From Predictive Algorithms to Automatic Generation of Anomalies*

Sendhil Mullainathan Ashesh Rambachan†

November 7, 2023

Abstract

Economic theories often progress through the discovery of "anomalies." Canonical examples of anomalies include the Allais Paradox and the Kahneman-Tversky choice experiments, which are constructed menus of lotteries that highlighted particular flaws in expected utility theory and spurred the development of new theories for decision-making under risk. In this paper, we develop algorithmic procedures to automatically generate such anomalies. Our algorithmic procedures take as inputs an existing theory and data it seeks to explain, and then generate examples on which we would likely observe violations of our existing theory if we were to collect data. As an illustration, we produce anomalies for expected utility theory using simulated lottery choice data from individuals who behave according to cumulative prospect theory. Our procedures recover known anomalies for expected utility theory in behavioral economics and discover novel anomalies based on the probability weighting function. We conduct incentivized experiments to collect choice data on our algorithmically generated anomalies, finding that participants violate expected utility theory at similar rates to the Allais Paradox and Common Ratio Effect. While this illustration is specific, our anomaly generation procedures are general and can be applied in any domain where there exists a formal theory and rich data that the theory seeks to explain.

*We thank Peter G. Chang for exceptional research assistance. We also thank audiences at Georgetown, Stanford, Yale, and the NBER Summer Institute Digital Economics and AI session, Roshni Sahoo, Suproteem Sarkar, Cassidy Shubatt, Keyon Vafa, and especially 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

Anomalies play a central role in improving economic theories. An anomaly is neither a hypothesis test nor a test statistic for whether an existing model is misspecified.[1] But rather it is a carefully constructed *example* that provides clues as to how or why a theory may fail empirically. In this paper, we ask whether machine learning can automatically generate anomalies for our existing economic theories.

As a concrete example, consider the field of decision-making under risk. Shortly after the axiomatization of expected utility theory (von Neumann and Morgenstern, 1944), questions arose surrounding its descriptive accuracy: how well does expected utility theory describe the risky choices of people? To illustrate a possible empirical weakness, Allais (1953) produced the now celebrated "Allais Paradox," which is a hypothetical pair of menus of lotteries and choices depicted in Table 1. The hypothetical menus in the Allais Paradox are crafted so

(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).

that expected utility theory restricts the possible choices across the two menus. Due to the independence axiom, the only choices that are consistent with expected utility theory are selecting lotteries $(A0, B0)$ or lotteries $(A1, B1)$. Allais conjectured that many individuals would, in fact, select lotteries $(A0, B1)$, and indeed researchers subsequently found this to be true empirically (e.g., Slovic and Tversky, 1974; Kahneman and Tversky, 1979; Huck and Müller, 2012).[2] This was only the beginning as researchers steadily accumulated more anomalies for expected utility theory.[3] Eventually, Tversky and Kahneman (1992) suggested

[1] 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.

[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 payoffs are real vs. hypothetical, etc.

[3] 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

that cumulative prospect theory could resolve many of these anomalies. Armed with a new theory, the cycle repeats itself: researchers have since crafted new anomalies, suggesting elements that were missing from cumulative prospect theory that in turn prompted the development of new theories of choice under risk.[4]

Indeed the field of decision-making under risk is not exceptional, as anomalies have played a crucial role in the development of game theory, asset pricing, and many other fields. Scientific discovery in economics often advances through the generation of anomalies that highlight possible inconsistencies between theories and nature. As anomalies are generated, researchers invest great effort in robustly evaluating them and developing improved theories to resolve them.

Anomalies are generated through a creative process that involves both empirical intuition and an existing theory. A researcher like Allais first builds an empirical intuition about how people behave and contrasts their intuitions with an existing theory. In order to highlight any discrepancies, the researcher then carefully crafts an anomaly, or a concrete example where the theory's predictions differ from what they believe to be the likely empirical patterns. We rely on researchers for each of these steps.

In this paper, we develop algorithmic procedures to automate this anomaly generation process. Like a researcher, our procedures take as input an existing theory. But additionally, our procedures take in rich data that the theory seeks to explain and use supervised machine learning algorithms to build a predictive model. The resulting black box, predictive model serves as our procedures' empirical intuition, exploiting the fact that supervised machine learning algorithms often uncover novel predictive signals that we may overlook ourselves and our existing theories may not capture.[5] Our procedures then automatically contrast the predictive model with the existing theory and return anomalies – small, generated datasets on which we would likely observe violations of our existing theory if we were to collect data.

To build these algorithmic procedures, we must first develop an econometric framework for anomaly generation that abstracts from any particular economic domain. As mentioned, anomalies play a key role, for example, in choice under risk, game theory, and asset pricing,

framing effects and preference reversals, and finally Kahneman and Tversky (1984); Tversky and Kahneman (1991) produced anomalies to highlight loss aversion.

[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]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.

yet theories across these economic domains share little resemblance. 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. An econometric framework for anomaly generation must therefore capture this immense diversity of economic theories.

To tackle this challenge, we abstractly model theories as *black box* mappings, which, when given any finite dataset, return correspondences that summarize their derived implications between some features and modeled outcomes. Expected utility theory, for example, derives implications about an individual's choice behavior from datasets of menus of lotteries and choice probabilities. We introduce four assumptions on such theory mappings and then establish two results. First, for any theory satisfying these assumptions, there exist *anomalies* or minimal datasets that would be incompatible with the theory if observed like the Allais Paradox. Second, any theory satisfies these assumptions if and only if it can be equivalently represented as an *allowable function class*. A theory's allowable function class summarizes all mappings between the features and the theory's modeled outcome that are consistent with its underlying structure (whatever that may be). Any theory can therefore be analyzed as if it searches for allowable functions that are consistent with any given dataset.

Given the tractable characterization of theories based on their allowable function classes, we next ask how we can search for anomalies. We observe that anomaly generation can be interpreted as an adversarial game between a falsifier and the theory. Given an estimated prediction function, the falsifier proposes conjectured datasets, or finite collections of features and the estimated prediction function evaluated on those features, and the theory attempts to explain those conjectured datasets by fitting an allowable function. The falsifier's payoff is increasing in the theory's average loss on the proposed dataset, and the theory's payoff is decreasing in its average loss. We show that anomalies can be characterized as conjectured datasets that induce a strictly positive, average loss for the theory in such a game.

Based on this characterization, 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 this feasible implementation of the falsifier's max-min program, establishing finite-sample bounds on how well it approximates its population analog. Furthermore, practically optimizing this max-min program may be challenging – the falsifier's maximization over proposed datasets will typically be non-concave, and so standard techniques may not apply (e.g., Rockafellar, 1970; Freund and Schapire, 1996). We therefore leverage recent results in adversarial learning and non-convex/concave min-max optimization to develop a gradient descent ascent procedure and analyze its conver-

gence properties (Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020). The resulting gradient descent ascent procedure generates anomalies by iteratively updating the falsifier's conjectured dataset to maximize the average loss of the theory's best-responding allowable function.

While this adversarial procedure exploits nothing beyond the theory's allowable functions, there in fact exists additional structure for anomaly generation. We introduce conditions under which any theory has a non-trivial, lower-dimensional representation of the features; 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. As a consequence, some anomalies, like the Allais Paradox, reveal what we call *representational* errors that the theory's lower dimensional representation has failed to capture some relevant dimension along which modeled outcomes systematically vary. We develop a dataset morphing procedure to generate representational anomalies for a theory, if they exist. Given an initial feature value, the dataset morphing procedure locally searches for nearby feature values that are representationally equivalent under the theory but across which the estimated prediction function varies.

As an illustration, we apply our anomaly generation procedures to the domain of choice under risk, returning to our motivating example of anomalies for expected utility theory. We explore what anomalies for expected utility theory would be uncovered by our procedures if given simulated lottery choice data from an individual whose choices are consistent with cumulative prospect theory. Since the properties of cumulative prospect theory have been well-studied by behavioral economists, we can compare and contrast our algorithmically generated anomalies against known anomalies for expected utility theory, such as those produced in Allais (1953), Kahneman and Tversky (1979), and others.

Our anomaly generation procedures recover known anomalies for expected utility theory, such as first-order stochastic dominance violations that are implied by particular parameterization of the probability weighting function. More importantly, thought, our anomaly generation procedures uncover novel anomalies for expected utility function that are implied by non-linearities in the probability weighting function. We categorize these novel anomalies and refer to them as a "dominated consequence effect," a "reverse dominated consequence effect," and a "strict dominance effect." These are all anomalies for expected utility theory that involve two menus of two lotteries and mixing lotteries with particular certain prospects. Provided the lotteries have only two monetary payoffs, the dominated consequence effect is a generalization of the well-known Common Ratio Effect. The other two anomaly categories cannot be cast as examples of either the Common Consequence Effect or Common Ratio Effect, and so these anomaly categories are genuine discoveries about the properties of the

probability weighting function. These categories, to our knowledge, have not been noticed before in Allais (1953), Kahneman and Tversky (1979), or elsewhere.

We emphasize that the primary contribution of our work is to develop procedures for the automatic generation of anomalies. Yet having algorithmically generated anomalies for expected utility theory that are implied by properties of the probability weighting function, one cannot help but wonder: do these anomalies for expected utility theory also hold empirically? Investigating this question is where the anomaly generation process ends, and the careful experimental work that is the hallmark of behavioral economics begins. Indeed, generating anomalies and then rigorously testing them by collecting new data is a valuable activity in evaluating our existing theories.

Fully establishing the empirical robustness of our novel categories of anomalies is obviously beyond the scope of the present paper. As a first step, we experimentally test our algorithmically generated anomalies, recruiting participants on Prolific to make incentivized choices between these lotteries. Participants' choices on our algorithmically generated anomalies violate expected utility theory at similar rates to 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; McGranaghan et al., Forthcoming). Our algorithmically generated anomalies hold empirically, at least to the same benchmark as existing anomalies for expected utility theory. Our data suggest these new anomalies merit the kind of rigorous experimental scrutiny that has given to other known anomalies for expected utility theory.

This specific application illustrates the broader potential for our anomaly generation procedures. Their success in generating novel anomalies in a well-trodden domain like choice under risk suggests they could be valuable in many other areas. Indeed, our algorithmic procedures are broadly applicable and can be used in any domain where there exists a formal theory and rich data that the theory seeks to explain. Furthermore, 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 generated datasets that may help researchers evaluate and improve theories.

Related work: This paper is part of a 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) provide recent reviews on the use of machine learning in physical sciences, such as biology, chemistry, and physics. Substantial progress has already been made in exploring how machine learning interacts with

economic theories. Several recent papers compare the out-of-sample predictive performance of black-box machine learning models against that of economic theories in domains, such as 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 properties (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); 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 observed play will differ from alternative theories of strategic behavior and then 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 adversarial learning and dataset morphing procedures are general-purpose procedures that enable researchers to search for anomalies given any formal theory.

2 Theories and the Anomaly Generation Problem

Theories derive implications about the relationship between some features and modeled outcomes by positing some underlying structure. Yet how theories mathematically model their underlying structure varies greatly, and an econometric framework for anomaly generation must somehow capture this diversity. In this section, we analyze theories as black box mappings that return correspondences between the features and modeled outcome from any finite dataset. We establish two results on the properties of these black boxes that serve as the foundation of our algorithmic procedures for anomaly generation.

2.1 Setting and theories

Let $x \in \mathcal{X}$ be some vector of features, $y^* \in \mathcal{Y}$ some modeled outcome, and $D = \{(x_1, y_1^*), \ldots, (x_n, y_n^*)\}$ a finite modeled dataset in a scientific domain. We let $\mathcal{D}$ denote the collection of all modeled datasets, $\mathcal{F}$ the collection of all mappings $f(\cdot)\colon \mathcal{X} \to \mathcal{Y}^*$, and $\mathcal{C}$ the collection of all correspondences $c(\cdot)\colon \mathcal{X} \rightrightarrows \mathcal{Y}^*$.

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 modeled datasets to correspondences between the features and modeled outcome, and $\mathcal{M}$ is some finite set with elements $m \in \mathcal{M}$.

Rather than focusing on any particular mathematical model, we define a theory as a reduced-form mapping. Given any modeled dataset $D \in \mathcal{D}$, 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 modeled dataset $D \in \mathcal{D}$, and $T(x; D) \subseteq \mathcal{Y}^*$ to be the theory's implications about the modeled outcome at feature $x \in \mathcal{X}$. All else about the scientific domain is collapsed into *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 focus on the behavior of its correspondence $T(\cdot)$.

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

Example: choice under risk Consider individuals making choices from menus of two lotteries over $J > 1$ monetary payoffs (e.g., Allais, 1953; Kahneman and Tversky, 1979; Erev et al., 2010, 2017; Peysakhovich and Naecker, 2017; Peterson et al., 2021, among many others). 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), the ordering of lotteries in the menu, or measures of the lottery's complexity (e.g., Enke and Graeber, 2023). 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 a modeled dataset D, expected utility theory searches for utility functions $u(\cdot)$ that "rationalize" the lottery choice probabilities, meaning $y^* \in \arg\max_{k \in \{0,1\}} \sum_{j=1}^{J} p_k(j) u(z_k(j))$

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 example, Harless and Camerer (1994) consider expected utility theory with idiosyncratic errors, which searches for utility functions $u(\cdot)$ and idiosyncratic error rate $\epsilon \in [0, 0.5]$ 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 $(x, y^*) \in D$. 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 and Shubatt (2023) consider expected utility theory with complexity-dependent noise. ▲

Example: play in normal-form games Consider individuals playing $J \times J$ normal-form games (e.g., Costa-Gomes, Crawford and Broseta, 2001; Wright and Leyton-Brown, 2010; Crawford, Costa-Gomes and Iriberri, 2013; Hartford, Wright and Leyton-Brown, 2016; Wright and Leyton-Brown, 2017; Hirasawa, Kandori and Matsushita, 2022, among many others). Let $\{1, \ldots, J\}$ denote the actions available to the row and column players, $\pi_{row}(j, j')$, $\pi_{col}(j, j')$ denote the payoff to the row player and column player respectively from action profile (j, j'). 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 a modeled dataset D, Nash equilibrium returns $T(x; D)$ satisfying $y^* = T(x; D)$ 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 $\sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} y^*(j)y^*_{col}(\tilde{j})\pi_{row}(j, \tilde{j}) \geq \sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} \tilde{y}^*(j)y^*_{col}(\tilde{j})\pi_{row}(j, \tilde{j})$ for all $\tilde{y}^* \in \Delta^{J-1}$ and $\sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} y^*(j)y^*_{col}(\tilde{j})\pi_{col}(j, \tilde{j}) \geq \sum_{j=1}^{J}\sum_{\tilde{j}=1}^{J} y^*(j)\tilde{y}^*(j)\pi_{col}(j, \tilde{j})$ for all $\tilde{y}^* \in \Delta^{J-1}$. Alternatively, for example, "level-0" behavior defined in Stahl and Wilson (1995) is a theory $T(\cdot)$ satisfying $y^* = T(x; D)$ for all $(x, y^*) \in D$ and $T(x; D) = (1/J, \ldots, 1/J)$ if and only if $y^* = (1/J, \ldots, 1/J)$ for all $(x, y^*) \in D$. Alternative theories of strategic behavior such as level-k behavior, the Poisson cognitive hierarchy model (Camerer, Ho and Chong, 2004), and level-$k(\alpha)$ behavior (Fudenberg and Liang, 2019) can also be similarly cast as particular theories $T(\cdot)$. ▲

2.2 Incompatible datasets and logical anomalies

A modeled dataset is incompatible with a theory $T(\cdot)$ if its underlying structure cannot accommodate the configuration of features and outcomes. Otherwise, a modeled dataset is

compatible with theory $T(\cdot)$.

Definition 2. A modeled dataset $D \in \mathcal{D}$ is

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

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 incompatible dataset. Researchers like Allais are not simply interested in characterizing all possible incompatible datasets of a theory. But rather they construct logical anomalies, which are incompatible datasets that satisfy an additional property and we define next.

Definition 3. A modeled dataset $D \in \mathcal{D}$ 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 "minimal" incompatible dataset in the sense that $T(\cdot)$ is compatible with any of its subsets. Consider again the Allais Paradox for expected utility theory in Table 1. We discuss logical anomalies for our other examples in Appendix C. The Allais Paradox is a particular modeled dataset that consists of the two menus x_A, x_B and associated modeled outcomes $y_A^* = 0, y_B^* = 1$. The independence axiom of expected utility theory implies that the choice on menu x_A determines the choice on menu x_B and vice versa – that is, $T(x_A; D) = T(x_B; D)$ for any $D \in \mathcal{D}$. The Allais Paradox is therefore an incompatible dataset for expected utility theory. Furthermore, any single choice (x_A, y_A^*) or (x_B, y_B^*) is compatible with expected utility theory, and so the Allais Paradox further satisfies Definition 3.

Whether a particular modeled dataset is a logical anomaly depends on the researcher's exact specification of theory $T(\cdot)$. As a simple example, a single observation (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, such as Harless and Camerer (1994); McGranaghan et al. (Forthcoming); Enke and Shubatt (2023).

2.3 Representation result and existence of logical anomalies

We next 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 modeled dataset. Assumption 2 states that whenever $T(\cdot)$ is compatible with a modeled dataset, it is consistent with all of its observations. Assumption 3 states that the theory can only refine its implications as more observations are collected. Finally, Assumption 4 states that there exists some modeled dataset and unseen feature at which theory $T(\cdot)$ derives non-trivial implications.

All of our previous examples of economic theories satisfy these assumptions. Consider expected utility theory. Appendix C discusses our other examples. First, expected utility theory satisfies Assumption 1 and Assumption 2. For any modeled dataset 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 $\alpha \in [0, 1)$.[6] Expected utility theory therefore satisfies Assumption 4.

For any theory $T(\cdot)$ satisfying Assumptions 1-4, we establish that there exists logical anomalies and it 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 modeled dataset $D \in \mathcal{D}$ if $f(x) = y^*$ for all $(x, y^*) \in D$. Modeled dataset 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 modeled dataset and satisfies, for all $x \in \mathcal{X}$ and $D \in \mathcal{D}$,

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

[6] 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 $(\alpha p + (1-\alpha)p', \alpha z + (1-\alpha)z')$.

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 any given modeled dataset $D \in \mathcal{D}$. At this level of abstraction, however, the functions in the allowable function class need not have any relationship to one another. Furthermore, the theory is not compatible with all possible datasets – in fact, there exists 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 of its scientific domain or mathematical structure, Proposition 2.1 serves as the launching point of our subsequent analysis.

We provide the complete proof in Appendix B but 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 dataset $D_{min} \in \mathcal{D}$ that is inconsistent with $\mathcal{F}^T$ (i.e., the fewest number of observations). 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 the collection of all incompatible datasets for $T(\cdot)$, which is non-empty by Assumption 4. We define $\mathcal{F}^{\neg T}$ to be the collection of all mappings that are consistent with any incompatible dataset $D \in \mathcal{D}^{\neg T}$. 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) at all $D \in \mathcal{D}$ and $x \in \mathcal{X}$. This proves part (i). To show part (ii), we establish that there exists a smallest incompatible dataset for theory $T(\cdot)$ and this must also be an anomaly by Definition 3.

Incompatible datasets and logical anomalies have a simple characterization in terms of a theory $T(\cdot)$'s allowable functions.

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 $y = y'$. Then,*

i. Modeled dataset $D \in \mathcal{D}$ 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 datasets of size strictly less than $n > 1$, then any incom-

patible dataset of size n is also an anomaly.

If given access to theory $T(\cdot)$'s allowable functions, searching for incompatible datasets is equivalent to searching for modeled datasets 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 datasets. This characterization of incompatible datasets (5) can be interpreted as an adversarial game between the theory (the min-player) and a falsifier. The falsifier proposes modeled datasets 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 over the modeled dataset, and the falsifier wishes to search for incompatible datasets that induce a positive loss for the theory's best-responding allowable function. We next 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. Harless and Camerer (1994) measures the predictive success of alternative theories for decision-making under uncertainty over three pairs of lotteries and proposes methods for aggregating evidence of predictive success across experiments. See also Beatty and Crawford (2011) for an application to consumer demand. Fudenberg, Gao and Liang (2020) measure the "restrictiveness" of economic theories, which generalizes Selten's definition. Our existence result for anomalies establishes that any black-box theory satisfying our axiomatization must be restrictive in the sense that there exist some minimal hypothetical datasets that it cannot explain.

2.4 Observable data and empirical anomalies

To this point, we analyzed the behavior of theory $T(\cdot)$ on modeled datasets $D \in \mathcal{D}$. Our goal is to ultimately contrast theory $T(\cdot)$ with nature in order to generate hypotheses about how it may be improved empirically.

We suppose 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 some observed outcome. Our main assumption is the observed outcome is statistically related to the theory's modeled outcome Y_i. We define the *empirical* modeled outcome of theory $T(\cdot)$ as

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

for some researcher-specified function $g(\cdot)$, where $\mathbb{E}_m[\cdot]$ denotes the expectation under $P_m(\cdot)$. The empirical modeled outcome of theory $T(\cdot)$ is some identified functional of each modeled context's underlying joint distribution.

At first glance, it may seem odd to label this an assumption since it is the starting point of much theoretically-motivated empirical work. 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. Yet 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 structure that the theory purports to model. We view this as a desirable attribute of our framework. The choice of how to map the observable data onto the theory's modeled datasets is an important input by the researcher.

For the rest of the paper, our goal will be to discover candidate *empirical anomalies* for theory $T(\cdot)$ in modeled context m, if they exist. Given modeled context $m \in \mathcal{M}$ with true function $f_m^*(\cdot)$, we search for empirical modeled datasets $D = \{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$ that are logical anomalies for theory $T(\cdot)$.

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 "average" empirical anomalies across multiple modeled contexts.

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 data that the theory seeks to explain.

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

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

be theory $T(\cdot)$'s loss over its allowable functions on the empirical modeled dataset $D = \{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$. Proposition 2.2 establishes D is incompatible with $T(\cdot)$ if and only if $\mathcal{E}_m^T(x_{1:n}) > 0$. Furthermore, it is also an empirical anomaly in modeled context m if there exists no smaller empirical dataset 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 empirical modeled datasets that are incompatible with $T(\cdot)$, or equivalently feature vectors $x_{1:n}$ satisfying $\mathcal{E}_m^T(x_{1:n}) > 0$; and second, iterating that search over successively larger dataset sizes n.

For any empirical dataset of size $n \geq 1$, we can directly optimize 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 datasets that generate large positive loss for the theory's best-responding allowable function (if they exist). We propose an iterative search for empirical anomalies based on this max-min program. For some maximal dataset size $\overline{n} \geq 1$, we iterate over $n = 1, \ldots, \overline{n}$ and solve the adversarial game (5), letting n^* denote the smallest dataset size for which the optimal value of the max-min program is strictly positive. Any empirical dataset $x_{1:n^*}$ with $\mathcal{E}_m^T(x_{1:n^*}) > 0$ 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 $\{x_{1:n^*} : \mathcal{E}_m^T(x_{1:n^*}) > 0\}$.

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 tackle both of these challenges and construct a feasible search procedure for empirical anomalies.

3.1 Statistical analysis of plug-in max-min optimization

Recall the true function $f_m^*(\cdot)$ is some identified functional of the joint distribution of the observable data in modeled context m – that is, $f_m^*(x) = \mathbb{E}_m[g(Y_i) \mid X_i = x]$ 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 construct an estimator $\widehat{f}_m^*(\cdot) \in \mathcal{F}$. 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 small 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 $f(\cdot; x_{1:n})$ 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

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

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 $n \geq 1$,*

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

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 in finite samples 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 error of $\widehat{f}_m^*(\cdot)$ for $f_m^*(\cdot)$ as measured by the supremum

norm ("sup-norm"). 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 $x \to f_m^*(x)$ in searching for incompatible datasets.

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.

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). It can be reinterpreted as a "data-poisoning" attack on the theory's allowable functions $\mathcal{F}^T$. Typical data-poisoning attacks fix a prediction function (e.g., an estimated neural network for image classification) and evaluate its worst-case empirical loss over a family of data perturbations that manipulate the features but leave the outcome fixed (e.g., manipulations of particular pixel values). In the plug-in max-min program (6), the theory moves after the falsifier, and so the falsifier must search for empirical datasets that simultaneously "poison" the performance of all allowable functions $f(\cdot) \in \mathcal{F}^T$. The falsifier's manipulation of the features therefore induces both variation in the theory's chosen allowable function and the true function $f_m^*(\cdot)$, making the outer maximization program difficult. We nonetheless exploit this connection to adversarial learning, using recent results on non-convex/concave max-min optimization (e.g., Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020) to develop a feasible gradient descent ascent (GDA)

optimization routine.

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 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. Therefore, we can 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 $\widehat{\mathcal{E}}_m^T(x_{1:n}, \theta) := n^{-1} \sum_{i=1}^{n} \ell \left(f_\theta(x_i), \widehat{f}_m^*(x_i) \right)$ and we assume $\widehat{\mathcal{E}}_m^T(x_{1:n}, \theta)$ is differentiable in $x_{1:n}$ for all $\theta \in \Theta$. For a collection of initial feature values $x_{1:n}^0$, some chosen step size sequence $\eta_t > 0$ and maximum number of iterations $T > 0$, we iterate over $t = 0, \ldots, T$ and calculate at each iteration

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

$$x_{1:n}^{t+1} = x_{1:n}^t + \eta_t \nabla \widehat{\mathcal{E}}_m^T(x_{1:n}^t; \theta^{t+1}). \tag{13}$$

At each step t, we construct an approximate solution to the inner minimization problem θ^{t+1}, and we then take a gradient ascent step on the feature values plugging in θ^{t+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 $\nabla \widehat{\mathcal{E}}_m^T(x_{1:n}, \theta) \approx 0$ 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: $\widehat{f}^*_m(\cdot)$, dataset size n, maximum iterations T, step size sequence η_t, initial feature vector $x^0_{1:n}$.

$t \leftarrow 0$;
while $t < T$ **do**
  $\theta^{t+1} \leftarrow \arg\min_{\theta \in \Theta} \widehat{\mathcal{E}}^T_m(x^t_{1:n}; \theta)$;
  $x^{t+1}_{1:n} \leftarrow x^t_{1:n} + \eta_t \nabla \widehat{\mathcal{E}}^T_m(x^t_{1:n}; \theta^{t+1})$;
  $t \leftarrow t + 1$;
return $\{(x^T_1, \widehat{f}^*_m(x^T_1)), \ldots, (x^T_n, \widehat{f}^*_m(x^T_n))\}$.

4 Representational Anomalies and Dataset Morphing

Our adversarial algorithm for anomaly generation exploits no structure about theory $T(\cdot)$ beyond its allowable functions. If a strengthened Assumption 4 ("non-trivial implications") 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 systematically vary. We propose a dataset morphing algorithm to generate such representational anomalies.

4.1 Representational equivalence and anomalies

To this point, we modeled theory $T(\cdot)$ as a reduced-form mapping that draws implications about the relationship between the features and modeled outcomes from any hypothetical dataset, 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 of terms of a theory's allowable functions – all allowable functions assign the same modeled outcome value to the two features.

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

Assumption 6 (Sharp implications)**.** There exists $x_1, x_2 \in \mathcal{X}$ 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 modeled dataset $D \in \mathcal{D}$ 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 potential observation (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 $T(\cdot)$. To make this more concrete, we return to some of our earlier examples to illustrate that Assumption 6 is often satisfied by leading economic theories.

Example: choice under risk Consider again individuals making choices from menus of two lotteries over $J > 1$ monetary payoffs and expected utility theory. Any utility function $u(\cdot)$ is associated with an allowable function $f(\cdot) \in \mathcal{F}^T$ under expected utility theory given by $f(x) = \arg\max \left\{ \sum_{j=1}^J p_{0j} u(z_{0j}), \sum_{j=1}^J p_{1j} u(z_{1j}) \right\}$ for menu $x_1 = (p_0, z_0, p_1, z_1)$. For any menu x_2 that consists of the compound lotteries $\alpha(p_0, z_0) + (1-\alpha)(\widetilde{p}, \widetilde{z})$ and $\alpha(p_1, z_1) + (1-\alpha)(\widetilde{p}, \widetilde{z})$, $f(x_1) = f(x_2)$ due to the linearity in probabilities of expected utility theory. Expected utility theory therefore satisfies Assumption 6. Proposition 4.2 implies that any pair of menus x_1, x_2 of this form are representationally equivalent under expected utility theory. ▲

If theory $T(\cdot)$ has a non-trivial representation of the features, then all logical anomalies for theory $T(\cdot)$ can be classified into two categories.

Observation 4.1. Consider any theory $T(\cdot)$ satisfying Assumptions 1, 2, 3 and 6. Any logical anomaly D for theory $T(\cdot)$ satisfies either

i. There exists $(x_1, y_1^*), (x_2, y_2^*) \in D$ such that x_1, x_2 are representationally equivalent under $T(\cdot)$ and $y_1^* \neq y_2^*$.

ii. There exists no pair $(x_1, y_1^*), (x_2, y_2^*) \in D$ such that x_1, x_2 are representationally equivalent.

We refer to anomalies satisfying Observation 4.1(i) 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 typically most interested in uncovering representational anomalies for theories as many classic examples of anomalies fall into this category. 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 hypothetical datasets 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. Indeed, other famous examples in decision-making under risk such as the Certainty Effect or Common Ratio Effect (e.g., Allais, 1953; Kahneman and Tversky, 1979) are also representational anomalies.

4.2 Dataset morphing for representational anomalies

Given modeled contexts $m \in \mathcal{M}$ with true functions $f_m^*(\cdot) = \mathbb{E}_m[g(Y_i) \mid X_i = x]$ for some researcher-specified $g(\cdot)$, we now 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 that 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 $\lambda \in (0, 1)$.

That is, given that two features $x_1, x_2 \in \mathcal{X}$ are representationally equivalent, any feature in their convex hull is also representationally equivalent. 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 $\mathcal{N}^T(x) = \{v \in \mathbb{R}^{dim(x)} \colon \nabla f(x)'v = 0 \text{ for all } f(\cdot) \in \mathcal{F}^T\}$ denote the subspace of directions that are orthogonal to the gradient of each allowable function. Under Assumption 7, $\mathcal{N}^T(x)$ is non-empty at any x for which there exists some representationally equivalent x'. For an initial feature value x^0, step size sequence η_t, and maximum number of iterations, we would iterate over $t = 0, \ldots, T$ and compute the update step

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

where $\text{Proj}\,(\cdot)$ is the projection operator and $\text{Proj}\left(\nabla f_m^*(x) \mid \mathcal{N}^T(x)\right)$ is the projection of the gradient of the true function $f_m^*(\cdot)$ onto the null space of the allowable functions. We therefore update 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.

This is, of course, not feasible since we do not directly 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-in this estimator into the morphing procedure and apply the update step

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

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

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

While Proposition 4.3 analyzes the statistical properties of plugging in the estimated gradient of the true function into the morphing procedure, it still leaves open the question of how to practically implement the projection operator.

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 then suggest to 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 the gradients $\nabla f_\theta(x)$. As B grows large, this better approximates the null space of the allowable function $\mathcal{N}^T(x)$. Algorithm 2 summarizes our practical implementation of the morphing procedure, which can be run over many randomly initialized feature values x^0.

Algorithm 2: Feasible dataset morphing for representational anomalies.

Input: $\nabla \widehat{f}_m^*(\cdot)$, $B > 0$, maximum iterations T, step size η, initial feature x^0.

$t \leftarrow 0$;

while $t < T$ **do**

  Sample $\theta_b \in \Theta$ for $b = 1, \ldots, B$;

  Construct $\mathcal{N}_\Theta^T(x^t) = \{v \in \mathbb{R}^{dim(x)} \text{ s.t. } \nabla f_{\theta_b}(x_0)^T v = 0 \text{ for all } b\}$;

  $x^{t+1} \leftarrow x^t - \eta \text{Proj}\left(\nabla \widehat{f}_m^*(x^t) \mid \mathcal{N}^T(x^t)\right)$;

  $t \leftarrow t + 1$;

return $\{(x^0, \widehat{f}_m^*(x^0)), (x^T, \widehat{f}_m^*(x^T))\}$.

5 Illustrative example: Generating Anomalies for Choice under Risk

In this section, we illustrate our procedures by generating logical anomalies for expected utility theory in simulated lottery choice data from individuals whose preferences are consistent with cumulative prospect theory. That is, we imagine ourselves in the 1950s, having access to the formal model of expected utility theory and wishing to generate candidate empirical anomalies in a hypothesized world where individuals make choices according to cumulative prospect theory. Since cumulative prospect theory has been well-studied by theorists, we compare and contrast the logical anomalies generated by our procedures against known anomalies for expected utility theory, such as those produced in Allais (1953), Kahneman and Tversky (1979), and others.

Our anomaly generation procedures recover known logical anomalies for expected utility theory based on the probability weighting function. Intriguingly, our procedures also uncover *novel* logical anomalies for expected utility theory that either generalize or differ from those that originally spurred the development of cumulative prospect theory. In an incentivized online experiment, participants' choices on our algorithmically generated, logical anomalies violated expected utility theory at similar rates as found in recent analyses of the Allais Paradox and Common Ratio Effect.

5.1 Simulation 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 curvature and level 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 probabilities close to one, and compressing intermediate probabilities. Figure 1 plots the resulting probability weighting functions associated with each choice of parameter values (δ, γ). Such non-linearity in the probability weighting function can generate several known logical anomalies for expected utility theory, such as the Allais Paradox (Table 1) or the Common Ratio Effect. 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., $\sum_{j=1}^{J} \pi_j(p; \delta, \gamma) > 1$). Such properties in the probability weighting function may lead the individual's choices to violate first-order stochastic dominance, meaning the individual may 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 algorithm (Algorithm 1) and our dataset morphing procedure (Algorithm 2) to the true choice probability function $f_m^*(\cdot)$. In Appendix F.4, we also 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 also 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 dataset 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). We report all resulting, numerically verified logical anomalies for expected utility theory.

5.2 Logical anomalies generated by the probability weighting function

Table 2 summarizes the logical anomalies for expected utility theory that are produced by our anomaly generation procedures at each calibrated parameter value (δ, γ). Our anomaly generation procedures uncover several distinct categories of logical anomalies.

5.2.1 First order stochastic dominance violations

All logical anomalies in the first row of Table 2 are menus of lotteries in which the individual's choice violates first-order stochastic dominance. As we show in the examples in Table 3, 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.[7] What is intriguing is that our anomaly generation procedures uncover these first-order stochastic dominance violations on its own.

5.2.2 The dominated consequence effect

The logical anomalies in the second row of Table 2 highlight what we refer to as a "dominated consequence effect." These are logical anomalies for expected utility theory that arise due to the non-linearity of the probability weighting function and are violations of the independence axiom. In Table 4, we provide three representative examples of pairs of menus of lotteries that

[7] Indeed, Kahneman and Tversky (1979) include an "editing phase" prior to choice that eliminates such first-order stochastic dominated lotteries prior. We refer the reader to Lattimore, Baker and Witte (1992); Wu and Gonzalez (1996) for further discussion.

were produced by our anomaly generation procedures and exhibit the dominated consequence effect.

To make this more concrete, consider the pair of menus in Table 4(a) generated by our dataset morphing algorithm for the individual with calibrated parameter values $(\delta, \gamma) = (0.726, 0309)$. 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 0.70; that is, $B0 = \alpha_0 A0 + (1 - \alpha_0)\delta_{0.70}$ for some $\alpha_0 \in (0, 1)$. Analogously, lottery B1 can be written as the compound lottery $B1 = \alpha_1 A1 + (1 - \alpha_1)\delta_{0.23}$ for some $\alpha_1 < \alpha_0$. The individual's choices therefore express that lottery $A0$ is preferred to lottery $A1$ and $\alpha_1 A1 + (1 - \alpha_1)\delta_{0.23}$ is preferred to $\alpha_0 A0 + (1 - \alpha_0)\delta_{0.70}$. This, however, contradicts the independence axiom of expected utility theory since it can be shown that $A0$ being preferred to $A1$ must imply that $\alpha_0 A0 + (1-\alpha_0)\delta_{0.70}$ is preferred to $\alpha_1 A1 + (1-\alpha_1)\delta_{0.23}$. We provide a formal proof in Appendix F.3.

More generally, all of the logical anomalies for expected utility theory in the second row of Table 2 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) 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 on 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 that exist over pairs of menus of two lotteries over two monetary payoffs. Our anomaly generation procedures uncovered this category of logical anomalies on its own.

5.2.3 The reverse dominated consequence effect and the strict dominance effect

In the third row of Table 2, all logical anomalies exhibit what we call a "reverse dominated consequence effect." We provide three illustrative examples in Table 5. Each of these logical anomalies have 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 highlight 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.

Finally, all logical anomalies in the fourth row of Table 2 exhibit what we call a "strict dominance effect," and we provide three illustrative examples in Table 6. 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 $\alpha_1\ell_1+(1-\alpha_1)\delta_{\overline{z}_1}$ 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 $\alpha_1\ell_1 + (1-\alpha_1)\delta_{\overline{z}_1}$ 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 highlight 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 in the first place. 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 that to our knowledge have not been noticed before.

5.3 Experimental test of algorithmically generated anomalies

Our anomaly generation 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 the anomaly generation process ends, and careful experimental work begins. While fully investigating their experimental robustness is beyond the scope of this paper, we next present 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 of the algorithmically generated logical anomalies for expected utility theory summarized in Table 2. 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. As a result, 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 pre-registered 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. 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. Altogether we recruited 258 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.

5.3.2 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. In Appendix A, we report the same results, dropping the top 10% of respondents who completed the surveys the fastest and finding similar results.

Figure 2 reports the fraction of respondents whose choices violate expected utility theory without noise 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. Appendix Figure A1 and Appendix Table A1 report the same estimates, organized by the calibrated parameter values (δ, γ) that we considered. 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.

Furthermore, these pooled estimates mask heterogeneity in the fraction of respondents' violating expected utility theory across logical anomalies. For example, we find that greater than 15% of respondents' choices violate expected utility theory on several dominated consequence 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 35 out of 36. Respondents' choices are therefore inconsistent with expected utility theory on each algorithmically generated, logical anomaly included in our surveys.

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. We explore this possibility in two ways.

First, we 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. As an example, consider a logical

anomaly that exhibits the dominated consequence effect such as the pair of menus depicted in Table 4(a). 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. On any choice, a respondent may erroneously deviate from their true preference with probability $\epsilon \geq 0$. Following in the spirit of Harless and Camerer (1994), we assume a single error rate for all choices since it is a parsimonious way to summarize observed choice fractions.[8] 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 $P(A0, B0), P(A1, B1), P(A1, B0), P(B0, A1)$.[9] 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).

Proceeding in this manner, 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 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.

Second, we 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

[8] Recent work argues that it may be empirically relevant to allow for choice-specific error rates that depend on the intensity of the individual's preference (McGranaghan et al., Forthcoming) or the complexity of the menu of lotteries (Enke and Shubatt, 2023). For our purposes, this simple model of noise 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.

[9] 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 $P(A1, B1) = (1-\epsilon)^2\pi(A0, B0) + \epsilon(1-\epsilon)\pi(A1, B0) + (1-\epsilon)^2 P(A1, B1)$.

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 many binary choice experiments that exhaustively test the Common Experiments across different payoffs and probabilities.

Table 8 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.[10]

Based on these experimental findings, our algorithmically generated, logical anomalies yield expected utility theory violation rates that are in line with those observed for celebrated logical anomalies like the Allais Paradox and the Common Ratio Effect. Altogether, this suggests that these new categories of anomalies may merit the same rigorous testing across a wide variety of experimental designs that have been given to other known anomalies for expected utility theory.

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;

[10] 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.

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

In this paper, we ask whether machine learning can accelerate the development of new theories through the automatic generation of anomalies. To tackle this problem, we developed an econometric framework for anomaly generation. We then proposed two algorithmic procedures for anomaly generation, one based on adversarial learning and another based on dataset morphing, that take as inputs *any* formal theory and data from a scientific domain, summarize the empirical relationship between some features and modeled outcomes using supervised learning, and then automatically generate anomalies, if they exist. 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.

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., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *The Quarterly Journal of Economics*, 119(3): 861–898.

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.

Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta. 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica*, 69(5): 1193–1235.

Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri. 2013. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." *Journal of Economic Literature*, 51(1): 5–62.

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."

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.

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.

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.

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

Huck, Steffen, and Wieland Müller. 2012. "Allais for all: Revisiting the paradox in a large representative sample." *Journal of Risk and Uncertainty*, 44(3): 261–293.

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.

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.

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.

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.

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

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

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.

Stahl, Dale O., and Paul W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior*, 10(1): 218–254.

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.

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.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princeton:Princeton University Press.

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."

Figures

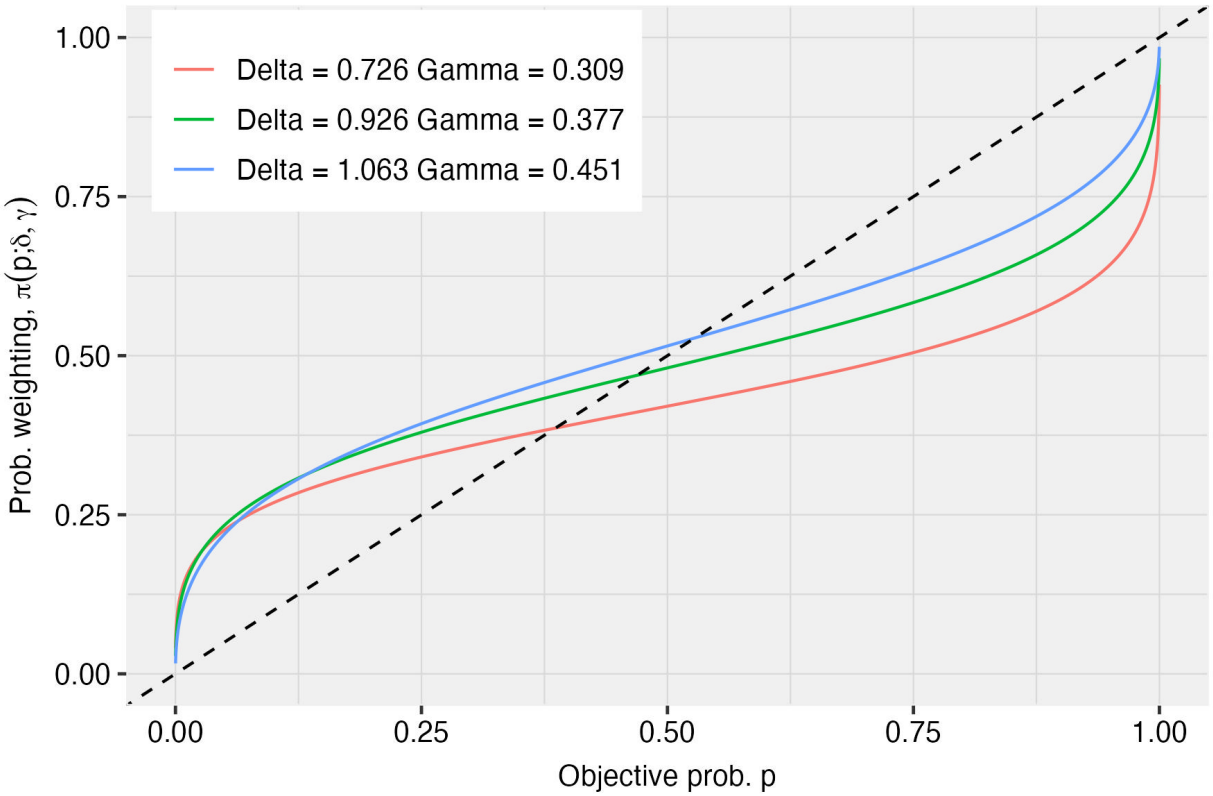


Figure 1: Probability weighting function for calibrated parameter values (δ, γ) in our illustration to choice under risk.

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.

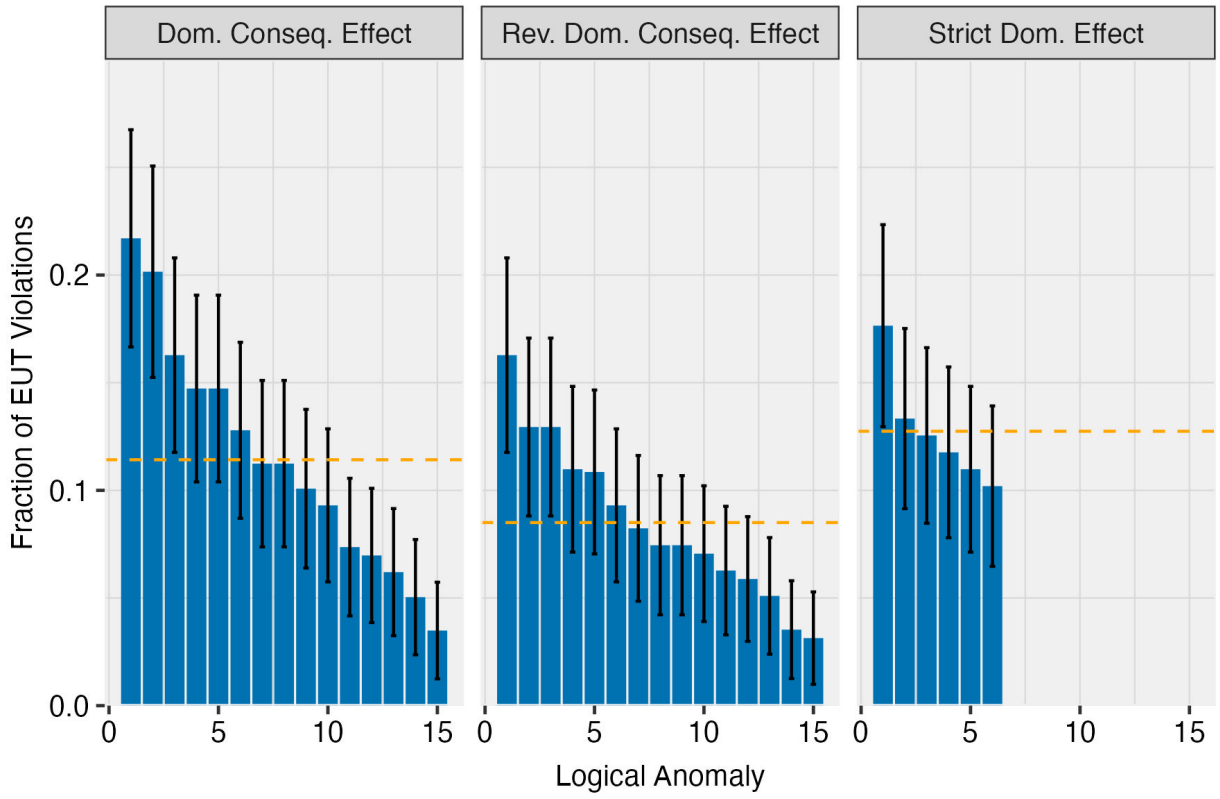


Figure 2: Fraction of respondents whose choices violate expected utility theory on algorithmically generated, logical anomalies.

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). We organize the estimates by category of logical anomaly (see Table 2). 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 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.

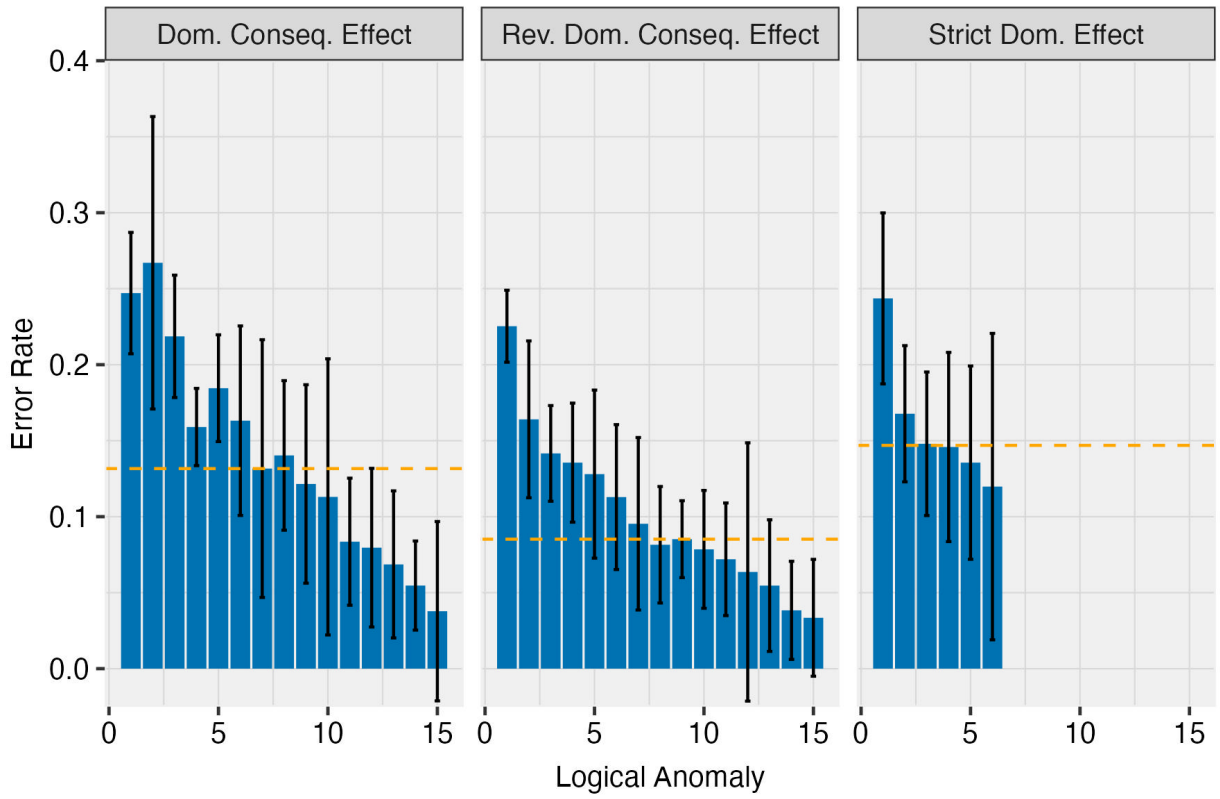


Figure 3: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated, logical anomalies.

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). We organize the estimates by category of logical anomaly (see Table 2). 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.

Tables

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

Table 2: 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 dataset moprhing 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.

(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%

Table 3: 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 $(\delta, \gamma) = (0.726, 0.309)$. Logical anomalies #1-2 are generated by our dataset morphing algorithm. Logical anomaly #3 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 point. See Section 5.2 for further discussion.

(a) Logical Anomaly #1

	Menu A (x_A, y_A^*)	
Lottery 0	0.70	5.96
	5%	95%
Lottery 1	0.23	7.48
	22%	78%
	Menu B (x_B, y_B^*)	
Lottery 0	0.70	5.96
	24%	76%
Lottery 1	0.23	7.48
	49%	51%

(b) Logical Anomaly #2

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

(c) Logical Anomaly #3

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

Table 4: 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 $\{(x_A, y_A^*), (x_B, y_B^*)\}$. Each algorithmically generated, logical anomalies depicted here is produced by our dataset morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$, logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$, and logical anomaly #3 on $(\delta, \gamma) = (1.063, 0.451)$. 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	2.59	8.87
	88%	12%
Lottery 1	3.51	8.65
	99%	1%
	Menu B (x_B, y_B^*)	
Lottery 0	2.59	8.87
	49%	51%
Lottery 1	3.51	8.65
	65%	35%

(b) Logical Anomaly #2

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

(c) Logical Anomaly #3

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

Table 5: 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 $\{(x_A, y_A^*), (x_B, y_B^*)\}$. Each algorithmically generated, logical anomalies depicted here is produced by our dataset morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$, logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$, and logical anomaly #3 on $(\delta, \gamma) = (1.063, 0.451)$. 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	6.71	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%

(b) Logical Anomaly #2

	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%

(c) Logical Anomaly #3

	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%

Table 6: 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 generated strict dominance effect anomaly $\{(x_A, y_A^*), (x_B, y_B^*)\}$ presented here is produced by our dataset morphing algorithm. Logical anomaly #1 is based on the probability weighting function $\pi(p; \delta, \gamma)$ for $(\delta, \gamma) = (0.726, 0.309)$, logical anomaly #2 on $(\delta, \gamma) = (0.926, 0.377)$, and logical anomaly #3 on $(\delta, \gamma) = (1.063, 0.451)$. 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.

	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 2). 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 8: 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.

From Predictive Algorithms to Automatic Generation of Anomalies

Online Appendix

Sendhil Mullainathan & Ashesh Rambachan

A Appendix Figures and Tables

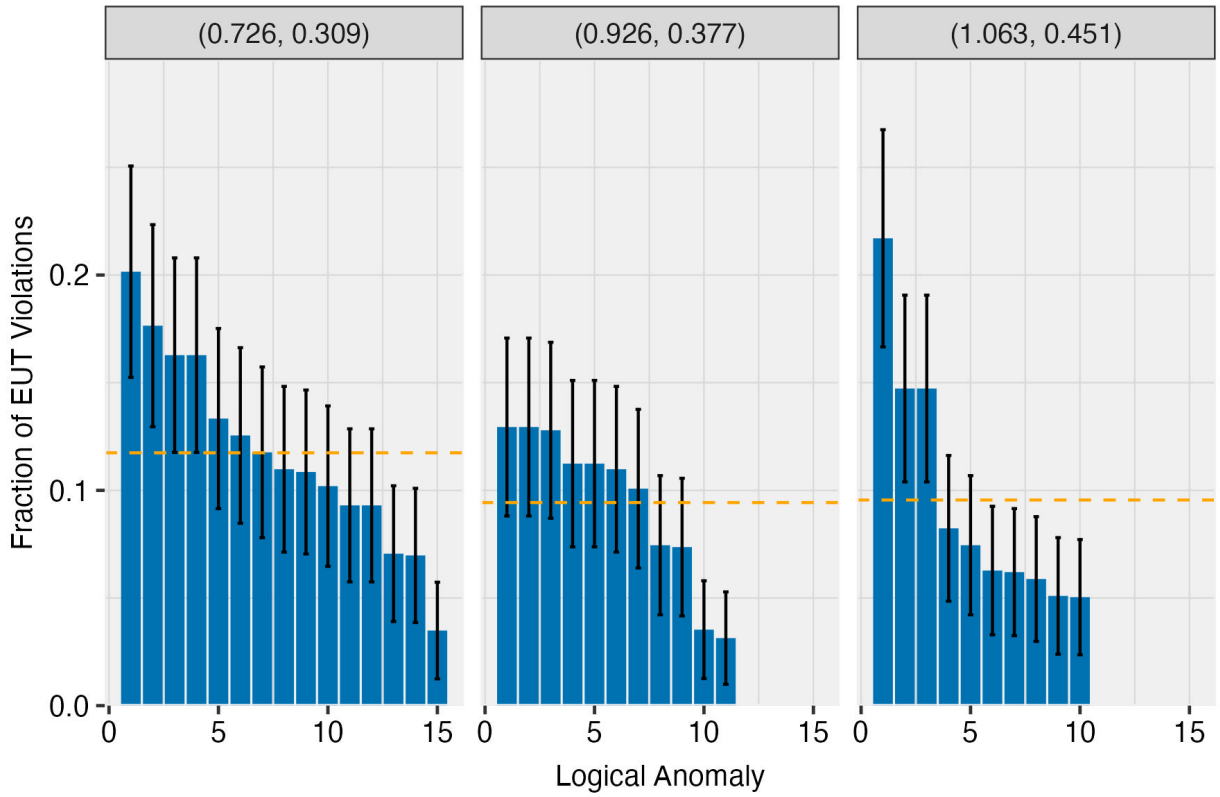


Figure A1: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, organized by calibrated parameter values (δ, γ).

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.

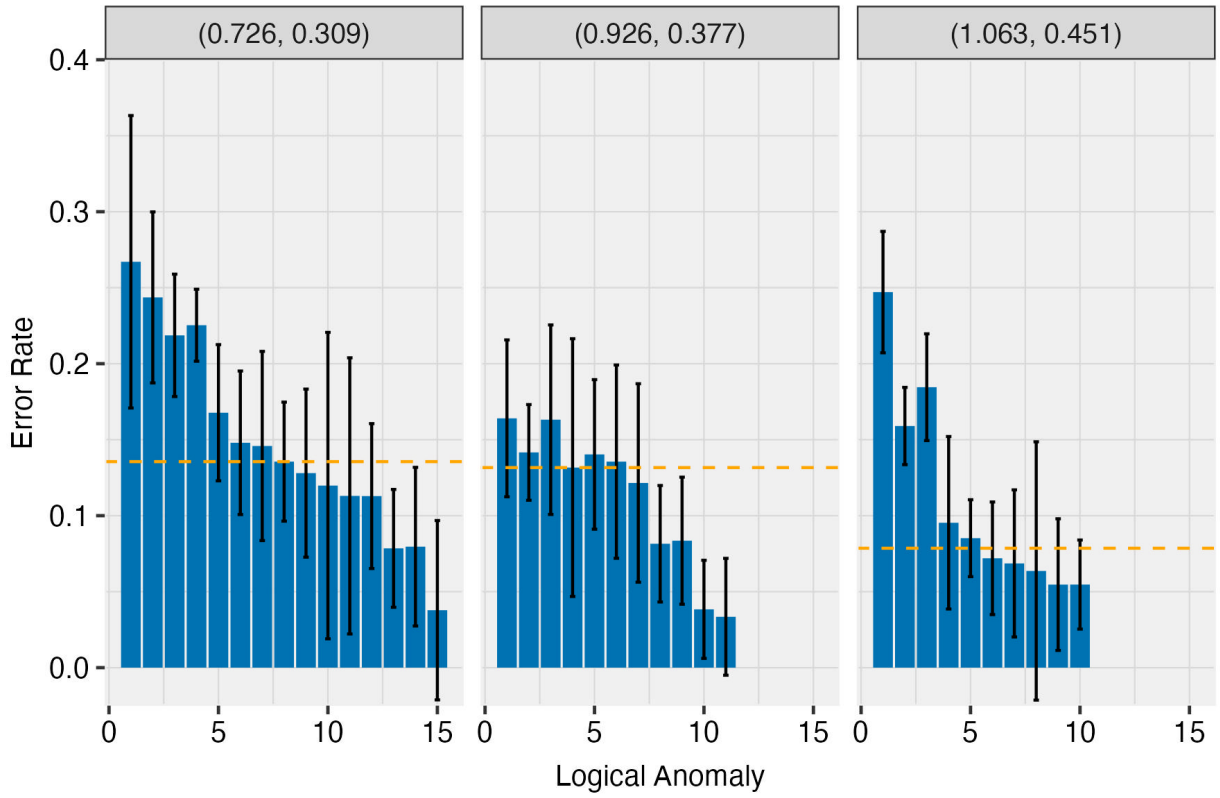


Figure A2: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated, logical anomalies, organized by calibrated parameter values (δ, γ).

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.

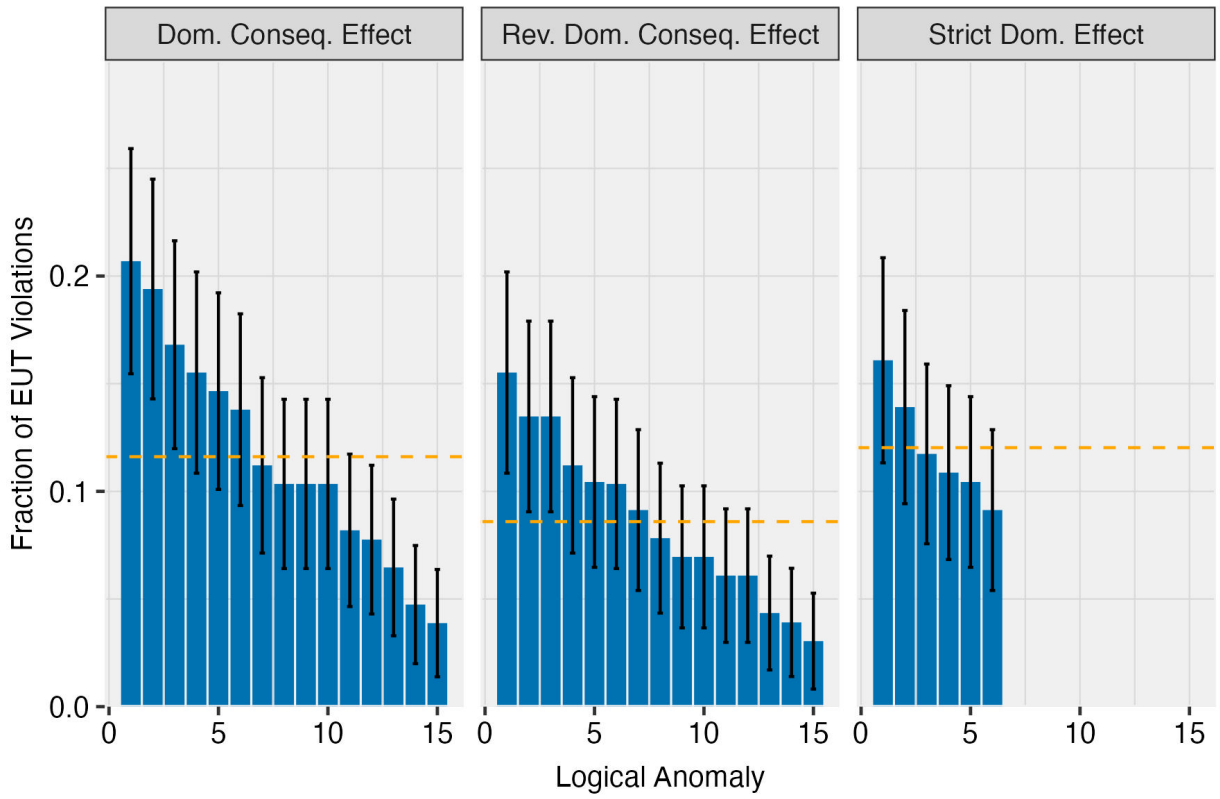


Figure A3: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, dropping the top 10% of respondents who completed the survey the fastest.

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 2). 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.

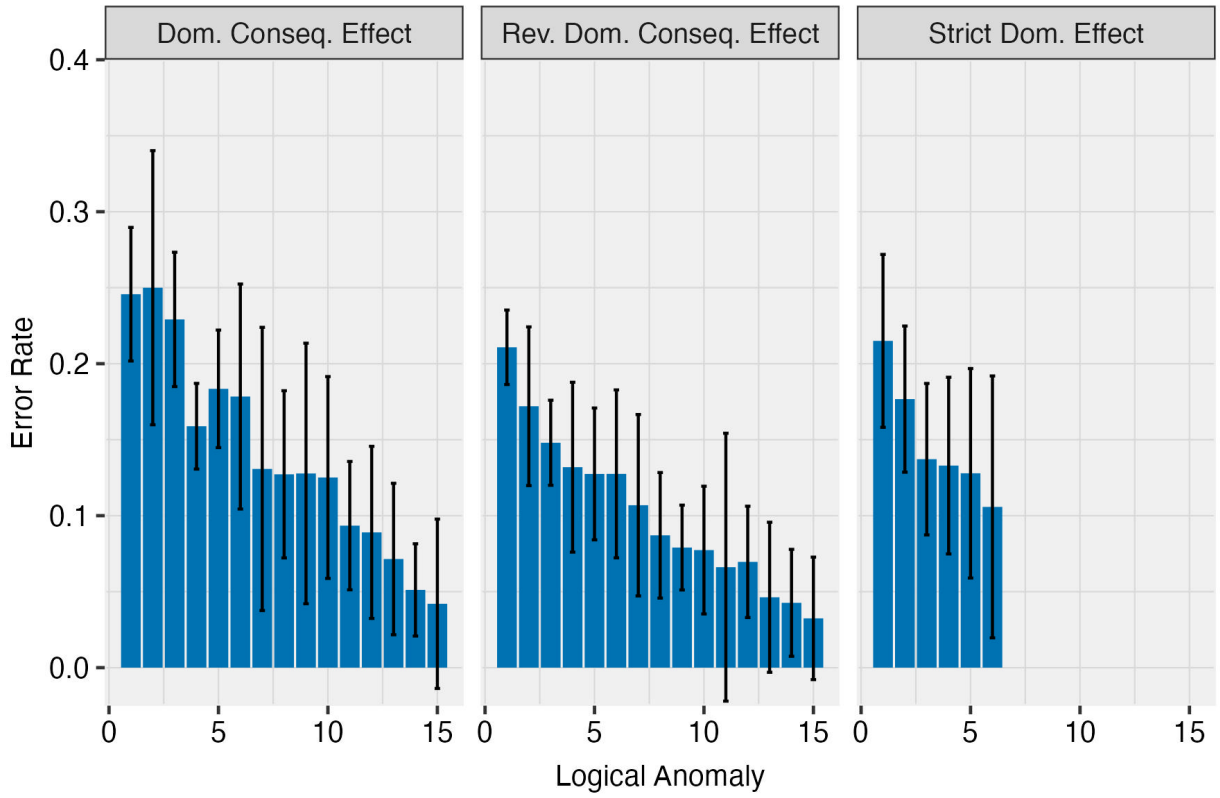


Figure A4: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated, logical anomalies, dropping the 10% of respondents that completed the survey the fastest.

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 2). 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.

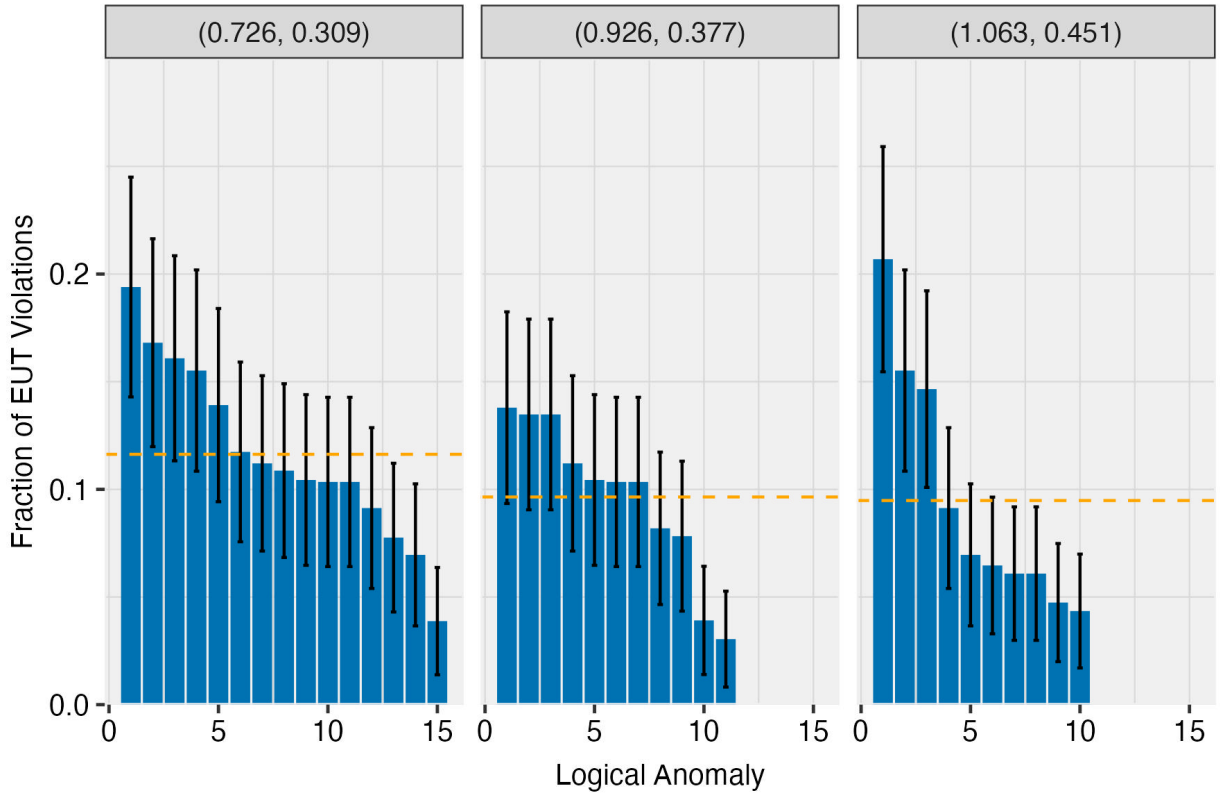


Figure A5: Fraction of respondents whose choices violate expected utility theory on logical anomalies of menus of two lotteries over two monetary payoffs, organized by the calibrated parameter values (δ, γ) and dropping the top 10% of respondents who completed the survey the fastest.

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.

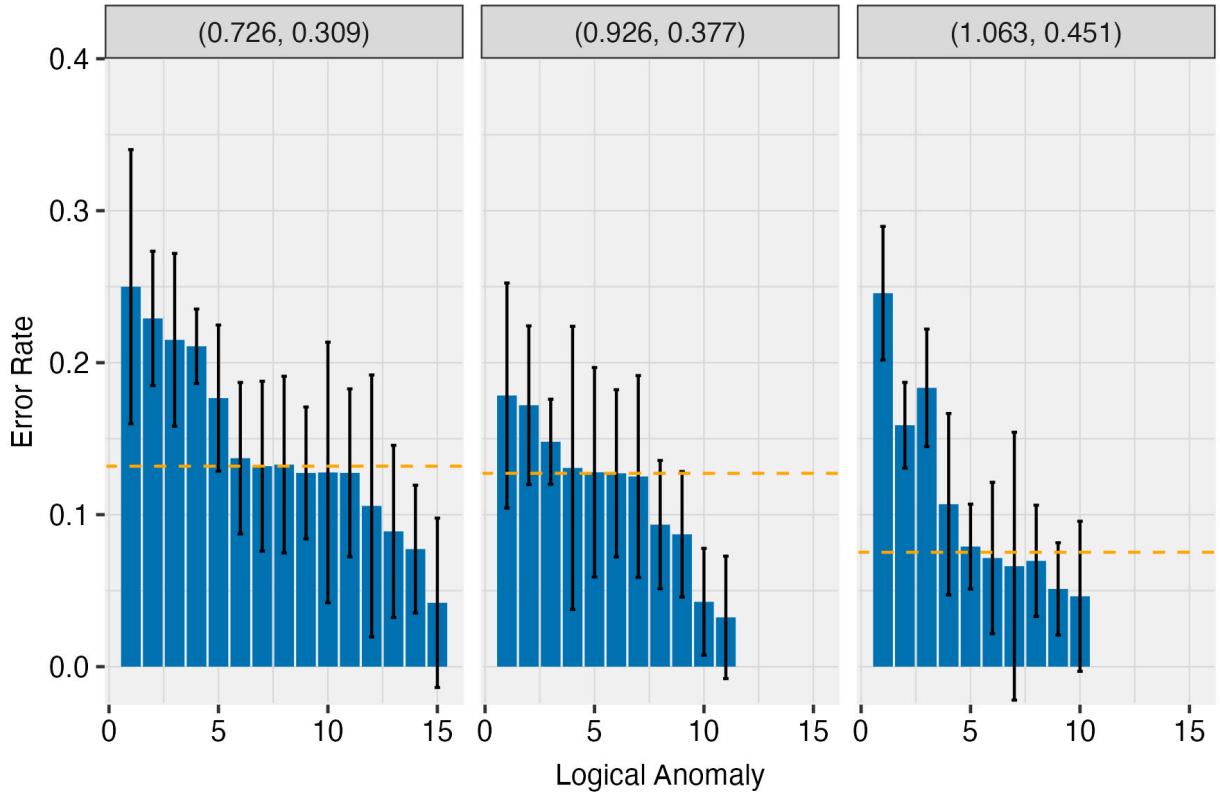


Figure A6: Estimated idiosyncratic error rate $\widehat{\epsilon}$ on algorithmically generated, logical anomalies, organized by calibrated parameter values (δ, γ) and dropping the 10% of respondents that completed the survey the fastest.

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.

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 A1: 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 2). 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.

B Omitted Proofs

B.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 falsifying datasets 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 a falsifying dataset for $T(\cdot)$. 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 dataset $D \in \mathcal{D}$ 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 a falsifying dataset of $T(\cdot)$, $D \notin \mathcal{D}^{\neg T}$ 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 dataset $D \in \mathcal{D}$ 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 a falsifying dataset for $T(\cdot)$ 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 dataset for $T(\cdot)$. Since there exists incompatible datasets, there must exist a smallest incompatible dataset $D \in \mathcal{D}$ for theory $T(\cdot)$. This must be an anomaly. If $|D| = 1$, then the definitions of an incompatible dataset and anaomly coincide. If $|D| > 1$ but $|D|$ is not an anomaly, then there exists a smaller incompatible dataset which is a contradiction. □.

B.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 an incompatible dataset for 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 an incompatible dataset by Proposition 2.1.

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

B.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,*

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

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 the 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. □

B.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 the modeled dataset $D = \{(x_1, f(x_1)), (x_2, f(x_2))\}$. Since $f(\cdot) \in \mathcal{F}^T$, $T(\cdot)$ must be consistent with this modeled dataset. 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. □

B.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 augemented dataset $\widetilde{D} = D \cup \{(x_1, y_1^*)\}$. By the allowable function representation (1), $\widetilde{D}$ is a compatible dataset. But Assumption 3 implies that $y_1^* \notin T(x_2; \widetilde{D})$, contradicting Assumption 6. □

B.6 Proof of Proposition 4.3

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

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

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

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

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

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

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. □

C Additional Examples for the Model of Theories

In this appendix section, we illustrate how additional examples map into our model of theories 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 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 a modeled dataset 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: 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 a modeled dataset 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 $y^* = \bar{y}_{risk\text{-}free} + \beta\left(\bar{y}_{market} - \bar{y}_{risk\text{-}free}\right)$. ▲

Example: 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 a hypothetical dataset 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 $y_j^* = P\left(j \in \arg\max_{\hat{j}} v_{\hat{j}}(x;\alpha,\beta)\right)$ 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 $y_j^* = P\left(j \in \arg\max_{\hat{j}} v_{\hat{j}}(x;\alpha,\beta)\right)$ for some (α, β) that matches D. ▲

C.1 Logical anomalies for other examples

Example: initial play in normal-form games Consider the normal-form game in Table A2. 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 hypothetical dataset $D \in \mathcal{D}$. Suppose instead the individual m was a level-1 thinker. In this case, she

	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 A2: An example anomaly for Nash equilibrium based on Level-1 thinking.

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

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 modeled dataset $D = \{(x_1, y_1^*), (x_2, y_2^*)\}$, where x_1, x_2 are such that the risk-free rate, market return and covariance structure are constant yet y_1^*, y_2^* vary. By construction, such a hypothetical dataset would be a logical anomaly for CAPM. For example, Barberis and Huang (2008) find that the skew (i.e., a higher moment) of an asset's returns influence asset returns in the cross-section. ▲

C.2 Assumptions for other examples

Example: initial 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 modeled datasets 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 inconsistent with CAPM, D is consistent with CAPM but D' is not, and both are consistent with CAPM in which case $\beta(D) = \beta(D')$. In all such cases, Assumption 3 is satisfied. Finally, Assumption 4 is satisfied for any modeled dataset 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 dataset 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 dataset, then it is also an incompatible dataset in some modeled context m. Furthermore, provided $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is a "systematically" incompatible dataset across modeled contexts, then it is also an average incompatible dataset.

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 dataset, 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 dataset.

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

$$\sum_{m \neq \tilde{m}} \left(n^{-1} \sum_{i=1}^{n} P(m \mid x) P(\tilde{m} \mid x) \left(f^T_m(x_i) - f^*_m(x_i) \right) \left(f^T_{\tilde{m}}(x_i) - f^*_{\tilde{m}}(x_i) \right) \right) \geq 0,$$

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

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 dataset, 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 dataset 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 these 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 in 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 a dataset 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 in 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 $\|\nabla\widehat{\bar{f}}^*(\cdot) - \nabla\bar{f}^*(\cdot)\|_2$. 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. Recall that for some parametrization of the theory's

allowable functions $\mathcal{F}^T = \{f_\theta(\cdot)\colon \theta \in \Theta\}$, initial feature values $x^0_{1:n}$, step size sequence $\eta_t > 0$ and maximum number of iterations $T > 0$, we iterate over $t = 0, \ldots, T$ and calculate

$$\theta^{t+1} = \arg\min_{\theta\in\Theta} \widehat{\mathcal{E}}^T_m(x^t_{1:n}; \theta)$$
$$x^{t+1}_{1:n} = x^t_{1:n} + \eta_t \nabla \widehat{\mathcal{E}}^T_m(x^t_{1:n}; \theta^{t+1})$$

at each iteration, where $\widehat{\mathcal{E}}^T_m(x_{1:n}, \theta) := n^{-1}\sum_{i=1}^n \ell\left(f_\theta(x_i), \widehat{f}^*_m(x_i)\right)$. 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 $\{x^0_{1:n}, \ldots, x^T_{1:n}\}$ and define $\widehat{\mathcal{E}}^T_m(x_{1:n}) = \min_{\theta\in\Theta} \widehat{\mathcal{E}}^T_m(x_{1:n}, \theta)$. To formally state the result, we define the *Moreau envelope* of $\widehat{\mathcal{E}}^T_m(x_{1:n})$ as

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

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 $\widehat{\mathcal{E}}^T_m(x_{1:n})$.

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

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

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 $\eta_t = \eta_0/\sqrt{T+1}$ for some $\eta_0 > 0$ satisfies*

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

where $\delta \ge 0$ 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 dataset 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 choice 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 $T = 50$. 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 $f_m^*(x^t)$ and the implied expected utility theory choice probabilities $f_\theta(x^t) = P\left(\sum_{j=1}^J p_{1j}^t u_\theta(z_{1j}) - \sum_{j=1}^J p_{0j} u_\theta(z_{0j}) + \xi\right)$ 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 $\widehat{\mathcal{E}}(x^t; \theta^{t+1})$ 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 $\log\left(\frac{f_m^*(x^t)}{1-f_m^*(x^t)}\right)\left(\sum_{j=1}^J p_{1j}^t u_{\theta^{t+1}}(z_{1j}) - \sum_{j=1}^J p_{0j} u_{\theta^{t+1}}(z_{0j})\right)$. This alternative loss function for the gradient ascent step applies the logit transformation to the choice probabilities so that $\log\left(\frac{f_m^*(x^t)}{1-f_m^*(x^t)}\right)$ is positive whenever $f_m^*(x^t) > 0.5$ and weakly negative otherwise. The overall loss function is therefore positive whenever the expected utility difference between the lotteries is positive but $f_m^*(x^t) < 0.5$ 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 Dataset morphing procedure

To implement the dataset 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 dataset morphing procedure 15,000 randomly initialized menus x^0. We set the maximum number of iterations to be $T = 50$. At each iteration t, we solve for the best-fitting allowable function θ^t and sample θ_b independent from a multivariate normal distribution with mean vector equal to $\bar{\theta}^t = \frac{1}{t}\sum_{s=1}^{t}\theta^s$ and variance matrix equal to $\frac{1}{t-1}\sum_{s=1}^{t}(\theta^s - \bar{\theta}^t)(\theta^s - \bar{\theta}^t)'$ 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 $(p_{0,0} - p_{0,1})^T u > 0$ and $(p_{1,0} - p_{1,1})^T u > 0$; (ii) $y_0^* = 1, y_1^* = 0$, we check whether there exists any vector u satisfying $(p_{0,0} - p_{0,1})^T u < 0$ and $(p_{1,0} - p_{1,1})^T u > 0$; 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

$$\ell_1 \succ \ell_0 \overset{(1)}{\implies} \alpha_1 \ell_1 + (1 - \alpha_1)\delta_{\underline{z}_1} \succ \alpha_1 \ell_0 + (1 - \alpha_1)\delta_{\underline{z}_1},$$
$$\alpha_1 \ell_0 + (1 - \alpha_1)\delta_{\underline{z}_1} \overset{(2)}{\succ} \alpha_1 \ell_0 + (1 - \alpha_1)\delta_{\underline{z}_0}$$
$$\alpha_1 \ell_0 + (1 - \alpha_1)\delta_{\underline{z}_0} \overset{(3)}{\succ} \alpha_0 \ell_0 + (1 - \alpha_0)\delta_{\underline{z}_0}$$

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 modeled dataset $\{((\ell_0, \ell_1), 1), ((\ell_0', \ell_1'), 0)\}$ is therefore a logical anomaly for expected utility theory.

In Table 4, we provide three examples of dominated consequence effect anomalies for expected utility theory. We now discuss how each example can be mapped into the dominated consequence effect. First, consider the logical anomaly presented in Table 4(a). This is a dominated consequence effect anomaly defining ℓ_1 as lottery $A0$, ℓ_0 as lottery $A1$, ℓ_1' as $B0$, and ℓ_0' as $B1$. Second, consider the logical anomaly in Table 4(b). This is a dominated consequence effect anomaly defining ℓ_1 as lottery $A1$, ℓ_0 as lottery $A0$, ℓ_1' as lottery $B1$, and ℓ_0' as lottery $B0$. Finally, consider the logical anomaly in Table 4(c). This is a dominated consequence effect anomaly defining ℓ_1 as lottery $A1$, ℓ_0 as lottery $A0$, ℓ_1' as lottery $B1$, and ℓ_0' as lottery $B0$.

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

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

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

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

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 modeled dataset $\{((\ell_0, \ell_1), 1), ((\ell'_0, \ell'_1), 0)\}$ is a logical anomaly for expected utility theory.

In Table 5, we provide three examples of reverse dominated consequence effect anomalies for expected utility theory. We now discuss how each example can be mapped into the reverse dominated consequence effect. First, consider the logical anomaly presented in Table 5(a). This is a reverse dominated consequence effect anomaly defining ℓ_1 as lottery $A0$, ℓ_0 as lottery $A1$, ℓ'_1 as lottery $B0$, and ℓ'_0 as lottery $B1$. Second, consider the logical anomaly in Table 5(b). This is a reverse dominated consequence effect anomaly defining ℓ_1 as lottery $A1$, ℓ_0 as lottery $A0$, ℓ'_1 as lottery $B0$, and ℓ'_0 as lottery $B1$. Finally, consider the logical anomaly in Table 5(c). This is a reverse dominated consequence effect anomaly defining ℓ_1 as lottery $A1$, ℓ_0 as lottery $A0$, ℓ'_1 as lottery $B1$, and ℓ'_0 as lottery $B0$.

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 modeled dataset $\{((\ell_0, \ell_1), 1), ((\ell_0', \ell_1'), 0)\}$ is a logical anomaly for expected utility theory.

In Table 6, we provide three examples of dominated consequence effect anomalies for expected utility theory. Each logical anomaly in Table 6 are strict dominance effect anomalies defining ℓ_1 as lottery $A1$, ℓ_0 as lottery $A0$, ℓ_1' as lottery $B1$, and ℓ_0' as lottery $B0$.

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 simulated dataset $\{(X_i, Y_i)\}_{i=1}^n$.

Using this simulated dataset, 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 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 dataset 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 A3 and Appendix Table A4 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	
First Order Stochastic Dominance	1	66	16	74	81
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
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	
First Order Stochastic Dominance	2	3	5	0	0
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
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	
First Order Stochastic Dominance	33	0	0	1	2
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
Other	7	0	0	0	1
# of Logical Anomalies	97	11	5	7	27

Table A3: 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 dataset 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 simulated dataset over $n = 1,000, 5,000, 10,000$ and $25,000$. For reference, the column "True Choice Prob." reproduces Table 2, 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	
First Order Stochastic Dominance	45	16	27	13	81
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
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	
First Order Stochastic Dominance	25	18	17	10	0
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
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	
First Order Stochastic Dominance	16	17	18	11	2
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
Other	3	0	3	4	1
# of Logical Anomalies	72	41	52	42	27

Table A4: 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 dataset 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 dataset over $n = 1,000, 5,000, 10,000$ and $25,000$. For reference, the column "True Choice Prob." reproduces Table 2, 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 dataset 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 dataset 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 A5 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)		
	$(0.726, 0.309)$	$(0.926, 0.377)$	$(1.063, 0.451)$
First Order Stochastic Dominance	16	5	11
Dominated Consequence Effect	12	4	1
Reverse Dominated Consequence Effect	12	2	3
Strict Dominance Effect	20	6	11
Other	0	0	0
# of Logical Anomalies	60	26	18

Table A5: 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 dataset 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 again uncover first-order stochastic dominance violations, in which the individual selects lotteries that are first-order stochastically dominated by the other lottery in the menu. We provide two representative examples in Table A6.

Second, our anomaly generation procedures uncover logical anomalies that exhibit (a generalization of) the dominated consequence effect. All of the logical anomalies in the second row of Table A5 have two possible, related structures. First, 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 $\alpha_1 \geq \alpha_0$ Second, 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 A7.

Third, our anomaly generation procedures uncover logical anomalies that exhibit the reverse dominated consequence effect. All of the logical anomalies in the third row of Table A5 have two possible, related structures. First, 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 $\alpha_1 \leq \alpha_0$. Second, 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 A8.

Finally, our anomaly generate procedures uncover logical anomalies that exhibit the strict dominance effect. All of the logical anomalies in the fourth row of Table A5 have two possible, related structures. First, 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''. Second, 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 A9.

(a) Logical anomaly #1

Lottery 0	5.86		
	100%		
Lottery 1	6.07	6.93	7.14
	5.1%	20.%8	74.1%

(b) Logical anomaly #2

Lottery 0	3.70	3.99	9.47
	38.2%	38.6%	23.2%
Lottery 1	2.74	9.45	
	81.3%	18.7%	

Table A6: 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 dataset 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.

(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 A7: 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 dataset 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	6.050	6.560	6.880
	0.000*	1.000*	0.000*
Lottery 1	4.620	7.360	9.360
	0.054	0.116	0.829
	Menu B (x_B, y_B^*)		
Lottery 0	6.050	6.560	6.880
	0.060	0.465	0.475
Lottery 1	4.620	7.360	9.360
	0.369	0.426	0.205

(b) Logical Anomaly #2

	Menu A (x_A, y_A^*)		
Lottery 0	2.150	5.370	8.950
	0.864	0.021	0.115
Lottery 1	3.770	4.450	8.930
	0.093	0.907	0.000*
	Menu B (x_B, y_B^*)		
Lottery 0	2.150	5.370	8.950
	0.414	0.145	0.440
Lottery 1	3.770	4.450	8.930
	0.182	0.589	0.229

Table A8: 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 dataset 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 A9: 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 dataset 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.

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 A5 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.

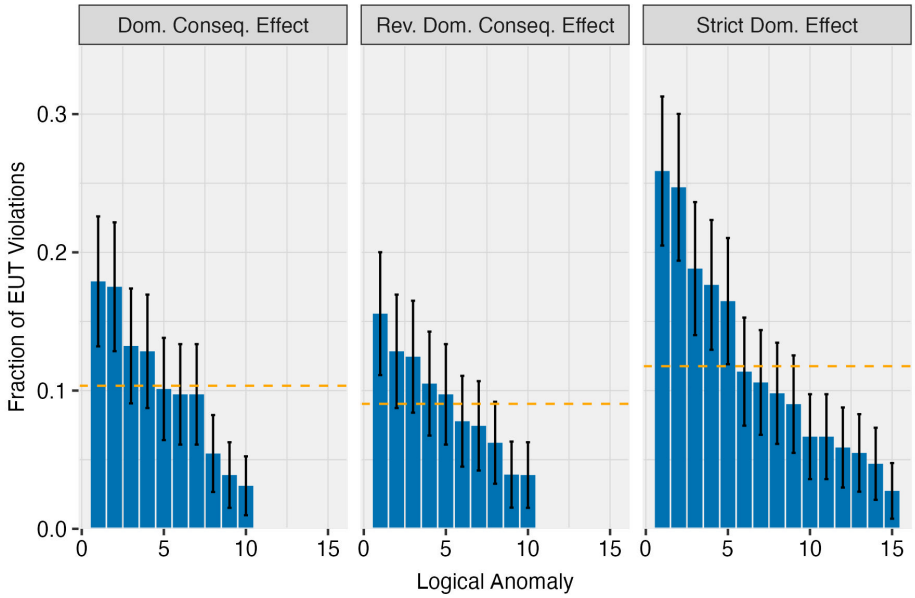
G.2.2 Experimental results

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

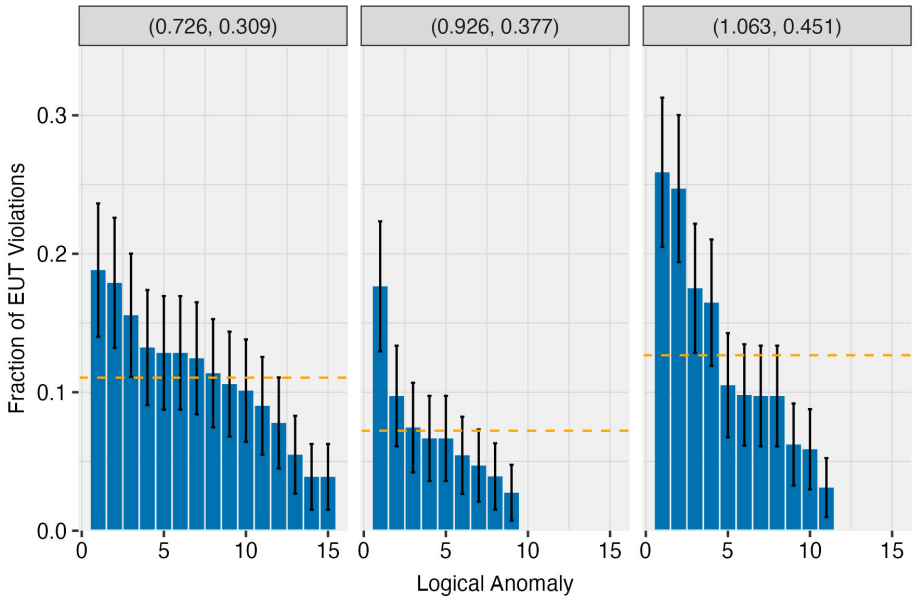
Appendix Figure A7(a) reports the fraction of respondents whose choices violate expected utility theory without noise on our algorithmically generated, logical anomalies ("expected utility theory violation rates"), organized by logical anomaly category. Appendix Figure A7(b) reports the same quantities 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 A10 and Appendix Table A11 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.



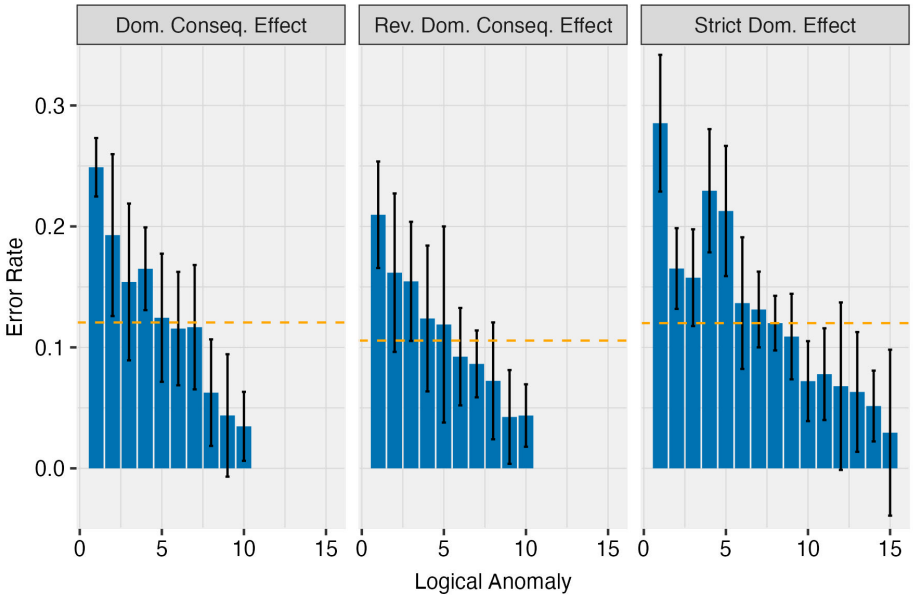
((A)) Estimates by logical anomaly category



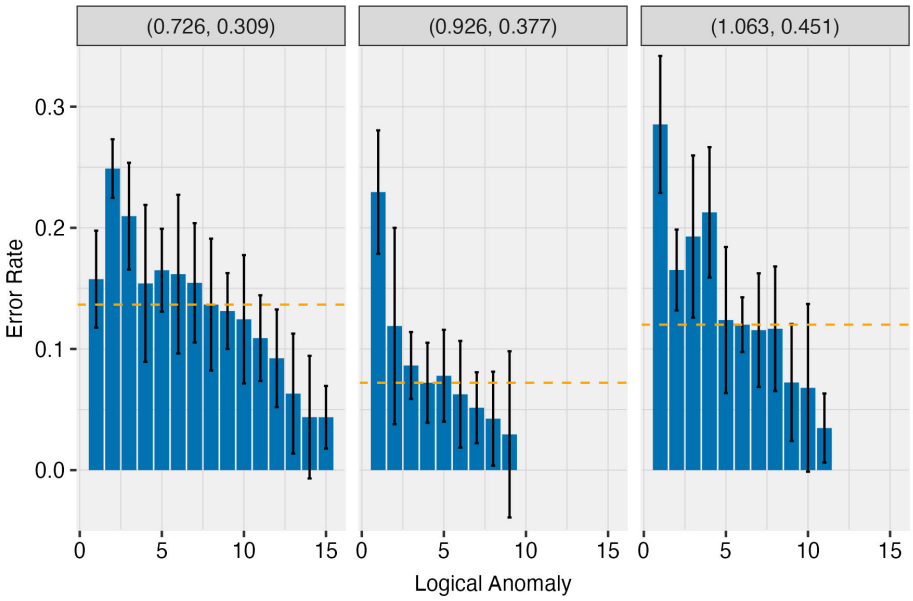
((B)) Estimates organized by calibrated parameter values (δ, γ)

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 A5) 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 A5) 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 A10: 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 A5). 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.

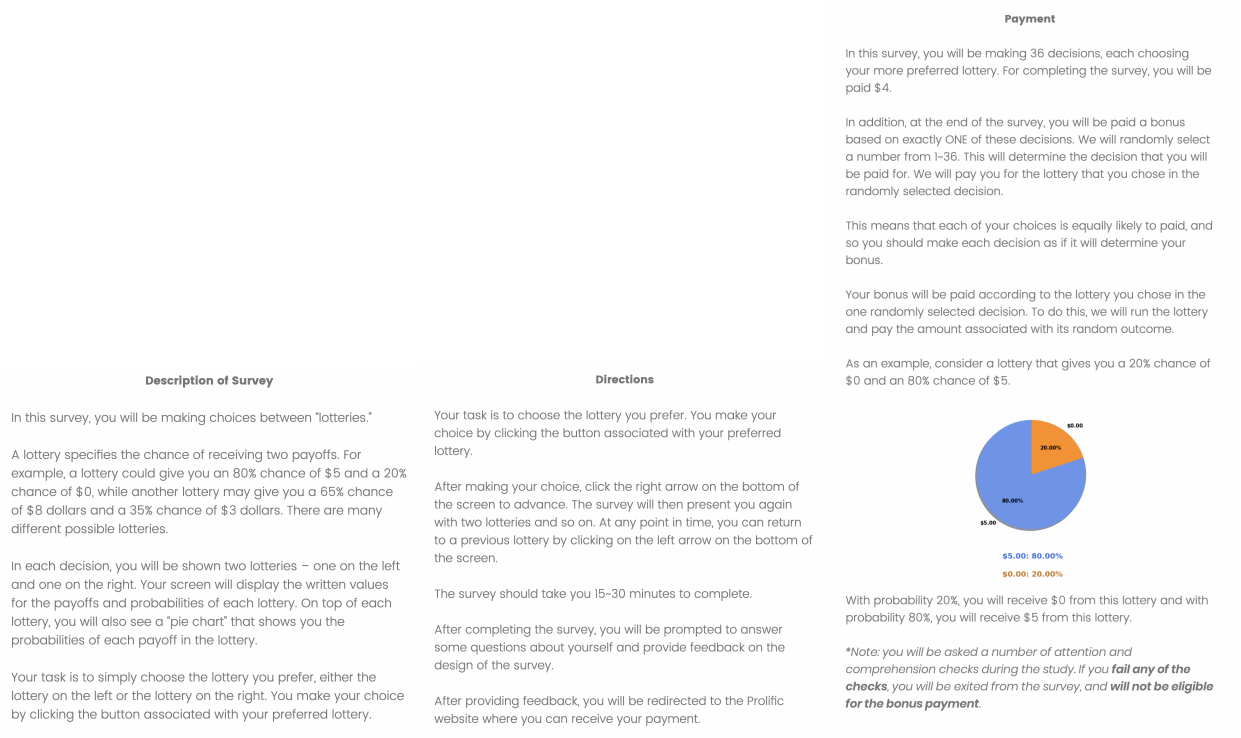
Prob. Weighting Function: (δ, γ)	Pooled Average	Median	First Quartile	Third Quartile
$(0.726, 0.309)$	0.110 (0.006)	0.113	0.084	0.130
$(0.926, 0.377)$	0.072 (0.006)	0.066	0.047	0.074
$(1.063, 0.451)$	0.126 (0.007)	0.098	0.079	0.169

Table A11: 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 2). 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.

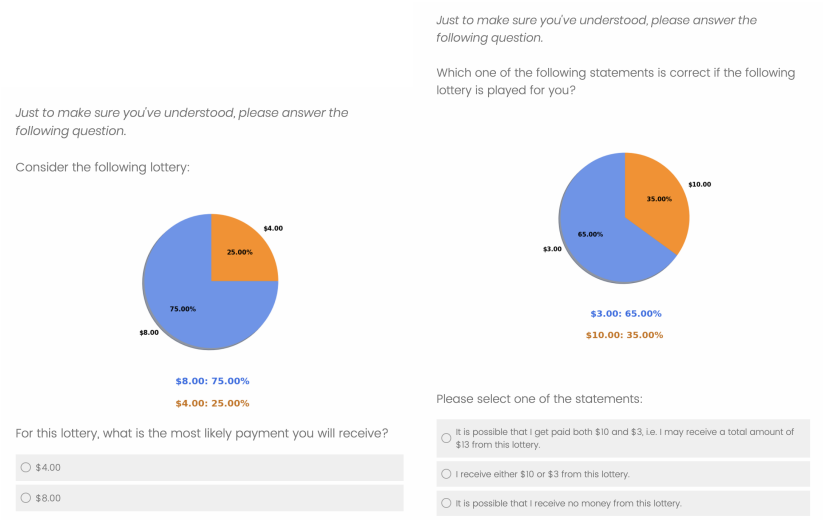


Figure H2: Screenshots of comprehension checks for the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

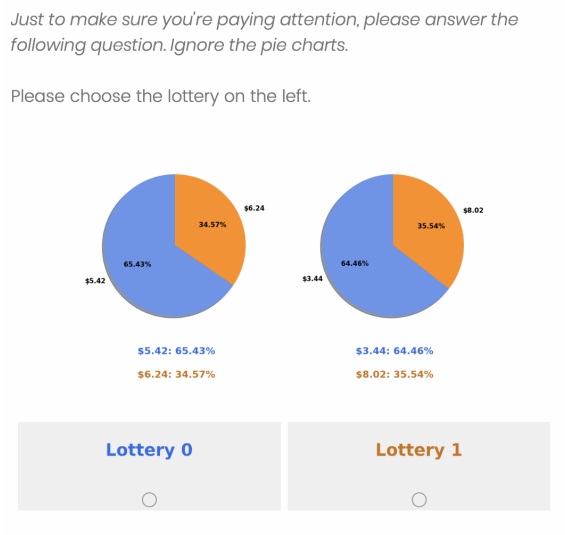


Figure H3: Screenshot of an attention check included in the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

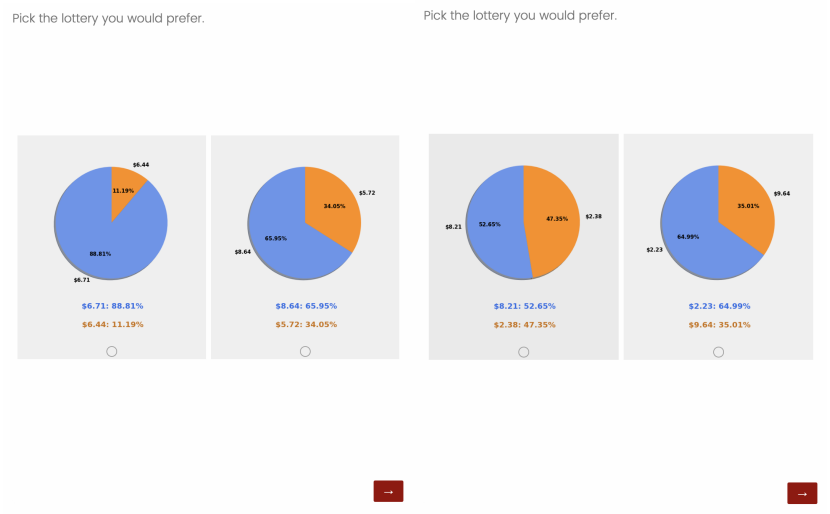


Figure H4: Screenshots of two main survey questions for the online surveys on choices from menus of two lotteries over two monetary payoffs. See Section 5.3 for further discussion.

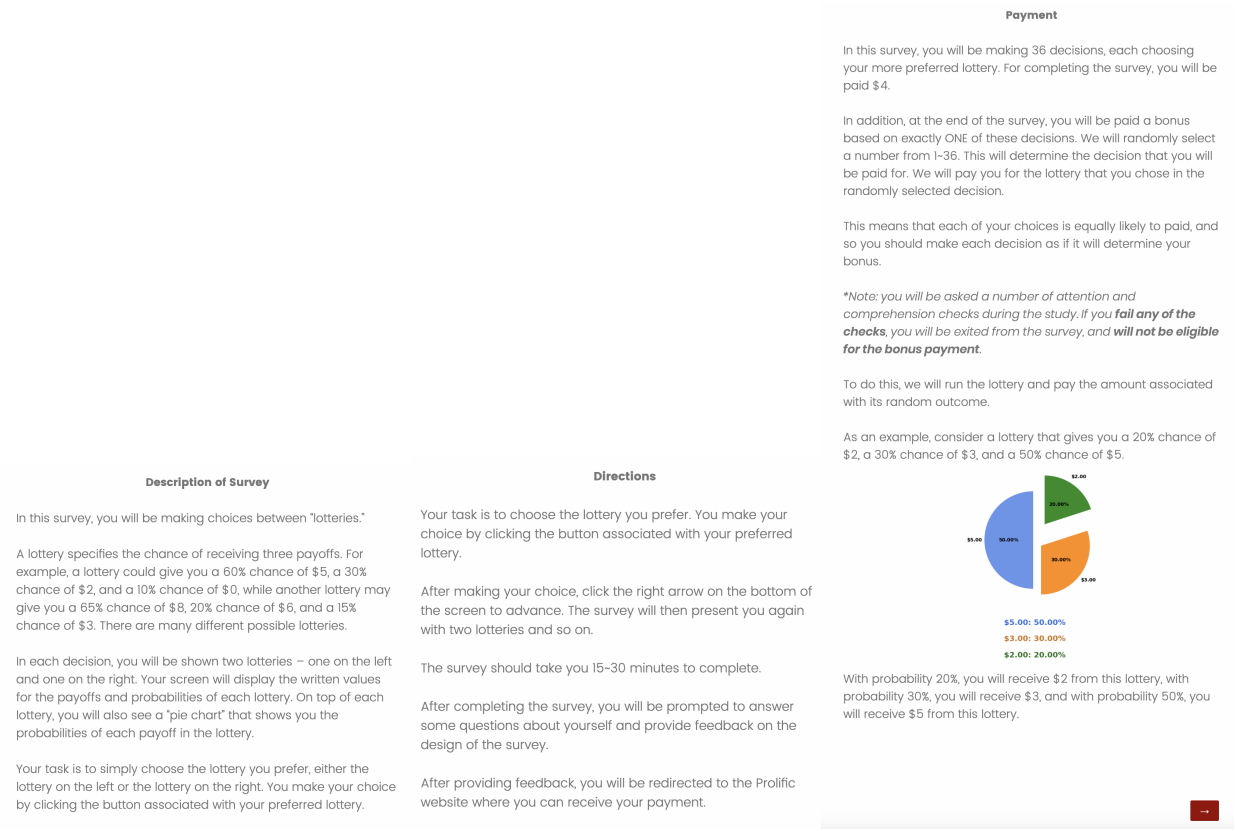


Figure H5: Screenshots of directions for the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

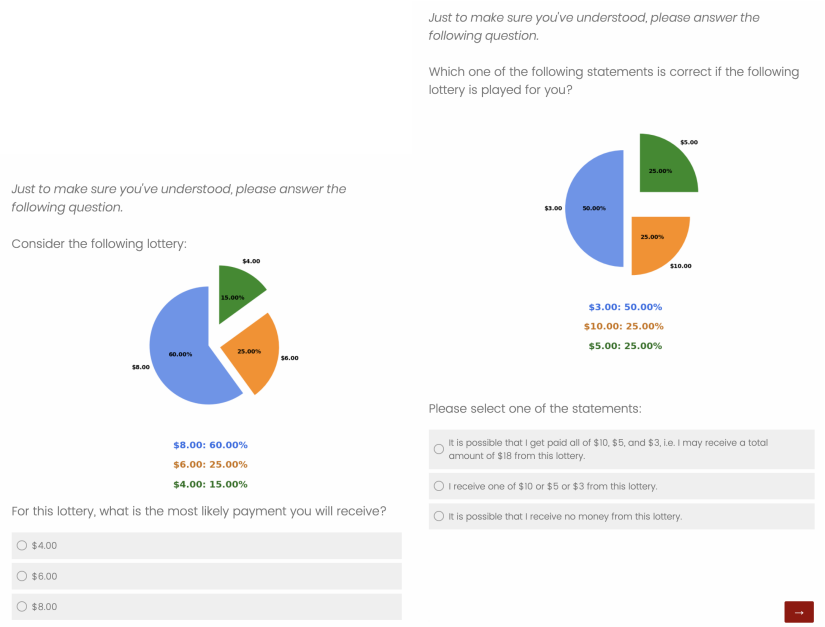


Figure H6: Screenshots of comprehension checks for the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

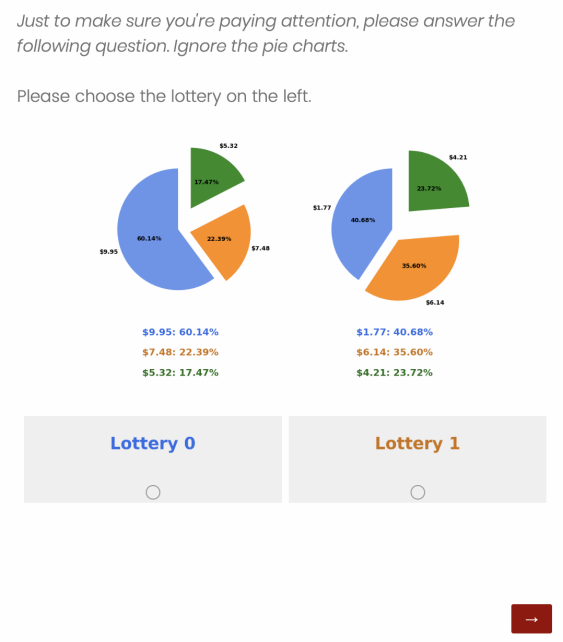


Figure H7: Screenshot of an attention check included in the online surveys on choices from menus of two lotteries over three monetary payoffs. See Appendix Section G for further discussion.

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.