \sum_{j=1}^{J} p_{1j} u(z_{1j}) \right\}$ for menu $x_1 = (p_0, z_0, p_1, z_1)$, and all payoffs and probabilities of the lotteries may influence choice. Yet there exists a representationally equivalent menu x_2 that consists of the compound lotteries $\alpha(p_0, z_0) + (1-\alpha)(\widetilde{p}, \widetilde{z})$ and $\alpha(p_1, z_1) + (1-\alpha)(\widetilde{p}, \widetilde{z})$. The pair of menus satisfies $f(x_1) = f(x_2)$ due to the independence axiom and the linearity of expected utility in probabilities.

We next strengthen Assumption 4 ("non-trivial implications"), and then we establish that any theory $T(\cdot)$ has a non-trivial, lower-dimensional representation of the features.

Assumption 6 (Sharp implications)**.** There exist $x_1, x_2 \in \mathcal{X}$ such that $T(x_k; D) = y_j^*$ for all $D \in \mathcal{D}$ compatible with theory $T(\cdot)$ and $(x_j, y_j^*) \in D$ for $j \neq k$.

Proposition 4.2. *Suppose theory $T(\cdot)$ satisfies Assumption 1, 2, 3 and 6. Then, there exists some pair $x_1, x_2 \in \mathcal{X}$ that are representationally equivalent under theory $T(\cdot)$.*

To prove the result, suppose that the pair $x_1, x_2 \in \mathcal{X}$ in Assumption 6 were not representationally equivalent under theory $T(\cdot)$ for sake of contradiction. There must then exist some $D \in \mathcal{D}$ at which $T(x_1; D) \neq T(x_2; D)$, and we can construct $\widetilde{D}$ satisfying $D \subset \widetilde{D}$ that is compatible with theory $T(\cdot)$ but violates Assumption 6. Assumption 6 states that there exists some pair of feature values $x_1, x_2 \in \mathcal{X}$ such that if theory $T(\cdot)$ is provided with either (x_1, y_1^*) or (x_2, y_2^*), then it sharply generalizes to the other feature value in the pair. Proposition 4.2 establishes that Assumption 6 is sufficient for there to exist a non-trivial representation of the features under theory $T(\cdot)$.

Representationally equivalent features, if they exist, provide more structure that can be exploited for anomaly generation. If theory $T(\cdot)$ has a non-trivial representation of the features, then particular logical anomalies highlight failures in the theory's representation.

Definition 5. Consider any theory $T(\cdot)$ satisfying Assumptions 1, 2, 3 and 6. A logical

anomaly D for theory $T(\cdot)$ is a *representational anomaly* if there exists $(x_1, y_1^*), (x_2, y_2^*) \in D$ such that x_1, x_2 are representationally equivalent under $T(\cdot)$ but $y_1^* \neq y_2^*$.

We refer to logical anomalies further satisfying Definition 5 as *representational anomalies*. A representational anomaly highlights that there exists some pair of features that are representationally equivalent under theory $T(\cdot)$ but across which the modeled outcome varies. In this sense, there is some variation in the modeled outcome across features that is not captured by the theory's allowable functions. Researchers are often most interested in uncovering representational anomalies for theories. Indeed, many classic examples of logical anomalies for expected utility theory are, in fact, representational anomalies.

Examples: the Allais paradox and Certainty effect Consider once again the Allais paradox for expected utility theory (Table 1). Due to the independence axiom, expected utility theory requires that $T(x_A; D) = T(x_B; D)$ for all collections and so the menus x_A, x_B are representationally equivalent. Yet the Allais paradox highlights that choices may vary across these two menus, and it is therefore a representational anomaly. Analogously, the Certainty effect for expected utility theory (Table 2) is also a representational anomaly by the same reasoning. ▲

Furthermore, building on the earlier discussion, Tversky and Kahneman (1981) construct other representational anomalies for expected utility theory that highlight whether lotteries are presented as two-stage lotteries versus simple lotteries may affect individuals' risky choices.

4.2 An example morphing algorithm

Given modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot) = \mathbb{E}_m[g(Y_i) \mid X_i = x]$ for some researcher-specified $g(\cdot)$, our goal is to search for empirical representational anomalies $\{(x_1, f_m^*(x_1)), (x_2, f_m^*(x_2))\}$ for theory $T(\cdot)$.

To motivate our procedure, we further assume the true function and all of theory $T(\cdot)$'s allowable functions are differentiable and that theory $T(\cdot)$'s representation is *local*.

Assumption 7 (Differentiability and local representational equivalence)**.**

1. $f_m^*(\cdot)$ and all $f(\cdot) \in \mathcal{F}^T$ are everywhere differentiable.

2. If features $x_1, x_2 \in \mathcal{X}$ are representationally equivalent, then so are $\lambda x_1 + (1-\lambda)x_2$ for any $\lambda \in (0, 1)$.

Under this assumption, representations are *local* in the sense that there exists a small deviation from x_1 or x_2 that is also representationally equivalent. Expected utility theory satisfies this assumption per our earlier discussion.

Under Assumption 7, we might hope to uncover representational anomalies by taking small gradient-based steps. Suppose we have oracle access to the true function $f_m^*(\cdot)$. Given an initial feature value x^0, we search for directions $v \in \mathbb{R}^{dim(x)}$ along which no allowable function $f(\cdot) \in \mathcal{F}^T$ changes but $f_m^*(\cdot)$ changes substantially, and we then update or *morph* x^0 in the direction v.

More precisely, let $\mathcal{N}(x) = \{v \in \mathbb{R}^{dim(x)} \colon \nabla f(x)'v = 0 \text{ for all } f(\cdot) \in \mathcal{F}^T\}$ denote the subspace of directions that are orthogonal to the gradient of each allowable function. Under Assumption 7, $\mathcal{N}(x)$ is non-empty at any x for which there exists some representationally equivalent x'. For an initial feature value x^0, maximum number of iterations S, and step size sequence $\{\eta_s\}_{s=0}^S$, we would iterate over $s = 0, \ldots, S$ and compute the update step

$$x^{s+1} = x^s - \eta_s \text{Proj}\left(\nabla f_m^*(x^s) \mid \mathcal{N}(x^s)\right), \tag{14}$$

where $\text{Proj}\left(\cdot\right)$ is the projection operator and $\text{Proj}\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right)$ is the projection of the gradient of the true function $f_m^*(\cdot)$ onto the null space of the allowable functions. We therefore move in descent directions of the true function $f_m^*(\cdot)$ that hold fixed the value of any allowable function $f(\cdot) \in \mathcal{F}^T$. We focus on descent directions, but we could instead apply an ascent step as well. Finally, if there are known directions v^s along which no allowable functions vary (e.g., the theory ignores some feature), then we could directly define the update direction as $\text{Proj}\left(\nabla f_m^*(x^s) \mid v^s\right)$.

This is, of course, not feasible since we do not observe the true function $f_m^*(\cdot)$. As a result, we again construct an estimator $\nabla \widehat{f}_m^*(\cdot)$ based on a random sample $(X_i, Y_i) \sim P_m(\cdot)$ i.i.d. for $i = 1, \ldots, n$. We then plug this estimator into the morphing procedure and apply the update step

$$x^{s+1} = x^s - \eta_s \text{Proj}\left(\nabla \widehat{f}_m^*(x^s) \mid \mathcal{N}(x^s)\right) \tag{15}$$

at each iteration $s = 0, \ldots, S$. Our next result establishes $\text{Proj}\left(\nabla \widehat{f}_m^*(x) \mid \mathcal{N}(x)\right)$ remains a descent direction for the true function $f_m^*(\cdot)$, provided the error in estimating the gradient $\nabla \widehat{f}_m^*(\cdot) - \nabla f_m^*(\cdot)$ is sufficiently small.

Proposition 4.3. *Under Assumption 7, $-Proj\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right)$ is a descent direction for $f_m^*(\cdot)$. Furthermore, $-Proj\left(\nabla \widehat{f}_m^*(x) \mid \mathcal{N}(x)\right)$ is also a descent direction for $f_m^*(\cdot)$ provided $\|\nabla \widehat{f}_m^*(x) - \nabla f_m^*(x)\|_2 \leq \|Proj\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right)\|_2$ is satisfied.*

While Proposition 4.3 analyzes the statistical properties of plugging the estimated gra-

dient of the true function into the morphing procedure, it still leaves open the question of how to practically implement the projection operator.

Algorithm 2: Feasible example morphing for representational anomalies.

Input: Estimated gradient $\nabla \widehat{f}_m^*(\cdot)$, $B > 0$, maximum iterations S, step size sequence $\{\eta_s\}_{s=0}^S$, initial feature x^0.

$s \leftarrow 0$;
while $s < S$ **do**
  Sample $\theta_b \in \Theta$ for $b = 1, \ldots, B$;
  Construct $\mathcal{N}_\Theta(x^s) = \{v \in \mathbb{R}^{dim(x)} \text{ s.t. } \nabla f_{\theta_b}(x^s)'v = 0 \text{ for all } b\}$;
  $x^{s+1} \leftarrow x^s - \eta_s \text{Proj}\left(\nabla \widehat{f}_m^*(x^s) \mid \mathcal{N}(x^s)\right)$;
  $s \leftarrow s + 1$;s
return $\{(x^0, \widehat{f}_m^*(x^0)), (x^S, \widehat{f}_m^*(x^S))\}$.

To do so, we will again assume that the theory's allowable functions can be flexibly parameterized, meaning $\mathcal{F}^T = \{f_\theta(\cdot) \colon \theta \in \Theta\}$ for some $\theta \in \Theta$ as in Section 3.2. We practically implement the projection operator by sampling $B > 0$ parameter values $\theta \in \Theta$ at each update step and directly orthogonalizing the gradient $\nabla \widehat{f}_m^*(x)$ with respect to each of the sampled gradients $\nabla f_\theta(x)$. As B grows large, this better approximates the null space of the allowable function $\mathcal{N}(x)$. Of course, if there are known directions v^s along which no allowable functions vary, then this gradient sampling step is not needed and we can directly orthogonalize the gradient $\nabla \widehat{f}_m^*(x)$ with respect to the known directions. Algorithm 2 summarizes our practical implementation of the morphing procedure, which can be run over many randomly initialized feature values x^0.

5 Algorithmically generating anomalies for choice under risk

Both of our anomaly generation procedures take two inputs: an existing theory represented by their allowable functions $\mathcal{F}^T$ and a black-box predictive algorithm $\widehat{f}_m(\cdot)$. The allowable functions, of course, depend on properties of the existing theory, but how may researchers construct the black-box predictive algorithm? The current researcher-driven process of anomaly generation suggests two possibilities.

One way researchers construct logical anomalies is by contrasting an existing theory with their intuitions about the world, as Allais likely did. By way of analogy, given a large collected dataset on outcomes modeled by the existing theory, we could fit a supervised learning algorithm to predict those outcomes based on the available features. The resulting black-box

predictive algorithm would then serve as the "intuition" for our anomaly generation procedures – an empirically derived object that summarizes statistical patterns. Alternatively, researchers construct logical anomalies by contrasting the existing theory with an alternative theory in mind. For instance, Kahneman and Tversky could have generated their celebrated logical anomalies for expected utility theory by contrasting it with particular implications of the cumulative prospect theory. We could likewise fit a predictive algorithm to a dataset simulated from an alternative theory, and the resulting black-box predictor could be given as input to our procedures. In this case, our procedures would produce logical anomalies for the existing theory that are implied by the alternative.

In this section, we apply our anomaly generation procedures to the second case. We algorithmically generate logical anomalies for expected utility theory that are implied by cumulative prospect theory, as Kahneman and Tversky may have done. While interesting in its own right, illustrating our procedure in this way serves an important purpose. Were we to apply our anomaly generation procedures on collected choice data, we would have no benchmark for what logical anomalies we should expect to find. By contrast, since cumulative prospect theory has been well-studied by theorists for decades, we can compare the logical anomalies generated by our algorithmic procedures against known anomalies for expected utility theory constructed by researchers, such as those produced in Allais (1953), Kahneman and Tversky (1979), and many others. We explore whether our algorithmic procedures reproduce known logical anomalies for expected utility theory implied by cumulative prospect theory. Our anomaly generation procedures clear this hurdle, and even go further generating novel logical anomalies that – to our knowledge – had not been noticed before by researchers.

5.1 Illustration design

We simulate lottery choice data from an individual who evaluates lotteries over $J > 1$ monetary payoffs according to the parametric probability weighting function

$$\pi_j(p;\delta,\gamma) = \frac{\delta p_j^\gamma}{\delta p_j^\gamma + \sum_{k\neq j} p_k^\gamma} \text{ for } j = 1,\ldots,J, \tag{16}$$

where $p \in \Delta^{J-1}$ and $\delta \geq 0, \gamma \geq 0$ are the parameters governing the level and curvature of the probability weighting function (Lattimore, Baker and Witte, 1992). We calibrate the parameters (δ,γ) using the pooled estimates based on the large-scale choice experiments in Bruhin, Fehr-Duda and Epper (2010) (reported in their Table V and Table IX), setting (δ,γ) to be equal to one of $(0.926, 0.377)$, $(0.726, 0.309)$, or $(1.063, 0.451)$.

For these parameter values of the probability weighting function (16), the individual distorts objective probabilities by over-weighting probabilities close to zero, under-weighting

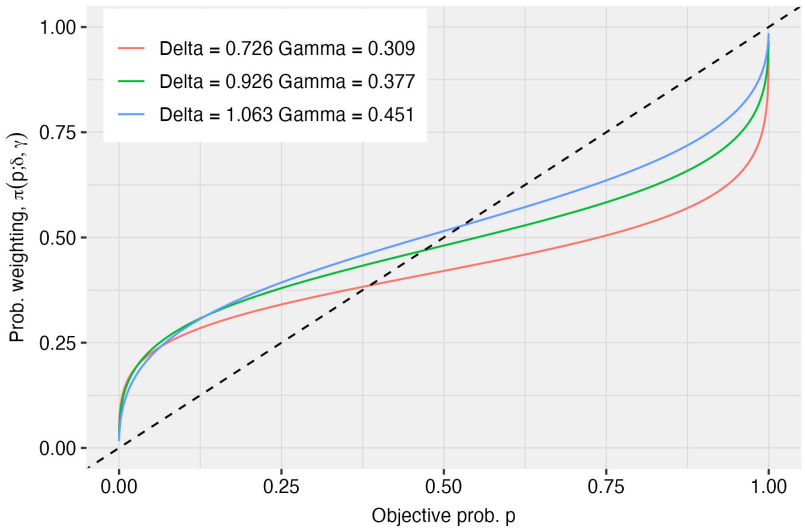


Figure 1: Probability weighting function for calibrated parameter values (δ, γ) in our illustration to choice under risk.

Notes: This figure plots the probability weighting function (16) for the calibrated parameter values (δ, γ) used in our illustration to choice under risk. We calibrate (δ, γ) to be equal to $(0.726, 0.309)$, $(0.926, 0.377)$, and $(1.063, 0.451)$ using the pooled estimates based on the large-scale choice experiments in Bruhin, Fehr-Duda and Epper (2010) (reported in their Table V and Table IX). See Section 5.1 for further discussion.

probabilities close to one, and compressing intermediate probabilities. Figure 1 plots the resulting probability weighting functions associated with each choice (δ, γ), depicting the canonical "s-shapes." Such non-linearities in the probability weighting function can generate several known logical anomalies for expected utility theory, such as the Allais paradox (Table 1), the Certainty effect (Table 2), and several others. These parameter values also introduce "outcome pessimism" when $\delta < 1$ as the individual's probability weights may sum to less than one (i.e., $\sum_{j=1}^{J} \pi_j(p; \delta, \gamma) < 1$), or "outcome optimism" when $\delta > 1$ as the individual's probability weights may sum to greater than one (i.e., $\sum_{j=1}^{J} \pi_j(p; \delta, \gamma) > 1$). These properties may lead the individual to select a lottery that is first-order stochastically dominated by another lottery in the menu. Expected utility maximization over any utility function that is weakly increasing in monetary payoffs cannot generate such first-order stochastic dominance violations.

We assume the individual has a linear utility function. For any payoff vector $z \in \mathbb{R}^J$ and associated probabilities $p \in \Delta^{J-1}$, the individual therefore evaluates the lottery (p, z) by $CPT(p, z; \delta, \gamma) := \sum_{j=1}^{J} \pi_j(p; \delta, \gamma) z_j$. On a menu of two lotteries, $x = (p_0, z_0, p_1, z_1)$, we simulate the individual's choice probability of selecting lottery 1 according to $f_m^*(x) = P\left(CPT(p_1, z_1; \delta, \gamma) - CPT(p_0, z_0; \delta, \gamma) + \xi \geq 0\right)$, where ξ is an i.i.d. logit shock. The individual's binary choice is given by the random variable $Y_i \mid X_i = x \sim Bernoulli(f_m^*(x))$.

To apply our anomaly generation procedures, we flexibly parametrize the allowable functions of expected utility theory and model the utility function as a linear combination of non-linear basis functions with $u_\theta(z) = \sum_{k=1}^{K} \theta_k b_k(z)$ for basis functions $b_1(\cdot), \ldots, b_K(\cdot)$ (e.g., polynomial bases or monotone I-splines), K finite, and parameter vector $\theta \in \Theta$. We then consider the parametrized allowable functions of expected utility theory as the collection $\{f_\theta(\cdot)\colon \theta \in \Theta\}$ for $f_\theta(x) = P\left(\sum_{j=1}^{J} p_1(j) u_\theta(z_1(j)) - \sum_{j=1}^{J} p_0(j) u_\theta(z_0(j)) + \xi \geq 0\right)$, where ξ is also an i.i.d. logit shock. We generate logical anomalies for expected utility theory over the space of menus of two lotteries on two monetary payoffs, applying our adversarial procedure (Algorithm 1) and our example morphing procedure (Algorithm 2) to the true choice probability function $f_m^*(\cdot)$. In Appendix F.4, we generate logical anomalies based on an estimated choice probability function $\widehat{f}_m(\cdot)$ from a random sample of binary choices. In Appendix G, we generate logical anomalies over the space of menus of two lotteries over three monetary payoffs.

For each parameter value (δ, γ), we apply our adversarial algorithm to 25,000 randomly initialized menus of two lotteries on two monetary payoffs x^0 and our example morphing algorithm to 15,000 randomly initialized menus. Appendix F.1 provides further details on our practical implementation. Each returned menu of lotteries over two monetary payoffs are logical anomalies for expected utility theory at our particular parametrization of the utility function $\{u_\theta(\cdot)\colon \theta \in \Theta\}$. Since these parametrized allowable functions are restrictive, we numerically verify whether the returned menu is a logical anomaly for expected utility theory at any increasing utility function and without noisy choices (see Appendix F.2 for details). We report all resulting numerically verified logical anomalies for expected utility theory.

5.2 What logical anomalies do the algorithms generate?

Table 3 summarizes the logical anomalies for expected utility theory that are produced by our anomaly generation procedures at each calibrated parameter value (δ, γ). In order to better interpret these generated anomalies, we categorize them ourselves based on the particular violation of expected utility theory they highlight. All together, our anomaly generation procedures uncover several distinct categories of logical anomalies for expected utility theory that are implied by the probability weighting function.

One of cumulative prospect theory's key insights is the role of reference points in decision making. In assessing probabilities, there are two natural reference points: zero (when an event is certain to not occur) and one (when an event is certain to occur). These reference points influence how we perceive probabilities, and we exhibit *diminishing sensitivity* to changes in probabilities as they move away from either zero or one (Tversky and Kahneman,

	Prob. Weighting Function: (δ, γ)		
	$(0.726, 0.309)$	$(0.926, 0.377)$	$(1.063, 0.451)$
Dominated Consequence Effect	85	34	10
Reverse Dominated Consequence Effect	17	15	14
Strict Dominance Effect	45	1	0
First Order Stochastic Dominance	81	0	2
Other	3	1	1
# of Logical Anomalies	231	51	27

Table 3: Logical anomalies for expected utility theory over the space of menus of two lotteries on two monetary payoffs.

Notes: This table summarizes all logical anomalies for expected utility theory over two lotteries on two monetary payoffs produced by our adversarial algorithm and our example morphing algorithm. The logical anomalies are organized by calibrated parameter values (δ, γ) of the probability weighting function and anomaly categories. See Section 5.2 for further discussion.

1992). Shifting a probability from either 1% to 10% or from 99% to 90% looms larger in our minds than shifting a probability from 41% to 50%.[13] Diminishing sensitivity in our perceptions of probabilities suggests that individuals do not treat probabilities linearly as modeled by expected utility theory, and the resulting probability distortions are captured by the well-known "s-shape" of the probability weighting function (see Figure 1).

Importantly, diminishing sensitivity in our perceptions of probabilities can produce apparent reversals in our choices across menus of lotteries, even when expected utility theory sharply predicts our choices should be unchanged. Several well-known logical anomalies for expected utility theory highlight exactly such choice reversals. For example, the Allais paradox (Table 1) highlights an extreme example in which risk attitudes appear to change across the menus: in menu A, many individuals select the lower expected value lottery, consistent with risk averse attitudes; yet, at the same time, individuals also select the higher expected value lottery in menu B, consistent with risk loving attitudes. Expected utility theory, by contrast, requires choices to be the same across the menus in the Allais paradox. Analogous choice reversals arise in the Certainty effect (Table 2) and in more complex logical anomalies like the Pseudo-certainty effect (Tversky and Kahneman, 1981) involving compound lotteries.

Our algorithmic procedures generate three categories of logical anomalies for expected utility theory that illustrate choice reversals over menus of two lotteries on two monetary payoffs due to the probability weighting function. We discuss each category next, focusing on particular illustrative examples produced by our algorithmic procedures.

[13] See Wu and Gonzalez (1996); Prelec (1998); Gonzalez and Wu (1999) for further discussion of properties of the probability weighting function and their roles in affecting decision-making.

5.2.1 The dominated consequence effect

The logical anomalies in the first row of Table 3 highlight what we refer to as a "dominated consequence effect." As an illustration, consider the algorithmically generated pair of menus in Table 4.[14] Begin by noting that the individual selects lottery A0. Since lottery A0 has a lower expected value than lottery A1, expected utility theory could rationalize this choice with an appropriate degree of risk aversion. Since the payoffs in the lotteries across menu A and menu B are the same, under expected utility theory the individual's risk attitudes should be fixed across these menus. Yet the individual also selects the lottery B1, which has a higher expected value than lottery B0. It appears as if the individual's risk attitudes have reversed on menu B. Put in another way, even though lottery B0 and lottery B1 both raise the probability of their lowest payoff, the movement from a 0% to 11% change of the lowest payoff across lottery A0 to lottery B0 looms larger in the individual's mind than the change from a 13% to 34% chance of the lowest payoff across lottery A1 to lottery B1.

Menu A (x_A, y_A^*)		
Lottery 0	\$6.44	\$6.71
	0%	100%
Lottery 1	\$5.72	\$8.64
	13%	87%

Menu B (x_B, y_B^*)		
Lottery 0	\$6.44	\$6.71
	11%	89%
Lottery 1	\$5.72	\$8.64
	34%	66%

Table 4: An illustrative example of an algorithmically generated logical anomaly for expected utility theory that illustrates the dominated consequence effect.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. This logical anomaly exhibiting the dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$, and it was produced by our example morphing algorithm applied to the choice probability function based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

Importantly, the change in probabilities across menu A and menu B is crafted exactly so that the choice predictions of expected utility theory are constant across these menus. Formally, each lottery in menu B can be expressed as a compound lottery over the corresponding lottery in menu A and some degenerate lotteries that yield certain payoffs. Lottery B0 can be expressed as a compound lottery over lottery A0 and a degenerate lottery that yields the certain payoff 6.44; that is, $B0 = \alpha_0 A0 + (1 - \alpha_0)\delta_{6.44}$ for some $\alpha_0 \in (0, 1)$. Analogously, lottery B1 can be written as the compound lottery $B1 = \alpha_1 A1 + (1 - \alpha_1)\delta_{5.72}$ for some $\alpha_1 < \alpha_0$. The individual's choices therefore express that lottery $A0$ is preferred to lottery $A1$ and $\alpha_1 A1 + (1 - \alpha_1)\delta_{5.72}$ is preferred to $\alpha_0 A0 + (1 - \alpha_0)\delta_{6.44}$. This contradicts the axioms

[14] In Appendix Table A1, we provide six additional illustrative examples of the dominated consequence effect that were produced by our algorithmic procedures.

of expected utility theory since it can be shown that $A0$ being preferred to $A1$ must imply that $\alpha_0 A0 + (1-\alpha_0)\delta_{6.44}$ is preferred to $\alpha_1 A1 + (1-\alpha_1)\delta_{5.72}$. We provide a formal proof in Appendix F.3.

More generally, all logical anomalies for expected utility theory in the second row of Table 3 have the following common structure. We define the appropriate pair of lotteries as $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ with $z_0 = (z_{0,1}, z_{0,2})$, $z_1 = (z_{1,1}, z_{1,2})$ and $\underline{z}_0 := \min_{j\in\{1,2\}} z_{0j} < \min_{j\in\{1,2\}} z_{1j} := \underline{z}_1$. Each of these logical anomalies can then be summarized as: for some $\alpha_0 \leq \alpha_1$, one menu consists of the choice between lottery ℓ_0 and lottery ℓ_1, and the other menu consists of the choice between the compound lotteries $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\underline{z}_0}$ and $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\underline{z}_1}$. Since the other menu mixes lotteries ℓ_0 and ℓ_1 with their minimal payoffs, selecting ℓ_1 over ℓ_0 implies that the individual also prefers $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\underline{z}_1}$ over $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\underline{z}_0}$. We therefore say these logical anomalies exhibit a "dominated consequence effect" as the pair of menus highlight a violation of the expected utility theory based on mixing each lottery with dominated certain consequences.

Furthermore, the Common ratio effect (e.g., Allais, 1953; Kahneman and Tversky, 1979) is a special case of the dominated consequence effect (see, for example, Machina (1987) for further discussion). It can be recovered from the dominated consequence effect by setting $\alpha_0 = \alpha_1$ and placing additional restrictions on how the probabilities p_0, p_1 relate to one another. The Common ratio effect is itself a generalization of the Certainty effect (Kahneman and Tversky, 1979) and the Bergen Paradox (Hagen, 1979). In this sense, the dominated consequence effect nests the most well-known logical anomalies for expected utility theory over pairs of menus of two lotteries over two monetary payoffs. Our anomaly generation procedures uncovered this category of logical anomalies on its own.

5.2.2 The reverse dominated consequence effect and the strict dominance effect

In the second row of Table 3, all logical anomalies exhibit what we call a "reverse dominated consequence effect." Consider again the algorithmically generated reverse dominated consequence effect anomaly in Table 5.[15] To build intuition, notice that the individual selects lottery A0. The lotteries in menu B again have the same payoffs as those in menu A, and the probabilities associated with the highest payoffs increase in both lotteries across these menus. Due to diminishing sensitivity of probabilities, the change from 12% to 51% of the highest payoff across lottery A0 to lottery B0 is less enticing than the change from 1% to 35% of the highest payoff across lottery A1 to lottery B1. This produces an apparent choice reversal as the individual now selects lottery B1. Once again, the change in probabilities

[15] We provide six additional illustrative examples of the reverse dominated consequence effect in in Appendix Table A2.

across these menus is crafted such that the choice predictions of expected utility theory do not change across these menus.

	Menu A (x_A, y_A^*)	
Lottery 0	\$2.59 88%	\$8.87 12%
Lottery 1	\$3.51 99%	\$8.65 1%

	Menu B (x_B, y_B^*)	
Lottery 0	\$2.59 49%	\$8.87 51%
Lottery 1	\$3.51 65%	\$8.65 35%

Table 5: An illustrative example of an algorithmically generated logical anomaly for expected utility theory that illustrates the reverse dominated consequence effect.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. This logical anomaly exhibiting the reverse dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$, and it was produced by our example morphing algorithm applied to the choice probability function based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

Each logical anomaly illustrating the reverse dominated consequence effect has a common structure. Again, we define the appropriate pair of lotteries as $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ with $z_0 = (z_{0,1}, z_{0,2})$, $z_1 = (z_{1,1}, z_{1,2})$ and $\overline{z}_0 := \min_{j\in\{1,2\}} z_{0j} < \min_{j\in\{1,2\}} z_{1j} := \overline{z}_1$. Each of these logical anomalies can be summarized as: for some $\alpha_1 \leq \alpha_0$, one menu consists of the choice between lottery ℓ_0 and lottery ℓ_1, and the other menu consists of the choice between the compound lotteries $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\overline{z}_0}$ and $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\overline{z}_1}$. Since the other menu mixes lotteries ℓ_0 and ℓ_1 with their maximal payoffs, selecting ℓ_1 over ℓ_0 implies that the individual also prefers $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\overline{z}_1}$ over $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\overline{z}_0}$ if their preferences are consistent with expected utility theory. In other words, the pair of menus highlights a violation of expected utility theory based on mixing each lottery with dominating certain consequences. We therefore refer to this category as a "reverse dominated consequence effect" due to its close parallel to the dominated consequence effect discussed earlier.

Next, all logical anomalies in the third row of Table 3 exhibit what we call a "strict dominance effect." We again provide an illustrative example in Table 6, and we list six additional examples in Appendix Table A3. In this case, we notice that the algorithmically generated logical anomaly shares a similar intuition as the original Allais paradox. The individual selects lottery A0 despite it having the lower expected payoff in the menu, demonstrating a degree of risk aversion. Since the payoffs are fixed across the menus, the individual's risk attitudes must be unchanged under expected utility theory. However, the individual selects lottery B1 in menu B which is the higher expected payoff, and it appears that the individuals' risk attitudes have reversed. This choice reversal is particularly transparent across this pair of menus. Lottery B0 raises the probability of the lowest payoff in lottery A0, whereas

lottery B1 raises the probability of the highest payoff in lottery 1. Yet despite selecting lottery A1 over lottery A0, the individual is now selects lottery B0 over lottery B1.

	Menu A (x_A, y_A^*)	
Lottery 0	\$6.71	\$8.98
	22%	78%
Lottery 1	\$7.17	\$8.04
	100%	0%

	Menu B (x_B, y_B^*)	
Lottery 0	\$6.71	\$8.98
	49%	51%
Lottery 1	\$7.17	\$8.04
	45%	55%

Table 6: An illustrative example of an algorithmically generated logical anomaly for expected utility theory that illustrates the strict dominance effect.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. This logical anomaly exhibiting the strict dominance effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$, and it was produced by our example morphing algorithm applied to the choice probability function based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

Once again, it can be shown these these menus are exactly crafted so that the choice predictions of expected utility theory do not change. For an appropriate choice of menu in these logical anomalies, menu A consists of the choice between lottery ℓ_0 and lottery ℓ_1, and menu B consists of the choice between the compound lotteries $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\bar{z}_1}$ and $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\underline{z}_0}$. Lottery B0 mixes lottery A0 with a certain lottery that yields its smallest payoff, and lottery B1 mixes lottery A1 with a certain lottery that yields its maximal payoff. If the individual selects lottery ℓ_1 over lottery ℓ_0, then the individual must also prefer $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\bar{z}_1}$ over $\alpha_0\ell_0 + (1-\alpha_0)\delta_{\underline{z}_0}$ if their preferences are consistent with expected utility theory. Yet we observe the opposite choice for the considered parameterizations of the probability weighting function. In this sense, the pair of menus highlights a violation of expected utility theory based on mixing lottery A1 with a certain consequence that strictly dominates the certain consequence that is mixed with lottery A0. Hence we refer to this category as a "strict dominance effect."

While sharing some similarities, these final two categories of logical anomalies for expected utility theory are importantly different than both the Common consequence Effect and Common ratio effect, which were important motivating logical anomalies for the development of the probability weighting function. These categories highlight violations of expected utility theory while using only two distinct payoffs in each lottery (like the Common ratio effect), but involve mixing each lottery with particular certain consequences. Our anomaly generation procedures uncovered categories of logical anomalies for expected utility theory that are implied by particular properties of the probability weighting function, but to our knowledge have not been noticed before.

5.2.3 First-order stochastic dominance violations

Finally, all logical anomalies in the last row of Table 3 are menus of lotteries in which the individual's choice violates first-order stochastic dominance. As we show in the examples in Appendix Table A4, these logical anomalies are all examples in which the individual selects lotteries that are first-order stochastically dominated by the other lottery in the menu. Such first-order stochastic dominance violations are generally viewed as an undesirable "bug" in particular specifications of the probability weighting function since we may believe they are unlikely to hold in real choices. Indeed, Kahneman and Tversky (1979) include an "editing phase" prior to choice that eliminates such first-order stochastic dominated lotteries. What is intriguing is that our anomaly generation procedures uncover these first-order stochastic dominance violations on their own.

5.3 Experimental test of algorithmically generated anomalies

Our procedures generate novel logical anomalies for expected utility theory that are implied by the probability weighting function. While these are interesting theoretically, a natural question nonetheless arises: are these logical anomalies also empirical anomalies for expected utility theory? Answering this question is where our anomaly generation procedures end, and careful experimental work begins. While fully investigating their experimental robustness is beyond the scope of this paper, we next present some experimental evidence suggesting that our algorithmically generated logical anomalies are also empirical anomalies for expected utility theory.

5.3.1 Experimental design

We selected 36 algorithmically generated logical anomalies for expected utility theory summarized in Table 3. These particular logical anomalies are chosen to span both the categories (i.e., the dominated consequence effect, the reverse dominated consequence effect, and the strict dominance effect) and the calibrated parameter values (δ, γ). We split these chosen 36 logical anomalies into two separate surveys, each containing 18 logical anomalies, which we deploy separately.

Each chosen logical anomaly consists of a pair of menus of two lotteries over two monetary payoffs. We present each logical anomaly as two separate binary choices on menus, and so each survey consists of 36 main questions. For a particular menu, we display the written probabilities and payoffs for each lottery in the menu, and we additionally depict each lottery as a color-coded pie chart. Each survey randomizes the order of questions and the left-right positioning of lotteries in a menu across respondents. We preregistered both of our surveys

on EGAP (see https://osf.io/2udca).

We recruited respondents for both surveys on Prolific. Each respondent received a base payment of $4 for completing a survey. We screened out inattentive respondents through comprehension questions and attention checks throughout the surveys.[16] Respondents that successfully completed a survey without failing comprehension and attention checks were eligible for a bonus payment based on a "random payment selection" mechanism (e.g., Azrieli, Chambers and Healy, 2018, 2020). We determined the bonus by randomly selecting a lottery that was chosen by a respondent on the survey. The respondent was paid the realization of the randomly selected lottery. The average bonus payment was $7.49 and $5.59 on each survey respectively, and respondents completed each survey in roughly 15 minutes on average. Respondents were therefore paid on average $45.96 and $38.36 per hour on survey respectively. Our financial incentives were unusually high by Prolific standards, which recommend that respondents be paid $12 per hour. We recruited 258 and 255 respondents on our two surveys respectively.

5.3.2 Experimental results

We analyze the choices of all respondents that completed the surveys without failing any attention and comprehension checks.[17] Figure 2 reports the fraction of respondents whose choices violate expected utility theory on our algorithmically generated logical anomalies. We organize the estimates by the category of logical anomaly, and we report 95% confidence intervals with standard errors clustered at the respondent level.[18] Table 7 provides summary statistics on the expected utility theory violation rates pooling across logical anomalies within the same category. The pooled expected utility theory violation rate is 11.4% (p-value < 0.001) on dominated consequence effect anomalies, 8.5% (p-value < 0.001) on reverse dominated consequence effect anomalies, and 12.7% (p-value < 0.001) on strict dominance effect anomalies. We therefore find strong evidence that the pooled respondents' choices are inconsistent with expected utility theory across our discovered categories of logical anomalies.

These pooled estimates mask heterogeneity in the fraction of respondents violating expected utility theory across logical anomalies. For example, we find that more than 15% of respondents' choices violate expected utility theory on several dominated consequence effect anomalies. Analyzing each logical anomaly separately and applying a conservative

[16] We include screenshots of the instructions, comprehension checks, attention checks, and main survey questions in Appendix H.

[17] In Appendix B, we report the same results, dropping the top 10% of respondents who completed the surveys the fastest and finding similar results.

[18] Appendix Figure A1 and Appendix Table A5 report the same estimates, organized by the calibrated parameter values (δ, γ) that we considered.

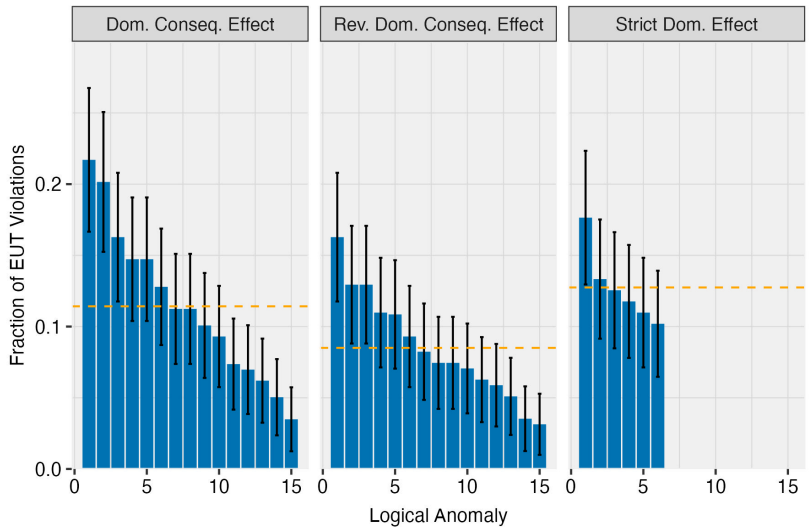


Figure 2: Fraction of respondents whose choices violate expected utility theory on algorithmically generated logical anomalies.

Notes: This figure summarizes the fraction of respondents whose choices violate expected utility theory on the logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors clustered at the respondent level). The orange dashed line reports the fraction of respondents whose choices violate expected utility theory pooling across all logical anomalies within the same category. The logical anomalies are sorted within each category in decreasing order based on the fraction of respondents whose choices violate expected utility theory and assign each logical anomaly an arbitrary numeric identifier. See Section 5.3 for further discussion.

Bonferroni correction for multiple hypotheses across all logical anomalies in our surveys, the expected utility theory violation rate is statistically different than zero at the 5% level for 35 out of 36. Respondents' choices are therefore inconsistent with expected utility theory on each algorithmically generated logical anomaly included in our surveys.

We can further compare the expected utility theory violation rates on our algorithmically generated logical anomalies against those of celebrated logical anomalies for expected utility theory in the behavioral economics literature. Several recent papers provide meta-analyses of past experiments and conduct comprehensive experimental designs to evaluate the empirical robustness of celebrated logical anomalies such as the Allais paradox and Common ratio effect. While the survey design and survey samples differ from our surveys, this work at least offers a rough benchmark to evaluate the magnitudes of the expected utility theory violation rates that we find on our algorithimcally generated, logical anomalies. In particular, we draw on Blavatskyy, Ortmann and Panchenko (2022) and Blavatskyy, Panchenko and Ortmann (2022), which conduct extensive meta-analyses of past experiments on the Allais paradox and the Common ratio effect respectively, as well as McGranaghan et al. (Forthcoming) and Jain and Nielsen (2023) which reported binary choice experiments that exhaustively test the

	Pooled Average	Median	First Quartile	Third Quartile
Dominated Consequence Effect	0.114 (0.006)	0.112	0.071	0.147
Reverse Dominated Consequence Effect	0.085 (0.007)	0.074	0.060	0.109
Strict Dominance Effect	0.127 (0.009)	0.121	0.111	0.131

Table 7: Summary statistics on the fraction of respondents whose choices violate expected utility theory on algorithmically generated logical anomalies.

Notes: This table reports summary statistics on the fraction of respondents whose choices violate expected utility theory ("expected utility theory violation rate") on algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs. We report summary statistics by category of logical anomaly (see Table 3). The "pooled average" column reports the expected utility theory violation rate, pooling together respondents' choices on all logical anomalies within the same category. Standard errors reported in parentheses are clustered at the respondent level. We also report the median, first quartile, and third quartile of the distribution of expected utility theory violation rates across logical anomalies within the same category. See Section 5.3 for further discussion.

Common Experiments across different payoffs and probabilities.

Appendix Table A6 summarizes the average expected utility theory violation rate as well as the median and interquartile range of the expected utility theory violation rate across experiments reported in these recent papers. There exists much variation in the expected utility theory violation rate on these celebrated logical anomalies across experiments. For example, Blavatskyy, Ortmann and Panchenko (2022) find that 16% of respondents' choices demonstrate the Allais paradox ("fanning out" choices) pooling together all experiments with real financial incentives, and the median experiment with real financial incentives only finds that 13.7% of respondents' choices do so. Similarly, in experiments conducted on Prolific with real financial incentives, McGranaghan et al. (Forthcoming) find that 15.6% of respondents' choices demonstrate the Common ratio effect and 12.9% demonstrate the Reverse Common ratio effect.[19]

Our algorithmically generated logical anomalies yield expected utility theory violation rates that are in line with these experimental findings on celebrated logical anomalies like the Allais paradox and the Common ratio effect. These new categories of anomalies may merit the same rigorous testing across a wide variety of experimental designs that have been

[19] McGranaghan et al. (Forthcoming) argue that the prior work included in Blavatskyy, Panchenko and Ortmann (2022)'s meta-analysis of the Common ratio effect select experimental designs that are more likely to induce the Common ratio effect.

given to other known anomalies for expected utility theory.

5.3.3 Could the algorithmically generated anomalies be explained by noisy choices?

Recent experimental work has suggested one simple extension to expected utility theory that could resolve many anomalies: specifically, incorporating a small amount of noise in individuals' choices.[20] We next examine whether small amounts of noise in individuals' choices could explain the empirical findings on our algorithmically generated logical anomalies.

We consider the expected utility theory with idiosyncratic errors, in which individuals mistakenly select the wrong lottery with some probability $\epsilon \in [0, 0.5]$. We estimate the idiosyncratic error rate ϵ from preferences consistent with expected utility theory that would be required to explain the observed choices of respondents on our algorithmically generated logical anomalies.

As an example, consider a logical anomaly that exhibits the dominated consequence effect such as the pair of menus depicted in Table 4. On this pair of menus, the only choices that are consistent with expected utility theory are $(A0, B0)$, $(A1, B1)$, and $(A1, B0)$, and let $\pi(A0, B0), \pi(A1, B1), \pi(A1, B0) \geq 0$ be the fraction of respondents associated with those true preferences. A respondent may erroneously deviate from their true preference with probability $\epsilon \geq 0$. Following Harless and Camerer (1994), we assume a single error rate for all choices since it is a parsimonious way to summarize observed choice fractions. For our purposes, this simple model of noisy choices serves to benchmark how frequently respondents must deviate from their true preferences in order to generate the observed choice fractions, whatever the source of those deviations may be. We may therefore search for the fraction of true preferences $\pi(A0, B0), \pi(A1, B1), \pi(A1, B0)$ and idiosyncratic error rate ϵ that could have generated the true choice fractions $P(A0, B0), P(A1, B1), P(A1, B0), P(B0, A1)$.[21] Given estimated choice fractions $\widehat{P}(A0, B0), \widehat{P}(A1, B0), \widehat{P}(A0, B1), \widehat{P}(A1, B1)$ from our surveys, we estimate the idiosyncratic error rate $\widehat{\epsilon}$ by a minimum distance estimator (Newey and McFadden, 1994).

Figure 3 reports the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on each algorithmically generated logical anomaly separately. We

[20] As mentioned earlier in Section 2, McGranaghan et al. (Forthcoming) explore whether a more general model of noisy expected utility theory could explain documented evidence of the Common ratio effect, and Enke and Shubatt (2023) consider whether expected utility theory with complexity-dependent noise could explain logical anomalies that arise due to the probability weighting function.

[21] In this example, the true choice fractions must satisfy $P(A0, B0) = (1-\epsilon)^2\pi(A0, B0)+\epsilon(1-\epsilon)\pi(A1, B0)+\epsilon^2 P(A1, B1)$, $P(A1, B0) = \epsilon(1-\epsilon)\pi(A0, B0) + (1-\epsilon)^2\pi(A1, B0) + \epsilon(1-\epsilon)P(A1, B1)$, $P(B0, A1) = \epsilon(1-\epsilon)\pi(A0, B0) + \epsilon^2\pi(A1, B0) + \epsilon(1-\epsilon)P(A1, B1)$, and $P(A1, B1) = (1-\epsilon)^2\pi(A0, B0) + \epsilon(1-\epsilon)\pi(A1, B0) + (1-\epsilon)^2 P(A1, B1)$.

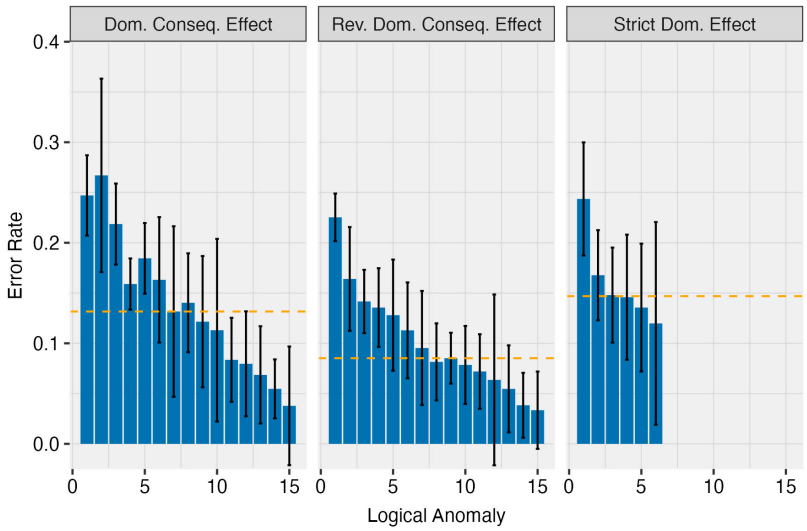


Figure 3: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated logical anomalies.

Notes: This figure summarizes the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on our algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors computed by the bootstrap). The orange dashed line reports the median estimated idiosyncratic error rate across all logical anomalies within the same category. Within each category, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

again organize the estimates by the category of logical anomaly, and we report 95% confidence intervals based on bootstrapped standard errors. Appendix Figure A2 reports the same estimates, organized by calibrated parameter values (δ, γ) that we considered. The median estimated idiosyncratic error rate $\widehat{\epsilon}$ across algorithmically generated logical anomalies is 13.1% for dominated consequence effect anomalies, 8.5% for reverse dominated consequence effect anomalies, and 14.7% for strict dominance effect anomalies. There again exists heterogeneity in these estimates across logical anomalies. For example, explaining the observed choice fractions on several specific logical anomalies across categories would require that respondents erroneously deviate from their true preferences at least 20% of the time.

6 Conclusion

By now, it is clear that machine learning has the capacity to change the way nearly every economic sector operates (e.g., Brynjolfsson and McAfee, 2014; Agarwal, Gans and Goldfarb, 2018). Why should economic research be any different? Of course, substantial progress has already been made in incorporating machine learning into many of the tasks performed by economic researchers, such as digitizing historical archives (e.g., Shen et al., 2021), processing novel data such as text and images for econometric analysis (e.g., Glaeser et al.,

2018; Gentzkow, Kelly and Taddy, 2019; Adukia et al., 2021), uncovering treatment effect heterogeneity (Athey and Wager, 2018; Chernozhukov et al., 2018) and empirical hypothesis generation (Ludwig and Mullainathan, 2023).

While machine learning algorithms can find predictive signals that researchers may fail to notice themselves, they are notoriously opaque black-boxes. How then can we use these predictive algorithms to improve economic theories? In this paper, we argue that anomalies provide a familiar solution to this novel problem. We develop two procedures to automatically generate anomalies for an existing theory from predictive algorithms. The resulting anomalies are minimal examples on which the theory cannot explain the black box's predictions. These algorithmically generated anomalies are natural places to search for possible inconsistencies between our theory and nature. Researchers can then collect further data on these anomalies and suggest improvements to existing theories based on them. While our illustration is specific to expected utility theory, our procedures are general and can be applied wherever there exists a formal theory and rich data that the theory seeks to explain. By extracting theoretical insights from machine learning algorithms, the automatic generation of anomalies can accelerate the development of new economic theories.

References

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2021. "What We Teach About Race and Gender: Representation in Images and Text of Children's Books." National Bureau of Economic Research Working Paper 29123.

Afriat, S. N. 1967. "The Construction of Utility Functions from Expenditure Data." *International Economic Review*, 8(1): 67–77.

Afriat, S. N. 1973. "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method." *International Economic Review*, 14(2): 460–472.

Agarwal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence.* Harvard Business Review Press.

Akhtar, Naveed, and Ajmal Mian. 2018. "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey." *IEEE Access*, 6: 14410–14430.

Allais, Maurice. 1953. "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine." *Econometrica*, 21(4): 503–546.

Andrews, Isaiah, Drew Fudenberg, Annie Liang, and Chaofeng Wu. 2022. "The Transfer Performance of Economic Models."

Athey, Susan. 2017. "Beyond prediction: Using big data for policy problems." *Science*, 355(6324): 483–485.

Athey, Susan, and Stefan Wager. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113: 1228–1242.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. "Incentives in Experiments: A Theoretical Analysis." *Journal of Political Economy*, 126(4): 1472–1503.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2020. "Incentives in experiments with objective lotteries." *Experimental Economics*, 23(1): 1–29.

Ballinger, T. Parker, and Nathaniel T. Wilcox. 1997. "Decisions, Error and Heterogeneity." *The Economic Journal*, 107(443): 1090–1105.

Barberis, Nicholas, and Ming Huang. 2008. "Stocks as Lotteries: The Implications of Probability Weighting for Security Prices." *American Economic Review*, 98(5): 2066–2100.

Beatty, Timothy K. M., and Ian A. Crawford. 2011. "How Demanding Is the Revealed Preference Approach to Demand?" *The American Economic Review*, 101(6): 2782–2795.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. 2015. "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics*, 186(2): 345–366.

Bernheim, B. Douglas, and Charles Sprenger. 2020. "On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting." *Econometrica*, 88(4): 1363–1409.

Blavatskyy, Pavlo, Andreas Ortmann, and Valentyn Panchenko. 2022. "On the Experimental Robustness of the Allais Paradox." *American Economic Journal: Microeconomics*, 14(1): 143–63.

Blavatskyy, Pavlo, Valentyn Panchenko, and Andreas Ortmann. 2022. "How common is the common-ratio effect?" *Experimental Economics*.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *The Quarterly Journal of Economics*, 127(3): 1243–1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics*, 14(1): 521–544.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, 78(4): 1375–1412.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* Norton & Company.

Bugni, Federico A., Ivan A. Canay, and Xiaoxia Shi. 2015. "Specification Tests for Partially Identified Models Defined by Moment Inequalities." *Journal of Econometrics*, 185(1): 259–282.

Camerer, Colin F. 2019. "Artificial Intelligence and Behavioral Economics." In *The Economics of Artificial Intelligence: An Agenda.* , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 587–608. University of Chicago Press.

Camerer, Colin F., and Richard H. Thaler. 1995. "Anomalies: Ultimatums, Dictators and Manners." *Journal of Economic Perspectives*, 9(2): 209–219.

Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. 2019. "Machine learning and the physical sciences." *Reviews of Modern Physics*, 91(4).

Cattaneo, Matias D., Max H. Farrell, and Yingjie Feng. 2020. "Large sample properties of partitioning-based series estimators." *The Annals of Statistics*, 48(3): 1718 – 1741.

Cerreia-Vioglio, Simone, David Dillenberger, and Pietro Ortoleva. 2015. "Cautious Expected utility and the Certainty Effect." *Econometrica*, 83(2): 693–728.

Cerreia-Vioglio, Simone, David Dillenberger, and Pietro Ortoleva. 2020. "An explicit representation for disappointment aversion and other betweenness preferences." *Theoretical Economics*, 15(4): 1509–1546.

Chen, Xiaohong. 2007. "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*. Vol. 6, , ed. James J. Heckman and Edward E. Leamer, 5549–5632. Elsevier.

Chen, Xiaohong, and Timothy M. Christensen. 2015. "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions." *Journal of Econometrics*, 188(2): 447–465.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." National Bureau of Economic Research Working Paper 24678.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who Is (More) Rational?" *American Economic Review*, 104(6): 1518–50.

Conlisk, John. 1989. "Three Variants on the Allais Example." *The American Economic Review*, 79(3): 392–407.

Davis, Damek, and Dmitriy Drusvyatskiy. 2018. "Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions."

Dembo, Aluma, Shachar Kariv, Matthew Polisson, and John K.-H. Quah. 2021. "Ever Since Allais."

Enke, Benjamin, and Cassidy Shubatt. 2023. "Quantifying Lottery Choice Complexity." National Bureau of Economic Research Working Paper 31677.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty." *The Quarterly Journal of Economics*, 138(4): 2021–2067.

Erev, Ido, Ert Eyal, Ori Plonsky, Doron Cohen, and Oded Cohen. 2017. "From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience." *Psychological Review*, 124(4): 369–409.

Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. 2010. "A choice prediction competition: Choices from experience and from description." *Journal of Behavioral Decision Making*, 23(1): 15–47.

Fischer, Simon, and Ingo Steinwart. 2020. "Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms." *J. Mach. Learn. Res.*, 21(1).

Freund, Yoav, and Robert E. Schapire. 1996. "Game Theory, On-line Prediction and Boosting." 325–332.

Froot, Kenneth A., and Richard H. Thaler. 1990. "Anomalies: Foreign Exchange." *Journal of Economic Perspectives*, 4(3): 179–192.

Fudenberg, Drew, and Annie Liang. 2019. "Predicting and Understanding Initial Play." *American Economic Review*, 109(12): 4112–4141.

Fudenberg, Drew, Annie Liang, Jon Kleinberg, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy*, 130(4): 956–990.

Fudenberg, Drew, Wayne Gao, and Annie Liang. 2020. "How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories."

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature*, 57(3): 535–74.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2018. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *Economic Inquiry*, 56(1): 114–137.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38(1): 129–166.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2018. "Empirical Asset Pricing via Machine Learning." National Bureau of Economic Research Working Paper 25398.

Hagen, Ole. 1979. "Towards a Positive Theory of Preferences under Risk." , ed. Maurice Allais and Ole Hagen, 271–302. Dordrecht:Springer Netherlands.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, 50(4): 1029–1054.

Harless, David W., and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica*, 62(6): 1251–1289.

Hartford, Jason S, James R Wright, and Kevin Leyton-Brown. 2016. "Deep Learning for Predicting Human Strategic Behavior." Vol. 29.

Hey, John D. 2005. "Why We Should Not Be Silent About Noise." *Experimental Economics*, 8(4): 325–345.

Hines, James R., and Richard H. Thaler. 1995. "Anomalies: The Flypaper Effect." *Journal of Economic Perspectives*, 9(4): 217–226.

Hirasawa, Toshihiko, Michihiro Kandori, and Akira Matsushita. 2022. "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies."

Imaizumi, Masaaki. 2023. "Sup-Norm Convergence of Deep Neural Network Estimator for Nonparametric Regression by Adversarial Training."

Jain, Ritesh, and Kirby Nielsen. 2023. "A Systematic Test of the Independence Axiom Near Certainty."

Jin, Chi, Praneeth Netrapalli, and Michael I. Jordan. 2019. "What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?"

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–291.

Kahneman, Daniel, and Amos Tversky. 1984. "Choices, values, and frames." *American Psychologist*, 39(4): 341–350.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1991. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *Journal of Economic Perspectives*, 5(1): 193–206.

Kelly, Bryan T, and Dacheng Xiu. 2023. "Financial Machine Learning." National Bureau of Economic Research Working Paper 31502.

Kitamura, Yuichi, and Jorg Stoye. 2018. "Nonparametric Analysis of Random Utility Models." *Econometrica*, 86(6): 1883–1909.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Kolter, Zico, and Alexander Madry. 2018. *Adversarial Robustness - Theory and Practice.* NeurIPS 2018 Tutorial. https://adversarial-ml-tutorial.org/.

Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao, and Alán Aspuru-Guzik. 2022. "On scientific understanding with artificial intelligence." *Nature Reviews Physics*, 4(12): 761–769.

Lamont, Owen A., and Richard H. Thaler. 2003. "Anomalies: The Law of One Price in Financial Markets." *Journal of Economic Perspectives*, 17(4): 191–202.

Lattimore, Pamela K., Joanna R. Baker, and Ann D. Witte. 1992. "The influence of probability on risky choice: A parametric examination." *Journal of Economic Behavior & Organization*, 17(3): 377–400.

Loewenstein, George, and Drazen Prelec. 1992. "Anomalies in Intertemporal Choice: Evidence and an Interpretation*." *The Quarterly Journal of Economics*, 107(2): 573–597.

Loewenstein, George, and Richard H. Thaler. 1989. "Anomalies: Intertemporal Choice." *The Journal of Economic Perspectives*, 3(4): 181–193.

Loomes, Graham. 2005. "Modelling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data." *Experimental Economics*, 8(4): 301–323.

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Scientific Discovery." NBER Working Paper Series No. 31017.

Machina, Mark J. 1987. "Choice under Uncertainty: Problems Solved and Unsolved." *Journal of Economic Perspectives*, 1(1): 121–154.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. "Towards Deep Learning Models Resistant to Adversarial Attacks."

McFadden, Daniel L. 1984. "Chapter 24 Econometric analysis of qualitative response models." In *Handbook of Econometrics*. Vol. 2, 1395–1457. Elsevier.

McGranaghan, Christina, Kirby Nielsen, Ted O'Donoghue, Jason Somerville, and Charles D. Sprenger. Forthcoming. "Distinguishing Common Ratio Preferences from Common Ratio Effects using Paired Valuation Tasks." *American Economic Review*.

Mullainathan, Sendhi, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *The Journal of Economic Perspectives*, 31(2): 87–106.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics*, 137(2): 679–727.

Newey, Whitney K., and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing." In . Vol. 4 of *Handbook of Econometrics*, 2111–2245. Elsevier.

Oprea, Ryan. 2022. "Simplicity Equivalents."

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. "Using large-scale experiments and machine learning to discover theories of human decision-making." *Science*, 372(6547): 1209–1214.

Peysakhovich, Alexander, and Jeffrey Naecker. 2017. "Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity." *Journal of Economic Behavior & Organization*, 133: 373–384.

Pion-Tonachini, Luca, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W. Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, Benjamin Nachman, Babetta L. Marrone, Nicola Falco, Prabhat, Daniel Arnold, Alejandro Wolf-Yadlin, Sarah Powers, Sharlee Climer, Quinn Jackson, Ty Carlson, Michael Sohn, Petrus Zwart, Neeraj Kumar, Amy Justice, Claire Tomlin, Daniel Jacobson, Gos Micklem, Georgios V. Gkoutos, Peter J. Bickel, Jean-Baptiste Cazier, Juliane Müller, Bobbie-Jo Webb-Robertson, Rick Stevens, Mark Anderson, Ken Kreutz-Delgado, Michael W. Mahoney, and James B. Brown. 2021. "Learning from learning machines: a new generation of AI technology to meet the needs of science."

Polisson, Matthew, John K.-H. Quah, and Ludovic Renou. 2020. "Revealed Preferences over Risk and Uncertainty." *American Economic Review*, 110(6): 1782–1820.

Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica*, 66(3): 497–527.

Puri, Indira. 2022. "Simplicity and Risk."

Raghu, Maithra, and Eric Schmidt. 2020. "A Survey of Deep Learning for Scientific Discovery."

Rambachan, Ashesh. 2022. "Identifying Prediction Mistakes in Observational Data."

Ramsay, J. O. 1988. "Monotone Regression Splines in Action." *Statistical Science*, 3(4): 425–441.

Razaviyayn, Meisam, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. 2020. "Nonconvex Min-Max Optimization: Applications, Challenges, and Recent Theoretical Advances." *IEEE Signal Processing Magazine*, 37(5): 55–66.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press.

Sargan, J. D. 1958. "The Estimation of Economic Relationships using Instrumental Variables." *Econometrica*, 26(3): 393–415.

Selten, Reinhard. 1991. "Properties of a measure of predictive success." *Mathematical Social Sciences*, 21(2): 153–167.

Selten, Reinhard, and Wilhelm Krischker. 1983. "Comparison of Two Theories for Characteristic Function Experiments." In *Aspiration Levels in Bargaining and Economic Decision Making.* , ed. Reinhard Tietz, 259–264. Springer.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis."

Slovic, Paul, and Amos Tversky. 1974. "Who accepts Savage's axiom?" *Behavioral Science*, 19(6): 368–373.

Slovic, Paul, and Sarah Lichtenstein. 1983. "Preference Reversals: A Broader Perspective." *American Economic Review*, 73(4): 596–605.

Strzalecki, Tomasz. 2022. *Stochastic Choice Theory.*

Sunstein, Cass R. 2022. "Governing by Algorithm? No Noise and (Potentially) Less Bias." *Duke Law Journal*, 71: 1175–1205.

Thaler, Richard H. 1988. "Anomalies: The Winner's Curse." *The Journal of Economic Perspectives*, 2(1): 191–202.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science*, 211(4481): 453–458.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *The Quarterly Journal of Economics*, 106(4): 1039–1061.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5(4): 297–323.

Tversky, Amos, and Richard H. Thaler. 1990. "Anomalies: Preference Reversals." *Journal of Economic Perspectives*, 4(2): 201–211.

Varian, Hal R. 1982. "The Nonparametric Approach to Demand Analysis." *Econometrica*, 50(4): 945–973.

Wang, Hanchen, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. 2023. "Scientific discovery in the age of artificial intelligence." *Nature*, 620(7972): 47–60.

Wright, James, and Kevin Leyton-Brown. 2010. "Beyond Equilibrium: Predicting Human Behavior in Normal-Form Games." *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1): 901–907.

Wright, James R., and Kevin Leyton-Brown. 2017. "Predicting human behavior in unrepeated, simultaneous-move games." *Games and Economic Behavior*, 106: 16–37.

Wu, George, and Richard Gonzalez. 1996. "Curvature of the Probability Weighting Function." *Management Science*, 42(12): 1676–1690.

Yang, Yun, Anirban Bhattacharya, and Debdeep Pati. 2017. "Frequentist coverage and sup-norm convergence rate in Gaussian process regression."

A Omitted proofs

A.1 Proof of Proposition 2.1

To prove part (i), we first note that the main text established that the allowable function representation (1) satisfies Assumptions 1-4. This establishes necessity. We prove sufficiency here. Consider any theory $T(\cdot)$ satisfying Assumptions 1-4. We construct an allowable function representation $\mathcal{F}^T$ satisfying (1).

Towards this, define $\mathcal{D}^{\neg T}$ to be the set of incompatible collections of examples for theory $T(\cdot)$. That is, $D \in \mathcal{D}^{\neg T}$ if and only if $T(x; D) = \varnothing$ for all $x \in \mathcal{X}$. By Assumption 4, there exists some $D \in \mathcal{D}$ such that $T(x; D) \subset \mathcal{Y}^*$ for some $x \notin D$. We can therefore define $D' = D \cup \{(x, \tilde{y}^*)\}$ for any $\tilde{y}^* \in \mathcal{Y}^* \setminus T(x; D)$. By construction, $T(x; D') = \varnothing$ for all $x \in D'$ since otherwise $T(\cdot)$ would violate Assumption 3. $\mathcal{D}^{\neg T}$ is therefore non-empty.

We next define $\mathcal{F}^{\neg T}$ to be the set of mappings $f(\cdot) \in \mathcal{F}$ that are consistent with $\mathcal{D}^{\neg T}$. That is, $f(\cdot) \in \mathcal{F}^{\neg T}$ if and only if $f(\cdot)$ is consistent with some $D \in \mathcal{D}^{\neg T}$. Finally, we define the allowable functions of $T(\cdot)$ as $\mathcal{F}^T = \mathcal{F} \setminus \mathcal{F}^{\neg T}$. We will next show that

$$T(x; D) = \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} \tag{17}$$

is satisfied for all $D \in \mathcal{D}$ and $x \in \mathcal{X}$.

By Assumptions 1-2, there are only two cases to consider. First, consider $D \in \mathcal{D}$ such that $T(x; D) = \varnothing$ for all $x \in \mathcal{X}$. By construction, $\{f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} = \varnothing$ since D is incompatible with $T(\cdot)$. We therefore focus on the second case in which $D \in \mathcal{D}$ satisfies $T(x; D) = y^*$ for all $(x, y^*) \in D$ and $T(x; D) \neq \varnothing$ for all $x \notin D$.

Observe that $\{f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} \neq \varnothing$ by construction. It therefore follows that $\{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} = y^*$ for all $(x, y^*) \in D$. All that remains to show is that $\{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} = T(x; D)$ for all $x \notin D$. As notation, for correspondence $c(\cdot) \colon \mathcal{X} \rightrightarrows \mathcal{Y}^*$ and mapping $f(\cdot) \colon \mathcal{X} \to \mathcal{Y}^*$, we write $f(\cdot) \in c(\cdot)$ if and only if $f(x) \in c(x)$ for all $x \in \mathcal{X}$.

Lemma 1. *For any $D \in \mathcal{D}$ such that $T(x; D) \neq \varnothing$ for all $x \in \mathcal{X}$, $f(\cdot) \in T(\cdot; D)$ implies that $f(\cdot) \in \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\}$.*

Proof. Suppose for sake of contradiction there exists some $f(\cdot) \in T(\cdot; D)$ such that $f(\cdot) \notin \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$. Since D is not incompatible with $T(\cdot)$, $D \notin \mathcal{D}^{\neg T}$ and therefore $f(\cdot) \notin \mathcal{F}^{\neg T}$ by construction. But this then implies that $f(\cdot) \in \mathcal{F}^T$, generating the desired contradiction. □

Lemma 2. *For any $D \in \mathcal{D}$ such that $T(x; D) \neq \varnothing$ for all $x \in \mathcal{X}$, $f(\cdot) \in \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\}$ implies $f(\cdot) \in T(\cdot; D)$.*

Proof. To prove this result, we will prove the contrapositive: $f(\cdot) \notin T(\cdot; D)$ implies $f(\cdot) \notin \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$.

Suppose for sake of contradiction there exists some $f(\cdot) \notin T(\cdot; D)$ with $f(\cdot) \in \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$. Since any $f(\cdot)$ that is not consistent with D cannot be an element of $\{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$ by construction, we focus on the case in $f(x) = y^*$ for all $(x, y^*) \in D$.

Pick any $x \in \mathcal{X}$ with $f(x) \notin T(x; D)$. Since D is consistent with $f(\cdot)$, define $D' = D \cup \{(x, f(x))\}$ and consider $T(\cdot; D')$. There are only two cases to consider by Assumption 2. First, if $T(\cdot; D') = \varnothing$, then D' is incompatible with $T(\cdot)$ and $f(\cdot) \notin \mathcal{F}^T$ by construction. This yields a contradiction. Second, if $T(\cdot; D') \neq \varnothing$, then $T(x; D') = f(x)$ by Assumption 2. But this then contradicts Assumption 3 since $T(x; D') \not\subseteq T(x; D)$. □

Lemma 1 implies $T(x; D) \subseteq \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$ for all $x \in \mathcal{X}$. Lemma 2 establishes that $\{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} \subseteq T(x; D)$. It therefore follows that $T(x; D) = \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\}$, and this proves the result. This proves part (i). To prove part (ii), consider $D \in \mathcal{D}$ such that $T(x; D) \subset \mathcal{Y}^*$ for some $x \notin D$ which must exist by Assumption 4. Define $D' = D \cup \{(x, \tilde{y}^*)\}$ for any $\tilde{y}^* \in \mathcal{Y}^* \setminus T(x; D)$. By construction, this is an incompatible with $T(\cdot)$. Since there exists incompatible collections, there must exist a smallest incompatible collection $D \in \mathcal{D}$ for theory $T(\cdot)$. This must be a logical anomaly. If $|D| = 1$, then the definitions of an incompatible collection and a logical anomaly coincide. If $|D| > 1$ but $|D|$ is not a logical anomaly, then there exists a smaller incompatible collection which is a contradiction. □.

A.2 Proof of Proposition 2.2

Part (i) is an immediate consequence of the allowable function representation in Proposition 2.1. First, suppose D is incompatible with theory $T(\cdot)$ and $T(x; D) = \varnothing$ for all $x \in \mathcal{X}$. Proposition 2.1 implies that there exists no $f(\cdot) \in \mathcal{F}^T$ consistent with D. It immediately follows that $\min_{f(\cdot) \in \mathcal{F}^T} |D|^{-1} \sum_{(x,y^*) \in D} \ell\left(f(x), y^*\right) > 0$. Next, suppose $\min_{f(\cdot) \in \mathcal{F}^T} |D|^{-1} \sum_{(x,y^*) \in D} \ell\left(f(x), y^*\right) > 0$. This implies that there exists no $f(\cdot) \in \mathcal{F}^T$ consistent with D, and so D must be incompatible by Proposition 2.1.

Part (ii) is an immediate consequence of Definition 3. If there exists no incompatible collection of size strictly less than n, any incompatible collection of size n must also be a logical anomaly as it must be the case that $D \setminus \{(x, y^*)\}$ is compatible with theory $T(\cdot)$ for all $(x, y^*) \in D$. □

A.3 Proof of Proposition 3.1

As a first step, we establish that the $\widehat{\mathcal{E}}_n$ approximately solves the plug-in max-min optimization program up to the optimization errors associated with the approximate inner minimization and outer maximization routines.

Lemma 3. *Under the same conditions as Proposition 3.1,*

$$\left\| \widehat{\mathcal{E}}_m - \max_{x_{1:n}} \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}_m^*(x_i)\right) \right\| \leq \delta + \nu.$$

Proof. As notation, let $\widehat{f}^T(\cdot; x_{1:n})$ denote the optimal solution to $\min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}_m^*(x_i)\right)$. Observe that

$$\left\| n^{-1} \sum_{i=1}^{n} \ell\left(\widetilde{f}(\widetilde{x}_i; \widetilde{x}_{1:n}), \widehat{f}_m^*(\widetilde{x}_i)\right) - \max_{x_{1:n}} \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}_m^*(x_i)\right) \right\| \overset{(1)}{\leq}$$

$$\left\| n^{-1} \sum_{i=1}^{n} \ell\left(\widetilde{f}(\widetilde{x}_i; \widetilde{x}_{1:n}), \widehat{f}^*_m(\widetilde{x}_i)\right) - \max_{x_{1:n}} n^{-1} \sum_{i=1}^{n} \ell\left(\widehat{f}^T(\cdot; x_{1:n}), \widehat{f}^*_m(x_i)\right) \right\| +$$

$$\left\| \max_{x_{1:n}} n^{-1} \sum_{i=1}^{n} \ell\left(\widehat{f}^T(\cdot; x_{1:n}), \widehat{f}^*_m(x_i)\right) - \max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) \right\| \overset{(2)}{\leq}$$

$$\nu + \left\| \max_{x_{1:n}} n^{-1} \sum_{i=1}^{n} \ell\left(\widehat{f}^T(\cdot; x_{1:n}), \widehat{f}^*_m(x_i)\right) - \max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) \right\| \overset{(3)}{\leq}$$

$$\nu + \left\| \max_{x_{1:n}} \left\{ n^{-1} \sum_{i=1}^{n} \ell\left(\widehat{f}^T(\cdot; x_{1:n}), \widehat{f}^*_m(x_i)\right) - \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) \right\} \right\| \overset{(4)}{\leq} \nu + \delta$$

where (1) adds/subtracts $\max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right)$ and applies the triangle inequality, (2) follows from properties of the approximate outer maximization routine, (3) uses sub-additivity of the maximum, and (4) follows from the properties of the approximate inner minimization routine. □

To analyze the convergence of the plug-in estimator, observe that

$$\left\| \widehat{\mathcal{E}}_m - \mathcal{E}_m \right\| \leq \left\| \widehat{\mathcal{E}}_m - \max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) \right\| + \left\| \max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) - \mathcal{E}_m \right\|.$$

Lemma 3 establishes that the first term is bounded by $\nu + \delta$. Therefore, we only need to establish a bound on the second term. Towards this, we rewrite the second term as

$$\left\| \max_{x_{1:n}} \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) - \mathcal{E}_m \right\| \leq$$

$$\left\| \max_{x_{1:n}} \left\{ \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) - \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), f^*_m(x_i)\right) \right\} \right\|.$$

Defining $\widehat{f}^T(\cdot; x_{1:n})$ to be the minimizer for $\min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right)$ and $f^T(\cdot; x_{1:n})$ as the minimizer for $\min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), f^*_m(x_i)\right)$, we rewrite

$$\min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right) - \min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), f^*_m(x_i)\right) =$$

$$n^{-1} \sum_{i=1}^{n} \ell\left(\widehat{f}^T(x_i; x_{1:n}), \widehat{f}^*_m(x_i)\right) - n^{-1} \sum_{i=1}^{n} \ell\left(f^T(x_i; x_{1:n}), f^*_m(x_i)\right) =$$

$$\underbrace{n^{-1} \sum_{i=1}^{n} \left\{ \ell\left(\widehat{f}^T(x_i; x_{1:n}), \widehat{f}^*_m(x_i)\right) - \ell\left(\widehat{f}^T(x_i; x_{1:n}), f^*_m(x_i)\right) \right\}}_{(a)} +$$

$$\underbrace{n^{-1}\sum_{i=1}^{n}\left\{\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(f^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)\right\}}_{(b)}.$$

Consider (a). Since $\ell(\cdot,\cdot)$ is convex in its second argument, (a) is bounded above by

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)\left(\widehat{f}_m^*(x_i)-f_m^*(x_i)\right)\right\}\leq$$

$$n^{-1}K\|\widehat{f}_m^*(x_{1:n})-f_m^*(x_{1:n})\|_1\leq K\|\widehat{f}_m^*(x_{1:n})-f_m^*(x_{1:n})\|_\infty$$

where we defined the shorthand notation $f(x_{1:n})=(f(x_1),\ldots,f(x_n))$, used that the loss function has bounded gradients, and the inequality $\|f(x_{1:n})\|_1\leq n\|f(x_{1:n})\|_\infty$. Next, we can rewrite (b) as being bounded by

$$n^{-1}\sum_{i=1}^{n}\left\{\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(f^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)\right\}=$$

$$n^{-1}\sum_{i=1}^{n}\left\{\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(f^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)\right\}-$$

$$n^{-1}\sum_{i=1}^{n}\left\{\ell\left(f^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)\right\}\overset{(1)}{\leq}$$

$$n^{-1}\sum_{i=1}^{n}\left\{\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)\right\}-$$

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell\left(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)(f_m^*(x_i)-\widehat{f}_m^*(x_i))\right\}\overset{(2)}{\leq}$$

$$n^{-1}\sum_{i=1}^{n}\left\{\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),f_m^*(x_i)\right)-\ell\left(\widehat{f}^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)\right\}-$$

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell\left(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)(f_m^*(x_i)-\widehat{f}_m^*(x_i))\right\}\overset{(3)}{\leq}$$

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell(\widehat{f}^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i))\left(\widehat{f}_m^*(x_i)-f_m^*(x_i)\right)\right\}-$$

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell\left(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i)\right)(f_m^*(x_i)-\widehat{f}_m^*(x_i))\right\}$$

where (1) uses that the loss is convex in its second argument, (2) uses $n^{-1}\sum_{i=1}^{n}\ell(f^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i))\geq n^{-1}\sum_{i=1}^{n}\ell(\widehat{f}^{T}(x_i;x_{1:n}),\widehat{f}_m^*(x_i))$, and (3) again uses that the loss is convex in it second argu-

ment. By the same argument as before, it follows that this is bounded by

$$\leq 2K \left\| \widehat{f}_m^*(x_{1:n}) - f_m^*(x_{1:n}) \right\|_\infty .$$

Combining the bound on (a), (b) yields the desired result. □

A.4 Proof of Proposition 4.1

To prove this result, we first observe that if $f(x_1) = f(x_2)$ for all $f(\cdot) \in \mathcal{F}^T$, then $T(x_1; D) = T(x_2; D)$ must be true for all $D \in \mathcal{D}$ by Proposition 2.1. Next suppose, for sake of contradiction, that there exist two features x_1, x_2 that are representationally equivalent but there exists some allowable function $f(\cdot) \in \mathcal{F}^T$ such that $f(x_1) \neq f(x_2)$. Consider $D = \{(x_1, f(x_1)), (x_2, f(x_2))\}$. Since $f(\cdot) \in \mathcal{F}^T$, $T(\cdot)$ must be consistent with D. Furthermore, by Assumption 2 ("consistency"), $T(\cdot)$ must also satisfy that $T(x_1; D) = f(x_1)$ and $T(x_2; D) = f(x_2)$, yielding the desired contradiction. □

A.5 Proof of Proposition 4.2

We first observe that Assumption 6 implies Assumption 4 and therefore there exists an allowable function representation $\mathcal{F}^T$ for theory $T(\cdot)$. Then, we will show that the pair $x_1, x_2 \in \mathcal{X}$ in Assumption 6 are representationally equivalent. There are three cases to consider. First, if $D \in \mathcal{D}$ is incompatible with $T(\cdot)$, then $T(x_1; D) = T(x_2; D) = \varnothing$. Second, if $D \in \mathcal{D}$ is such that $(x_j, y_j^*) \in D$ for $j \neq k$, the $T(x_k; D) = y_j^*$ by Assumption 6. Finally, suppose for sake of contradiction $x_1, x_2 \notin D$ but $T(x_1; D) \neq T(x_2; D)$. If there exists some $y_1^* \in T(x_1; D)$ with $y_1^* \notin T(x_2; D)$, construct the collection $\widetilde{D} = D \cup \{(x_1, y_1^*)\}$. By the allowable function representation (1), $\widetilde{D}$ is a compatible collection. But Assumption 3 implies that $y_1^* \notin T(x_2; \widetilde{D})$, contradicting Assumption 6. □

A.6 Proof of Proposition 4.3

To prove the first result, let us define the shorthand notation $g^* = \nabla f_m^*(x)$, $g = \text{Proj}\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right)$, and $g^\perp = g^* - g$. Observe that

$$\left\langle -\text{Proj}\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right), \nabla f_m^*(x) \right\rangle = \langle -g, g^* \rangle = \langle -g, g^\perp + g \rangle = -\|g\|^2 \leq 0,$$

and so $-\text{Proj}\left(\nabla f_m^*(x) \mid \mathcal{N}(x)\right)$ is a descent direction for $f_m^*(\cdot)$.

To prove the second result, let Ω to be the orthogonal projection matrix onto $\mathcal{N}(x)$ and define $\widehat{g}^* = \nabla \widehat{f}_m^*(x)$, $\widehat{g} = \text{Proj}\left(\nabla \widehat{f}_m^*(x) \mid \mathcal{N}(x)\right)$ and $\widehat{g}^\perp = \widehat{g}^* - \widehat{g}$. Observe that

$$\left\langle -\text{Proj}\left(\nabla \widehat{f}_m^*(x) \mid \mathcal{N}(x)\right), \nabla f_m^*(x) \right\rangle = \langle -\widehat{g}, g^* \rangle = \langle -\widehat{g}, g + g^\perp \rangle = \langle -\widehat{g}, g \rangle =$$

$$\langle -\widehat{g} + g - g, g \rangle = -\|g\|^2 + \langle g - \widehat{g}, g \rangle \leq -\|g\|^2 + \|g - \widehat{g}\|\|g\|,$$

where the last inequality follows by the Cauchy-Schwarz inequality. The stated condition implies that

$$\|g - \widehat{g}\| \leq \|g\|$$

since $\|g - \widehat{g}\| = \|\Omega(g^* - \widehat{g}^*)\| \leq \|\Omega\|_{op}\|g^* - \widehat{g}^*\|$ and $\|\Omega\|_{op} \leq 1$. But the previous display can

be equivalently rewritten as

$$-\|g\|^2 + \|g - \widehat{g}\|\|g\| \leq 0$$

thus proving the result. $\square$

B Appendix figures and tables

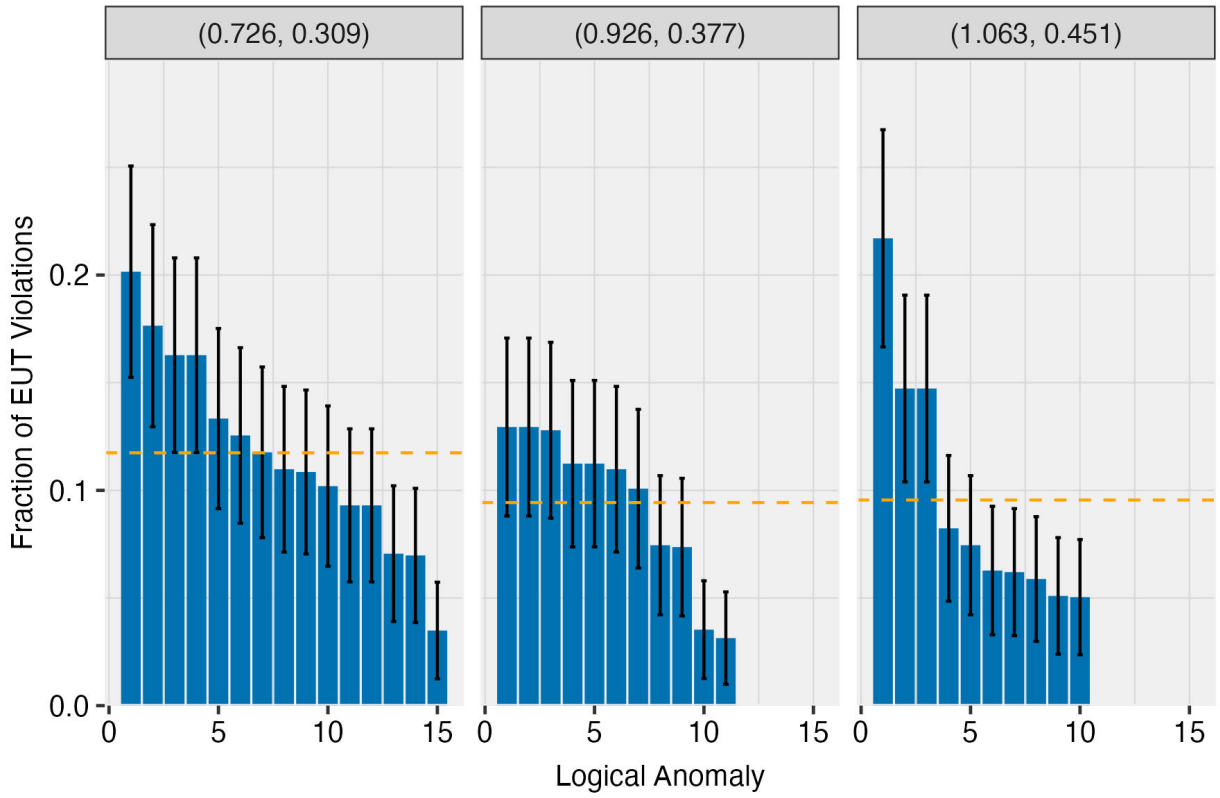


Figure A1: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, organized by calibrated parameter values (δ, γ).

Notes: This figure summarizes the fraction of respondents whose choices violate expected utility theory on the logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors clustered at the respondent level). We organize the estimates by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the fraction of respondents whose choices violate expected utility theory pooling across all logical anomalies within the same parameter values. Within the same parameter values, we sort the logical anomalies and assign each anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

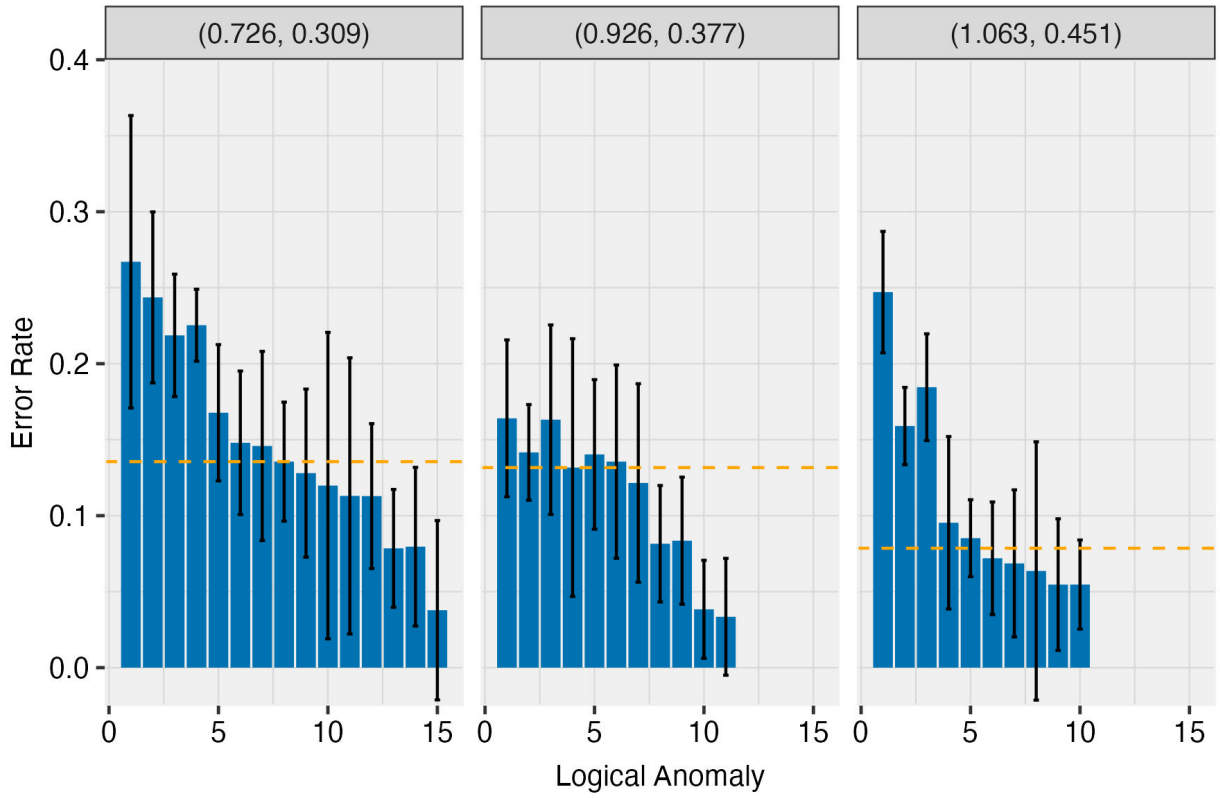


Figure A2: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated logical anomalies, organized by calibrated parameter values (δ, γ).

Notes: This figure summarizes the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on our algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors computed by the bootstrap). We organize the estimates by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the median estimated idiosyncratic error rate across all logical anomalies within the same parameter values. Within the same parameter values, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

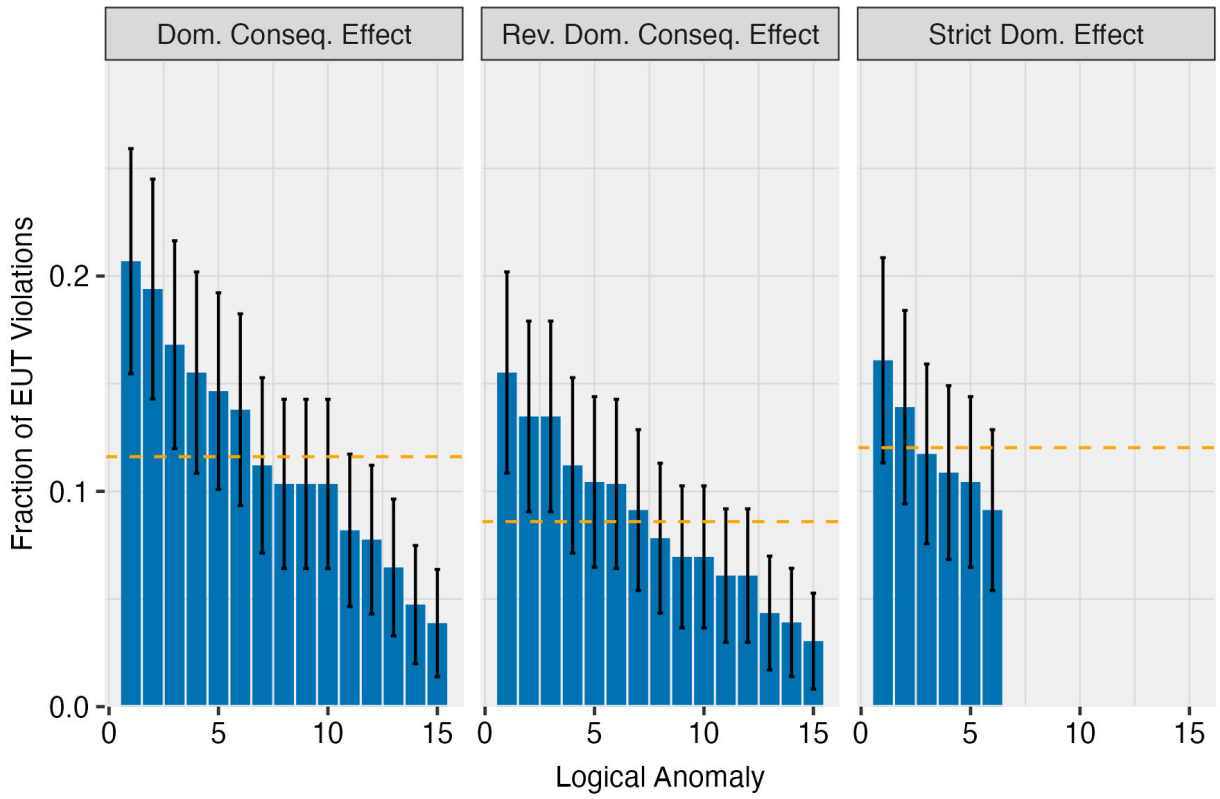


Figure A3: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, dropping the top 10% of respondents who completed the survey the fastest.

Notes: This figure summarizes the fraction of respondents whose choices violate expected utility theory on the logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors clustered at the respondent level), dropping the top 10% of respondents who completed the survey the fastest. We organize the estimates by category of logical anomaly (see Table 3). The orange dashed line reports the fraction of respondents whose choices violate expected utility theory pooling across all logical anomalies within the same category. Within each category, we sort the logical anomalies and assign each anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

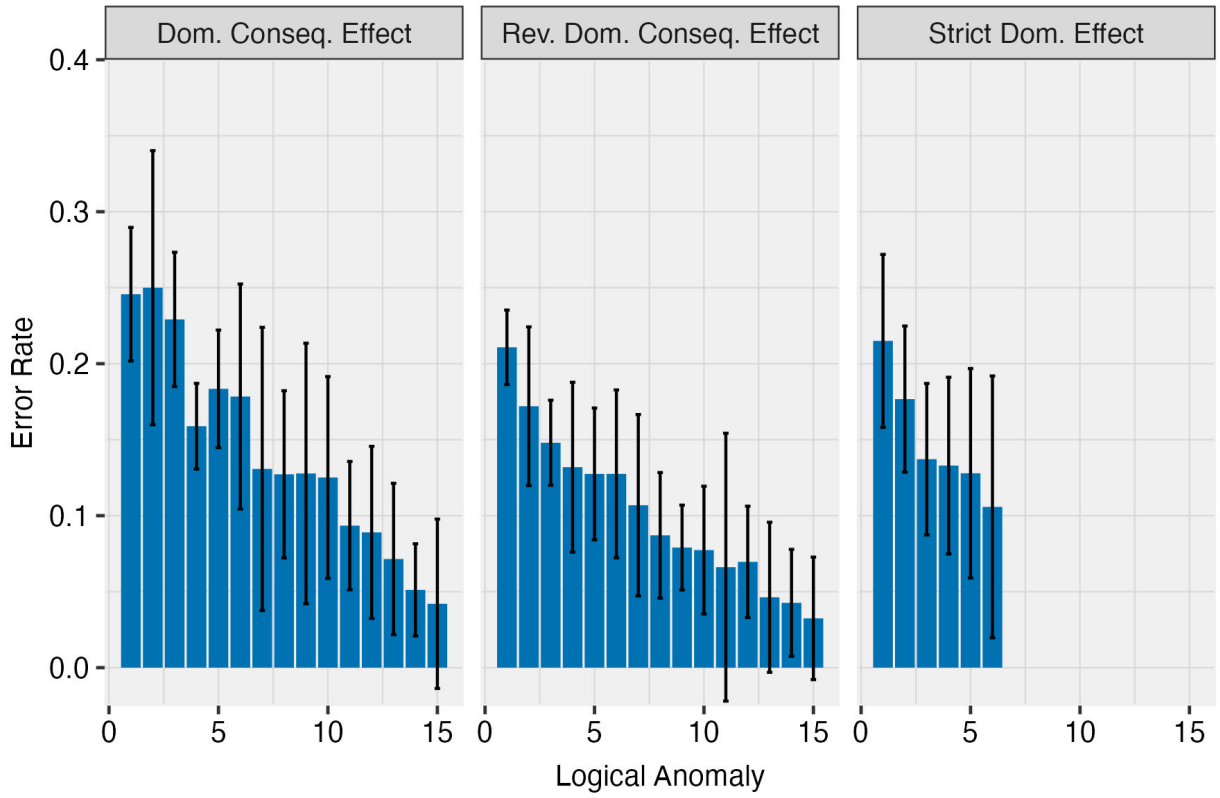


Figure A4: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated logical anomalies, dropping the 10% of respondents that completed the survey the fastest.

Notes: This figure summarizes the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on our algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors computed by the bootstrap), dropping the top 10% of respondents who completed the survey the fastest. We organize the estimates by category of logical anomaly (see Table 3). The orange dashed line reports the median estimated idiosyncratic error rate across all logical anomalies within the same category. Within each category, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

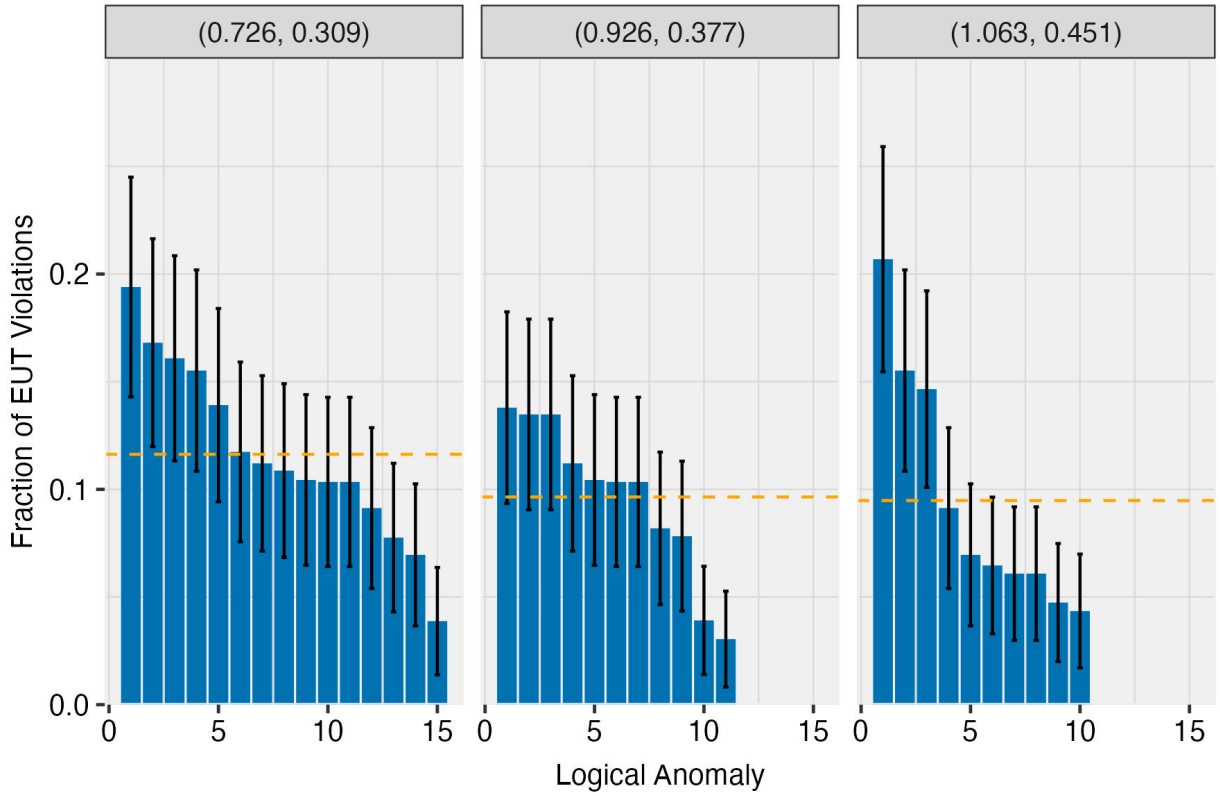


Figure A5: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, organized by the calibrated parameter values (δ, γ) and dropping the top 10% of respondents who completed the survey the fastest.

Notes: This figure summarizes the fraction of respondents whose choices violate expected utility theory on the logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors clustered at the respondent level), dropping the top 10% of respondents who completed the survey the fastest. We organize the estimates by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the fraction of respondents whose choices violate expected utility theory pooling across all logical anomalies within the same parameter values. Within the same parameter values, we sort the logical anomalies and assign each anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

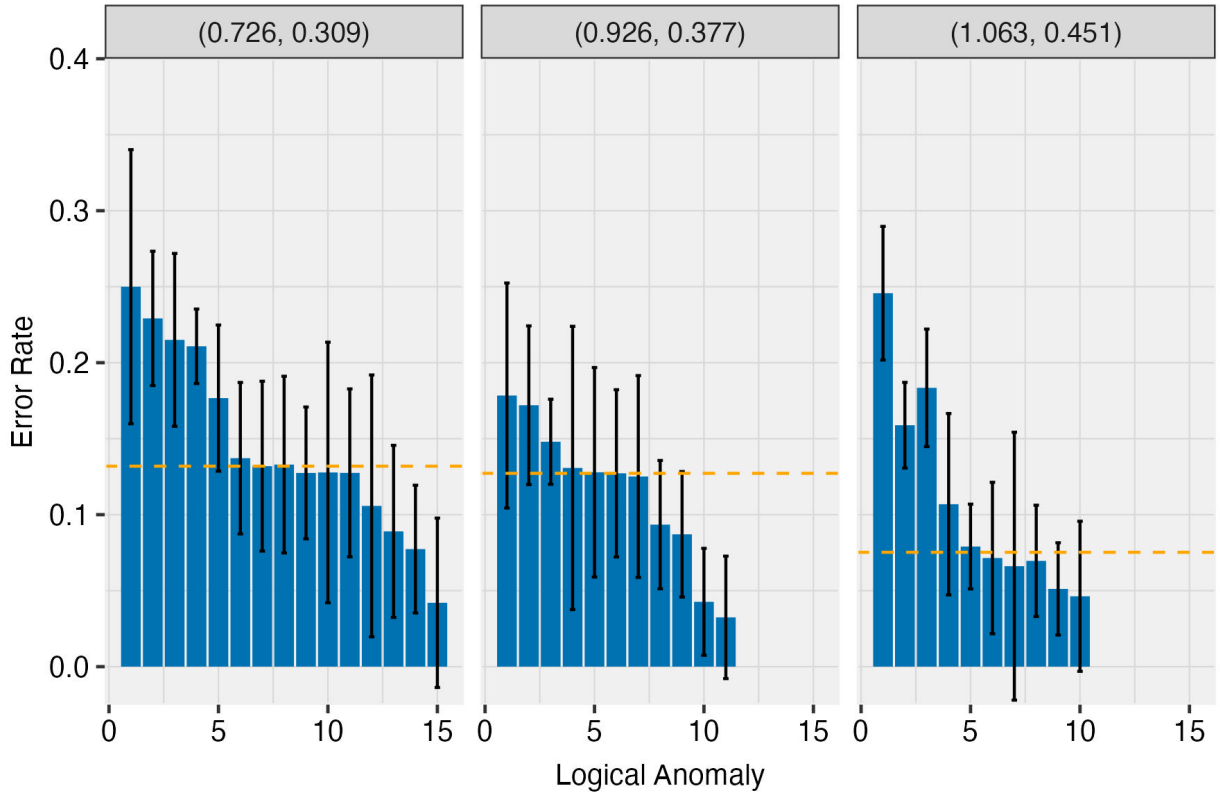


Figure A6: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated logical anomalies, organized by calibrated parameter values (δ, γ) and dropping the 10% of respondents that completed the survey the fastest.

Notes: This figure summarizes the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on our algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors computed by the bootstrap), dropping the top 10% of respondents who completed the survey the fastest. We organize the estimates by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the median estimated idiosyncratic error rate across all logical anomalies within the same parameter values. Within the same parameter values, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Section 5.3 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)	
Lottery 0	1.10 15%	7.48 85%
Lottery 1	1.50 1%	5.94 99%
	Menu B (x_B, y_B^*)	
Lottery 0	1.10 45%	7.48 55%
Lottery 1	1.50 18%	5.94 82%

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)	
Lottery 0	0.08 34%	9.26 66%
Lottery 1	0.76 0%	5.54 100%
	Menu B (x_B, y_B^*)	
Lottery 0	0.08 63%	9.26 37%
Lottery 1	0.76 13%	5.54 87%

(c) Logical Anomaly #3

	Menu A (x_A, y_A^*)	
Lottery 0	2.52 39%	7.64 61%
Lottery 1	3.10 21%	5.78 79%
	Menu B (x_B, y_B^*)	
Lottery 0	2.52 7%	7.64 93%
Lottery 1	3.10 0%	5.78 100%

(d) Logical Anomaly #4

	Menu A (x_A, y_A^*)	
Lottery 0	6.17 9%	7.60 91%
Lottery 1	5.72 27%	8.61 73
	Menu B (x_B, y_B^*)	
Lottery 0	6.17 0%	7.60 100%
Lottery 1	5.72 0%	8.61 92%

(e) Logical Anomaly #5

	Menu A (x_A, y_A^*)	
Lottery 0	1.74 10%	5.21 90%
Lottery 1	1.83 7%	4.71 93%
	Menu B (x_B, y_B^*)	
Lottery 0	0.70 2%	5.96 98%
Lottery 1	0.23 0%	7.48 100%

(f) Logical Anomaly #6

	Menu A (x_A, y_A^*)	
Lottery 0	1.98 48%	9.21 52%
Lottery 1	2.49 34%	7.69 66%
	Menu B (x_B, y_B^*)	
Lottery 0	1.98 5%	9.21 95%
Lottery 1	2.49 0%	7.69 100%

Table A1: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the dominated consequence effect.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each logical anomaly exhibiting the dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. All payoffs are denominated in dollars. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)	
Lottery 0	4.44 100%	7.76 0%
Lottery 1	3.65 95%	7.83 5%
	Menu B (x_B, y_B^*)	
Lottery 0	4.44 36%	7.76 64%
Lottery 1	3.65 23%	7.83 77%

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)	
Lottery 0	1.36 100%	5.91 0%
Lottery 1	0.05 0.93%	6.05 7%
	Menu B (x_B, y_B^*)	
Lottery 0	1.36 68%	5.91 32%
Lottery 1	0.05 56%	6.05 44%

(c) Logical Anomaly #3

	Menu A (x_A, y_A^*)	
Lottery 0	2.23 62%	7.69 38%
Lottery 1	0.75 38%	7.77 62%
	Menu B (x_B, y_B^*)	
Lottery 0	2.23 99%	7.69 1%
Lottery 1	0.75 83%	7.77 17%

(d) Logical Anomaly #4

	Menu A (x_A, y_A^*)	
Lottery 0	3.02 80%	8.12 20%
Lottery 1	0.29 49%	9.43 51%
	Menu B (x_B, y_B^*)	
Lottery 0	3.02 100%	8.12 0%
Lottery 1	0.29 79%	9.43 21%

(e) Logical Anomaly #5

	Menu A (x_A, y_A^*)	
Lottery 0	0.84 51%	9.88 49%
Lottery 1	3.32 76%	9.25 24%
	Menu B (x_B, y_B^*)	
Lottery 0	0.84 85%	9.88 15%
Lottery 1	3.32 100%	9.25 0%

(f) Logical Anomaly #6

	Menu A (x_A, y_A^*)	
Lottery 0	0.93 18%	6.82 82%
Lottery 1	2.02 28%	6.78 72%
	Menu B (x_B, y_B^*)	
Lottery 0	0.93 95%	6.82 5%
Lottery 1	2.02 100%	6.78 0%

Table A2: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the reverse dominated consequence effect.

Notes: In each menu, we color the lottery in the menu that is selected by the individual with probability at least 0.50 in green. Each logical anomaly exhibiting the reverse dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. All payoffs are denominated in dollars. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)	
Lottery 0	6.28	6.91
	65%	35%
Lottery 1	5.94	7.77
	53%	47%
	Menu B (x_B, y_B^*)	
Lottery 0	6.28	6.91
	100%	0%
Lottery 1	5.94	7.77
	24%	76%

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)	
Lottery 0	3.93	7.26
	39%	61%
Lottery 1	5.02	5.71
	100%	0%
	Menu B (x_B, y_B^*)	
Lottery 0	3.93	7.26
	41%	59%
Lottery 1	5.02	5.71
	98%	2%

(c) Logical Anomaly #3

	Menu A (x_A, y_A^*)	
Lottery 0	3.63	4.32
	61% 39%	
Lottery 1	3.72	4.21
	51%	49%
	Menu B (x_B, y_B^*)	
Lottery 0	3.63	4.32
	100%	0%
Lottery 1	3.72	4.21
	13%	87%

(d) Logical Anomaly #4

	Menu A (x_A, y_A^*)	
Lottery 0	6.89	8.24
	64%	36%
Lottery 1	7.01	7.18
	46%	54%
	Menu B (x_B, y_B^*)	
Lottery 0	6.89	8.24
	34%	66%
Lottery 1	7.01	7.18
	100%	0%

(e) Logical Anomaly #5

	Menu A (x_A, y_A^*)	
Lottery 0	6.03	6.31
	41%	59%
Lottery 1	3.49	8.99
	35%	65%
	Menu B (x_B, y_B^*)	
Lottery 0	6.03	6.31
	100%	00%
Lottery 1	3.49	8.99
	34%	66%

(f) Logical Anomaly #6

	Menu A (x_A, y_A^*)	
Lottery 0	6.33	6.51
	42%	58%
Lottery 1	6.31	6.61
	52%	48%
	Menu B (x_B, y_B^*)	
Lottery 0	6.33	6.51
	100%	00%
Lottery 1	6.31	6.61
	24%	76%

Table A3: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the strict dominance effect.

Notes: We color the lottery in the menu that is selected by the individual with probability at least 0.50 in green. Each logical anomaly exhibiting the strict dominance effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. All payoffs are denominated in dollars. For simplicity, we round each payoff to the nearest cent and each probability to the nearest percentage. See Section 5.2 and Appendix F.3 for further discussion.

(a) Logical Anomaly #1

Lottery 0	5.72	6.19
	19%	81%
Lottery 1	5.26	
	100%	

(b) Logical Anomaly #2

Lottery 0	8.17	
	100%	
Lottery 1	9.03	9.70
	23%	77%

(c) Logical Anomaly #3

Lottery 0	7.97	
	100%	
Lottery 1	8.85	9.88
	59%	41%

(d) Logical Anomaly #4

Lottery 0	7.20	7.61
	33%	67%
Lottery 1	6.99	7.50
	98%	2%

(e) Logical Anomaly #5

Lottery 0	8.07	9.05
	21%	79%
Lottery 1	7.84	
	100%	

(f) Logical Anomaly #6

Lottery 0	6.89	8.88
	46%	54%
Lottery 1	6.87	
	100%	

(g) Logical Anomaly #7

Lottery 0	6.30	6.85
	18%	82%
Lottery 1	6.09	
	100%	

(h) Logical Anomaly #8

Lottery 0	4.90	
	100%	
Lottery 1	5.04	5.27
	31%	69%

(i) Logical Anomaly #9

Lottery 0	7.67	
	100%	
Lottery 1	8.31	8.57
	43%	57%

(j) Logical Anomaly #10

Lottery 0	7.74	9.55
	95%	5%
Lottery 1	7.64	
	100%	

(k) Logical Anomaly #11

Lottery 0	7.11	9.50
	99%	1%
Lottery 1	7.02	
	100%	

(l) Logical Anomaly #12

Lottery 0	5.640	
	100%	
Lottery 1	5.84	7.35
	50%	50%

Table A4: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate first-order stochastic dominance violations.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each generated first-order stochastic dominance violation presented here (x, y^*) is based on the probability weighting function $\pi(p; \delta, \gamma)$ with $(\delta, \gamma) = (0.726, 0.309)$. All payoffs are denominated in dollars. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage point. See Section 5.2 for further discussion.

Prob. Weighting Function: (δ, γ)	Pooled Average	Median	First Quartile	Third Quartile
$(0.726, 0.309)$	0.117 (0.005)	0.109	0.093	0.148
$(0.926, 0.377)$	0.094 (0.006)	0.109	0.074	0.120
$(1.063, 0.451)$	0.095 (0.006)	0.068	0.059	0.131

Table A5: Summary statistics on the fraction of respondents whose choices violate expected utility theory on algorithmically generated logical anomalies, organized by calibrated parameter values (δ, γ).

Notes: This table reports summary statistics on the fraction of respondents whose choices violate expected utility theory ("expected utility theory violation rate") on algorithmically generated logical anomalies of menus of two lotteries over two monetary payoffs. We report summary statistics by calibrated parameter values of probability weighting function (δ, γ) (see Table 3). The "pooled average" column reports the expected utility theory violation rate, pooling together respondents' choices on all logical anomalies within the same category. Standard errors reported in parentheses are clustered at the respondent level. We also report the median, first quartile, and third quartile of the distribution of expected utility theory violation rates across logical anomalies within the same category. See Section 5.3 for further discussion.

	Pooled Average	Median	First Quartile	Third Quartile
Blavatskyy, Ortmann and Panchenko (2022)				
Allais paradox: Fan-Out	0.160	0.137	0.087	0.184
Allais paradox: Fan-In	0.194	0.173	0.093	0.244
Blavatskyy, Panchenko and Ortmann (2022)				
Common ratio effect	0.268	0.256	0.129	0.366
Reverse common ratio effect	0.099	0.085	0.043	0.153
McGranaghan et al. (Forthcoming)				
Common ratio effect	0.155	0.133	0.095	0.190
Reverse common ratio effect	0.128	0.113	0.0782	0.179

Table A6: Summary statistics on the fraction of respondents whose choices violate expected utility theory on the Allais paradox and Common ratio effect in recent meta-analyses and comprehensive experiments.

Notes: This table reports summary statistics on the fraction of respondents whose choices violate expected utility theory on logical anomalies like the Allais paradox and Common ratio effect in recent large-scale meta-analyses and comprehensive experiments. The summary statistics for Blavatskyy, Ortmann and Panchenko (2022) are based on the experiments included in their meta-analysis of the Allais paradox as reported in their Table 1. The summary statistics for Blavatskyy, Panchenko and Ortmann (2022) are based on the experiments included in their meta-analysis of the Common ratio effect as reported in their Table 1. The summary statistics for McGranaghan et al. (Forthcoming) are based on the choice experiments reported in their Table D.11 and Table D.12.

C Additional Examples for the Model of Theories

In this appendix section, we illustrate how additional examples map into our framework described in Section 2 of the main text.

Example: choice under risk As in the main text, consider individuals evaluating a lottery over $J > 1$ monetary payoffs. The features are a complete description of the lottery $x = (z, p)$, where $z \in \mathbb{R}^J$ is the lottery's payoffs and $p \in \Delta^{J-1}$ is the lottery's probabilities. The modeled outcome is now the certainty equivalent $y^* \in \mathbb{R}$ for the lottery (e.g., Tversky and Kahneman, 1992; Bruhin, Fehr-Duda and Epper, 2010; Bernheim and Sprenger, 2020; Fudenberg et al., 2022; Andrews et al., 2022, among many others), and the modeled contexts $m \in \mathcal{M}$ are each individual. Given D, expected utility theory searches for utility functions $u(\cdot)$ that rationalize the certainty equivalents of the lotteries, meaning $y^* = u^{-1}\left(\sum_{j=1}^{J} p(j)u(z(j))\right)$ for all $(x, y^*) \in D$. On any new lottery, expected utility theory returns $T(x; D)$, where $y^* \in T(x; D)$ if and only if $y^* = u^{-1}\left(\sum_{j=1}^{J} p(j)u(z(j))\right)$ for some utility function $u(\cdot)$ rationalizing D. Alternative behavioral models such as cumulative prospect theory can be cast as particular theories $T(\cdot)$ of certainty equivalents. ▲

Example: multi-attribute discrete choice Consider individuals making choices from menus of J items (e.g., McFadden, 1984; Strzalecki, 2022). The features are a complete description of each item in the menu $x = (z_1, p_1, \ldots, z_J, p_J)$, where z_j are the attributes of item j and p_j is its price. The features may even include information about how items are presented in the menu or their ordering. The modeled outcomes are menu choice probabilities $y^* \in \Delta^{J-1}$, and the modeled contexts $m \in \mathcal{M}$ may either be interpreted as individuals or distinct groups of individuals (e.g., see the discussion in Ch. 1 of Strzalecki, 2022).

A popular class of parametric additive random utility models, such as the multinomial logit, specify the indirect utility of item j as $v_j(x; \alpha, \beta) = z_j\beta - \alpha p_j + \epsilon_j$, where (α, β) are parameters and ϵ_j is a random taste shock with some known distribution. Given D of menus and choice probabilities, such a parametric additive random utility model searches for parameter values (α, β) that match the choice probabilities, meaning $y^*_j = P\left(j \in \arg\max_{\widehat{j}} v_{\widehat{j}}(x; \alpha, \beta)\right)$ for all $j = 1, \ldots, J$ and $(x, y^*) \in D$. On any new menu of items x, it returns $T(x; D)$, where $y^* \in T(x; D)$ if and only if $y^*_j = P\left(j \in \arg\max_{\widehat{j}} v_{\widehat{j}}(x; \alpha, \beta)\right)$ for some (α, β) that matches D. ▲

C.1 Logical anomalies for other examples

Example: play in normal-form games Consider the normal-form game in Table A7. In our framework, such a normal-form game is a particular feature $x \in \mathcal{X}$. The iterated elimination of strictly dominated strategies implies that $(Top, Left)$ is the unique Nash equilibrium of the game. Therefore, $T(x; D) = \varnothing$ or $T(x; D) = (1, 0, 0)$ for any $D \in \mathcal{D}$. Suppose instead the individual m was a level-1 thinker. In this case, she would eliminate Bottom since it is strictly dominated but would fail to recognize the Right is now strictly dominated for her opponent by the iterated elimination of strictly dominated strategies. She would then play the game as if her opponent randomizes across all of her actions, and we

	Left	Center	Right
Top	(10, 4)	(5,3)	(3,2)
Middle	(0,1)	(4,6)	(6,0)
Bottom	(2,1)	(3,5)	(2,8)

Table A7: An example anomaly for Nash equilibrium based on Level-1 thinking.

may observe her strategy profile y^* placing positive probability on both Top and Middle. By construction, a collection of examples that consisted of only this normal-form game and such a strategy profile would be a logical anomaly for Nash equilibrium (Definition 3). ▲

Example: asset pricing As mentioned in the main text, CAPM models the expected return of an asset as $\bar{y}_{\text{risk-free}} + \beta\left(\bar{y}_{\text{market}} - \bar{y}_{\text{risk-free}}\right)$ based on the expected returns of all assets and their covariance structure. Consider the example $\{(x, y^*)\}$, where there exists some asset that does not satisfy the asset pricing equation. By construction, this is a logical anomaly for CAPM (Definition 3). For example, Barberis and Huang (2008) find that skew (i.e., a higher moment) affects asset returns in the cross-section. ▲

C.2 Assumptions for other examples

Example: play in normal-form games Nash equilibrium is the correspondence $T(\cdot)$ satisfying: (i) if for all $(x, y^*) \in D$ there exists some $y^*_{col} \in \Delta^{J-1}$ such that $\sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} y^*(j)y^*_{col}(\tilde{j})\pi_{row}(j,\tilde{j}) \geq \sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} \tilde{y}^*(j)y^*_{col}(\tilde{j})\pi_{row}(j,\tilde{j})$ for all $\tilde{y}^* \in \Delta^{J-1}$, then $T(x; D)$ is defined as in the main text for all $x \in \mathcal{X}$; (ii) otherwise, $T(x; D) = \varnothing$ for all $x \in \mathcal{X}$. We immediately observe that Assumption 1, Assumption 2, and Assumption 4 are satisfied by construction. Assumption 3 is also satisfied as $T(x; D') \subseteq T(x; D)$ for all D, D' with $D \subseteq D'$. ▲

Example: asset pricing We observe that CAPM as described in the main text immediately satisfies Assumption 1 and Assumption 2 on examples consisting of moments of historical asset prices. Second, consider any D, D' satisfying $D \subseteq D'$. There are only three cases to consider – either both D, D' are incompatible with CAPM, D is compatible with CAPM but D' is not, and both are compatible with CAPM in which case $\beta(D) = \beta(D')$. In all such cases, Assumption 3 is satisfied. Finally, Assumption 4 is satisfied for any D that either point or partially identifies the assets' parameter β_j.

More specifically, CAPM provides a procedure for calculating the expected market return $\bar{y}_{market}$, risk-free rate $\bar{y}_{risk\text{-}free}$, and the asset's covariance with the market return β from any feature x_1 consisting of the expected returns of all assets and higher moments. As a result, the allowable functions of CAPM can be written as $f(x_1) = \bar{y}_{risk\text{-}free} + \beta(\bar{y}_{market} - \bar{y}_{risk\text{-}free})$. For any other feature x_2 that leads to the same expected market return, risk-free rate and asset's covariance with the market return, we have that $f(x_1) = f(x_2)$. CAPM therefore satisfies Assumption 6. Any pair of features x_1, x_2 of this form are representationally equivalent under CAPM. ▲

D Average Anomalies across Modeled Contexts

In the main text, our anomaly generation procedures focused on searching for anomalies in a single modeled context, whereas we may be empirically interested in generating anomalies that hold across many modeled contexts $m \in \mathcal{M}$. Our algorithmic procedures can be directly applied across modeled contexts.

D.1 Adversarial algorithm

Suppose we observe a random sample $(M_i, X_i, Y_i) \sim P(\cdot)$ for $i = 1, \ldots, N$ across modeled contexts. Under this joint distribution, define $\bar{f}^*(x) := \mathbb{E}[g(Y_i) \mid X_i = x]$ as the average relationship between features and the modeled outcome across all modeled contexts. Define $P(m \mid x) := P(M_i = m \mid X_i = x)$ and $f_m^*(x) := \mathbb{E}_m[g(Y_i) \mid X_i = x]$ in each modeled context $m \in \mathcal{M}$ as before. An *average* incompatible collection of examples is a collection of features $x_{1:n}$ such that $D = \{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is incompatible with theory $T(\cdot)$. An *average* empirical anomaly is defined analogously.

If $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is an average incompatible collection, then it is also an incompatible collection in some modeled context m. Provided $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is a "systematically" incompatible collection across modeled contexts, then it is also an average incompatible collection.

Proposition D.1. *Suppose theory $T(\cdot)$ satisfies Assumptions 1-4. Then,*

i. If $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is an average incompatible collection, then there exists some modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot)$ such that $\{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$ is an incompatible collection.

ii. Provided $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is incompatible in some modeled context and satisfies

$$\sum_{m \neq \tilde{m}} \left(n^{-1} \sum_{i=1}^{n} P(m \mid x) P(\tilde{m} \mid x) \left(f_m^T(x_i) - f_m^*(x_i) \right) \left(f_{\tilde{m}}^T(x_i) - f_{\tilde{m}}^*(x_i) \right) \right) \geq 0,$$

for all $f_m(\cdot), f_{\tilde{m}}(\cdot) \in \mathcal{F}^T$, then $x_{1:n}$ is also an average incompatible collection.

Proof. To prove this result, it suffices to focus on the squared loss function $\ell(y, y') = (y - y')^2$. To show (i), we define $\bar{f}^T(x; x_{1:n}) := \sum_{m \in \mathcal{M}} P(m \mid x) f_m^T(x; x_{1:n})$. We then observe that

$$n^{-1} \sum_{i=1}^{n} \left(\bar{f}^T(x_i; x_{1:n}) - \bar{f}^*(x_i) \right) = n^{-1} \sum_{i=1}^{n} \left(\sum_{m \in \mathcal{M}} P(m \mid x_i) \left(f_m^T(x_i; x_{1:n}) - f_m^*(x_i) \right) \right)^2$$

$$\leq 2n^{-1} \sum_{i=1}^{n} \sum_{m \in \mathcal{M}} P(m \mid x_i)^2 \left(f_m^T(x_i; x_{1:n}) - f_m^*(x_i) \right)^2 \leq 2 \sum_{m \in \mathcal{M}} \left(n^{-1} \sum_{i=1}^{n} P(m \mid x_i) \left(f_m^T(x_i; x_{1:n}) - f_m^*(x_i) \right)^2 \right)$$

Then, since $x_{1:n}$ is an average incompatible collection, this implies

$$0 < \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^{n} (f(x_i) - \bar{f}^*(x_i))^2 \leq 2 \sum_{m \in \mathcal{M}} \left(n^{-1} \sum_{i=1}^{n} P(m \mid x_i) \left(f_m^T(x_i; x_{1:n}) - f_m^*(x_i) \right)^2 \right),$$

which in turn implies that $n^{-1}\sum_{i=1}^{n} P(m \mid x_i)\left(f_m^T(x_i; x_{1:n}) - f_m^*(x_i)\right)^2 > 0$ for some modeled context $m \in \mathcal{M}$. To show (ii), observe that

$$\min_{f(\cdot)\in\mathcal{F}^T} n^{-1}\sum_{i=1}^{n}\left(f(x_i) - \bar{f}^*(x_i)\right)^2 \geq \min_{f_m(\cdot)\in\mathcal{F}^T} n^{-1}\sum_{i=1}^{n}\left(\sum_{m\in\mathcal{M}} P(m \mid x_i)(f_m(x_i) - f_m^*(x_i))\right)^2,$$

where

$$n^{-1}\sum_{i=1}^{n}\left(\sum_{m\in\mathcal{M}} P(m \mid x_i)(f_m(x_i) - f_m^*(x_i))\right)^2 =$$

$$n^{-1}\sum_{m\in\mathcal{M}}\sum_{i=1}^{n} P(m \mid x_i)^2(f_m(x_i) - f_m^*(x_i))^2 +$$

$$n^{-1}\sum_{m\neq\tilde{m}}\sum_{i=1}^{n} P(m \mid x_i)P(\tilde{m} \mid x_i)(f_m(x_i) - f_m^*(x_i))(f_{\tilde{m}}(x_i) - f_{\tilde{m}}(x_i)).$$

Then, under the assumption that $x_{1:n}$ is systematically incompatible with theory $T(\cdot)$ across modeled contexts, it follows that

$$\min_{f(\cdot)\in\mathcal{F}^T} n^{-1}\sum_{i=1}^{n}\left(f(x_i) - \bar{f}^*(x_i)\right)^2 \geq \sum_{m\in\mathcal{M}}\left\{\sum_{i=1}^{n} P(m \mid x_i)^2(f_m(x_i) - f_m^*(x_i))^2\right\}.$$

The result then follows as $x_{1:n}$ is also an incompatible collection for some modeled context m. □

The condition in Proposition D.1(ii) requires that $x_{1:n}$ be "systematically" incompatible with theory $T(\cdot)$ across modeled contexts in the sense that the errors of the theory's best fitting allowable functions across modeled contexts do not cancel out on average.

Proposition D.1 suggests that we can search for empirical anomalies across modeled contexts by plugging an estimator $\widehat{\bar{f}}^*(\cdot)$ into our adversarial search procedure. Our same theoretical analysis applies, except now the difference between the plug-in optimal value and the population optimal value now depends on the estimation error $\|\widehat{\bar{f}}^*(\cdot) - \bar{f}^*(\cdot)\|_\infty$. By pooling data across modeled contexts, we may hope to obtain better control of this estimation error in finite samples.

D.2 Average representational anomalies across modeled contexts

In Section 4, we developed an example morphing procedure to generate representational anomlaies in a single modeled context. We may also be interested in generating representational anomalies across many modeled contexts $m \in \mathcal{M}$.

Suppose we again observe a random sample $(M_i, X_i, Y_i) \sim P(\cdot)$ for $i = 1, \ldots, N$ across modeled contexts, letting $\bar{f}^*(x) := \mathbb{E}[g(Y_i) \mid X_i = x]$ and $P(m \mid x) = P(M_i = m \mid X_i = x)$ as before. We define an empirical *average* representational anomaly as a pair of features x_1, x_2 such that $\bar{f}^*(x_1) \neq \bar{f}^*(x_2)$. If there are no compositional changes in modeled contexts across these features, then x_1, x_2 is an empirical average representational anomaly if and only if it is an empirical representational anomaly in some modeled context m.

Proposition D.2. *Consider features $x_1, x_2 \in \mathcal{X}$ and suppose $P(m \mid x_1) = P(m \mid x_2)$ for all $m \in \mathcal{M}$. Then,if $\{(x_1, \bar{f}^*(x_1)), (x_2, \bar{f}^*(x_2))\}$ is an average representational anomaly, then there exists some modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot)$ such that $\{(x_1, f_m^*(x_1), (x_2, f_m^*(x_2))\}$ is a representational anomaly.*

Proof. To prove this result, observe that

$$\bar{f}^*(x_1) - \bar{f}^*(x_2) = \sum_{m \in \mathcal{M}} P(m \mid x_1) f_m^*(x_1) - \sum_{m \in \mathcal{M}} P(m \mid x_2) f_m^*(x_2)$$

$$= \sum_{m \in \mathcal{M}} P(m \mid x_1) \left(f_m^*(x_1) - f_m^*(x_2)\right) + \sum_{m \in \mathcal{M}} \left(P(m \mid x_1) - P(m \mid x_2)\right) f_m(x_2).$$

Assuming that $P(m \mid x_1) = P(m \mid x_2)$ for all $m \in \mathcal{M}$ implies that the second term in the previous display equals zero. The result is then immediate. □

The condition in Proposition D.2 requires that there exists the same composition of modeled context across features x_1, x_2. If not, there could exist variation in $\bar{f}^*(\cdot)$ across these features even though there exists no empirical representational anomaly in any modeled context.

Proposition D.2 suggests that we can search for empirical average representational anomalies across modeled contexts by simply plugging an estimator $\widehat{\bar{f}}^*(\cdot)$ into our morphing procedure. Our same theoretical analysis applies, except now the difference between the plug-in gradient and the population gradient depends on the error $\|\nabla \widehat{\bar{f}}^*(x) - \nabla \bar{f}^*(x)\|_2$. By pooling data across modeled contexts, we may hope to obtain better control of this estimation error.

E Analysis of Gradient Descent Ascent Optimization over Allowable Functions

In Section 3.2 of the main text, we proposed a gradient descent ascent (GDA) procedure to optimize the plug-in max-min program (6). Recall that for some parametrization of the theory's allowable functions $\mathcal{F}^T = \{f_\theta(\cdot) \colon \theta \in \Theta\}$, initial feature values $x_{1:n}^0$, maximum number of iterations $S > 0$, and step size sequence $\{\eta_s\}_{s=0}^S > 0$, we iterate over $s = 0, \ldots, S$ and calculate

$$\theta^{s+1} = \arg\min_{\theta \in \Theta} \widehat{\mathcal{E}}_m(x_{1:n}^s; \theta)$$
$$x_{1:n}^{s+1} = x_{1:n}^t + \eta_t \nabla \widehat{\mathcal{E}}_m(x_{1:n}^s; \theta^{s+1})$$

at each iteration, where $\widehat{\mathcal{E}}_m(x_{1:n}, \theta) := n^{-1} \sum_{i=1}^n \ell\left(f_\theta(x_i), \widehat{f}_m^*(x_i)\right)$. We apply recent results from Jin, Netrapalli and Jordan (2019) on non-convex/concave max-min optimization to establish that this GDA procedure converges to an approximate stationary point of the outer maximization problem

Define $\bar{x}_{1:n}$ to be the random variable drawn uniformly over $\{x_{1:n}^0, \ldots, x_{1:n}^S\}$ and define $\widehat{\mathcal{E}}_m(x_{1:n}) = \min_{\theta \in \Theta} \widehat{\mathcal{E}}_m(x_{1:n}, \theta)$. To formally state the result, we define the *Moreau envelope* of $\widehat{\mathcal{E}}_m(x_{1:n})$ as

$$\phi_\lambda(x_{1:n}) = \min_{x'_{1:n}} \widehat{\mathcal{E}}_m(x'_{1:n}) + \frac{1}{2\lambda} \|x_{1:n} - x'_{1:n}\|_2^2$$

For non-convex functions, the Moreau envelope is a smooth, convex approximation that is often used to analyze the properties of gradient descent algorithms (e.g, see Davis and Drusvyatskiy, 2018). Our analysis of the GDA procedure provides a bound on the gradient of the Moreau envelope $\phi_\lambda(\cdot)$. Standard results in convex optimization establish that a bound on the gradient of the Moreau envelope implies a bound on the subdifferentials of $\widehat{\mathcal{E}}_m(x_{1:n})$.

Lemma 4 (Lemma 30 in Jin, Netrapalli and Jordan (2019))**.** *Suppose $\widehat{\mathcal{E}}_m(x_{1:n})$ is b-weakly convex. For an $\lambda < \frac{1}{b}$ and $\widetilde{x}_{1:n} = \arg\min_{x'_{1:n}} \widehat{\mathcal{E}}_m(x'_{1:n}) + \frac{1}{2\lambda}\|x_{1:n} - x'_{1:n}\|_2^2$, $\|\nabla\phi_\lambda(x_{1:n})\| \leq \epsilon$ implies*

$$\|\widetilde{x}_{1:n} - x_{1:n}\| = \lambda\epsilon \text{ and } \min_{g \in \partial\widehat{\mathcal{E}}_m(\widetilde{x}_{1:n})} \|g\| \leq \epsilon,$$

where ∂ denotes the subdifferential of a weakly convex function.

Proposition E.1. *Suppose $\ell(\cdot,\cdot), \widehat{f}_m^*(\cdot)$ and $\{f_\theta(\cdot)\colon \theta \in \Theta\}$ are k-times continuously differentiable with K-bounded gradients. Then, the output $\bar{x}_{1:n}$ of the gradient descent ascent algorithm with step size sequence given by $\eta_s = \eta_0/\sqrt{S+1}$ for some $\eta_0 > 0$ satisfies*

$$\mathbb{E}\left[\|\nabla\phi_{0.5b}(\bar{x}_{1:n})\|_2^2\right] \leq 2\frac{\left(\phi_{0.5b}(x^0_{1:n}) - \min_{x_{1:n}} \widehat{\mathcal{E}}_m(x_{1:n})\right) + bK^2\eta_0^2}{\eta_0\sqrt{S+1}} + 4b\delta,$$

where $\delta \geq 0$ is the error associated with the approximate inner minimization oracle in Assumption 5(i).

Proof. This result is an immediate consequence of Theorem 31 in Jin, Netrapalli and Jordan (2019). □

F Additional Implementation Details and Results for Choice under Risk with Lotteries over Two Monetary Payoffs

F.1 Implementation details of anomaly generation procedures

In this section, we describe the implementation details of our anomaly generation procedures in the illustration to choice under risk in Section 5.1.

For both the adversarial procedure and example morphing procedure, we constructed randomly initialized menus of two independent lotteries in the following manner. We simulated each payoff in a lottery independently from a uniform distribution on $[0, 10]$. We simulated the probabilities in a lottery by drawing each lottery probability uniformly from the unit interval, and then normalizing the draws so they lie on the unit simplex.

F.1.1 Adversarial procedure

To implement the adversarial procedure based on gradient descent ascent described in Section 3.2, we must first specify a parametric basis for the allowable functions of expected utility theory. We parametrize the utility function of the individual $u_\theta(\cdot)$ as a linear combination of polynomials up to order K or I-splines with some number of knot points q and degree

K (see Ramsay, 1988). We experimented with both choices of basis functions, varying the maximal degree of the polynomial bases as well as the number of knot points and degree of the I-spline bases. Since we found qualitatively similar results, we focus on presenting anomalies generated by a polynomial utility function basis with order $K = 6$. We set the learning rate to be $\eta = 0.01$.

For any particular choice of utility function basis and learning rate, we ran the gradient descent ascent procedure for 25,000 randomly initialized menus x^0. We set the maximum number of iterations to be $S = 50$. For a particular choice of utility basis functions, we solve the inner minimization problem (12) by minimizing the cross-entropy loss between the true choice probabilities on the menus $f_m^*(x^s)$ and the implied expected utility theory choice probabilities $f_\theta(x^s) = P\left(\sum_{j=1}^J p_{1j}^s u_\theta(z_{1j}) - \sum_{j=1}^J p_{0j}^s u_\theta(z_{0j}) + \xi\right)$ for ξ an i.i.d. logistic shock. We then implement the outer gradient ascent step (13) directly. After each gradient ascent step, we project the updated lottery probability vectors back into the unit simplex.

A subtle numerical issue arises as the gradients of the cross-entropy loss $\widehat{\mathcal{E}}(x^s; \theta^{s+1})$ vanish whenever expected utility theory can exactly match the choice probabilities. To avoid this vanishing gradients problem, we instead implement the outer gradient ascent step (13) by following the gradient of $\log\left(\frac{f_m^*(x^s)}{1-f_m^*(x^s)}\right)\left(\sum_{j=1}^J p_{1j}^s u_{\theta^{s+1}}(z_{1j}) - \sum_{j=1}^J p_{0j} u_{\theta^{s+1}}(z_{0j})\right)$. This alternative loss function for the gradient ascent step applies the logit transformation to the choice probabilities so that $\log\left(\frac{f_m^*(x^s)}{1-f_m^*(x^s)}\right)$ is positive whenever $f_m^*(x^s) > 0.5$ and weakly negative otherwise. The overall loss function is therefore positive whenever the expected utility difference between the lotteries is positive but $f_m^*(x^s) < 0.5$ and vice versa. We take gradient ascent steps on only the probabilities of the lotteries in the menu, meaning that only the probabilities of the lotteries in the menu are modified over the gradient descent ascent algorithm. We then collect together the anomalies produced across all runs of the adversarial procedure.

F.1.2 Example morphing procedure

To implement the example morphing procedure described in Algorithm 2, we again must specify a parametric basis for the allowable functions of expected utility theory. Like the adversarial procedure, we experimented with both polynomial bases up to order K and I-spline bases varying the number of knot points q and degree K. Since we found qualitatively similar results, we focus on presenting anomalies generated by the I-spline basis with $q = 10$ knot points and degree $K = 3$. We set the learning rate to be $\eta = 10$.

For any particular utility function basis and learning rate, we ran the example morphing procedure 15,000 randomly initialized menus x^0. We set the maximum number of iterations to be $S = 50$. At each iteration s, we solve for the best-fitting allowable function θ^s and sample θ_b independent from a multivariate normal distribution with mean vector equal to $\bar{\theta}^s = \frac{1}{s}\sum_{s'=1}^s \theta^{s'}$ and variance matrix equal to $\frac{1}{s-1}\sum_{s'=1}^s (\theta^{s'} - \bar{\theta}^s)(\theta^{s'} - \bar{\theta}^s)'$ for $b = 1, \ldots, B$. We set $B = 200,000$. We take gradient ascent steps on only the probabilities of the lotteries in the menu, meaning that only the probabilities of the lotteries in the menu are modified by the dataset morphing procedure. We then collect together the anomalies produced across all runs of the dataset morphing procedure.

F.2 Numerical verification of logical anomalies for expected utility theory

As discussed in Section 5.1 of the main text, each returned menu of lotteries over two monetary payoffs by our anomaly generation procedures are logical anomalies for expected utility theory at our particular parametrization of the utility function $\{u_\theta(\cdot)\colon \theta \in \Theta\}$. Given any such returned menus of lotteries over two monetary payoffs, we numerically verify whether the dataset of returned menus is a logical anomaly for expected utility theory at any increasing utility function and without noisy choices. In the main text, we report all resulting numerically verified logical anomalies for expected utility theory at any increasing utility function.

Concretely, consider a modeled dataset $\{(x_0, f_m^*(x_0)), (x_1, f_m^*(x_1))\}$ returned by our anomaly generation procedures, where $x_0 = (z_{0,0}, p_{0,0}, z_{0,1}, p_{0,1})$ and $x_1 = (z_{0,0}, p_{1,0}, z_{0,1}, p_{1,1})$. For ease of exposition, we assume the monetary payoffs are the same across the two menus. Define $y_0^* = 1\{f_m^*(x_0) \geq 0.50\}$ and $y_1^* = 1\{f_m^*(x_1) \geq 0.50\}$, and the ordered monetary payoffs as

$$z_{(1)} < z_{(2)} < z_{(3)} < z_{(4)}.$$

We check whether there exists any increasing utility function $u(z)$ satisfying $u(z_{(1)}) < u(z_{(2)}) < u(z_{(3)}) < u(z_{(4)})$ that could rationalize the given configuration of binary choices (y_0^*, y_1^*). Abusing notation, let us redefine $p_{0,0} \in \Delta^4$ as the vector of probabilities associated with the ordered monetary payoffs, and $p_{0,1}, p_{1,0}, p_{1,1}$ analogously. Let $u = (u_1, u_2, u_3, u_4)$ denote the vector of utility values assigned to the ordered monetary payoffs. Checking whether there exists any increasing utility function that could rationalize the given configuration of binary choices is equivalent to checking whether there exists a solution to a system of linear inequalities. In particular, if (i) $y_0^* = y_1^* = 0$, we check whether there exists any vector u satisfying $(p_{0,0} - p_{0,1})'u > 0$ and $(p_{1,0} - p_{1,1})'u > 0$; (ii) $y_0^* = 1, y_1^* = 0$, we check whether there exists any vector u satisfying $(p_{0,0} - p_{0,1})'u < 0$ and $(p_{1,0} - p_{1,1})'u > 0$; and so on.

F.3 Proofs of logical anomalies for expected utility theory

In this section, we prove that pairs of menus of two lotteries over two monetary payoffs exhibiting the dominated consequence effect, reverse dominated consequence effect, and strict dominance effect are logical anomalies for expected utility theory.

F.3.1 Dominated consequence effect anomalies

Consider the first menu defined over the pair of lotteries $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ and the second menu defined over the lotteries $\ell_0' = (p_0', z_0)$, $\ell_1' = (p_1', z_1)$. Let $\underline{z}_0 = \min_j z_0(j)$ and $\underline{z}_1 = \min_j z_1(j)$.

Suppose that the lotteries in the second menu can be written as

$$\ell_0' = \alpha_0 \ell_0 + (1 - \alpha_0)\delta_{\underline{z}_0}$$
$$\ell_1' = \alpha_1 \ell_1 + (1 - \alpha_1)\delta_{\underline{z}_1}$$

for some $\alpha_0, \alpha_1 \in [0, 1]$. Further assume (i) $\underline{z}_0 < \underline{z}_1$, (ii) ℓ_1 is preferred to ℓ_0 – that is, $\ell_1 \succ \ell_0$, (iii) ℓ_0' is preferred to ℓ_1' – that is, $\ell_0' \succ \ell_1'$, and (iv) $\alpha_1 \geq \alpha_0$. To see why this is a logical

anomaly for expected utility theory, observe

$$\begin{aligned}
&\ell_1 \succ \ell_0 \overset{(1)}{\Longrightarrow} \alpha_1\ell_1 + (1-\alpha_1)\delta_{\underline{z}_1} \succ \alpha_1\ell_0 + (1-\alpha_1)\delta_{\underline{z}_1},\\
&\alpha_1\ell_0 + (1-\alpha_1)\delta_{\underline{z}_1} \overset{(2)}{\succ} \alpha_1\ell_0 + (1-\alpha_1)\delta_{\underline{z}_0}\\
&\alpha_1\ell_0 + (1-\alpha_1)\delta_{\underline{z}_0} \overset{(3)}{\succ} \alpha_0\ell_0 + (1-\alpha_0)\delta_{\underline{z}_0}
\end{aligned}$$

where (1) follows by the independence axiom, (2) follows by utility must be increasing in monetary payoffs and the independence axiom, and (3) follows by preservation of first-order stochastic dominance. An application of the transitivity axiom then yields that ℓ_1 being preferred to ℓ_0 must imply ℓ_1' is preferred to ℓ_0'. The collection $\{((\ell_0, \ell_1), 1), ((\ell_0', \ell_1'), 0)\}$ is therefore a logical anomaly for expected utility theory.

In Appendix Table A1, we provide eight examples of dominated consequence effect anomalies for expected utility theory. Each of these examples can be mapped into the dominated consequence effect through an appropriate choice of ℓ_0, ℓ_1 and ℓ_0', ℓ_1' in the algorithmically generated menus of lotteries.

F.3.2 Reverse dominated consequence effect anomalies

Consider the first menu defined over the pair of lotteries $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ and the second menu defined over the pair of lotteries $\ell_0' = (p_0', z_0)$, $\ell_1' = (p_1', z_1)$. Let $\overline{z}_0 = \max_j z_0(j)$ and $\overline{z}_1 = \max_j z_1(j)$.

Suppose that the lotteries in the second menu can be written as

$$\begin{aligned}
\ell_0' &= \alpha_0\ell_0 + (1-\alpha_0)\delta_{\overline{z}_0}\\
\ell_1' &= \alpha_1\ell_1 + (1-\alpha_1)\delta_{\overline{z}_1}
\end{aligned}$$

for some $\alpha_0, \alpha_1 \in [0, 1]$. Further assume (i) $\overline{z}_1 > \overline{z}_0$, (ii) ℓ_1 is preferred to ℓ_0 – that is, $\ell_1 \succ \ell_0$, (iii) ℓ_0' is preferred to ℓ_1' – that is, $\ell_0' \succ \ell_1'$, and (iv) $\alpha_0 \geq \alpha_1$. To see why this is a logical anomaly for expected utility theory, observe

$$\begin{aligned}
&\ell_1 \succ \ell_0 \overset{(1)}{\Longrightarrow} \alpha_1\ell_1 + (1-\alpha_1)\delta_{\overline{z}_1} \succ \alpha_1\ell_0 + (1-\alpha_1)\delta_{\overline{z}_1},\\
&\alpha_1\ell_0 + (1-\alpha_1)\delta_{\overline{z}_1} \overset{(2)}{\succ} \alpha_1\ell_0 + (1-\alpha_1)\delta_{\overline{z}_0}\\
&\alpha_1\ell_0 + (1-\alpha_1)\delta_{\overline{z}_0} \overset{(3)}{\succ} \alpha_0\ell_0 + (1-\alpha_0)\delta_{\overline{z}_0}
\end{aligned}$$

where (1) follows by the independence axiom, (2) follows by utility must be increasing in monetary payoffs and the independence axiom, and (3) follows by preservation of first-order stochastic dominance. An application of the transitivity axiom then yields that ℓ_1 being preferred to ℓ_0 must imply that ℓ_1' is preferred to ℓ_0'. Therefore, the collection $\{((\ell_0, \ell_1), 1), ((\ell_0', \ell_1'), 0)\}$ is a logical anomaly for expected utility theory.

In Table A2, we provide eight examples of reverse dominated consequence effect anomalies for expected utility theory. Each of these examples can be mapped into the reverse dominated consequence effect through an appropriate choice of ℓ_0, ℓ_1 and ℓ_0', ℓ_1' in the algo-

rithmically generated menus of lotteries.

F.3.3 Strict dominance effect anomalies

Consider the first menu defined over the pair of lotteries $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ and the second menu defined over the pair of lotteries $\ell'_0 = (p'_0, z_0)$, $\ell'_1 = (p'_1, z_1)$. Let $\underline{z}_0 = \max_j z_0(j)$ and $\overline{z}_1 = \max_j z_1(j)$.

Suppose that the lotteries in the second menu can be written as

$$\ell'_0 = \alpha_0 \ell_0 + (1 - \alpha_0)\delta_{\underline{z}_0}$$
$$\ell'_1 = \alpha_1 \ell_1 + (1 - \alpha_1)\delta_{\overline{z}_1}$$

for some $\alpha_0, \alpha_1 \in [0, 1]$. Further assume that (i) ℓ_1 is preferred to ℓ_0 – that is, $\ell_1 \succ \ell_0$, and (ii) ℓ'_0 is preferred to ℓ'_1 – that is, $\ell'_0 \succ \ell'_1$. To see why this is a logical anomaly for expected utility theory, observe that

$$\ell'_1 \overset{(1)}{\succ} \ell_1$$
$$\ell_0 \overset{(2)}{\succ} \ell'_0,$$

where (1) and (2) follow by preservation of first-order stochastic dominance. An application of the transitivity axiom therefore means that the ℓ_1 being preferred to ℓ_0 must imply that ℓ'_1 is preferred to ℓ'_0. Therefore, the collection $\{((\ell_0, \ell_1), 1), ((\ell'_0, \ell'_1), 0)\}$ is a logical anomaly for expected utility theory.

In Table A3, we provide eight examples of dominated consequence effect anomalies for expected utility theory. Each of these examples can be mapped into the strict dominance effect through an appropriate choice of ℓ_0, ℓ_1 and ℓ'_0, ℓ'_1 in the algorithmically generated menus of lotteries.

F.4 Anomaly generation from an estimated choice probability function

In this section, we generate logical anomalies based on an estimated choice probability function $\widehat{f}_m(\cdot)$ using a random sample of binary choices.

Concretely, for each calibrated parameter value (δ, γ), we simulate a dataset of menus of two lotteries over two monetary payoffs and the individual's binary choice on each menu. For $i = 1, \ldots, n$, we simulate menus of two lotteries over two monetary payoffs X_i by drawing each payoff in the lotteries independently from a uniform distribution on $[0, 10]$, and simulating the probabilities in each lottery by drawing uniformly from the unit interval $[0, 1]$ and normalizing the draws so they lie on the unit simplex. For a particular choice of parameter values (δ, γ), we draw the individual's binary choice according to $Y_i \mid X_i \sim Bernoulli(f^*_m(X_i))$. This yields the binary choice dataset $\{(X_i, Y_i)\}_{i=1}^n$.

We then approximate the individual's true choice probability function

$$f^*_m(x) = P\left(CPT(p_1, z_1; \delta, \gamma) - CPT(p_0, z_0; \delta, \gamma) + \xi \geq 0\right)$$

in two ways. First, we consider the class of correctly specified choice probability functions, and estimate the parameter values $(\widehat{\delta}, \widehat{\gamma})$ that minimize the average cross-entropy loss between the individual's observed choices Y_i and the implied choice probabilities

$$(\widehat{\delta}, \widehat{\gamma}) = \arg\min_{\tilde{\delta}, \tilde{\gamma}} n^{-1} \sum_{i=1}^{n} -Y_i \log(f_{(\tilde{\delta}, \tilde{\gamma})}(X_i)) - (1 - Y_i) \log(1 - f_{(\tilde{\delta}, \tilde{\gamma})}(X_i)) \tag{18}$$

for $f_{(\tilde{\delta}, \tilde{\gamma})}(x) = \frac{e^{CPT(p_1, z_1; \tilde{\delta}, \tilde{\gamma}) - CPT(p_0, z_0; \tilde{\delta}, \tilde{\gamma})}}{1 + e^{CPT(p_1, z_1; \tilde{\delta}, \tilde{\gamma}) - CPT(p_0, z_0; \tilde{\delta}, \tilde{\gamma})}}$. This yields the estimated choice probability function $\widehat{f}_m(\cdot) = f_{(\widehat{\delta}, \widehat{\gamma})}(\cdot)$.

Second, we consider the class of choice probability functions that can be characterized by deep neural networks. We specifically consider over-parametrized deep neural networks with four hidden layers and 500 hidden nodes each with rectified linear unit (ReLU) activation functions. We minimize the average cross-entropy loss between the individual's observed choices Y_i and the implied choice probabilities

$$f_m^{DNN}(\cdot) = \arg\min_{\tilde{f} \in \mathcal{F}^{DNN}} n^{-1} \sum_{i=1}^{n} -Y_i \log(\tilde{f}(X_i)) - (1 - Y_i) \log(1 - \tilde{f}(X_i)) \tag{19}$$

using mini-batch gradient descent with a batch size of 256 observations over 2,000 epochs. For both the estimated probability weighting parameters and the deep neural network, the resulting estimated choice probability function $\widehat{f}_m(\cdot)$ is differentiable in the payoffs and probabilities of the lotteries in the menu. We can therefore directly apply our anomaly generation procedures.

For each calibrated parameter value (δ, γ), we simulate one binary choice dataset $\{(X_i, Y_i)\}_{i=1}^n$, and approximate the individual's true choice probability function $f_m^*(\cdot)$ using both the estimated probability weighting parameters (18) and the deep neural network (19). We apply our anomaly generation procedures on the estimated choice probability function $\widehat{f}_m^*(\cdot)$. As described in Section 5.1 of the main text and Appendix F.1, we flexibly parametrize the utility function as a linear combination of non-linear basis functions, and we apply our adversarial algorithm to 25,000 randomly initialized menus of two lotteries on two monetary payoffs and our example morphing algorithm to 15,000 randomly initialized menus. Each returned menu of lotteries over two monetary payoffs and the implied choices based on $\widehat{f}_m^*(\cdot)$ is a logical anomaly for expected utility theory at our particular parameterization of the utility function. We therefore again numerically verify whether the returned menu and implied choices based on $\widehat{f}_m^*(\cdot)$ is a logical anomaly for expected utility theory at any increasing utility function and without noisy choices, as discussed in Appendix F.2.

Appendix Table A8 and Appendix Table A9 summarize the logical anomalies for expected utility theory that are produced by our anomaly generation procedures at each calibrated parameter value (δ, γ) by approximating the individual's true choice probability function using the estimated probability weighting parameters and the deep neural network respectively. We vary the size of the simulated dataset over $n = 1,000, 5,000, 10,000$ and $25,000$. Using estimated choice probability functions, our anomaly generation procedures uncover the same categories of logical anomalies for expected utility theory as we found in Section 5.2 of the main text.

(a) $(\delta, \gamma) = (0.726, 0.309)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	7	25	2	17	85
Reverse Dominated Consequence Effect	1	4	0	3	17
Strict Dominance Effect	10	77	9	57	45
First Order Stochastic Dominance	1	66	16	74	81
Other	1	4	0	3	3
# of Logical Anomalies	20	176	27	154	231

(b) $(\delta, \gamma) = (0.926, 0.377)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	2	3	9	5	34
Reverse Dominated Consequence Effect	9	2	4	5	15
Strict Dominance Effect	17	5	1	1	1
First Order Stochastic Dominance	2	3	5	0	0
Other	2	2	0	0	1
# of Logical Anomalies	32	15	19	11	51

(c) $(\delta, \gamma) = (1.063, 0.451)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	5	7	2	0	10
Reverse Dominated Consequence Effect	13	4	3	5	14
Strict Dominance Effect	39	0	0	1	0
First Order Stochastic Dominance	33	0	0	1	2
Other	7	0	0	0	1
# of Logical Anomalies	97	11	5	7	27

Table A8: Logical anomalies for expected utility theory over two lotteries on two monetary payoffs, generated using an estimated choice probability function $\widehat{f}_m^*(\cdot) = f_{(\widehat{\delta},\widehat{\gamma})}(\cdot)$.

Notes: This table summarizes all logical anomalies for expected utility theory over the space of two lotteries on two monetary payoffs produced by applying our adversarial algorithm and our example morphing algorithm on an estimated choice probability function $\widehat{f}_m^*(\cdot)$. For each calibrated parameter values (δ, γ), we estimate the choice probability function by simulating a dataset $\{(X_i, Y_i)\}_{i=1}^n$ of menus of lotteries and binary choices and estimating the parameter values (δ, γ) that minimize average cross-entropy loss (18). We vary the size of the binary choice dataset over $n = 1,000, 5,000, 10,000$ and $25,000$. For reference, the column "True Choice Prob." reproduces Table 3, which generated logical anomalies using the true choice probability function $f_m^*(\cdot)$. See Appendix F.4 for further discussion.

(a) $(\delta, \gamma) = (0.726, 0.309)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	21	18	17	13	85
Reverse Dominated Consequence Effect	14	3	3	0	17
Strict Dominance Effect	35	7	2	1	45
First Order Stochastic Dominance	45	16	27	13	81
Other	3	0	1	3	3
# of Logical Anomalies	118	44	50	30	231

(b) $(\delta, \gamma) = (0.926, 0.377)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	16	17	22	15	34
Reverse Dominated Consequence Effect	17	6	4	5	15
Strict Dominance Effect	33	5	1	0	1
First Order Stochastic Dominance	25	18	17	10	0
Other	1	2	2	3	1
# of Logical Anomalies	92	48	46	33	51

(c) $(\delta, \gamma) = (1.063, 0.451)$

	Sample Size: n				True Choice Prob.
	1,000	5,000	10,000	25,000	
Dominated Consequence Effect	19	15	22	23	10
Reverse Dominated Consequence Effect	8	7	6	4	14
Strict Dominance Effect	26	2	3	0	0
First Order Stochastic Dominance	16	17	18	11	2
Other	3	0	3	4	1
# of Logical Anomalies	72	41	52	42	27

Table A9: Logical anomalies for expected utility theory over two lotteries on two monetary payoffs, generated using an estimated choice probability function $\widehat{f}_m^*(\cdot) = f^{DNN}(\cdot)$.

Notes: This table summarizes all logical anomalies for expected utility theory over the space of two lotteries on two monetary payoffs produced by applying our adversarial algorithm and our example morphing algorithm on an estimated choice probability function $\widehat{f}_m(\cdot)$. For each calibrated parameter values (δ, γ), we estimate the choice probability function by simulating a dataset $\{(X_i, Y_i)\}_{i=1}^n$ of menus of lotteries and binary choices and fitting a deep neural network to minimize average cross-entropy loss (19). We vary the size of the simulated binary choice dataset over $n = 1,000, 5,000, 10,000$ and $25,000$. For reference, the column "True Choice Prob." reproduces Table 3, which generated logical anomalies using the true choice probability function $f_m^*(\cdot)$. See Appendix F.4 for further discussion.

G Additional Results for Choice under Risk with Lotteries over Three Payoffs

In this Appendix, we extend our illustrative application to generate logical anomalies for expected utility theory over the space of menus of two lotteries over three monetary payoffs. We follow the same set-up as in Section 5.2 of the main text, applying our adversarial algorithm and example morphing algorithm to the true choice probability functions $f_m^*(\cdot)$ and setting the parameters (δ, γ) of the probability weighting function to be equal to the same calibrated parameter values $(0.726, 0.309), (0.926, 0.377), (1.063, 0.451)$.

For each calibrated parameter value (δ, γ), we apply our adversarial algorithm to 25,000 randomly initialized menus of three lotteries over three monetary payoffs x^0 and our example morphing algorithm to $15,000$ randomly initialized menus. We take gradient steps only updating the probabilities of the lotteries in the menu. We numerically verify whether the returned menus are logical anomalies for expected utility theory at any increasing utility function without noisy choices using the same procedure as described in Appendix F.2. We report all resulting, numerically verified logical anomalies for expected utility theory.

G.1 Logical anomalies generated by the probability weighting function

Appendix Table A10 summarizes the logical anomalies for expected utility theory that are produced by our anomaly generation procedures at each calibrated parameter values (δ, γ). Our anomaly generation procedures uncover analogous categories of logical anomalies as we found in the Section 5.2 of the main text over menus of lotteries over two monetary payoffs. We briefly discuss each category in turn.

	Prob. Weighting Function: (δ, γ)		
	$(0.726, 0.309)$	$(0.926, 0.377)$	$(1.063, 0.451)$
Dominated Consequence Effect	12	4	1
Reverse Dominated Consequence Effect	12	2	3
Strict Dominance Effect	20	6	11
First Order Stochastic Dominance	16	5	11
Other	0	0	0
# of Logical Anomalies	60	26	18

Table A10: Logical anomalies for expected utility theory over the menus of two lotteries on three monetary payoffs.

Notes: This table summarizes all logical anomalies for expected utility theory over the space of menus of two lotteries on three monetary payoffs produced by our adversarial algorithm and our example moprhing algorithm, organized by calibrated parameter values (δ, γ) of the probability weighting function and anomaly categories. See Appendix G.1 for further discussion.

First, our anomaly generation procedures uncover logical anomalies that exhibit (a generalization of) the dominated consequence effect. All of the logical anomalies in the first row of Table A10 have two possible, related structures. In the first case, for an appropriate choice of menu, menu A consists of the choice between lottery $\ell_0 = (p_0, z_0)$ and $\ell_1 = (p_1, z_1)$

each with support over three monetary payoffs. Furthermore, we can express the other pair of lotteries in menu B as

$$\ell_0' = \alpha_0 \ell_0 + (1 - \alpha_0)\ell_0'' \tag{20}$$

$$\ell_1' = \alpha_1 \ell_1 + (1 - \alpha_1)\ell_1'', \tag{21}$$

where ℓ_0'' is first order stochastically dominated by ℓ_0, ℓ_1'' first order stochastically dominates ℓ_0'', and $\alpha_1 \geq \alpha_0$ In the second case, for an appropriate choice of menu, menu A consists of the choice between

$$\ell_0 = \alpha_{0,A}\ell_0' + (1 - \alpha_{0,A})\ell_0'' \text{ and } \ell_1 = \alpha_{1,A}\ell_1' + (1 - \alpha_{1,A})\ell_1'', \tag{22}$$

where ℓ_0', ℓ_0'' and ℓ_1', ℓ_1'' have support over two or fewer monetary payoffs. We can analogously express menu B as

$$\ell_0' = \alpha_{0,B}\ell_0' + (1 - \alpha_{0,B})\ell_0'' \text{ and } \ell_1' = \alpha_{1,B}\ell_1' + (1 - \alpha_{1,B})\ell_1'' \tag{23}$$

where ℓ_0'' is first order stochastically dominated by ℓ_0' and ℓ_1'', and further $\alpha_{0,A} > \alpha_{0,B}$, $\alpha_{1,A} > \alpha_{1,B}$. In both cases, we observe (i) ℓ_1 is chosen over ℓ_0, and (ii) ℓ_0' is chosen over ℓ_1'. These logical anomalies exhibit a "dominated consequence effect" as the pair of menus highlight a violation of expected utility theory based on mixing each lottery with dominated lottery. We provide two illustrative examples in Table A11.

Second, our anomaly generation procedures uncover logical anomalies that exhibit the reverse dominated consequence effect. All of the logical anomalies in the second row of Table A10 have two possible, related structures. In the first case, for an appropriate choice of menu, menu A consists of the choice between lottery $\ell_0 = (p_0, z_0)$ and $\ell_1 = (p_1, z_1)$. We can express the other pair of lotteries in menu B as ℓ_0', ℓ_1' as (20) and (21) respectively, where now ℓ_1'' first order stochastically dominates ℓ_1, ℓ_1'' first order stochastically dominates ℓ_0'', and $\alpha_1 \leq \alpha_0$. In the second case, for an appropriate choice of menu, the lotteries in menu A can be written as (22), where ℓ_0', ℓ_0'' and ℓ_1', ℓ_1'' have support over two or fewer monetary payoffs. Menu B can be analogously expressed as (23), where now ℓ_1'' first order stochastically dominates ℓ_1' and ℓ_0'' as well as $\alpha_{1,B} < \alpha_{1,A}$, $\alpha_{0,B} < \alpha_{0,A}$. In both cases, we observe (i) ℓ_1 is chosen over ℓ_0; and (ii) ℓ_0' is chosen over ℓ_1'. These logical anomalies exhibit a "reverse dominated consequence effect" as the pair of menus highlight a violation of expected utility theory based on mixing each lottery with another dominating lottery. We provide two representative examples in Table A12.

Third, our anomaly generate procedures uncover logical anomalies that exhibit the strict dominance effect. All of the logical anomalies in the third row of Table A10 have two possible, related structures. In the first case, for an appropriate choice of menu, menu A consists of the choice between lottery $\ell_0 = (p_0, z_0)$ and $\ell_1 = (p_1, z_1)$. We can express the other pair of lotteries in menu B as ℓ_0', ℓ_1' as (20) and (21) respectively, where now ℓ_1'' dominates ℓ_1 and ℓ_0 first order stochastically dominates ℓ_0''. In the second case, for an appropriate choice of menu, the lotteries in menu A can be written as (22), where ℓ_0', ℓ_0'' and ℓ_1', ℓ_1'' have support over two or fewer monetary payoffs. Menu B can be analogously expressed as (23), where now ℓ_1'' first order stochastically dominates ℓ_0'' as well as $\alpha_{1,B} < \alpha_{1,A}, \alpha_{0,B} < \alpha_{0,A}$. These logical anomalies exhbit a "strict dominance effect" as the pair of menus highlight a violation of

expected utility theory based on mixing lottery ℓ_1 with a lottery that strictly dominates the lottery that is mixed with lottery ℓ_0. We provide two representative examples in Table A13.

Finally, our anomaly generation procedures again uncover first-order stochastic dominance violations, in which the individual selects lotteries that are first-order stochastically domianted by the other lottery in the menu. We provide two representative examples in Table A14.

(a) Logical Anomaly #1

Menu A (x_A, y_A^*)			
Lottery 0	\$6.56	\$6.92	\$7.40
	36%	27%	37%
Lottery 1	\$5.75	\$5.95	\$9.44
	39%	33%	28%
Menu B (x_B, y_B^*)			
Lottery 0	\$6.56	\$6.92	\$7.40
	100%	0%	0%
Lottery 1	\$5.75	\$5.95	\$9.44
	13%	16%	71%

(b) Logical Anomaly #2

Menu A (x_A, y_A^*)			
Lottery 0	\$1.03	\$4.90	\$6.64
	0%	96%	4%
Lottery 1	\$0.71	\$5.46	\$7.48
	13%	1%	86%
Menu B (x_B, y_B^*)			
Lottery 0	\$1.03	\$4.90	\$6.64
	37%	40%	23%
Lottery 1	\$0.71	\$5.46	\$7.48
	50%	27%	23%

Table A11: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the dominated consequence effect over menus of lotteries on three monetary payoffs.

Notes: In the menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each algorithmically generated, logical anomaly exhibiting the dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. Each algorithmically generated, logical anomaly presented here is produced by our example morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$ and logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Appendix G.1 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)		
Lottery 0	\$6.05	\$6.56	\$6.88
	0%	100%	0%
Lottery 1	\$4.620	\$7.360	\$9.360
	5%	12%	83%
	Menu B (x_B, y_B^*)		
Lottery 0	\$6.05	\$6.56	\$6.88
	6%	46%	48%
Lottery 1	\$4.62	\$7.36	\$9.36
	37%	43\$	20%

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)		
Lottery 0	\$2.15	\$5.37	\$8.95
	86%	2%	12%
Lottery 1	\$3.77	\$4.45	\$8.93
	9%	91%	0%
	Menu B (x_B, y_B^*)		
Lottery 0	\$2.15	\$5.37	\$8.95
	41%	15%	44%
Lottery 1	\$3.77	\$4.45	\$8.93
	18%	59%	23%

Table A12: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the reverse dominated effect over menus of lotteries on three monetary payoffs.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each logical anomaly exhibiting the revese dominated consequence effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. Each algorithmically generated, logical anomaly depicted here is produced by our example morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$ and logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Appendix G.1 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)		
Lottery 0	\$4.41	\$7.28	\$7.98
	7%	11%	82%
Lottery 1	\$5.89	\$6.53	\$7.41
	100%	0%	0%
	Menu B (x_B, y_B^*)		
Lottery 0	\$4.41	\$7.28	\$7.98
	27%	24%	49%
Lottery 1	\$5.89	\$6.53	\$7.41
	69%	29%	2%

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)		
Lottery 0	\$1.37	\$1.67	\$6.44
	93%	2%	5%
Lottery 1	\$1.87	\$2.30	\$5.56
	14%	85%	1%
	Menu B (x_B, y_B^*)		
Lottery 0	\$1.37	\$1.67	\$6.44
	48%	27%	25%
Lottery 1	\$1.87	\$2.30	\$5.56
	10%	75%	15%

Table A13: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate the strict dominance effect over menus of lotteries on three monetary payoffs.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each logical anomaly exhibiting the strict dominance effect consists of two menus $\{(x_A, y_A^*), (x_B, y_B^*)\}$. Each algorithmically generated, logical anomaly depicted here is produced by our example morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$, and logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Appendix G.1 for further discussion.

(a) Logical anomaly #1

Lottery 0	\$5.86		
	100%		
Lottery 1	\$6.07	\$6.93	\$7.14
	5%	21%	74%

(b) Logical anomaly #2

Lottery 0	\$3.70	\$3.99	\$9.47
	38%	39%	23%
Lottery 1	\$2.74	\$9.45	
	81%	19%	

Table A14: Representative examples of algorithmically generated logical anomalies for expected utility theory that illustrate first-order stochastic dominance violations over menus of lotteries on three monetary payoffs.

Notes: In each menu, we color the lottery that is selected by the individual with probability at least 0.50 in green. Each generated first-order stochastic dominance violation presented here (x, y^*) is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$. Logical anomaly #1 are generated by our example morphing algorithm. Logical anomaly #2 is generated by our adversarial algorithm. For ease of interpretation, we round each payoff to the nearest cent and each probability to the nearest percentage. See Appendix G.1 for further discussion.

G.2 Experimental test of algorithmically generated anomalies

As in Section 5.3 of the main text, we empirically test our algorithmically generated logical anomalies over menus of lotteries over three monetary payoffs in incentivized online experiments.

G.2.1 Experimental design

We selected 35 logical anomalies for expected utility theory over menus of two lotteries over three monetary payoffs in Table A10 that span both the categories (dominated consequence, reverse dominated consequence, and strict dominance effect) as well as the calibrated parameter values (δ, γ) that we analyzed. We then split these 35 logical anomalies into two separate surveys, one containing 18 logical anomalies and another containing 17 logical anomalies.

Each chosen logical anomaly consists of a pair of menus of two lotteries over three monetary payoffs. We therefore present each logical anomaly as two separate binary choices on menus, and so the surveys consists of 36 main questions and 34 main questions respectively. For a particular menu, we display the written probabilities and payoffs for each lottery in the menu, and we additionally depict each lottery as a color-coded pie chart. Each survey randomizes the order of questions and the left-right positioning of lotteries in a menu across respondents. We pre-registered both our surveys on `EGAP` (see `https://osf.io/tjg2p`).

We recruited respondents for both surveys on Prolific. Each respondent received a base payment of \$4 for completing a survey. As in the main text, we screened out inattentive respondents through comprehension questions and attention checks throughout the surveys. Respondents that successfully completed a survey without failing any of the comprehension questions and attention checks were eligible for a randomized bonus payment based on a "random payment selection" mechanism (Azrieli, Chambers and Healy, 2018, 2020). The average bonus payment was \$8.37 and \$6.63 on each survey respectively, and respondents completed each survey in roughly 15 minutes on average. Respondents were therefore paid on average \$49.48 and \$42.52 per hour on survey respectively. Altogether, we recruited 257 and 255 respondents on our two surveys respectively.

We include screenshots of the instructions, comprehension checks, attention checks, and main survey questions in Appendix H.

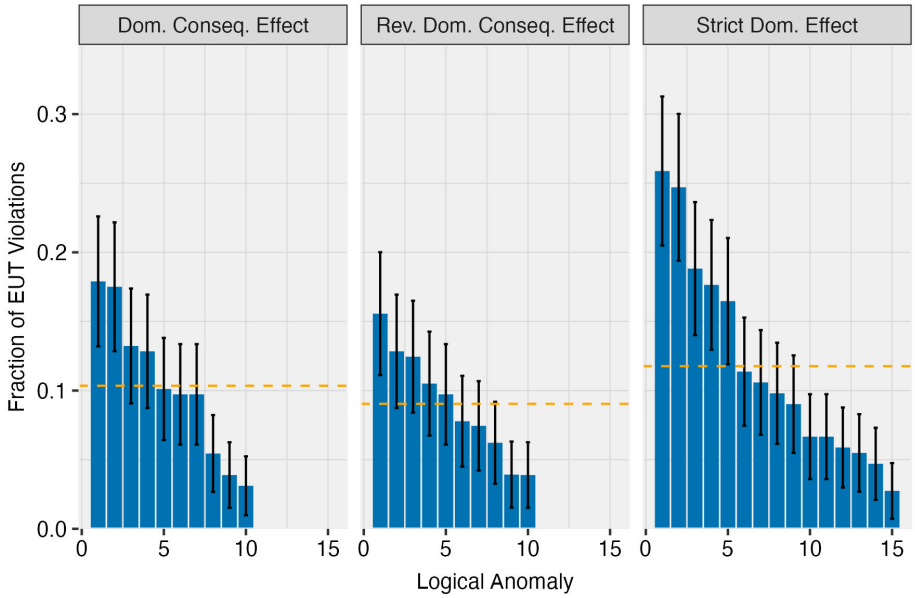
G.2.2 Experimental results

We analyze the choices on our algorithmically generated logical anomalies of all respondents that completed the surveys without failing any attention and comprehension checks.

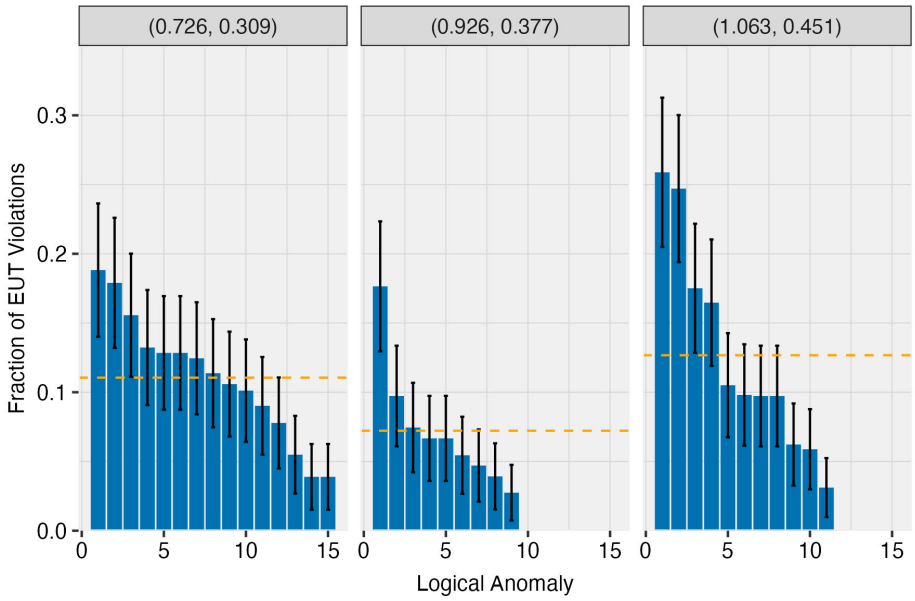
Appendix Figure A7(a) reports the fraction of respondents whose choices violate expected utility theory without noise on our algorithmically generated logical anomalies ("expected utility theory violation rates"), organized by logical anomaly category. Appendix Figure A7(b) reports the same quantities organized by the calibrated parameter values (δ, γ) that we considered. We report 95% confidence intervals with standard errors clustered at the respondent level. Appendix Table A15 and Appendix Table A16 provide summary statistics on the expected utility theory violation rates pooling across logical anomalies within the same category and same calibrated parameter values respectively. We find that the pooled expected utility theory violation rate is 10.3% (p-value < 0.001) on dominated consequence

effect anomalies, 9.0% (p-value < 0.001) on reverse dominated consequence effect anomalies, and 11.7% (p-value < 0.001) on strict dominance effect anomalies. Analyzing each logical anomaly separately and applying a conservative Bonferroni correction for multiple hypotheses across all logical anomalies in our surveys, the expected utility theory violation rate is statistically different than zero at the 5% level for 33 out of 35. We therefore find strong evidence that the pooled respondents' choices are inconsistent with expected utility theory across our discovered categories of logical anomalies over lotteries on three monetary payoffs.

Of course, if there exists enough idiosyncratic noise in respondents' choices, we would expect to find non-zero expected utility theory violation rates on our algorithmically generated logical anomalies. As in Section 5.3 of the main text, we therefore estimate the probability of erroneous deviations from preferences consistent with expected utility theory that would be required to explain the observed choices of respondents on our algorithmically generated logical anomalies. Appendix Figure A8(a) reports the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on each algorithmically generated, logical anomaly separately and organized by logical anomaly category. Appendix Figure A7(b) reports the same quantities organized by the calibrated parameter values (δ, γ) that we considered. We report 95% confidence intervals based on bootstrapped standard errors. Appendix Figure A8 reports the same estimates, organized by calibrated parameter values (δ, γ) that we considered. The median estimated idiosyncratic error rate $\widehat{\epsilon}$ across algorithmically generated logical anomalies is 12.0% for dominated consequence effect anomalies, 10.5% for reverse dominated consequence effect anomalies, and 12.0% for strict dominance effect anomalies. We again find substantial heterogeneity in these estimates across logical anomalies. For example, explaining the observed choice fractions on several specific logical anomalies across categories would require that respondents erroneously deviate from their true preferences at least 20% of the time.



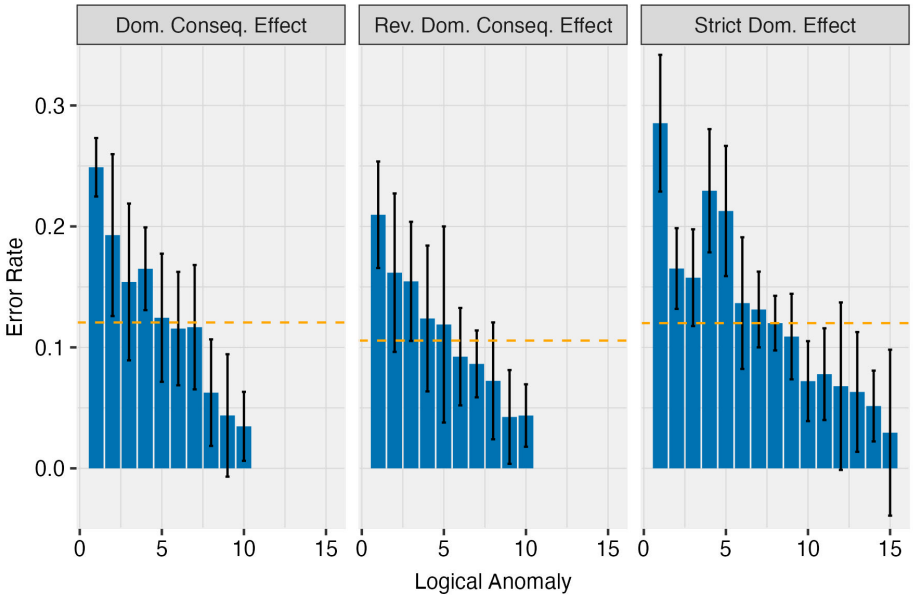
((A)) Estimates by logical anomaly category



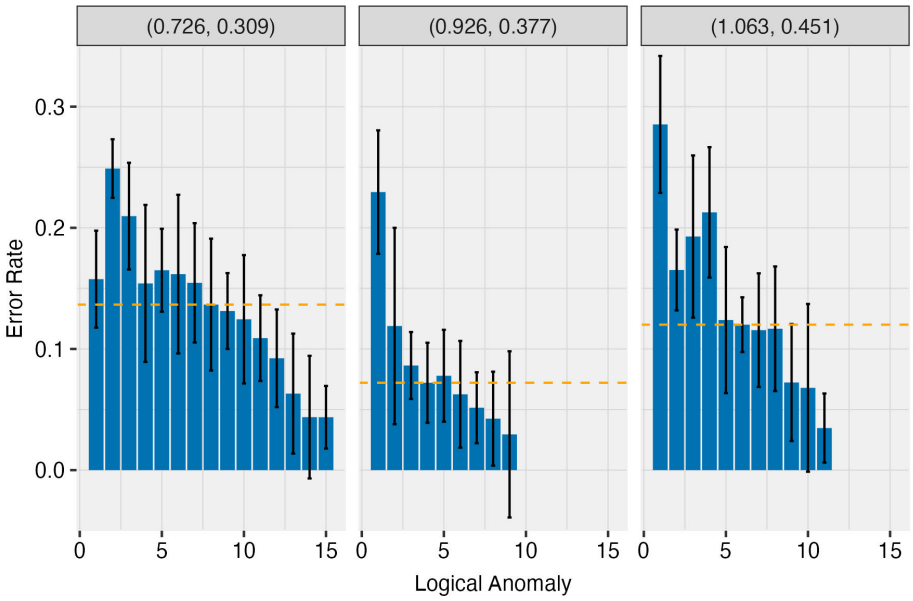
((B)) Estimates organized by calibrated parameter values (δ, γ)

Figure A7: Fraction of respondents whose choices violate expected utility theory on algorithmically generated logical anomalies over menus of lotteries on three monetary payoffs.

Notes: This figure summarizes the fraction of respondents whose choices violate expected utility theory on the logical anomalies of menus of two lotteries over two monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors clustered at the respondent level). We organize the estimates by category of logical anomaly (see Table A10) and by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the fraction of respondents whose choices violate expected utility theory pooling across all logical anomalies within the same grouping. Within each grouping, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Appendix G.2 for further discussion.



((A)) Estimates by logical anomaly category



((B)) Estimates organized by calibrated parameter values (δ, γ)

Figure A8: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated logical anomalies over menus of lotteries on three monetary payoffs.

Notes: This figure summarizes the estimated idiosyncratic error rate $\widehat{\epsilon}$ required to explain the observed choices of respondents on our algorithmically generated logical anomalies of menus of lotteries over three monetary payoffs (blue bars) and 95% confidence intervals (black error bars; standard errors computed by the bootstrap). We organize the estimates by category of logical anomaly (see Table A10) and by the calibrated parameter values (δ, γ) of the probability weighting function (16). The orange dashed line reports the median estimated idiosyncratic error rate across all logical anomalies within the same grouping. Within each grouping, we sort the logical anomalies and assign each logical anomaly an arbitrary numeric identifier in decreasing order based on the fraction of respondents whose choices violate expected utility theory. See Appendix G.2 for further discussion.

	Pooled Average	Median	First Quartile	Third Quartile
Dominated Consequence Effect	0.103 (0.006)	0.099	0.065	0.131
Reverse Dominated Consequence Effect	0.090 (0.006)	0.087	0.065	0.119
Strict Dominance Effect	0.117 (0.006)	0.098	0.062	0.170

Table A15: Summary statistics on the fraction of respondents whose choices violate expected utility theory on algorithmically generated logical anomalies over menus of lotteries on three monetary payoffs.

Notes: This table reports summary statistics on the fraction of respondents whose choices violate expected utility theory ("expected utility theory violation rate") on algorithmically generated logical anomalies of menus of two lotteries over three monetary payoffs. We report summary statistics by category of logical anomaly (see Table A10). The "pooled average" column reports the expected utility theory violation rate, pooling together respondents' choices on all logical anomalies within the same category. Standard errors reported in parentheses are clustered at the respondent level. We also report the median, first quartile, and third quartile of the distribution of expected utility theory violation rates across logical anomalies within the same category. See Appendix G.2 for further discussion.

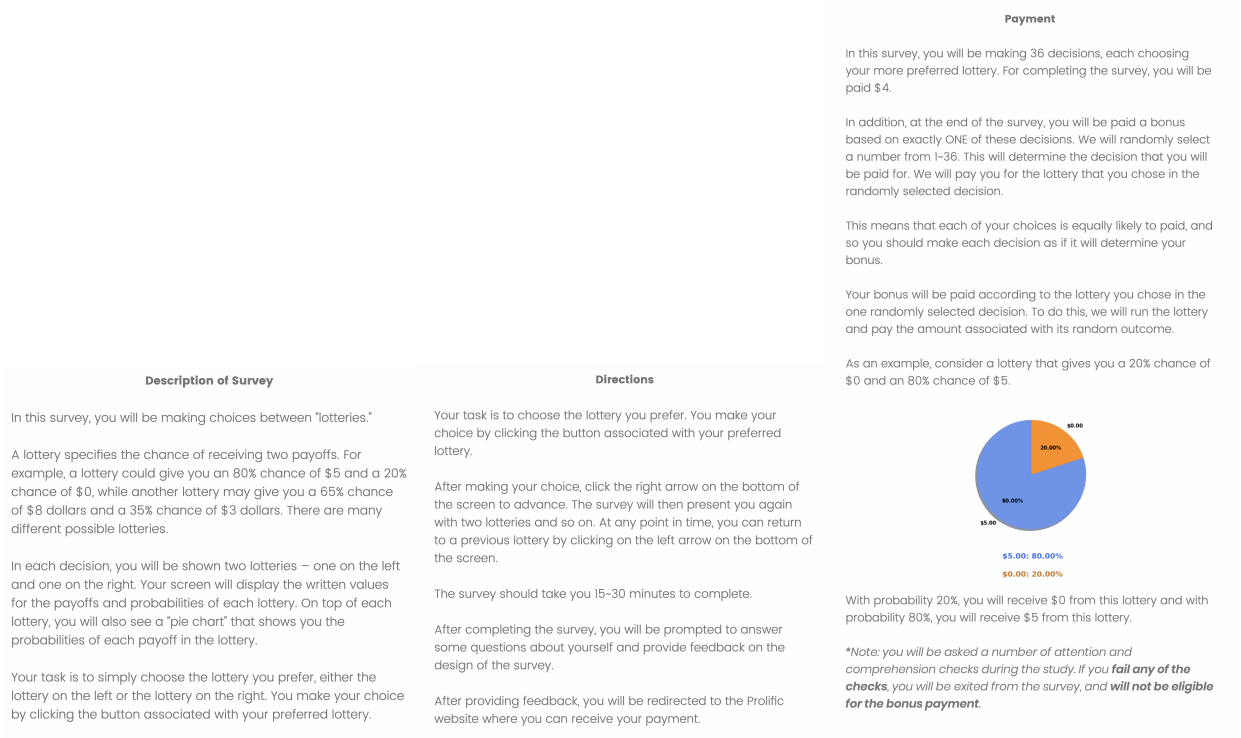
Prob. Weighting Function: (δ, γ)	Pooled Average	Median	First Quartile	Third Quartile
$(0.726, 0.309)$	0.110 (0.006)	0.113	0.084	0.130
$(0.926, 0.377)$	0.072 (0.006)	0.066	0.047	0.074
$(1.063, 0.451)$	0.126 (0.007)	0.098	0.079	0.169

Table A16: Summary statistics for anomalous fractions on logical anomalies over menus of two lotteries over menus of lotteries on three monetary payoffs, organized by calibrated parameter values of probability weighting function (δ, γ).

Notes: This table reports summary statistics on the fraction of respondents whose choices violate expected utility theory ("expected utility theory violation rate") on algorithmically generated logical anomalies of menus of two lotteries over three monetary payoffs. We report summary statistics by calibrated parameter values of probability weighting function (δ, γ) (see Table 3). The "pooled average" column reports the expected utility theory violation rate, pooling together respondents' choices on all logical anomalies within the same category. Standard errors reported in parentheses are clustered at the respondent level. We also report the median, first quartile, and third quartile of the distribution of expected utility theory violation rates across logical anomalies within the same category. See Appendix G.2 for further discussion.

H Experimental Instructions and Control Questions for Online Surveys

In this section, we provide screenshots of the instructions, attention and comprehension checks, and survey questions of the online surveys of our algorithmically generated logical anomalies for expected utility theory over menus of two lotteries on two monetary payoffs and menus of two lotteries on three monetary payoffs.



Description of Survey

In this survey, you will be making choices between "lotteries."

A lottery specifies the chance of receiving two payoffs. For example, a lottery could give you an 80% chance of $5 and a 20% chance of $0, while another lottery may give you a 65% chance of $8 dollars and a 35% chance of $3 dollars. There are many different possible lotteries.

In each decision, you will be shown two lotteries – one on the left and one on the right. Your screen will display the written values for the payoffs and probabilities of each lottery. On top of each lottery, you will also see a "pie chart" that shows you the probabilities of each payoff in the lottery.

Your task is to simply choose the lottery you prefer, either the lottery on the left or the lottery on the right. You make your choice by clicking the button associated with your preferred lottery.

Directions

Your task is to choose the lottery you prefer. You make your choice by clicking the button associated with your preferred lottery.

After making your choice, click the right arrow on the bottom of the screen to advance. The survey will then present you again with two lotteries and so on. At any point in time, you can return to a previous lottery by clicking on the left arrow on the bottom of the screen.

The survey should take you 15~30 minutes to complete.

After completing the survey, you will be prompted to answer some questions about yourself and provide feedback on the design of the survey.

After providing feedback, you will be redirected to the Prolific website where you can receive your payment.

Payment

In this survey, you will be making 36 decisions, each choosing your more preferred lottery. For completing the survey, you will be paid $4.

In addition, at the end of the survey, you will be paid a bonus based on exactly ONE of these decisions. We will randomly select a number from 1~36. This will determine the decision that you will be paid for. We will pay you for the lottery that you chose in the randomly selected decision.

This means that each of your choices is equally likely to paid, and so you should make each decision as if it will determine your bonus.

Your bonus will be paid according to the lottery you chose in the one randomly selected decision. To do this, we will run the lottery and pay the amount associated with its random outcome.

As an example, consider a lottery that gives you a 20% chance of $0 and an 80% chance of $5.

With probability 20%, you will receive $0 from this lottery and with probability 80%, you will receive $5 from this lottery.

Note: you will be asked a number of attention and comprehension checks during the study. If you **fail any of the checks, you will be exited from the survey, and **will not be eligible for the bonus payment**.*

Figure H1: Screenshots of directions for the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

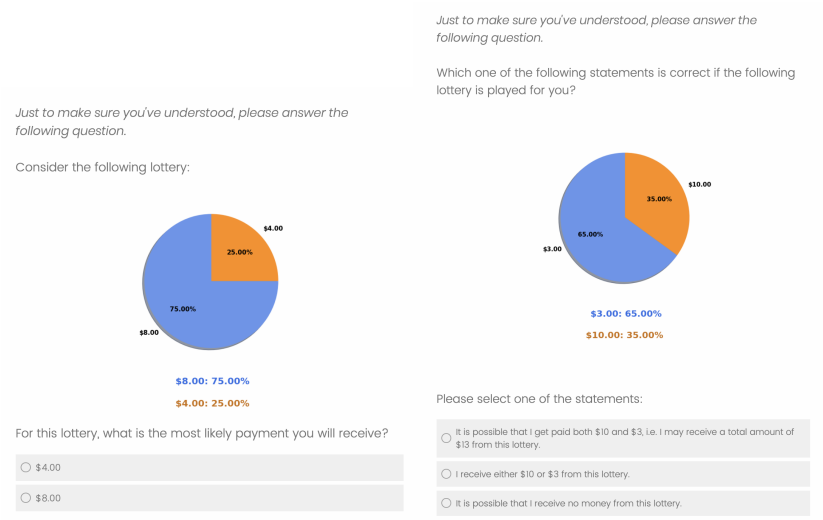


Figure H2: Screenshots of comprehension checks for the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

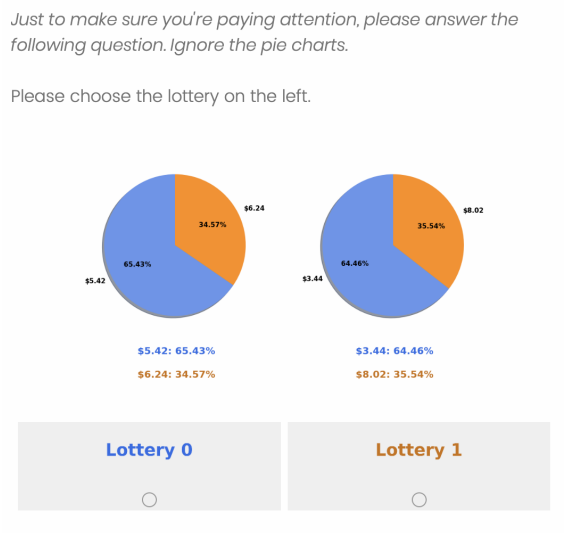


Figure H3: Screenshot of an attention check included in the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

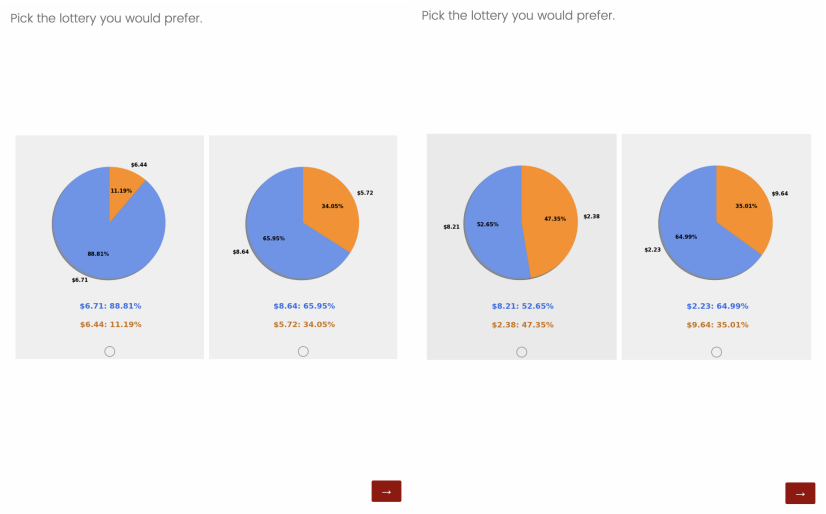


Figure H4: Screenshots of two main survey questions for the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

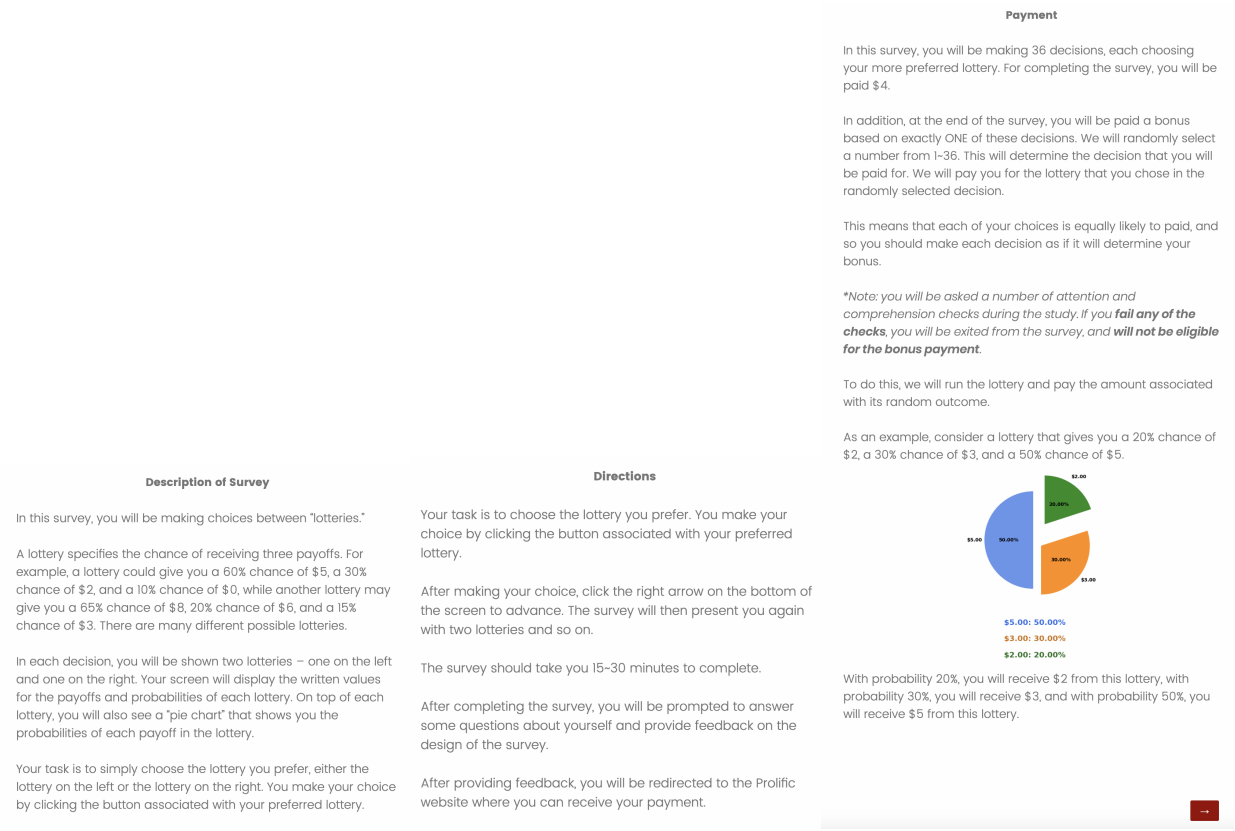


Figure H5: Screenshots of directions for the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

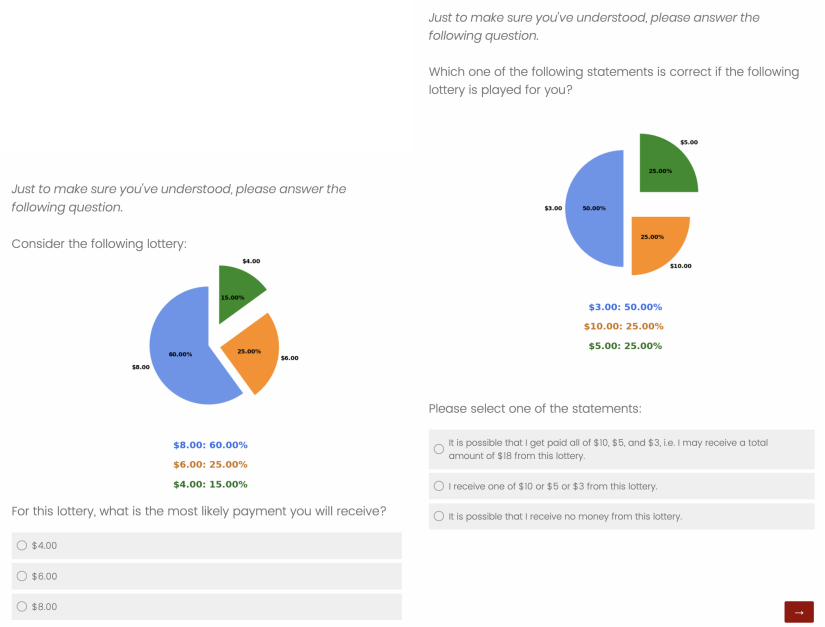


Figure H6: Screenshots of comprehension checks for the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

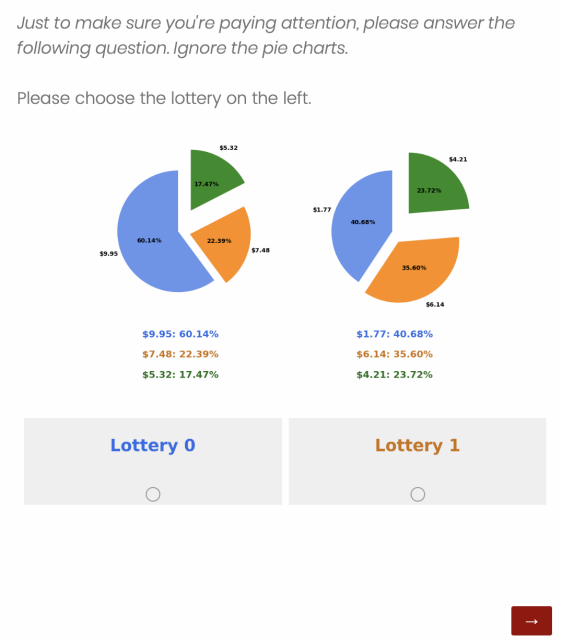


Figure H7: Screenshot of an attention check included in the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

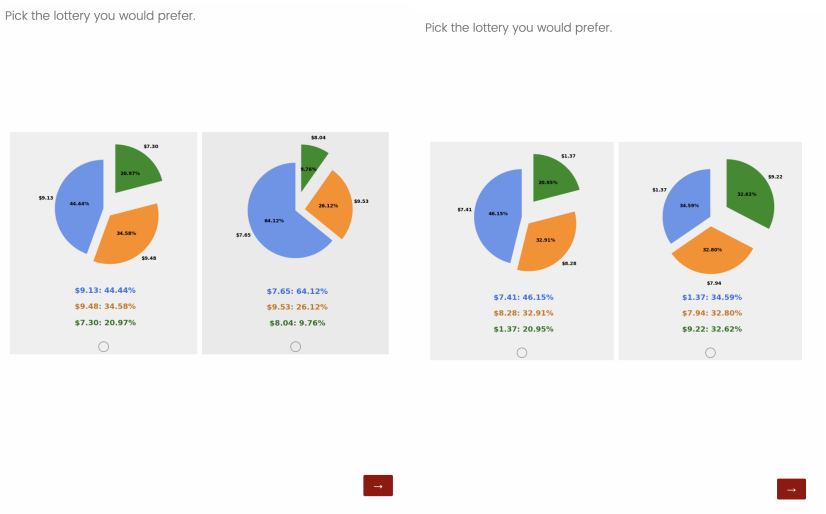


Figure H8: Screenshots of two main survey questions for the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.