From Predictive Algorithms to Automatic Generation of Anomalies[*]

Sendhil Mullainathan Ashesh Rambachan[†]

May 5, 2023

Abstract

We ask how machine learning can change a crucial step of the scientific process in economics: the advancement of theories through the discovery of "anomalies." Canonical examples of anomalies include the Allais Paradox and the Kahneman-Tversky choice experiments, which are concrete examples of menus of lotteries that highlighted flaws in expected utility theory and spurred the development of new theories for decision-making under uncertainty. We develop an econometric framework for anomaly generation and develop two algorithmic procedures to generate anomalies (if they exist) when provided a formal theory and data that the theory seeks to explain. Our algorithmic procedures are general since anomalies play an important role across a wide variety of fields in economics. As an illustration, we apply our procedures to generate anomalies for expected utility theory using simulated lottery choice data by an individual who behaves according to cumulative prospect theory. We produce novel anomalies for the independence axiom based on the probability weighting function that to our knowledge have not been noticed before. While this illustration is specific, it is our view that automatic anomaly generation can accelerate the development of new theories.

[*] We thank Peter G. Chang for exceptional research assistance and participants at the Stanford econometrics seminar for helpful comments. Mullainathan thanks the Center for Applied Artificial Intelligence for invaluable support and funding. All errors are our own.

[†] Mullainathan: University of Chicago and NBER (`Sendhil.Mullainathan@uchicago.edu`). Rambachan: Microsoft Research and MIT (`asheshr@mit.edu`).

1 Introduction

Anomalies play a central role in developing new 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 fails empirically. In this paper, we ask how machine learning can change the scientific process in economics by automatically generating anomalies for existing theories.

As a concrete example, consider the field of decision-making under uncertainty. 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 its empirical weakness, Allais (1953) produced the now celebrated "Allais Paradox," a hypothetical pair of menus of lotteries depicted in Table 1. The hypothetical menus in the Allais Paradox are crafted so that expected utility theory restricts the possible choices across the two menus. Due to the independence axiom, choices that are consistent with expected utility theory must either select lotteries (A0, B0) or lotteries (A1, B1) (e.g., Machina, 1987). In contrast, individuals tend to empirically select lotteries (A0, B1) (e.g., Slovic and Tversky, 1974). This was only the beginning as researchers

(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 typically observed choices made by subjects when presented these two menus (e.g. Slovic and Tversky, 1974). Allais (1953) originally denominated the payoffs in French Francs, and we reproduce the version of the Allais Paradox used in Slovic and Tversky (1974).

[1] Constructing test statistics and hypothesis tests for model mis-specification is a celebrated and foundational literature in econometrics and economic theory. See, for example, Sargan (1958); Afriat (1967, 1973); Hansen (1982); Varian (1982, 1990); Choi et al. (2014); Bugni, Canay and Shi (2015); Kitamura and Stoye (2018); Polisson, Quah and Renou (2020) among many others.

steadily accumulated more anomalies for expected utility theory.[2] Eventually, Tversky and Kahneman (1992) suggested that cumulative prospect theory could resolve many of these anomalies. Armed with a new theory, the cycle repeated itself: researchers crafted new anomalies, suggesting elements that were missing from cumulative prospect theory that in turn prompted the development of new theories of choice under uncertainty.[3]

Indeed the field of decision-making under uncertainty is not exceptional.[4] Economics often advances through the discovery of anomalies that highlight inconsistencies between our current theories and nature. As anomalies accumulate, researchers eventually develop new theories to resolve them, and the cycle repeats itself. Scientific discovery in economics therefore iterates between theory development and anomaly generation.

While theory development inherently relies on abstraction, anomaly generation is an empirical activity at its core. To generate an anomaly, a researcher like Allais reflects on "mental" data about how individuals make choices between lotteries, contrasts these patterns against the theoretical predictions made by an existing theory like expected utility theory, and then generates a concrete *example* where the theory's predictions will differ from what they believe are the likely empirical patterns. We rely on the creativity and intuition of researchers for all of these steps in generating anomalies.

Yet machine learning algorithms can process far more domain-specific data than any one person. They can uncover novel predictive signals, ones that our existing theories do

[2] For example, Allais (1953); Kahneman and Tversky (1979) produced the certainty effect or common ratio effect choice experiments, Slovic and Lichtenstein (1983); Tversky and Kahneman (1986); Tversky, Slovic and Kahneman (1990) produced choice experiments to highlight framing effects and preference reversals, and finally Kahneman and Tversky (1984); Tversky and Kahneman (1991) produced choice experiments to highlight loss aversion.

[3] Some more recent examples include salience theory (Bordalo, Gennaioli and Shleifer, 2012), betweenness preferences and certainty independence (Dekel, 1986; Cerreia-Vioglio, Dillenberger and Ortoleva, 2015, 2020), preferences for simplicity (Oprea, 2022; Puri, 2022), and cognitive uncertainty (Enke and Graeber, 2023) among many others.

[4] Anomalies have played a vital role throughout economics. For example, Richard H. Thaler's series of articles entitled "Anomalies" in *The Journal of Economic Perspectives* highlighted anomalies in asset pricing (e.g., Lamont and Thaler, 2003), game theory (e.g., Camerer and Thaler, 1995), international finance (e.g., Froot and Thaler, 1990), public finance (Hines and Thaler, 1995), decision-making under uncertainty (e.g., Kahneman, Knetsch and Thaler, 1991), intertemporal choice (Loewenstein and Thaler, 1989), and auction theory (Thaler, 1988).

not model and we may overlook ourselves.[5,6] How then can we go from these predictive algorithms to the automatic generation of anomalies?

Our main contribution is to develop algorithmic procedures that take as inputs *any* formal theory and data from a scientific domain that it seeks to explain, applies a supervised learning algorithm to that data, and then automatically generates anomalies, if they exist. As an illustration, we apply these algorithmic procedures to generate anomalies for expected utility theory in simulated lottery choice data from an individual behaves according to cumulative prospect theory. Our procedures recover known anomalies and intriguingly uncover novel Allais Paradox-like anomalies for expected utility theory based on the probability weighting function that to our knowledge have not been noticed before.

In order to develop algorithmic procedures for anomaly generation, we must first develop a common econometric framework for analyzing theories that abstracts from any particular scientific domain. Anomalies play a key role in choice under uncertainty, game theory and asset pricing, yet theories across these scientific domains share little resemblance. Expected utility theory is expressed as a collection of axioms that restrict preferences over lotteries, Nash equilibrium is an equilibrium condition on choices in normal-form games, and the capital asset pricing model in finance is a model of homogeneous investors optimizing portfolios in a frictionless market. Even more, the same theory often has multiple equivalent representations: expected utility theory as a collection of Bernoulli utility function over payoffs; Nash equilibrium as the fixed point of players' best response functions; or the capital asset pricing model as a procedure for calculating assets' covariances with the market return and an asset pricing equation. Any econometric framework for anomaly generation must therefore capture this immense diversity of theories.

To tackle this challenge, we abstractly model theories as *black-boxes* that derive implications between some features and modeled outcomes from any hypothetical dataset.

[5] See, for example, Varian (2014), Athey (2017), Mullainathan and Spiess (2017), Athey (2019) among many others.

[6] Outside of economics, machine learning algorithms are increasingly used as aids in the process of scientific discovery. See, for example, Raghu and Schmidt (2020) and Pion-Tonachini et al. (2021) for recent reviews.

For example, expected utility theory derives implications about choice behavior from hypothetical datasets of menus of lotteries and choice probabilities, Nash equilibrium derives implications about strategic behavior from hypothetical datasets of normal-form games and strategy profiles, and the capital asset pricing model models expected returns from hypothetical datasets on the cross-section of asset price returns. We introduce four intuitive axioms that restrict the properties of a black-box theories such that it behaves as-if it has some underlying structure (whatever that may be) and establish two results. First, for any theory satisfying these axioms, we show that there exists *anomalies* or minimal datasets that are logically incompatible with the theory like the Allais Paradox. Second, we show that any theory satisfying these axioms can be equivalently represented as an implicit *allowable function class*. The allowable function class summarizes all mappings between the features and modeled outcomes that are consistent with the theory's underlying structure (whatever that may be). Any theory can therefore be cast into an empirical risk minimization framework over some hypothetical datasets, and all anomalies can be interpreted through the lens of the theory's allowable function class. These results therefore serve as the basis of our procedures for automatic anomaly generation.

Given the existence of anomalies and the tractable characterization of theories based on their allowable function classes, we next ask how can we efficiently search for anomalies given data drawn from some joint distribution over the scientific domain. Using this framework, we next develop two algorithmic procedures that take as input an existing theory and data it seeks to explain, estimate a supervised machine learning algorithm to summarize the empirical relationship between the features and modeled outcome, and then generate anomalies if they exist.

Our first procedure is based on an adversarial characterization of anomalies over a theory's allowable function class. To capture the intuition, consider the following zero-sum game between a theory and a falsifier. The falsifier proposes datasets to the theory, and the theory then attempts to explain those datasets by fitting an allowable function via empirical risk minimization. The falsifier's payoff is increasing in the theory's expected loss over the proposed dataset, and the theory's payoff is decreasing its is expected loss. We show that anomalies can be characterized as datasets that induce a strictly positive expected loss for the theory in such a game, or equivalently datasets that cannot be explained by any of the theory's allowable functions. We therefore build our first procedure for anomaly generation based on a feasible implementation of this game as a max-min optimization program over the theory's allowable functions. We analyze the statistical properties of this feasible implementation and establish conditions under which it approximates the population version. Practically solving this max-min optimization problem may be challenging as the outer maximization will typically be non-concave, and so standard optimization techniques that aim to find a saddle point may not be applied (e.g., Rockafellar, 1970; Freund and Schapire, 1996). Instead, we leverage recent breakthroughs in non-convex/concave min-max optimization in computer science (e.g., Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020) to solve the feasible implementation via a gradient descent ascent procedure and analyze its convergence properties.

While this adversarial procedure exploits nothing beyond the theory's allowable functions, there in fact exists additional structure that can be used for anomaly generation. We show that any theory satisfying our axiomatization 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. It is as-if the theory collapses these feature values together. As a consequence, some anomalies reveal what we call *representational errors* that the theory's implicit lower-dimensional representation has failed to capture some relevant dimension of nature. In economics, we are often most interested in such representational anomalies, and canonical anomalies like the Allais Paradox are in fact representational anomalies for our existing theories. We therefore 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 another feature value that is representationally equivalent under the theory but across which nature varies.

Finally, as an illustration, we apply our algorithmic procedures to choice under uncertainty, returning to our motivating example of the Allais Paradox and other anomalies for expected utility theory. We explore what anomalies for expected utility theory would be uncovered if our procedures are provided with simulated lottery choice data from an individual whose choices are consistent with cumulative prospect theory. We find that our algorithms recover known anomalies for probability weighting functions and generates novel anomalies for the independence axiom. These novel anomalies cannot be cast as examples of the common consequence or common ratio effects. We dub this (to our knowledge) new anomaly a "dominated consequence effect." It is a violation of the independence axiom that uses only two possible payoffs but involves mixing lotteries with dominated, certain prospects. Even in simulated data, our anomaly generation procedures therefore appear to have made a genuine discovery about the probability weighting function. In ongoing work, we are collecting real lottery choice data in order to uncover empirical anomalies that can inform our theories of decision-making under uncertainty.

Substantial progress has already been made in exploring how machine learning interacts with economic theories. Several 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 uncertainty and initial play in normal form games, measuring the "completeness" of economic theories (Fudenberg et al., 2022). When a supervised machine learning algorithm predicts more accurately out-of-sample than an existing theory, this line of research then attempts to "open" the black-box prediction function and understand its properties (e.g., Peysakhovich and Naecker, 2017; Peterson et al., 2021; Hirasawa, Kandori and Matsushita, 2022). By contrast, we use supervised machine learning algorithms as a stepping stone to automatically generate anomalies or concrete examples that are logically incompatible with the theory. Fudenberg, Gao and Liang (2020) measure the "restrictiveness" of economic theories. Our existence result for anomalies can be interpreted as establishing that any black-box theory satisfying our axiomatization must be "restrictive" in the sense that there exists some minimal hypothetical datasets that it cannot explain (albeit this is weaker than the definition of restrictiveness offered in Fudenberg, Gao and Liang (2020)). Further afield, Andrews et al. (2022) develops procedures based on conformal inference to measure the transfer performance or out-of-distribution predictive performance of economic theories.

More closely related to our work, Fudenberg and Liang (2019) use machine-learning based models to predict on which normal-form games observed play will differ from the predictions of a particular model of strategic behavior and generate new normal-form games where the particular model will predict poorly. From the perspective of our analysis, Fudenberg and Liang (2019) can be formally re-interpreted as a particular heuristic for our adversarial characterization of anomalies tailored to the model of strategic behavior they study. Ludwig and Mullainathan (2023) develop a morphing procedure for images in order to uncover novel, implicit characteristics of defendant mug-shots that affect the pretrial release decisions of judges. Our adversarial learning and dataset morphing procedures enable researchers to search for anomalies given any formal theory.

2 A Model of Scientific Theories

By positing some underlying structure, theories derive logical implications about the relationship some features and modeled outcomes from any hypothetical dataset. For example, by inferring preference relations, expected utility theory derives logical implications about an individual's choices from menus of lotteries from hypothetical datasets of menus and choice probabilities. How exactly theories model their underlying structure varies greatly, and any econometric framework for anomaly generation must capture this diversity.

In this section, we analyze theories as black-box mappings that return correspondences between the features and modeled outcome given any hypothetical dataset. We introduce four intuitive axioms on the properties of these black-box mappings and establish two main results. First, we show that there exists *anomalies*, or minimal logically incompatible datasets, for any such theory. Second, we show that any such theory can be equivalently represented by an *allowable function class*, which summarizes all mappings between the features and modeled outcomes that are consistent with theory's underlying structure. These two results serve as the foundation of our algorithmic procedures for anomaly generation.

2.1 Scientific domains and theories

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

In such a scientific domain, a theory posits an underlying structure that enables it to derive novel implications about the relationship between the features and modeled outcome from any hypothetical dataset. Rather than focusing on any particular model, we abstractly model a theory as a reduced-form black-box.

Definition 2.1. A *theory* consists of the pair $(T(\cdot), \mathcal{M})$, where $T(\cdot) \colon \mathcal{D} \to \mathcal{C}$ is a mapping from hypothetical datasets to correspondences between the features and modeled outcome and $\mathcal{M}$ is some finite set with elements $m \in \mathcal{M}$.

Given any hypothetical dataset $D \in \mathcal{D}$, theory $T(\cdot)$ returns a correspondence that summarizes all implications that it can draw about the relationship between the features and the modeled outcome. We write $T(\cdot; D) \in \mathcal{C}$ to be the theory's correspondence when applied to hypothetical dataset $D \in \mathcal{D}$ and $T(x; D) \subseteq \mathcal{Y}^*$ to be the theory's implications about the modeled outcome at $x \in \mathcal{X}$. All else about the scientific domain is left unmodeled and collapsed into *modeled contexts* $m \in \mathcal{M}$. The modeled contexts summarize the theory's own scope constraints – the theory refines its underlying structure within a modeled context and does not extrapolate across modeled contexts. We take a theory's modeled contexts as a primitive in this paper. We instead focus on the behavior of the theory's correspondence $T(\cdot)$.

Definition 2.1 is necessarily abstract in order to capture the diversity of theories across scientific domains. To make it concrete, we illustrate how several leading economic theories map into this framework.

Example: choice under uncertainty Consider individuals making choices from menus of two lotteries over $J > 1$ monetary payoffs (e.g., Allais, 1953; Kahneman and Tversky, 1979; Harless and Camerer, 1994). 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 include information about how each lottery is presented (e.g., presented as a two-stage lottery) or the ordering of lotteries in the menu. The modeled outcome is the choice probability $y^* \in [0, 1]$, and the modeled contexts $m \in \mathcal{M}$ are each individual. Given a dataset D of hypothetical menus and choices, 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 $u(\cdot)$ that rationalizes D. ▲

Example: initial play in normal-form games Consider individuals playing $J \times J$ normal-form games (e.g., Wright and Leyton-Brown, 2010; Hartford, Wright and Leyton-Brown, 2016; Wright and Leyton-Brown, 2017; Fudenberg and Liang, 2019; Hirasawa, Kandori and Matsushita, 2022). 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 $x = (\pi_{row}(1, 1), \pi_{col}(1, 1), \ldots, \pi_{row}(J, J), \pi_{col}(J, J)) \in \mathbb{R}^{2 \cdot J \cdot J}$. 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 dataset D of hypothetical normal-form games and strategies for the row player, 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 $\tilde{y}^* \in \Delta^{J-1}$. ▲

Example: asset pricing Consider the evolution of asset prices 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 hypothetical dataset D of expected returns and their variance-covariance matrix in several periods, 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 $y^* = \bar{y}_{risk\text{-}free} + \beta \left(\bar{y}_{market} - \bar{y}_{risk\text{-}free} \right)$. ▲

2.2 Incompatible datasets and anomalies

A hypothetical dataset is logically incompatible with theory $T(\cdot)$ if its underlying structure cannot make sense of the configuration of features and modeled outcomes. For example, a hypothetical collection of lottery menus and choice probabilities may not be rationalizable under expected utility theory. Otherwise, a hypothetical dataset is compatible with theory $T(\cdot)$.

Definition 2.2. A hypothetical 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 $x \in \mathcal{X}$.

Many incompatible datasets are opaque and it may be difficult for researchers to exactly understand what drives the failure of the theory's underlying structure. For this reason, researchers such as Allais are not simply interested in characterizing the collection of incompatible datasets of a theory. But rather they construct minimal incompatible datasets that highlight how or why the theory's underlying structure fails. We call these anomalies and offer a formal definition.

Definition 2.3. Hypothetical dataset $D \in \mathcal{D}$ is an *anomaly* for theory $T(\cdot)$ if D is incompatible with theory $T(\cdot)$ and $D \setminus \{(x, y^*)\}$ is compatible with theory $T(\cdot)$ for all $(x, y^*) \in D$.

An anomaly is a minimal incompatible dataset in the sense that $T(\cdot)$ is compatible with any of its subsets. To make the definition more concrete, let us return to our earlier example of the Allais Paradox for expected utility theory (Table 1). In Appendix B, we discuss anomalies for our other examples. The Allais Paradox is a hypothetical dataset that consists of two menus x_A, x_B and associated outcomes $y^*_A = 0$, $y^*_B = 1$. The independence axiom of expected utility theory implies that the implies that the choice on menu x_A uniquely pins down the choice on menu x_B and vice versa. As a result, $T(x_A; D) = T(x_B, D)$ for all hypothetical datasets $D \in \mathcal{D}$ and the Allais Paradox dataset is incompatible with expected utility theory. Yet any choice on any single menu x_A or x_B is compatible with expected utility theory. The Allais Paradox therefore satisfies Definition 2.3.

2.3 Axiomatization

We next introduce four axioms on the properties of theory's correspondence $T(\cdot)$. These axioms place restrictions on the behavior of $T(\cdot)$ such that it behaves as-if it has some underlying structure, whatever that may be. We establish that there exists anomalies for any theory satisfying these axioms and that any theory satisfying these axioms can be equivalently represented by an allowable function class.

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

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

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

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

Axiom 1 states that theory $T(\cdot)$ is either compatible or incompatible with any hypothetical dataset but not both. Axiom 2 states that whenever the theory is compatible with a hypothetical dataset, it is consistent with all observations in the dataset. Axiom 3 states that the theory can only refine its implications as more hypothetical observations are collected. Finally, Axiom 4 states that there exists some hypothetical dataset and unseen feature at which theory $T(\cdot)$ derives non-trivial implications.

All of our previous examples of leading economic theories satisfy these axioms. Consider expected utility theory. Appendix B discusses our other examples. First, expected utility theory satisfies Axiom 1 and Axiom 2. For any dataset D of hypothetical 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 of hypothetical datasets D, D' satisfying $D \subseteq D'$, the rationalizing utility functions for dataset D' must be a subset of the rationalizing utility functions for dataset D. This implies expected utility theory satisfies Axiom 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 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 $\alpha \in [0, 1)$.[7] Expected utility theory therefore satisfies Axiom 4.

2.4 Representation result and existence of anomalies

For any theory $T(\cdot)$ satisfying Axioms 1-4, we establish that there exists anomalies and it can be equivalently represented by an allowable function class that summarizes all logical implications $T(\cdot)$ may draw from any hypothetical dataset.

To state this result, we say a mapping $f(\cdot) \in \mathcal{F}$ is *consistent* with hypothetical dataset

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

$D \in \mathcal{D}$ if $f(x) = y^*$ for all $(x, y^*) \in D$. Hypothetical 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 Axioms 1-4 if and only if there exists a function class $\mathcal{F}^T \subset \mathcal{F}$ that is inconsistent with some hypothetical datasets and satisfies, for all $x \in \mathcal{X}$ and $D \in \mathcal{D}$,

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

ii. There exists anomalies for any theory $T(\cdot)$ satisfying Axioms 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, whatever that may be. As a result, theory $T(\cdot)$ can be analyzed as-if it applies empirical risk minimization given hypothetical dataset $D \in \mathcal{D}$ over the allowable function class $\mathcal{F}^T$. Furthermore, the theory's underlying structure is not compatible with all possible datasets – in fact, there exists anomalies for any theory $T(\cdot)$ satisfying Axioms 1-4. By establishing the existence of anomalies and placing all theories into a common allowable function representation irrespective of its scientific domain or underlying structure, Proposition 2.1 serves as the launching point of our subsequent analysis.

We provide the complete proof of Proposition 2.1 in Appendix A but we briefly sketch our proof strategy here. It is clear that the allowable function representation (1) satisfies Axioms 1-3. To show it also satisfies Axiom 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 $(x, y^*) \in D_{min}$, Axiom 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 have 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 Axioms 1-4. To do so, we define $\mathcal{D}^{\neg T}$ as the collection of all falsifying datasets for $T(\cdot)$, which is non-empty by Axiom 4. We define $\mathcal{F}^{\neg T}$ to be the collection of all mappings that are consistent with any falsifying 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 2.3.

2.5 Observable data and theories

To this point, we only modeled the logical content of theory $T(\cdot)$. Our goal is to contrast theory $T(\cdot)$'s underlying structure with nature in order to understand how it may be improved.

Towards this, 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. The observed outcome is related to the theory's modeled outcome statistically. For example, in choice under uncertainty, we only observe an individual's binary choices on a menu but not their choice probabilities. In initial play in normal-form games, we only observe an individual's action but not their chosen strategy profile. We capture this by defining the empirical modeled outcome of the theory $T(\cdot)$ as

$$f^*_m(x) := \mathbb{E}_m \left[g(Y_i) \mid X_i = x \right] \tag{2}$$

for some known function $g(\cdot)$, where $\mathbb{E}_m[\cdot]$ denotes the expectation under $P_m(\cdot)$. We interpret the modeled outcome of theory $T(\cdot)$ as some identified functional of each modeled context's underlying joint distribution.

3 An Adversarial Algorithm for Anomalies

We established that there exists anomalies among hypothetical datasets $D \in \mathcal{D}$ for any theory $T(\cdot)$ satisfying Axioms 1-4. In this section, we now ask whether there exists *empirical* anomalies for theory $T(\cdot)$, and if so how to efficiently find them. That is, given modeled contexts $m \in \mathcal{M}$ with true functions $f^*_m(\cdot)$, we search for empirical anomalies $D = \{(x_1, f^*_m(x_1)), \ldots, (x_n, f^*_m(x_n))\}$ for theory $T(\cdot)$? We develop an adversarial learning algorithm to generate candidate empirical anomalies when given access to the theory's allowable functions $\mathcal{F}^T$ and data that the theory seeks to explain.

3.1 Adversarial algorithm

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

Proposition 3.1. *Suppose theory $T(\cdot)$ satisfies Axioms 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. Dataset $D \in \mathcal{D}$ is incompatible with theory $T(\cdot)$ if and only if

$$\min_{f(\cdot)\in\mathcal{F}^T} |D|^{-1} \sum_{(x,y^*)\in D} \ell\left(f(x), y^*\right) > 0. \tag{3}$$

ii. If there exists no incompatible datasets of size strictly less than $n > 1$, then any incompatible 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 hypothetical datasets that induce a strictly positive loss for the theory's allowable functions. Furthermore, we can search for anomalies by iteratively searching for larger incompatible datasets. We next build on Proposition 3.1 in order to develop a procedure for automatic anomaly generation.

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

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

to be the empirical loss over the theory $T(\cdot)$'s allowable functions. Proposition 3.1 establishes that the empirical dataset $\{(x_1, f^*_m(x_1)), \ldots, (x_n, f^*_m(x_n))\}$ is incompatible with theory $T(\cdot)$

if and only if $\mathcal{E}^T(x_{1:n}) > 0$. Furthermore, it is also an anomaly if there exists no smaller empirical datasets that are incompatible with theory $T(\cdot)$.

If we had oracle access to the true function $f_m^*(\cdot)$, we could therefore search for anomalies by (i) searching for empirical incompatible datasets, or equivalently $x_{1:n}$ such that $\mathcal{E}^T(x_{1:n}) > 0$; and (ii) iterating that search over successively larger dataset sizes n. Searching for empirical incompatible datasets $x_{1:n}$ can be accomplished by searching for solutions to the following optimization program

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

for dataset size $n \geq 1$. This max-min optimization program (5) can be interpreted as an adversarial game between the theory (the min-player) and a falsifier (the max-player). The falsifier proposes empirical datasets $\{(x_1, f_m^*(x_1)), \ldots, (x_n, f_m^*(x_n))\}$ to the theory and the theory attempts to explain them using its allowable functions. The theory's payoffs are decreasing its average loss over those empirical datasets and the falsifier wishes to find empirical datasets that generate large, positive loss for the theory. The max-min optimization program also has connections to a recent literature in computer science on adversarial learning (e.g., Madry et al., 2017; Akhtar and Mian, 2018; Kolter and Madry, 2018).[8] We exploit these connections to develop an algorithmic procedure for solving a feasible implementation of the max-min program.

We base our iterative search for anomalies on solving the max-min optimization program (5). For some maximal dataset size $\overline{n} \geq 1$, we iterate over $n = 1, \ldots, \overline{n}$ and solve (5). Let n^* denote the smallest dataset size for which the optimal value of the max-min optimization

[8] In particular, the max-min optimization program can be loosely interpreted as a "data-poisoning" attack on the theory $T(\cdot)$'s allowable functions with two key differences. 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. By contrast, in (5), 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 also both induce variation in the theory's allowable functions and the true function $f_m^*(\cdot)$.

program is strictly positive. Proposition 3.1 implies that any empirical dataset $x_{1:n^*}$ with $\mathcal{E}^T(x_{1:n^*}) > 0$ is an anomaly. We therefore can search for anomalies by searching for other elements in the set $\{x_{1:n^*} \colon \mathcal{E}_m^T(x_{1:n^*}) > 0\}$. We summarize this oracle search procedure in Algorithm 1.

Algorithm 1: Oracle search for anomalies based on max-min optimization.

```
Input: $f_m^*(\cdot)$.
$n \leftarrow 1$;
while $n < \overline{n}$ do
    $\mathcal{E}(n) \leftarrow \max_{x_{1:n}} \min_{f(\cdot) \in \mathcal{F}^T} n^{-1} \sum_{i=1}^n \ell\left(f(x_i), f^*(x_i)\right)$;
    if $\mathcal{E}(n) > 0$ then
        $n^* \leftarrow n$;
        Break.
    $n \leftarrow n + 1$;
return $n^*$, $\{x_{1:n^*} \colon \mathcal{E}_m^T(x_{1:n^*}) > 0\}$.
```

Of course, directly implementing Algorithm 1 is not feasible. First, we do not directly observe the true function $f_m^*(\cdot)$. Second, practically solving the max-min optimization program may be quite difficult. Both the inner minimization of the theory's allowable functions and the outer maximization over the features may be intractable. We tackle both of these challenges next in order to construct a feasible implementation of Algorithm 1.

3.1.1 Statistical analysis of plug-in max-min optimization

Recall that 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 known function $g(\cdot)$. We now suppose that 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 $\widehat{f}_m^*(\cdot)$. We plug-in this estimator into the max-min optimization program

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

and analyze its resulting properties.

In order to analyze the plug-in max-min optimization program, we assume that we have access to approximate inner minimization and outer maximization routines.

Assumption 3.1.

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 $\delta > 0$.

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 $\nu > 0$.

This high-level assumption allows us to separate out the effects of optimization errors that arise from solving the inner minimization and outer maximization from the statistical error introduced by estimating $f^*_m(\cdot)$. Our analysis of the plug-in max-min optimization program will depend on the optimization errors associated with the approximate optimization routines.

Define $\widetilde{f}^T(\cdot; x_{1:n})$ to be the allowable function returned when the approximate inner minimization routine is applied to solve $\min_{f(\cdot)\in\mathcal{F}^T} n^{-1} \sum_{i=1}^{n} \ell\left(f(x_i), \widehat{f}^*_m(x_i)\right)$ 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 is applied to solve $\max_{x_{1:n}} n^{-1} \sum_{i=1}^{n} \ell\left(\widetilde{f}^T(x_i; x_{1:n}), \widehat{f}^*_m(x_i)\right)$. Finally, consider the optimal values of the plug-in and population optimal values of the max-min optimization program

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

respectively. We analyze the difference between these two quantities.

Proposition 3.2. *Suppose the loss function $\ell(\cdot,\cdot)$ is differentiable with gradients bounded by some $K < \infty$ and α-strongly convex in its second argument. Then, for any $n \geq 1$,*

$$\left\| \widehat{\mathcal{E}}_n - \mathcal{E}^*_n \right\| \leq (\delta + \nu) + 3 \left(K + \frac{\alpha}{2} \right) \| \widehat{f}^*_m(\cdot) - f^*_m(\cdot) \|_\infty, \tag{10}$$

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

The difference between the plug-in and population optimal values of the max-min optimization program is bounded by the optimization error introduced by the approximate optimization routines and the estimation error of $\widehat{f}^*_m(\cdot)$ for the true function $f^*_m(\cdot)$. Furthermore, 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 sup-norm. Equivalently, ignoring optimization error, the rate at which the plug-in optimal value converges to the population optimal value is bounded the rate at which $\widehat{f}^*_m(\cdot)$ converges uniformly to the true function $f^*_m(\cdot)$. While strong, it is perhaps unsurprising that this strong form of convergence is necessary for the plug-in optimal value to approximate the population optimal value well as the max-min optimization program explores $f^*_m(\cdot)$ in searching for incompatible datasets.

3.1.2 Gradient descent ascent optimization

While Proposition 3.2 analyzes the statistical properties of the plug-in max-min optimization program, this still leaves open the question of how to practically solve the inner minimization and outer maximization. To do so, we leverage recent results on non-convex/concave max-min optimization (e.g., Jin, Netrapalli and Jordan, 2019; Razaviyayn et al., 2020) and propose a feasible gradient descent ascent (GDA) optimization routine.

We first simplify the inner minimization problem over the theory's allowable functions. We assume that 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. For example, we may construct such a parametrization using a flexible sieve basis or class of neural networks. With this parametrization, the inner minimization over the theory's allowable functions becomes $\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \ell\left(f_\theta(x_i), \widehat{f}_m^*(x_i)\right)$. 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 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 Assumption 3.1(i).

By contrast, the outer maximization over features remains difficult as varying the feature values 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 nonetheless propose a gradient-based optimization procedure. As notation, let $\widehat{\mathcal{E}}^T(x_{1:n}, \theta) := n^{-1} \sum_{i=1}^n \ell\left(f_\theta(x_i), \widehat{f}_m^*(x_i)\right)$ and we assume $\widehat{\mathcal{E}}^T(x_{1:n}, \theta)$ 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

$$\theta^{t+1} = \arg\min_{\theta \in \Theta} \widehat{\mathcal{E}}(x_{1:n}^t; \theta) \tag{11}$$

$$x_{1:n}^{t+1} = x_{1:n}^t + \eta \nabla \widehat{\mathcal{E}}(x_{1:n}^t; \theta^{t+1}). \tag{12}$$

At each step t of the optimization procedure, we construct a solution to the inner minimization problem θ^{t+1} by either convex optimization or gradient descent and then take a gradient ascent step on the feature values plugging in θ^{t+1}. Recent results in non-convex/concave max-min optimization implies 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 $\nabla \widehat{\mathcal{E}}(x_{1:n}, \theta) \approx 0$ at the returned feature values and parameter vectors. We state this result formally in Appendix C.

3.2 Average anomalies across modeled contexts

Our adversarial search procedure so far focused on searching for anomalies in a single modeled context, whereas we may be most empirically interested in generating anomalies that hold across many modeled contexts $m \in \mathcal{M}$. The Allais Paradox is after all an anomaly for expected utility theory over the choices made by a large fraction of individuals. Our algorithmic procedure can be directly applied across modeled contexts to search for "average" anomalies and incompatible datasets.

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. We write $P(m \mid x) = P(M_i = m \mid X_i = x)$ and define $f^*_m(x) = \mathbb{E}_m[g(Y_i) \mid X_i = x]$ in each modeled context $m \in \mathcal{M}$ as before. An *average* incompatible dataset is a collection of features $x_{1:n}$ such that $\{(x_1, \bar{f}^*(x_i)), \ldots, (x_n, \bar{f}^*(x_n))\}$ is incompatible with theory $T(\cdot)$. An *average* anomaly is defined analogously. We next show that if $x_{1:n}$ is an average incompatible dataset, then it is also an incompatible dataset in some modeled context m. Furthermore, provided $x_{1:n}$ is a "systematically" incompatible dataset across modeled contexts, then it is also an average incompatible dataset.

Proposition 3.3. *Suppose theory $T(\cdot)$ satisfies Axioms 1-4. Then,*

i. If $x_{1:n}$ 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 $x_{1:n}$ is incompatible in some modeled context and satisfies

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

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

The condition in Proposition 3.3(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 3.3 suggests that we can search for 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.

4 Representational Anomalies and Dataset Morphing

The adversarial learning algorithm for anomaly generation exploits no structure about theory $T(\cdot)$ beyond its allowable functions. In this section, we next show that there often exists additional structure that can be exploited for anomaly generation. If a strengthened Axiom 4 ("non-trivial implications") is satisfied, then we show that theory $T(\cdot)$ must have a non-trivial, lower-dimensional representation of the features, meaning the theory always behaves as-if it pools together it pools some feature values. In this case, we may be particularly interested in uncovering "representational anomalies" that highlight ways in which the theory's lower dimensional representation has failed to capture some part of nature, and so we propose a dataset morphing algorithm to generate representational anomalies.

4.1 Implicit representation of theories

To this point, we modeled theory $T(\cdot)$ as a reduced-form black-box that can draw implications about the relationship between the features and modeled outcomes from any hypothetical dataset and placed no restrictions on the behavior of this black-box across feature values. However, a theory may draw similar implications across feature values x, x' and we capture this with the following definition.

Definition 4.1. Features $x_1, x_2 \in \mathcal{X}$ are *representationally equivalent* under theory $T(\cdot)$ if $T(x; D) = T(x; D')$ for all $D \in \mathcal{D}$.

Two feature values are representationally equivalent if theory $T(\cdot)$ draws the same implications on these values for all possible hypothetical datasets. The theory always behaves as-if it groups together these two features and derives the same implications across them.

Furthermore, representational equivalence has a simple statement in terms of a theory's allowable functions.

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

This is an immediate consequence of Proposition 2.1. That is, two feature values are representationally equivalent under theory $T(\cdot)$ if and only if all allowable functions assign the same modeled outcome values to them. It need not be the case that each allowable function assigns the same modeled outcome value.

We next introduce a strengthening of Axiom 4 (”non-trivial implications”) that implies theory $T(\cdot)$ has a non-trivial, lower-dimensional representation of the features.

Axiom 5 (Sharp implications)**.** There exists pair $x_1, x_2 \in \mathcal{X}$ 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 4.1. *Suppose theory $T(\cdot)$ satisfies Axiom 1, 2, 3 and 5. Then, there exists some pair $x_1, x_2 \in \mathcal{X}$ that are representationally equivalent under theory $T(\cdot)$.*

To prove the result, we establish that the pair $x_1, x_2 \in \mathcal{X}$ in Axiom 5 must be representationally equivalent under theory $T(\cdot)$ by contradiction. If not, there exists some hypothetical dataset $D \in \mathcal{D}$ at which $T(x_1; D) \neq T(x_2; D)$ and we can construct an augmented dataset $\widetilde{D}$ satisfying $D \subset \widetilde{D}$ that is compatible with theory $T(\cdot)$ but violates Axiom 5.

Proposition 4.1 establishes that Axiom 5 is sufficient for there to exist a non-trivial representation of the features under theory $T(\cdot)$. Axiom 5 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. To make this more concrete, we return to some of our earlier examples to illustrate that Axiom 5 is often satisfied in economic theories.

Example: choice under uncertainty 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 that is 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 $\lambda(p_0, z_0) + (1-\lambda)(\widetilde{p}, \widetilde{z})$ and $\lambda(p_1, z_1) + (1-\lambda)(\widetilde{p}, \widetilde{z})$, $f(x_1) = f(x_2)$ due to the linearity in probabilities of each allowable function. Expected utility theory therefore satisfies Axiom 5. Proposition 4.1 implies that any pair of menus x_1, x_2 of this form are representationally equivalent under expected utility theory. ▲

Example: asset pricing Consider again the evolution of asset prices over time and the capital asset pricing model (CAPM). 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 Axiom 5. Any pair of features x_1, x_2 of this form are representationally equivalent under CAPM. ▲

4.2 Taxonomy of anomalies

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

Observation 4.1. Consider any anomaly $T(\cdot)$ satisfying Axioms 1, 2, 3 and 5. Any 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 $(x_1, y_2^*), (x_1, y_2^*) \in D$ 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 behave differently in nature. A representational anomaly therefore highlights that there is some dimension of nature that is not captured by the theory's allowable functions. By contrast, we refer to anomalies satisfy Observation 4.1(ii) as *specification anomalies.* Specification anomalies highlight that while theory $T(\cdot)$ models variation across the features, it does so incorrectly.

Researchers in economics 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 vary across these two menus. In this sense the Allais Paradox is an example of a representational anomaly and satisfies Observation 4.1(i). Indeed, other famous examples in decision-making under uncertainty such as the certainty effect or common ratio experiments (e.g., Allais, 1953; Kahneman and Tversky, 1979) are also representational anomalies.

4.3 Dataset morphing for representational anomalies

Our previous result establish that any theory $T(\cdot)$ satisfying Axioms 1, 2, 3 and 5 has a non-trivial representation of the features. We next ask whether there exists *empirical* representational anomalies and if so how we might find them. That is, given modeled contexts $m \in \mathcal{M}$ with true functions $f_m^*(\cdot)$, we 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 theory $T(\cdot)$'s representation is *local.*

Assumption 4.1 (Diffentiability and local representational equivalence)**.**

1. $f_m^*(\cdot)$ and all $f(\cdot) \in \mathcal{F}^T$ are 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 $\lambda \in (0, 1)$ (i.e., $f(x_1) = f(x_2) = f(\lambda x_1 + (1 - \lambda) x_2)$ for any allowable function $f(\cdot) \in \mathcal{F}^T$.

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 4.1, we might hope to uncover representational anomalies by taking local steps. Suppose we have oracle access to the true function $f_m^*(\cdot)$. Given an initial feature values x^0, we would like to 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 then *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 4.1, $\mathcal{N}^T(x)$ is non-empty at any x for which there exists some representationally equivalent x'. For an initial feature value x^0, step size η and maximum number of iterations, we would iterate over $t = 0, \dots, T$ and compute the update step

$$x^{t+1} = x^t - \eta \text{Proj}\left(\nabla f_m^*(x^t) \mid \mathcal{N}^T(x^t)\right), \tag{13}$$

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 the same idea applies if we instead constructed an ascent step.

Of course, this is not feasible since we do not directly observe the true function $f_m^*(\cdot)$. But recall that $f_m^*(\cdot)$ is an identified functional of the joint distribution of the observable data in modeled context m – that is, $f_m^*(x) = \mathbb{E}_m\left[g(Y_i) \mid X_i = x\right]$ for some known function $g(\cdot)$. We now construct an estimator $\widehat{f}_m^*(\cdot)$ based on a random sample $(X_i, Y_i) \sim P_m(\cdot)$ i.i.d. for $i = 1, \dots, n$. We then plug-in this estimator into the morphing procedure and apply the update step

$$x^{t+1} = x^t - \eta \text{Proj}\left(\nabla \widehat{f}_m^*(x^t) \mid \mathcal{N}^T(x^t)\right). \tag{14}$$

Our next result establishes that this remains a descent direction for the true function $f_m^*(\cdot)$ provided the error $\nabla \widehat{f}_m^*(x^t) - \nabla \widehat{f}_m^*(x^t)$ is sufficiently small.

Proposition 4.2. *Under Assumption 4.1, $-$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 $\|\nabla \widehat{f}_m^*(x) - \nabla f_m^*(x)\| \leq \|$Proj$\left(\nabla f_m^*(x) \mid \mathcal{N}^T(x)\right)\|$ is satisfied.*

While Proposition 4.2 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 3.1.2. With this parametrization, we 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.

Algorithm 2: Feasible dataset morphing for representational anomalies.

```
Input: $\widehat{f}^*_m(\cdot)$, $B > 0$, maximum iterations $T$, learning rate $\eta$, initial feature $x^0$.
$t \leftarrow 0$;
while $t < T$ do
    Sample $\theta_b \in \Theta$ for $b = 1, \ldots, B$;
    Construct $\mathcal{N}^T_\Theta(x^t) = \{v \in \mathbb{R}^{dim(x)}$ s.t. $\nabla f_{\theta_b}(x_0)^T v = 0$ for all $b\}$;
    $x^{t+1} \leftarrow x^t - \eta \mathrm{Proj}\left(\nabla \widehat{f}^*_m(x^t) \mid \mathcal{N}^T(x^t)\right)$;
    $t \leftarrow t + 1$;
return $n^*$, $\{x_{1:n^*} \colon \mathcal{E}^T_m(x_{1:n^*}) > 0\}$.
```

4.4 Average representational anomalies across modeled contexts

Our procedure so far searches for representational anomalies in a single modeled context, whereas we may be most empirically interested in generating representational anomalies across many modeled contexts $m \in \mathcal{M}$. Our morphing procedure can be extended across modeled contexts to search for "average" representational anomalies.

As in Section 3.2, suppose we 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 *average* representational anomaly as a pair of features x_1, x_2 such that $\bar{f}^*(x_1) \neq \bar{f}^*(x_2)$. We next show that if there are no compositional changes in modeled contexts across these features, then x_1, x_2 is an average representational anomaly if and only if it is a representational anomaly in some modeled context m.

Proposition 4.3. *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 $x_{1:n}$ 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.*

The condition in Proposition 4.3 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 representational anomaly in any modeled context. Proposition 4.3 suggests that we can search for 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 now depends on the quantity $\|\nabla \widehat{\bar{f}}^*(\cdot) - \nabla \bar{f}^*(\cdot)\|_2$. By pooling data across modeled contexts, we may hope to obtain better control of this estimation error.

5 Simulations for Choice under Uncertainty

In this section, we explore the behavior of our procedures by generating anomalies for expected utility theory in simulated lottery choice data from an individual whose preferences 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 the anomalies generated by our procedures against known anomalies for expected utility theory such as Allais (1953), Kahneman and Tversky (1979) and others. Our procedures recover known anomalies for probability weighting functions. Intriguingly, we also uncover *novel* anomalies for expected utility theory that differ from those which spurred the development of cumulative prospect theory. This suggests that lottery choices consistent with cumulative prospect theory imply the existence of new Allais Paradox-like anomalies. In future drafts, we will apply our anomaly generation procedures to uncover anomalies for expected utility theory using real lottery choice data collected from subjects.

5.1 Simulation design

We simulate lottery choice data from an individual m with CRRA utility function parameterized by $\rho \geq 0$. For $\rho \neq 1$, we define the utility function as

$$u(z; \rho) = \begin{cases} \frac{z^{1-\rho}-1}{1-\rho} \text{ if } z \geq 0 \\ -\frac{(-z)^{1-\rho}-1}{1-\rho} \text{ if } z < 0. \end{cases} \tag{15}$$

For $\rho = 1$, we define $u(z) = \log(z)$ for $z > 0$ and $-\log(-z)$ for $z < 0$. Throughout, we set $\rho = 0.5$ when generating lottery choices by this individual.

Importantly, the individual evaluates lotteries with the parametric probability weighting function

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

for $p \in \Delta^{J-1}$ the vector of lottery probabilities and $\delta \geq 0$, $\gamma \geq 0$ parameters governing the curvature and level of the probability weighting function (Lattimore, Baker and Witte, 1992). We set $\delta = 0.1, \gamma = 0.1$ throughout, plotting the resulting probability weighting function in Figure 1 below. For this choice of parameter values, the individual distorts probabilities by over-weighting probabilities close to zero and under-weighting probabilities close to one. This non-linearity in the probability weighting function can generate several known anomalies for the independence axiom of expected utility, such as the Allais Paradox (Table 1) or the certainty effect (e.g., Kahneman and Tversky, 1979). In addition to non-linearity in the probability weighting function, this choice of parameter values additionally introduces "subcertainty" or "prospect pessimism," meaning that the individual's probability weights do not sum to one (i.e., $\sum_{j=1}^{J} \pi_j(p; \delta, \gamma) < 1$). Subcertainty in the probability weighting function implies that the individual's choices may violate first-order stochastic dominance, meaning the individual may select a lottery in the menu that is first-order stochastically dominated by the other lottery. This is another anomaly that we may hope to find since expected utility maximization over any utility function that is weakly increasing in payoffs

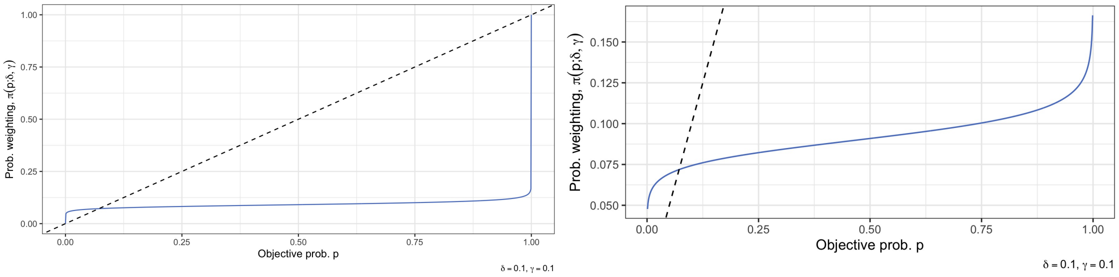

Figure 1: Probability weighting function (16) for level parameter $\delta = 0.1$ and curvature parameter $\gamma = 0.1$

cannot violate first-order stochastic dominance.

For some payoff vector $z \in \mathbb{R}^J$, the individual evaluates lottery (p, z) by its subjective expected utility $EU(p, z; \delta, \gamma, \rho) := \sum_{j=1}^{J} \pi_j(p; \delta, \gamma) u(z_j; \rho)$ 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 $f_m^*(x) = P\left(EU(p_1, z_1; \delta, \gamma, \rho) - EU(p_0, z_0; \delta, \gamma, \rho) + \xi \geq 0\right)$, where ξ is some i.i.d. logistic shock.

We then generate anomalies for expected utility theory by applying our adversarial procedure and dataset morphing procedure to the true choice probability function $f_m^*(\cdot)$. To do so, we flexibly parametrize the allowable functions of expected utility theory $\mathcal{F}^T = \{f_\theta(\cdot) \colon \theta \in \Theta\}$ by specifying the utility function must be some linear combination of a polynomial basis or monotone I-splines (e.g., Ramsay, 1988) – that is, we set $u_\theta(z) = \sum_{k=1}^{K} \theta_k b_k(z)$ for some basis functions $b_1(\cdot), \ldots, b_K(\cdot)$ and the parameters $\theta \in \Theta$ specify the weights on each basis function. We then generate anomalies for expected utility theory over lotteries on two payoffs ("binary lotteries") and lotteries on three payoffs ("ternary lotteries"). In Appendix D, we provide more details on our practical implementation.

5.2 Anomalies generated by the probability weighting function with subcertainty

First consider the individual with probability weighting function (16) parameterized by $\delta = 0.1, \gamma = 0.1$. Our adversarial procedure and dataset morphing procedure uncover two distinct categories of anomalies in this case.

We provide several examples of menus of binary lotteries and an example of a menu of ternary lotteries from the first category in Table 2 generated by our adversarial procedure. All anomalies in the first category are violations of first-order stochastic dominance – the individual is selecting a lottery that is first-order stochastically dominated in the menu. As mentioned earlier, it is well-known that probability weighting functions that demonstrate subcertainty can produce this behavior. Furthermore, it is generally viewed that such first-order stochastic dominance violations may be an undesirable "bug" in the particular specification of the probability weighting function since we may be unlikely to hold in real choices.[9] In Appendix D, we report first-order stochastic dominance anomalies that were generated by our dataset morphing procedure. What is intriguing is that our anomaly generation procedures uncovered these first-order stochastic dominance violations on its own.

(a) Generated Anomaly #1 (x_1)

Lottery 0	12.453	15.295
	0.882	0.118
Lottery 1	1.704	4.470
	ε	$1-\varepsilon$

(b) Generated Anomaly #2 (x_2)

Lottery 0	7.260	12.124
	0.747	0.253
Lottery 1	1.687	4.760
	ε	$1-\varepsilon$

(c) Generated Anomaly #3 (x_3)

Lottery 0	7.444	14.943
	0.626	0.374
Lottery 1	3.212	5.501
	ε	$1-\varepsilon$

(d) Generated Anomaly #4 (x_4)

Lottery 0	0.764	1.735	2.850
	ε	ε	$1-\varepsilon$
Lottery 1	9.778	14.593	15.232
	0.306	0.092	0.602

Table 2: Examples of generated first-order stochastic dominance anomalies for the probability weighting function with subcertainty ($\delta = 0.1$, $\gamma = 0.1$).

Notes: We color the lottery selected by the individual with probability at least 0.5 in green. Since the gradient of the probability weighting function $\pi(p; \delta, \gamma)$ in (16) diverges as $p \to 0$ and $p \to 1$, we clip the probabilities to be bounded below by ε and $1 - \varepsilon$ for $\varepsilon = 1 \times 10^{-6}$. These anomalies are produced using our adversarial procedure and polynomial basis function parametrization of expected utility theory. See Appendix D for additional discussion.

The second category of anomalies, however, appears to be a genuine discovery. Table

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

3 provides several examples of pairs of menus of binary lotteries from this second category. These are anomalies for expected utility theory that arise due to the non-linearity of the probability weighting function and therefore are violations of the independence axiom. To make this more concrete, consider the pair of menus of binary lotteries in Table 3(a) generated by our dataset morphing procedure. We first observe that the lotteries in menu B can be expressed as a compound lottery over the lotteries in menu A and some degenerate lottery that yields a payoff with certainty. That is, lottery B0 can be expressed as a compound lottery over lottery A0 and a degenerate lottery that yields payoff 9.196 with certainty; that is, $B0 = \alpha_0 A0 + (1-\alpha_0)\delta_{9.196}$ for some $\alpha_0 \in [0,1]$. Analogously, lottery $B1$ can be expressed as $B1 = \alpha_1 A1 + (1-\alpha_1)\delta_{4.114}$ for some $\alpha_1 \in [0,1]$. The individual's choices in these menus express the preference relation

$$A0 \succ A1 \text{ and } \alpha_1 A1 + (1-\alpha_1)\delta_{4.114} \succ \alpha_0 A0 + (1-\alpha_0)\delta_{9.196}. \tag{17}$$

However, this contradicts the independence axiom of expected utility theory since it can be shown that $A0 \succ A1$ must imply that $\alpha_0 A0 + (1-\alpha_0)\delta_{9.196} \succ \alpha_1 A1 + (1-\alpha_1)\delta_{4.114}$ (see Appendix D for the proof). The only choices that can be expressed over this pair of menus is $\{(f(x_A), f(x_B)) \colon f(\cdot) \in \mathcal{F}^T\} = \{(A0, B0), (A1, B1)\}$. A similar argument can be applied to show that the pair of menus in Table 3(b) is an anomaly as well.

These anomalies for the independence axiom have a common structure. In particular, defining the pair of lotteries $\ell_0 = (p_0, z_0)$, $\ell_1 = (p_1, z_1)$ with $z_0 = (z_{0,1}, z_{0,2})$ and $z_1 = (z_{1,1}, z_{1,2})$ with $\overline{z}_0 := \max_{j \in \{1,2\}} z_{0j} > \min_{j \in \{1,2\}} z_{1j} := \underline{z}_1$, both menus in Table 3 can be summarized as a more general pattern for this parameterization of the probability weighting function: for $\alpha_0 > \alpha_1$, 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 $\alpha_0 \ell_0 + (1-\alpha_0)\delta_{\overline{z}_0}$ and $\alpha_1 \ell_1 + (1-\alpha_1)\delta_{\underline{z}_1}$. Lottery B0 mixes lottery A0 with a certain payoff equal to its maximal payoff and lottery B1 mixes lottery A1 with a certain payoff equal to its minimal payoff. We therefore refer to this as

(a) Generated Anomaly #1

	Menu A (x_A)	
Lottery 0	1.751	9.196
	0.470	0.530
Lottery 1	4.114	5.304
	0.875	0.125
	Menu B (x_B)	
Lottery 0	1.751	9.196
	0.439	0.561
Lottery 1	4.114	5.304
	0.989	0.011

(b) Generated Anomaly #2

	Menu A (x_A)	
Lottery 0	4.779	10.143
	0.080	0.920
Lottery 1	5.901	7.833
	0.036	0.964
	Menu B (x_B)	
Lottery 0	4.779	10.143
	0.009	0.991
Lottery 1	5.901	7.833
	0.001	0.999

Table 3: Examples of generated independence axiom anomalies for the probability weighting function with subcertainty ($\delta = 0.1$, $\gamma = 0.1$) over binary lotteries.

Notes: We color the lottery selected by the individual with probability at least 0.5 in green. These anomalies are produced using our dataset morphing procedure and I-spline basis function parametrization of expected utility theory. See Appendix D for additional discussion.

a "dominated consequence" anomaly. While sharing some similarities, this general pattern is importantly different than the common consequence and common ratio effects (e.g., see Machina, 1987), which were important motivating anomalies for the development of the probability weighting function in cumulative prospect theory. It has the distinct feature of highlighting violations of the independence axiom while only using two possible payoffs (like the common ratio effect) but still involving mixing lotteries with particular certain prospects (like the common consequence effect). Our algorithm has therefore uncovered interesting violations of the independence axiom of expected utility theory that solely arise due to the nonlinearity in the probability weighting function.

6 Conclusion

By now, it is clear that machine learning has the capacity to change the way nearly every sector operates (e.g., Brynjolfsson and McAfee, 2014; Agarwal, Gans and Goldfarb, 2018). Why should economic research be any different? Of course, substantial progress is already being made in incorporating machine learning into many of the tasks performed by economic researchers, such as digitizing historical archives (e.g., Shen et al., 2021), processing novel data such as text and images for econometric analysis (e.g., Glaeser et al., 2018; Gentzkow, Kelly and Taddy, 2019; Adukia et al., 2021), uncovering treatment effect heterogeneity (Athey and Wager, 2018; Chernozhukov et al., 2018) and 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, summarizes the empirical relationship between some features and modeled outcomes using supervised learning, and then automatic generates anomalies, if they exist. The framework and procedures apply to a wide variety of theories across scientific domains. While our illustration is specific to expected utility theory, we believe these procedures can be applied in any place 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. 2019. "The Impact of Machine Learning on Economics." *The Economics of Artificial Intelligence: An Agenda*, , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 507–547. Chicago:University of Chicago Press.

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.

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

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

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

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.

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.

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

Dekel, Eddie. 1986. "An axiomatic characterization of preferences under uncertainty: Weakening the independence axiom." *Journal of Economic Theory*, 40(2): 304–318.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty."

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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."

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

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.

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.

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

Tversky, Amos, and Daniel Kahneman. 1986. "Rational Choice and the Framing of Decisions." *The Journal of Business*, 59(4): S251–S278.

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, Paul Slovic, and Daniel Kahneman. 1990. "The Causes of Preference Reversal." *The American Economic Review*, 80(1): 204–217.

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

Varian, Hal R. 1990. "Goodness-of-fit in optimizing models." *Journal of Econometrics*, 46(1): 125–140.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, 28(2): 3–28.

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.

From Predictive Algorithms to Automatic Generation of Anomalies

Online Appendix

Sendhil Mullainathan & Ashesh Rambachan

A Omitted Proofs

Proof of Proposition 2.1

To prove part (i), we first note that the main text established that the allowable function representation (1) satisfies Axioms 1-4. This establishes necessity. We prove sufficiency here. Consider any theory $T(\cdot)$ satisfying Axioms 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 Axiom 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 Axiom 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

$$T(x; D) = \{f(x) \colon f(\cdot) \in \mathcal{F}^T \text{ consistent with } D\} \tag{18}$$

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

By Axioms 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 A.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 A.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 $f(\cdot) \notin \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\}$.

Suppose for sake of contradiction there exists some $f(\cdot) \notin T(\cdot; D)$ with $f(\cdot) \in \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\}$. Since any $f(\cdot)$ that is not consistent with D cannot be an element of $\{f(x) \colon f(\cdot) \in \mathcal{F}^T$ 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 Axiom 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 Axiom 2. But this then contradicts Axiom 3 since $T(x; D') \not\subseteq T(x; D)$. □

Lemma A.1 implies $T(x; D) \subseteq \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\}$ for all $x \in \mathcal{X}$. Lemma A.2 establishes that $\{f(x) \colon f(\cdot) \in \mathcal{F}^T$ consistent with $D\} \subseteq T(x; D)$. It therefore follows that $T(x; D) = \{f(x) \colon f(\cdot) \in \mathcal{F}^T$ 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 Axiom 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. □.

Proof of Proposition 3.1

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 2.3. If there exists no incompatible datasets 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 $(x, y^*) \in D$. □

Proof of Proposition 3.2

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 A.3. *Under the same conditions as Proposition 3.2,*

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

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

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

Lemma A.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

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

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

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 α-strongly convex in its second argument, (a) is bounded above by

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

$$n^{-1}K\|\widehat{f}^*_m(x_{1:n})-f^*_m(x_{1:n})\|_1-\frac{\alpha n^{-1}}{2}\|\widehat{f}^*_m(x_{1:n})-f^*_m(x_{1:n})\|_2^2\leq(K+\frac{\alpha}{2})\|\widehat{f}^*_m(x_{1:n})-f^*_m(x_{1:n})\|_\infty$$

where we defined the shorthand notation $f(x_{1:n})=(f(x_1),\ldots,f(x_n))$, (1) uses that the loss function has bounded gradients, and (2) uses the inequalities $\|f(x_{1:n})\|_1\leq n\|f(x_{1:n})\|_\infty$ and $\|f(x_{1:n})\|_2^2\leq n\|f(x_{1:n})\|_\infty$. Next, we can rewrite (b) as being bounded by

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

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

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

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

$$n^{-1}\sum_{i=1}^{n}\left\{\nabla_2\ell\left(f^T(x_i;x_{1:n}),\widehat{f}^*_m(x_i)\right)(f^*_m(x_i)-\widehat{f}^*_m(x_i))-\frac{\alpha}{2}\|f^*_m(x_i)-\widehat{f}^*_m(x_i)\|^2\right\}\overset{(3)}{\leq}$$

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

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

where (1) uses that the loss is strongly convex in its second argument, (2) uses $n^{-1}\sum_{i=1}^{n} \ell(f^T(x_i; x_{1:n}), \widehat{f}^*_m(x_i)$ $n^{-1}\sum_{i=1}^{n} \ell(\widehat{f}^T(x_i; x_{1:n}), \widehat{f}^*_m(x_i))$, and (3) again uses that the loss is strongly convex in it second argument. By the same argument as before, it follows that this is bounded by

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

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

Proof of Proposition 3.3

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

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

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

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

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

which in turn implies that $n^{-1}\sum_{i=1}^{n} P(m \mid x_i) \left(f^T_m(x_i; x_{1:n}) - f^*_m(x_i)\right)^2 > 0$ 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 a 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. $\square$

Proof of Proposition 4.1

We first observe that Axiom 5 implies Axiom 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 Axiom 5 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 Axiom 5. 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 Axiom 3 implies that $y_1^* \notin T(x_2; \widetilde{D})$, contradicting Axiom 5. $\square$

Proof of Proposition 4.2

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. $\square$

Proof of Proposition 4.3

To prove this result, we 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. □

B Additional Examples for the Model of Scientific Theories

In this section, we introduce anomalies for other theories in economics and show how these theories satisfy Axioms 1-4.

B.1 Anomalies for other examples

Initial play in normal-form games Consider the normal-form game in Table 4. 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, $T(x; D) = \emptyset$ 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 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, such a hypothetical dataset would be an anomaly for Nash equilibrium. ▲

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 covariances. Consider a dataset D with two hypothetical observations $(x_1, y_1^*), (x_2, y_2^*)$, where x_1, x_2 are such that the risk-free rate, market return and covariances are constant yet y_1^*, y_2^* vary. By construction, such a hypothetical dataset would be an 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. ▲

B.2 Axiomatization for other examples

Initial play in normal-form games We define Nash equilibrium as 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, $T(x; D) = \emptyset$ for all $x \in \mathcal{X}$. We immediately observe that Axiom 1, Axiom 2, and Axiom 4 are satisfied by construction.

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

Axiom 3 is also satisfied as $T(x; D') \subseteq T(x; D)$ for all pairs of datasets D, D' with $D \subseteq D'$. ▲

Asset pricing We observe that CAPM as described in the main text immediately satisfies Axiom 1 and Axiom 2 on hypothetical datasets of moments of historical asset prices. Second, consider any pair of datasets 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, Axiom 3 is satisfied. Finally, Axiom 4 is satisfied for any dataset D that either point or partially identifies the assets' parameter β_j.

C Analysis of Gradient Descent Ascent Optimization over Allowable Functions

In Section 3.1.2 of the main text, we proposed a gradient descent ascent (GDA) procedure to optimize the max-min optimization 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

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

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

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

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 properties of the GDA procedure will be stated in terms of 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 $\widehat{\mathcal{E}}(x_{1:n})$.

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

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

where ∂ denotes the subdifferential of a weakly convex function.

Proposition C.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*

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

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

D Implementation Details and Additional Results for Choice under Uncertainty Simulations

In this section of the Appendix, we describe the implementation details for our adversarial learning algorithm and dataset morphing algorithm in the choice under uncertainty simulations. We also present anomalies uncovered by our dataset morphing procedure. The main text focuses on anomalies generated by the adversarial procedure.

D.1 Implementation details of anomaly generation procedures

In this section, we describe the implementation details of our anomaly generation procedures in more detail.

For both the adversarial procedure and dataset morphing procedure, we clip the lottery probabilities at either ε or $1 - \varepsilon$ for $\varepsilon = 1 \times 10^{-6}$ during each gradient descent step (that is, if the gradient descent step pushes some probability p_j to be less than ε or greater than $1 - \varepsilon$, we round this value back to ε or $1 - \varepsilon$ respectively). We implement this clipping because the gradient of the parametric probability function $\pi_j(p; \delta, \gamma) = \frac{\delta p_j^\gamma}{\delta p_j^\gamma + \sum_{k \neq j} p_k^\gamma}$ for $j = 1, \ldots, J$ diverges to infinity as $p_j \to 0$ or $p_j \to 1$ for $\gamma < 1$.

For both the adversarial procedure and dataset morphing procedure, we constructed randomly initialized menus of lotteries in the following manner. We simulated the payoffs of the lottery from a truncated normal distribution with mean zero and variance equal to ten. We simulated the probabilities by drawing each lottery probability uniformly from the unit interval, and then normalizing the draws so they lie on the simplex.

D.1.1 Adversarial procedure

To implement the adversarial procedure based on gradient descent ascent described in Section 3.1.2, we must first specify a parametric basis for the allowable functions of expected utility theory. In the simulations, we parametrize the utility function of the individual $u_\theta(\cdot)$ as either 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 also experimented with varying the choice of learning rate $\eta \in \{0.5, 0.1, 0.05, 0.01, 0.005, 0.0005\}$.

For any particular choice of utility function basis and learning rate, we ran the gradient descent ascent procedure for 500 randomly initialized menus x^0. We set the maximum number of iterations to be $T = 30$. For either choice of utility basis function, we solve the inner minimization problem (11) 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 (12) directly. A subtlety 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 (12) 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 the expected utility difference between the lotteries is positive but $f_m^*(x^t) < 0.5$ and vice versa. We consider either taking gradient ascent steps on all of the payoffs and probabilities in the menu or only taking gradient ascent steps on the probabilities in the menu. We then collect together the anomalies produced across all runs of the adversarial procedure and report only a handful in the main text of the paper.

D.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 a 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 $q = 10$ know points and degree $K = 3$. We also experimented with varying the choice of learning rate over $\eta \in \{0.1, 0.5, 1.0, 5.0, 10.0\}$. We presented anomalies generated with $\eta = 10$ in the main text.

For any particular utility function basis and learning rate, we ran the dataset morphing procedure for 500 randomly initialized menus x^0. We set the maximum number of iterations to be $T = 30$. We consider either taking gradient ascent steps on all of the payoffs and probabilities in the menu or only taking gradient ascent steps on the probabilities in the menu. We then collect together the anomalies produced across all runs of the dataset morphing procedure and report only a handful in the main text of the paper.

D.2 Proofs of anomalies for independence axiom

In this section, we now prove that the pairs of menus of lotteries presented in Table 3 are anomalies for expected utility theory and inconsistent with the independence axiom.

First, consider the pair of menus of binary lotteries depicted in Table 3(a). As mentioned in the main text, we first observe that the lotteries in menu B can be expressed as a compound lottery over the corresponding lottery in menu A and some degenerate lottery that yields a payoff with certainty. In particular, $B0 = \alpha_0 A0 + (1 - \alpha_0)\delta_{9.196}$ for $\alpha_0 = 0.439/0.470$ and $B1 = \alpha_1 A1 + (1 - \alpha_1)\delta_{4.114}$ for $\alpha_1 = 0.011/0.125$, where $\alpha_1 < \alpha_0$. The individual's choices on these menus express the preference relation $A0 \succ A1$ and $\alpha_1 A1 + (1-\alpha_1)\delta_{4.114} \succ \alpha_0 A0 + (1 - \alpha_0)\delta_{9.196}$. We next observe that $A0 \succ A1$ implies that $\alpha_0 A0 + (1 - \alpha_0)\delta_{9.196} \succeq \alpha_0 A1 + (1 - \alpha_0)\delta_{9.196}$ by the independence axiom. A further application of the independence axiom and that the utility function must be weakly increasing in payoffs implies that $\alpha_0 A_1 + (1 - \alpha_0)\delta_{9.196} \succeq \alpha_0 A_1 + (1 - \alpha_0)\delta_{4.114}$. Finally, we use the fact that $A_1 \succeq \delta_{4.114}$ by first-order stochastic dominance and $\alpha_0 > \alpha_1$ to conclude that $\alpha_0 A_1 + (1 - \alpha_0)\delta_{4.114} \succeq \alpha_1 A_1 + (1 - \alpha_1)\delta_{4.114}$. By tying these preference relations together through transitivity, we have shown that if $A0 \succ A1$ then it must be the case that $\alpha_0 A0 + (1 - \alpha_0)\delta_{9.196} \succeq \alpha_1 A1 + (1 - \alpha_1)\delta_{4.114}$ for any preference consistent with expected utility theory at some weakly increasing utility function. It therefore follows that Table 3(a) is an anomaly.

Second, consider the pair of menus of binary lotteries depicted in Table 3(b). To show that this is an anomaly, we observe that the lotteries in menu A can be written as a compound lottery over the corresponding lotteries and some degenerate lottery that yields a payoff with certainty. In particular, $A0 = \alpha_0 B0 + (1 - \alpha_0)\delta_{4.779}$ for $\alpha_0 = 0.92/0.991$ and $A1 = \alpha_1 B1 + (1 - \alpha_1)\delta_{5.901}$ for $\alpha_1 = 0.964/0.999$ with $\alpha_0 < \alpha_1$. The individual's choices on these menus express the preference relation $B1 \succ B0$ and $\alpha_0 B0 + (1 - \alpha_0)\delta_{4.779} \succ \alpha_1 B1 + (1 - \alpha_1)\delta_{5.901}$. We next observe that $B1 \succ B0$ implies that $\alpha_1 B1 + (1 - \alpha_1)\delta_{5.901} \succeq \alpha_1 B0 + (1 - \alpha_1)\delta_{5.901}$ by the independence axiom. A further application of the independence axiom and that the utility function must be weakly increasing in payoffs implies that $\alpha_1 B0 + (1 - \alpha_1)\delta_{5.901} \succeq \alpha_1 B0 + (1 - \alpha_1)\delta_{4.779}$. Finally, we use the fact that $B0 \succeq \delta_{4.779}$ and $\alpha_0 < \alpha_1$ to conclude that $\alpha_1 B0 + (1 - \alpha_1)\delta_{4.779} \succeq \alpha_0 B0 + (1 - \alpha_0)\delta_{4.779}$. It therefore follows that Table 3(b) is an anomaly.

D.3 Additional anomalies generated by the probability weighting function with subcertainty

Our dataset morphing procedure also uncovered anomalies that are violations of the first-order stochastic dominance for the probability weighting function with $\delta = 0.1, \gamma = 0.1$. We provide several such examples in Table 5 below.

(a) Generated Anomaly #1 (x_1)

Lottery 0	5.883	15.903
	ε	$1 - \varepsilon$
Lottery 1	19.597	19.842
	0.943	0.057

(b) Generated Anomaly #2 (x_2)

Lottery 0	6.555	7.951
	ε	$1 - \varepsilon$
Lottery 1	11.870	12.563
	0.396	0.604

Table 5: Examples of generated first-order stochastic dominance anomalies for the probability weighting function with subcertainty ($\delta = 0.1$, $\gamma = 0.1$).

Notes: We color the lottery selected by the individual with probability at least 0.5 in green. Since the gradient of the probability weighting function $\pi(p; \delta, \gamma)$ in (16) diverges as $p \to 0$ and $p \to 1$, we clip the probabilities to be bounded below by ε and $1 - \varepsilon$ for $\varepsilon = 1 \times 10^{-6}$. These anomalies are produced using our dataset morphing procedure and I-spline basis function parametrization of expected utility theory. See Appendix D for additional discussion.