# Implicit Preferences*

Tom Cunningham† Jonathan de Quidt‡

May 1, 2022

Latest version here

**Abstract**

We show how simple decisions can, by themselves, reveal two layers of preference. Consider a hiring manager who always chooses a woman over a man with the same qualifications, but always chooses the man if their qualifications differ. Intuitively, these intransitive choices reveal an *explicit* preference for women, but an *implicit* preference for men. More generally, we define an implicit preference for an attribute as one whose influence increases as the attribute is mixed with a superset of other attributes ("dilution"). We show that implicit preferences arise under a diverse set of psychological foundations: rule-based decision-making, signaling motives, and implicit associations. We prove a representation theorem for the model and show how implicit preferences can be identified from binary choices, or joint evaluation data. We apply the model to two published datasets, finding evidence for implicit risk preferences, implicit selfishness, and implicit discrimination.

*We thank for comments, among others, Ingvild Almås, Roland Benabou, Colin Camerer, Ed Glaeser, Karin Hederos, Sendhil Mullainathan, Pietro Ortoleva, Antonio Rangel, Alex Rees-Jones, Anna Sandberg, Sebastian Schweighofer-Kodritsch, Tomasz Strzalecki, Florian Zimmermann, and seminar audiences at USC, Columbia, Caltech, Stanford, Facebook, Harvard, Princeton, Santa Cruz, Gothenburg, Waseda, Berlin, VIBES, and the Stockholm School of Economics. First circulated in 2015 as "Implicit Preferences Inferred from Choice."

†Twitter. tom.cunningham@gmail.com

‡Institute for International Economic Studies, Stockholm, CAGE, CEPR, CESifo, ThReD. jonathan.dequidt@iies.su.se

# 1 Introduction

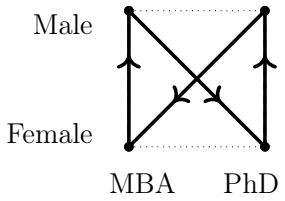
> *"However we may conceal our passions under the veil, there is always some place where they peep out"* - La Rochefoucauld.

Inconsistencies in decision-making are often described as arising from a conflict between opposing motives. In this paper we a formalize a common intuition about how motives interact, and fully characterize the testable implications. Our theory is consistent with a variety of psychological foundations for the underlying conflict.

Suppose you observe a hiring manager's choices between pairs of job applicants who differ in gender, and have either an MBA or a PhD. You notice that:

1. They choose the woman when the candidates' qualifications are the same,
2. They choose the man when the candidates' qualifications differ.

Using $A \succ B$ to represent the choice of $A$ from $\{A, B\}$, we can visualize these choices:




The choices are intransitive and therefore inconsistent with standard utility maximization. But they form an intuitive "figure 8" pattern, and seem to reveal two distinct attitudes towards gender: favoring women in the vertical choice sets, and favoring men in the diagonals.

We generalize this observation of decisions revealing two distinct preferences. We study preferences over bundles of binary attributes (male/female, Black/White, aisle/window), and we assume that the decision maker has an implicit preference for each attribute (positive, neutral, or negative). We identify implicit preferences from behavior with a "dilution" assumption: the implicit preference for an attribute has more influence when the choice set mixes that attribute with a superset of other attributes. In the example above the diagonal choice sets mix gender with qualification, strengthening the influence of the decision maker's implicit preference for men over women, causing the intransitivity.

The model can also be applied to *evaluation* data, such as willingness to pay, teachers' grading decisions, or judges' sentencing decisions, when the evaluation involves a comparison. Suppose our manager is setting wages for pairs of new hires, one male and the other female. In our model, the manager's implicit preference for men over women will make the *man's* wage sensitive to the *woman's* attributes. For example, we would predict that the man's

wage is always lower when the woman has the same qualification as him, compared to when she has a different qualification.

Our formal model presupposes a comparative utility function $u(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{x}$ is the bundle under consideration (which we will call the "target"), and $\boldsymbol{z}$ is its "comparator." We assume that implicit preferences over attributes are separable and their influence increases when the comparison becomes more "opaque," where opacity is derived from a partial order over the set of all possible comparisons. Our core result is a representation theorem stating that a dataset, consisting of inequalities between comparative utilities, can be rationalized by a given set of implicit preferences if and only if there does not exist a matching between inequalities such that bundles with implicitly-preferred attributes are ranked higher when opacity is lower, and vice versa.

The theorem is general with respect to the partial order applied to opacity. To apply the model we assume (1) opacity depends only on which attributes are shared and which differ between target and comparator ($|\boldsymbol{x} - \boldsymbol{z}|$); (2) implicit preferences obey "dilution," meaning that opacity with respect to an attribute is higher in comparisons that mix that attribute with a superset of other attributes. For evaluation data we add a third assumption which permits additional identification.

The "figure 8" pattern in choice unambiguously identifies an implicit preference for men over women. We describe a number of other intuitive patterns in choice ("right triangle," "parallel triangles," "square") and in evaluation ("scissor," "parallel scissor"), and show what they reveal about the decision maker's implicit preferences. The examples we provide are easy to test for in empirical applications, and lend themselves to implementation in experiments.

Our definition of "implicitness" is behavioral, similar to decision-theoretic concepts like "complementarity" or "elasticity" which are defined without reference to the underlying psychology. But why might decisions exhibit implicit preferences? We next provide three foundational models that exhibit implicit preferences, and satisfy all of our opacity assumptions.

The first foundation (*ceteris paribus*) is a decision maker who is subject to a set of rules that apply in "all else equal" situations, and incurs a utility penalty if they break a rule. When the penalty is infinitely large it represents a hard constraint, a special case of models in which the decision maker chooses from a subset of elements that are maximal by some other set of rankings (e.g. Manzini and Mariotti (2007), Masatlioglu et al. (2012), Cherepanov et al. (2013), Ridout (2021)). Under this model our manager prefers men over women. He can follow his preference when gender is mixed with other attributes, but a rule constrains him from choosing a man over an equally-qualified women, generating a cycle.

The second foundation is a *signaling* model: the decision maker has intrinsic preferences over bundles but also cares about others' perceptions of those preferences. This foundation

relates to work on signaling motives (including self-signaling), excuse-driven behavior and "moral wiggle room" (e.g. Bodner and Prelec (2003), Benabou and Tirole (2003, 2006), Norton et al. (2004), Dana et al. (2007), Andreoni and Bernheim (2009), Exley (2016), Bursztyn et al. (2022)). Choosing a male over a female candidate reveals less about the decision maker's gender preferences as gender is mixed with other attributes, giving them more scope to express their intrinsic preference. Thus, under a signaling interpretation, the manager prefers men but wishes people to believe he has a preference for women.

The third foundation is an *implicit associations* decision maker for whom some knowledge is tacit. This foundation relates to psychological theories of implicit bias and unconscious judgment (e.g. Greenwald and Krieger (2006), Greenwald et al. (1998), Kahneman (2011), Rand et al. (2012)). Under this model the hiring manager holds an unconscious positive association towards men, which gives them a "good feeling" about male candidates. Feelings are typically a good signal of quality, but if they have reason to distrust feelings about gender then this can generate inconsistent behavior. When the candidates differ only in gender the decision maker can diagnose the source of their good feeling and overrule it. As gender is mixed with other attributes, it becomes harder to distinguish the influence of gender from other associations, so their judgment is swayed.

The implicit preference in favor of male candidates revealed by the "figure 8" cycle can thus be interpreted in three ways: (1) a sincere preference for men that is sometimes constrained by rules; (2) a sincere preference for men that is sometimes obscured by signaling motives; (3) an unconscious positive association in favor of men that loses its power when it becomes accessible to conscious awareness.

Finally, we apply the model to two existing datasets. Our first application identifies implicit selfishness and implicit risk attitudes in choice data from Exley (2016), our second identifies implicit racism in evaluation data from DeSante (2013).

We do not know of any prior theory which identifies implicit preferences from multiattribute choice. Existing theories of menu-dependent preferences do not predict the figure 8 pattern.[1] Nevertheless we think that the idea of implicit preferences being revealed by more or less "dilute" decisions taps into a commonsense understanding of decision-making, and most of our formal results correspond to natural intuitions.

A set of related theories are proposed by Manzini and Mariotti (2012) (MM) ("choice by

[1]E.g. "salience" (Bordalo et al. (2013)), "relative thinking" (Bushong et al. (2020)), "magnitude effects" (Cunningham (2013)), or "focusing" (Kőszegi and Szeidl (2012)). To the best of our knowledge the only paper besides Cunningham (2016) which identifies a figure 8 pattern in choice is Cubitt et al. (2018) which finds a figure 8 in cross-modal intertemporal tradeoffs. They propose that "the weight put on each attribute of an option is inversely related to how many attributes differ between those options." We discuss in Section 7 how this type of model will not generally exhibit implicit preferences in our sense; Cubitt et al. (2018)'s model does so due to an assumption that money does not have a variable weight.

lexicographic semiorder"), Cherepanov et al. (2013) (CFS) ("rationalization"), and Ridout (2021) (R) ("justification"). In these models the decision maker facing a given choice set chooses the element which maximizes their true preference, out of the subset which are "justifiable," meaning that the element is undominated relative to at least one of a set of given relations. The models differ on the nature of the relations: MM assume a single semiorder, CFS assume multiple binary relations, and R assumes multiple complete orders. We regard this class of models as complementary to ours. The most important difference is that these models treat outcomes as "atomic" while we treat outcomes as bundles of binary attributes. Models with atomic outcomes are more parsimonious, and those models give unambiguous predictions for choice sets with 3 or more elements, which ours does not.[2]

An advantage of using bundles of attributes is that we can infer implicit preferences from binary choice.[3] Additionally, linking implicit preferences to attributes instead of atomic outcomes allows us to make out-of-sample predictions: our hiring manager's gender bias can be predicted to carry over to choice between new candidates with entirely different profiles. Finally, our model can be readily applied to data on joint evaluation, as well as choice. This set of features makes our framework particularly well suited to applied work, and in Section 3 we provide an extensive collection of identification tools that can be implemented in existing datasets and new experiments. We demonstrate this with our own applications. Our tools have also been adopted by others: Barron et al. (2022) apply our approach and find evidence of implicit gender bias.

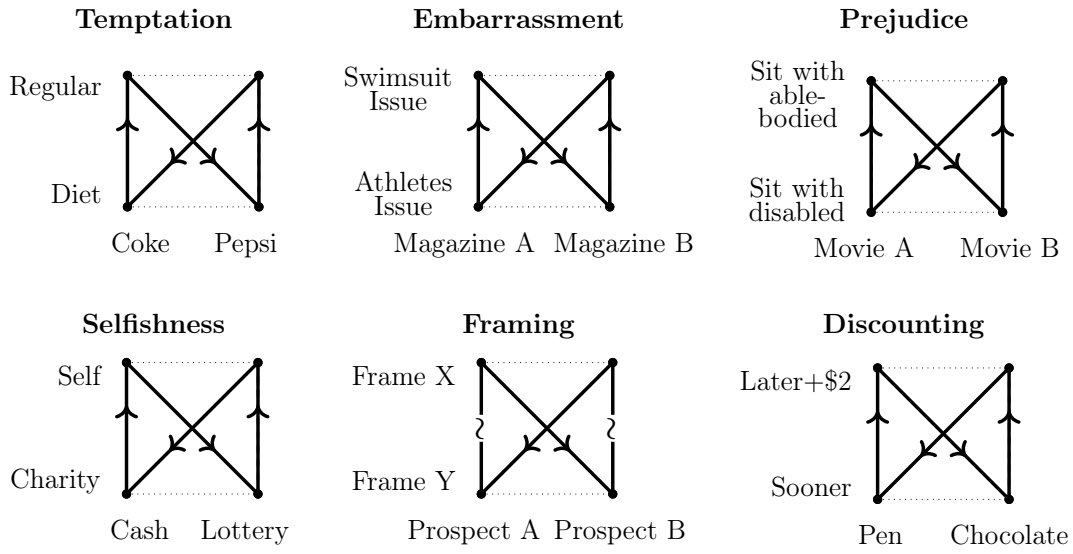
In psychology, the term "implicit" is usually reserved for cognition, attitudes, judgments, preferences, or knowledge that are "outside conscious attentional focus" (Greenwald and Krieger, 2006), often described as "automatic," "unconscious," "associative." In dual-process theories (e.g., Kahneman (2011)) they are associated with the fast "System 1." In contrast, explicit attitudes are those that are stated or revealed deliberately. Psychologists have developed an array of *non-choice* techniques, most notably, the Implicit Association Test (IAT) (Greenwald et al., 1998), which uses response time to measure implicit associations. IATs have been widely adopted, including within economics (e.g. Glover et al. (2017), Corno et al. (2018), Carlana (2019)), however their interpretation, and ability to predict real-world choices remains controversial (Oswald et al., 2013; Greenwald et al., 2015). In contrast our model defines implicit preferences directly from and with reference to real decisions.

A number of prior empirical studies rely on related intuition to our formal model, that underlying motives can be revealed by comparing decisions that vary in their opacity or

[2]Extending our notion of implicit preferences to 3-element comparisons requires additional assumptions.

[3]With atomic elements binary choice will generally be uninformative: a cycle of the form $a \succ b \succ c \succ a$ implies only that there must exist some constraint on choice.

directness: Snyder et al. (1979) on discrimination against the disabled, Norton et al. (2004), Uhlmann and Cohen (2005), and Bohnet et al. (2016) on gender discrimination, Hodson et al. (2002) on racial discrimination, Caruso et al. (2009) on body weight discrimination, Exley (2016) on excuses for selfish behavior, Cubitt et al. (2018) on time discounting. Each paper uses identification approaches tailored to their setting. We provide a formal foundation for this intuition, and empirical tools that can be applied in these and many more settings.




**Temptation.** The decision maker chooses between diet and full-sugar sodas. They explicitly prefer diet soda, but reveal an implicit preference for the less healthy option.

**Embarrassment.** The decision maker chooses between magazines, which may have a swimsuit issue or a special issue covering famous athletes (Chance and Norton, 2009). They reveal an explicit preference for the athletes issue but an implicit preference for the swimsuit issue.

**Prejudice.** The decision maker chooses between movies, which will be watched with an able-bodied or a disabled person (Snyder et al., 1979). They explicitly prefer to sit with the disabled person, but reveal an implicit preference for sitting with the able-bodied person.

**Selfishness.** The decision maker chooses between a lottery and a safe amount, where the beneficiary is themself or charity (Exley, 2016). They explicitly prefer to give to charity, but reveal an implicit preference for self, i.e. they are implicitly selfish.

**Framing.** The decision maker chooses between prospects (A and B) framed in different ways (X and Y). They are indifferent between differently-framed versions of the same prospect, but strictly prefer frame X when the prospects differ. This reveals an implicit preference for frame X, but no explicit preference.

**Discounting** The decision maker chooses between a pen or a box of chocolates, either now, or with a financially-compensated delay (Cubitt et al., 2018). They reveal an explicit preference for sooner rewards, but an implicit preference for later; i.e. they are implicitly patient.

Figure 1: Figure 8 intransitivities applied to various domains.

Our introductory example shows how we can identify implicit discrimination, a topic of great recent interest.[4] There are many other possible applications, in principle we can

[4]Bertrand et al. (2005) and Bertrand and Duflo (2017) discuss the economic importance of implicit

detect implicit preferences over any binary attribute, and there are many contexts in which we might expect them. Figure 1 shows a variety of figure 8 cycles in different domains to illustrate implicit preferences that we believe are plausibly detectible. For simplicity and intuition we show comparisons that differ in either one or two attributes. But our approach is more general, and an implicit preference can be identified even when multiple attributes vary in every comparison.

# 2 Model

The primitives of utility will be **bundles** of $n$ binary attributes: $\boldsymbol{x} \in \mathcal{X} = \{-1, 1\}^n$ (e.g. male/female, PhD/MBA, aisle/window).[5] For example, attribute $i$ might correspond to gender, taking value $x_i = -1$ for females, and $x_i = 1$ for males.

We will begin by introducing a *comparative utility function*, $u(\boldsymbol{x}, \boldsymbol{z}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Utility is comparative in the sense that the utility of bundle $\boldsymbol{x}$ depends on another bundle, $\boldsymbol{z}$, to which $\boldsymbol{x}$ is being compared. We will refer to $\boldsymbol{x}$ as the "target" and $\boldsymbol{z}$ as the "comparator."

For the formal analysis we do not need to take a stand on where the comparator comes from, but for applications we highlight two types of decision that are inherently comparative. The first is binary choice, between $\boldsymbol{x}$ and $\boldsymbol{z}$. The second is pairwise evaluation, under which the decision maker simultaneously reports values for $\boldsymbol{x}$ and for $\boldsymbol{z}$.

We will model utility a two separable components. First, an *explicit value*, that depends only on $\boldsymbol{x}$ and is otherwise unrestricted. Second, an *implicit value* that depends on the comparison between $\boldsymbol{x}$ and $\boldsymbol{z}$. Implicit value is separable in each attribute, and obeys certain restrictions. Specifically, for each attribute $i$ the decision-maker has an implicit preference $\kappa_i \in \{-1, 0, 1\}$. The weight on the implicit preference varies between comparisons, according to the comparison's *opacity* for that attribute. A positive implicit preference on the gender

discrimination, and the difficulty of measuring it. They mention that implicit discrimination will be more pronounced in more "ambiguous" situations: our paper can be seen as formalizing this notion.

The economics literature highlights the distinction between taste-based (Becker, 1957) and statistical (Phelps, 1972; Arrow, 1973) discrimination (of which the latter may be inaccurate: Bohren et al. (2019)). Either type of discrimination can be implicit.

Bohren et al. (2022) decompose discrimination into *direct* and *systemic* components. For example, direct discrimination early in a woman's career contributes to systemic discrimination later on, as she ends up with a weaker resume than an equally-able man. Our notion of implicit discrimination is a form of direct discrimination, and systemic effects can amplify its impact.

[5] All vectors will be column vectors, indicated with a bold font, and $\boldsymbol{x}^T$ will refer to the transpose of $\boldsymbol{x}$. The absolute value of a vector will be written as $|\boldsymbol{x}|^T = \begin{pmatrix} |x_1| & \dots & |x_n| \end{pmatrix}$. Inequalities between vectors will be defined as:

$$\begin{aligned} \boldsymbol{x} \geq \boldsymbol{z} &\iff x_i \geq z_i \text{ for } i = 1, \dots, n. \\ \boldsymbol{x} > \boldsymbol{z} &\iff x_i \geq z_i \text{ for } i = 1, \dots, n, \text{ and } \boldsymbol{x} \neq \boldsymbol{z}. \\ \boldsymbol{x} \gg \boldsymbol{z} &\iff x_i > z_i \text{ for } i = 1, \dots, n. \end{aligned}$$

attribute ($\kappa_i = 1$) means that the utility of bundles with $x_i = 1$ (men) increases as opacity about gender increases, and the utility of bundles with $x_i = -1$ (women) decreases. It is the comparison-dependent conflict between explicit and implicit values that generates intransitivities, such as in our leading gender-discrimination example.

Our analysis will show when we can learn the sign of the decision-maker's implicit preference for an attribute ($\kappa_i \in \{-1, 0, 1\}$). The formal analysis is quite abstract. We will define a generic *opacity dominance* relation between comparisons that can be used to identify variation in implicit preferences, and we will specify what is in principle observable in data on behavior. Our representation theorem then tells us, given an opacity dominance relation, what can be learned about implicit preferences.

To apply the model empirically, in section 2.1 we will propose three restrictions on the opacity dominance relation, and in section 2.2 we will define two classes of dataset, corresponding to choice and evaluation data. Then, in section 3, we derive precise implications for several canonical patterns of behavior that can be easily tested for in applications.

We begin by introducing the comparative utility function.

**Assumption 1.** *The utility of bundle $\boldsymbol{x} \in \mathcal{X}$ with comparator $\boldsymbol{z} \in \mathcal{X}$, is:*

$$u(\boldsymbol{x}, \boldsymbol{z}) = \overbrace{v(\boldsymbol{x})}^{\text{explicit value}} + \overbrace{\sum_{i=1}^{n} x_i \cdot \underbrace{\kappa_i}_{\substack{\text{implicit} \\ \text{preference} \\ \text{for } i}} \cdot \underbrace{\theta_i(\delta(\boldsymbol{x}, \boldsymbol{z}))}_{\substack{\text{opacity of} \\ \text{comparison} \\ \text{for } i}}}^{\text{implicit value}}, \tag{1}$$

*with $v : \mathcal{X} \to \mathbb{R}$, $\kappa_i \in \{-1, 0, 1\}$, $\delta : \mathcal{X} \times \mathcal{X} \to \Delta$, $\theta_i : \Delta \to \mathbb{R}$.*

The *explicit value* of the target bundle, $v(\boldsymbol{x})$, does not depend on the comparison and represents an ordinary, transitive preference.

The *implicit value* of $\boldsymbol{x}$ does depend on the comparator, $\boldsymbol{z}$. The contribution of attribute $i$ to implicit value is increased or decreased according to (1) $\boldsymbol{x}$'s realization of $i$, $x_i \in \{-1, 1\}$; (2) the implicit preference on $i$, $\kappa_i \in \{-1, 0, 1\}$; and (3) the *opacity* of $i$, denoted $\theta_i$. Opacity is a function of the *comparison*, which we label $\delta(\boldsymbol{x}, \boldsymbol{z}) \in \Delta$. The function $\theta(\cdot)$ maps comparisons to scalar weights, so we can say that a comparison $\delta$ has higher or lower opacity with respect to attribute $i$, causing a higher or lower weight on $i$'s implicit preference $\kappa_i$.

Variation in opacity will derive from a partial order $\sqsupseteq_i$ over the set of comparisons $\Delta$. Given a pair of comparisons $\delta, \delta' \in \Delta$ we describe $\delta \sqsupseteq_i \delta'$ as $\delta$ **opacity dominates** $\delta'$ **on attribute** $i$. We assume that $\theta_i(\cdot)$ obeys this partial order:

**Definition 1** (Opacity Dominance)**.** *For any $\delta, \delta' \in \Delta$, $\delta \sqsupseteq_i \delta' \implies \theta_i(\delta) \geq \theta_i(\delta')$.*

Our representation theorem relies only on the fact that $\sqsupseteq_i$ is a partial order, so is quite general. In subsection 2.1 we introduce three restrictions on $\sqsupseteq_i$, motivated by our psychological foundations. The key assumption will be *dilution*, formalizing the idea that opacity about an attribute increases when the comparison mixes that attribute with more other attributes.

With the utility function in hand, and a language to describe how comparisons affect utility, we next define what is observed in data. We will assume that what is in principle observable is a set of inequalities between pairs of comparative utilities.[6] Our definition is chosen to be able to accommodate both binary choice and pairwise evaluation.

A **dataset** $D$ is a collection of $m$ 4-tuples, $(\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j)_{j=1}^m$, with $\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j \in \mathcal{X}$, each of which represents an inequality between two comparative utilities: $u(\boldsymbol{x}^j, \boldsymbol{z}^j) \gtrsim u(\boldsymbol{x}'^j, \boldsymbol{z}'^j)$. By convention we order inequalities such that the first $\bar{m}$ (at least one) are strict, and the remainder are weak. We say a comparative utility function $u$ **rationalizes** a dataset $D$ if all $m$ inequalities are satisfied by $u$, i.e. for every $j = 1, \dots m$,

$$\begin{aligned} u(\boldsymbol{x}^j, \boldsymbol{z}^j) &> u(\boldsymbol{x}'^j, \boldsymbol{z}'^j) \quad , 1 \leq j \leq \bar{m} \qquad \text{(strict inequalities)} \\ u(\boldsymbol{x}^j, \boldsymbol{z}^j) &\geq u(\boldsymbol{x}'^j, \boldsymbol{z}'^j) \quad , \bar{m} < j \leq m. \quad \text{(weak inequalities)} \end{aligned}$$

In general a dataset will contain some inequalities that are informative about implicit preferences, some which are uninformative, and some which our analysis will rely on repeatedly. So we next define the refinement of a dataset to which our analysis will be applied.

**Definition 2** (Cyclical Selection)**.** *Given a dataset $D = \{\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j\}_{j=1}^m$ a **cyclical selection** is a vector of non-negative integer weights $\boldsymbol{s} \in \mathbb{N}^m$ that select inequalities such that each bundle appears equally often on the left- and right-hand sides. I.e., for every $\boldsymbol{x} \in \mathcal{X}$,*

$$\underbrace{\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\}}_{\text{appearances of } \boldsymbol{x} \text{ on LHS}} = \underbrace{\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}}_{\text{appearances of } \boldsymbol{x} \text{ on RHS}},$$

*with $s_j > 0$ for at least one $j \in 0, \dots, \bar{m}$ (i.e., at least one strict inequality is included).*

A cyclical selection defines a weighted subset of a given dataset, consisting of a one or more sequences of inequalities that begin and end with the same target (i.e., $u(\boldsymbol{x}, .) > \dots > u(\boldsymbol{x}, .)$).[7] We will show (Corollary 1) that a dataset can be rationalized by a menu-

[6] Since we only use inequalities between utilities, we could without loss of generality a) wrap (1) in a strictly increasing function (i.e. linearity is not critical), and b) normalize $\theta_i$ to, e.g. take only non-negative values.

[7] E.g., an $\boldsymbol{s}$ that selects the single inequality $u(\boldsymbol{a}, \boldsymbol{b}) > u(\boldsymbol{a}, \boldsymbol{c})$, constitutes a cyclical selection, as does an $\boldsymbol{s}$ that selects the three inequalities $u(\boldsymbol{a}, \boldsymbol{b}) > u(\boldsymbol{b}, \boldsymbol{a})$, $u(\boldsymbol{b}, \boldsymbol{c}) > u(\boldsymbol{c}, \boldsymbol{b})$, $u(\boldsymbol{c}, \boldsymbol{a}) > u(\boldsymbol{a}, \boldsymbol{c})$.

*independent* utility function $u(\boldsymbol{x}, \boldsymbol{z}) = v(\boldsymbol{x})$ if and only if it is not possible to construct a cyclical selection.[8]

Next we define three terms "wins," "losses," and "score." These are used to keep track of which comparisons favor, or disfavor, each attribute. The intuition will be that implicitly preferred bundles tend to "win" when opacity is high, and "lose" when opacity is low. Our theorem essentially checks whether there exists a $u$ that can rationalize the observed pattern of wins and losses.

Given an inequality $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}', \boldsymbol{z}')$ we count it as a "win" for attribute $i$ if the preferred bundle, $\boldsymbol{x}$, has a positive realization of that attribute ($x_i = 1$), and/or the dispreferred bundle, $\boldsymbol{x}'$, has a negative realization ($x'_i = -1$). We define a "loss" as the reverse.[9] For a given attribute $i$ and comparison $\delta$, the "score" $c_{i,\delta}$ equals the wins minus the losses.

**Definition 3** (Score)**.** *Given a dataset $D = \{x^j, z^j, x'^j, z'^j\}_{j=1}^m$ and a cyclical selection $\boldsymbol{s} \in \mathbb{N}^m$ the **score** vector, $\boldsymbol{c} \in \mathbb{Z}^{n|\Delta|}$ represents for each $i \in \{1, \ldots, n\}$ and $\delta \in \Delta$, the net number of times that the attribute wins in $\boldsymbol{s}$:*

$$c_{i,\delta} = \underbrace{\sum_{\{j:\delta(\boldsymbol{x}^j,\boldsymbol{z}^j)=\delta\}} s_j x_i^j}_{\text{bundles on LHS of inequality}} - \underbrace{\sum_{\{j:\delta(\boldsymbol{x}'^j,\boldsymbol{z}'^j)=\delta\}} s_j x_i'^j}_{\text{bundles on RHS of inequality}} .$$

In a cyclical selection each bundle appears equally often on the LHS and RHS, so for each attribute $i$ the score must sum to zero: $\sum_{\delta \in \Delta} c_{i,\delta} = 0$.

Keeping track of scores allows us to check if a candidate vector of implicit preferences is consistent with the data. Roughly speaking, a positive implicit preference for attribute $i$ cannot explain the data if more opaque comparisons have negative scores, and less opaque comparisons have positive scores. Our theorem essentially checks whether we can match comparisons with positive and negative scores in such a way that the partial order $\sqsupseteq_i$ allows us to rank their opacity.

**Definition 4** (wins opacity dominate losses)**.** *Given a vector of scores for attribute $i$, $\boldsymbol{c}_i \in \mathbb{Z}^{|\Delta|}$ we say **wins opacity dominate losses** for attribute $i$ if there exists a matrix of non-*

[8] Intuitively, a cyclical selection implies there exists an intransitive cycle over target bundles.

[9] Thus, if gender is encoded as Male = 1, it is a win if a man appears on the LHS and/or a woman on the RHS, and a loss if a woman appears on the LHS and/or a man on the RHS.

*negative integers* $M_i \in \mathbb{N}^{|\Delta|\times|\Delta|}$ *with:*

$$\forall \bar{\delta}, \bar{\delta}' \in \Delta, \quad (M_{i,\bar{\delta},\bar{\delta}'} > 0) \implies (\bar{\delta} \sqsupseteq_i \bar{\delta}') \qquad \textit{(matches obey dominance)}$$

$$\forall \delta \in \Delta, \qquad c_{i,\delta} = \underbrace{\sum_{\bar{\delta}' \in \Delta} M_{i,\delta,\bar{\delta}'}}_{\substack{\textit{outgoing matches} \\ (\delta \textit{ dominating})}} - \underbrace{\sum_{\bar{\delta} \in \Delta} M_{i,\bar{\delta},\delta}}_{\substack{\textit{incoming matches} \\ (\delta \textit{ dominated})}} \qquad \textit{(all scores are accounted for)}$$

The first condition says that comparison $\bar{\delta}$ is only matched to $\bar{\delta}'$ if $\bar{\delta}$ opacity dominates $\bar{\delta}'$. The second condition checks that all scores are matched: each positive score, $c_{i,\delta} > 0$ (where wins exceed losses) will have a net outflow equal to $c_{i,\delta}$, and each negative score, $c_{i,\delta} < 0$ will have a net inflow equal to $|c_{i,\delta}|$. It shows that matrix $M_i$ can be thought of as a *matching* between scores in $\boldsymbol{c}_i$.[10]

We likewise say that **losses opacity dominate wins** if there is an $M$ that satisfies the same conditions but the last line in the definition sums to $-c_{i,\delta}$ instead of $c_{i,\delta}$.

We are now ready to state our representation theorem. It tells us that, given an opacity dominance relation $\sqsupseteq_i$ and a candidate vector of implicit preferences $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$, there exists a utility function satisfying (1) that can rationalize the data if and only if there exists no matching between wins and losses that contradicts the proposed signs of *every* $\kappa_i$.

**Theorem 1** (Representation)**.** *Given an opacity dominance relation* $\sqsupseteq_i, i = 1, ..., n$*, and a candidate vector of implicit preferences* $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$*, a dataset* $D$ *can be rationalized by a comparative utility function* $u$ *if and only if* $D$ *does not contain a cyclical selection* $\boldsymbol{s}$ *such that, (i) for every attribute with a positive implicit preference (*$\kappa_i = 1$*), losses opacity dominate wins, and (ii) for every attribute with a negative implicit preference (*$\kappa_i = -1$*), wins opacity dominate losses.*

Thus, the set of $\boldsymbol{\kappa}$ vectors that are consistent with the data are those not ruled out by this condition. Section 3 demonstrates with many examples how we can identify individual elements, e.g. $\kappa_i = 1$, by ruling out all candidate $\boldsymbol{\kappa}$s with $\kappa_i \neq 1$.

To apply the theorem, one must search over *all* cyclical selections that can be constructed from the dataset to verify that the matching condition holds. For simple datasets this is straightforward and can often be done by simple visual inspection (see Section 3). For larger datasets the matching procedure can be laborious. However, the proof of Theorem 1 uses

[10] **Worked example.** Consider a cyclical selection with one inequality: $u\left(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}-1\\-1\end{smallmatrix}\right]\right) > u\left(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right]\right)$. Call the LHS comparison $\bar{\delta}$ and the RHS comparison $\bar{\delta}'$. Now take attribute 1. Let us assume $\bar{\delta} \sqsupseteq_1 \bar{\delta}'$. We have $c_{1,\bar{\delta}} = 1$ (a single win) and $c_{1,\bar{\delta}'} = -1$ (a single loss). There exists a matching matrix for attribute 1 that matches the positive score on $\bar{\delta}$ to the negative score on $\bar{\delta}'$: $M_1 = \left[\begin{smallmatrix}0 & 1\\0 & 0\end{smallmatrix}\right]$. Hence, wins opacity dominate losses for attribute 1.

an equivalency result between the matching procedure and a matrix representation of the dataset, that can then be solved numerically.

An immediate corollary characterizes when standard preferences can rationalize the data.

**Corollary 1** (Rationalization by a menu-independent utility function)**.** *A dataset $D$ can be rationalized by a* menu-independent *utility function (i.e. $\boldsymbol{\kappa} = \mathbf{0}$) if and only if it does not contain a cyclical selection.*[11]

Our theory is falsified if and only if the data cannot be rationalized by a comparative utility function for any $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$. The next corollary provides a useful sufficient condition.[12]

**Corollary 2** (Falsification)**.** *A dataset $D$ cannot be rationalized by a comparative utility function $u$ if it contains a cyclical selection $\boldsymbol{s}$ such that for* every *attribute, losses opacity dominate wins* and *wins opacity dominate losses.*

Section 3 gives examples where the vector of scores equals zero ($c_i = \mathbf{0}, \forall i$). Then, for each decision in the dataset that favors a given attribute (a win), there exists another decision with the same comparison that disfavors it (a loss). No $\boldsymbol{\kappa}$ that can generate such a pattern, because for all attributes losses opacity dominate wins *and* wins opacity dominate losses.[13]

## 2.1 Assumptions on Opacity

We now add assumptions on the opacity dominance relation ($\sqsupseteq_i$) to tailor our representation theorem to applications. Our assumptions are motivated by the foundational models that we present in section 4. However we note that Theorem 1 is more general, and can easily be applied under alternative assumptions about $\sqsupseteq_i$.[14]

First, we assume that the opacity of a comparison $\delta(\boldsymbol{x}, \boldsymbol{z})$ depends only on which attributes are *shared* ($|x_i - z_i| = 0$) and which are *non-shared* ($|x_i - z_i| = 2$).

**Assumption 2** (Equivalence)**.** *For any $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}' \in \mathcal{X}$:*

$$|\boldsymbol{x} - \boldsymbol{z}| = |\boldsymbol{x}' - \boldsymbol{z}'| \Rightarrow \delta(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i \delta(\boldsymbol{x}', \boldsymbol{z}'), \forall i$$

[11] Proof: if $\boldsymbol{\kappa} = \mathbf{0}$ the matching condition is trivially satisfied for every attribute in any cyclical selection.

[12] The condition in Corollary 2 is sufficient because it implies a single cyclical selection that rules out every possible $\boldsymbol{\kappa}$ vector. But it is not necessary: one can construct examples where no single cyclical selection falsifies the model, but multiple cyclical selections exist that collectively do. See Web Appendix B.1.1.

[13] Corollary 2 is also relevant when multiple comparisons can have identical opacity. The model is falsified if all positive scores ($c_{i,\delta} > 0$) can be matched to negative scores with the same opacity ($c_{i,\delta'} < 0, (\delta \sqsupseteq_i \delta') \wedge (\delta' \sqsupseteq_i \delta)$). Again this is because, for all attributes, losses opacity dominate wins *and* wins opacity dominate losses.

[14] For example, our signaling-choice foundation in section 4 assumes a naïve observer. A sophisticated observer would imply modifications to assumptions 2 and 3.

Equivalence means that opacity is identical between any two comparisons with the same sets of shared and non-shared attributes. For example, $\delta\left(\binom{\text{Male}}{\text{MBA}}, \binom{\text{Female}}{\text{PhD}}\right) = \delta\left(\binom{\text{Female}}{\text{MBA}}, \binom{\text{Male}}{\text{PhD}}\right)$. Equivalence reduces the set of comparisons to $\Delta = \{0, 2\}^n$, and the comparison function to $\delta(\boldsymbol{x}, \boldsymbol{z}) = |\boldsymbol{x} - \boldsymbol{z}|$. From this point we will treat the comparison $\boldsymbol{\delta}$ as a vector (and so print it in bold) with elements $\delta_i = |x_i - z_i|$. We will use the term **status** to refer to whether an attribute is shared ($\delta_i = 0$) or non-shared ($\delta_i = 2$) in a comparison.

Our primary assumption on opacity is that an attribute $i$ becomes more opaque as additional other attributes share status with $i$. We call the assumption "dilution."

**Assumption 3** (Dilution)**.** *For any* $i \in \{1, \ldots, n\}, \boldsymbol{\delta}, \boldsymbol{\delta}' \in \Delta$*,*

$$\underbrace{(\delta_i = \delta'_i)}_{\substack{\text{attribute } i \text{ has} \\ \text{same status in } \delta \text{ \& } \delta'}} \quad \wedge \quad \underbrace{\{j : \delta_j = \delta_i\} \supseteq \{j : \delta'_j = \delta'_i\}}_{\substack{\text{more attributes share status with } i \\ \text{in } \delta \text{ than in } \delta'}} \quad \Longrightarrow \quad \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\text{attribute } i \text{ is more} \\ \text{opaque under } \delta \text{ than } \delta'}}$$

Suppose $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on gender. Dilution implies that an implicit preference for one gender over the other will have the weakest influence when $\boldsymbol{x}$ and $\boldsymbol{z}$ differ *only* on gender, becoming progressively stronger as $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on more other attributes in addition to gender. This is the key property that allows us to identify implicit preferences from choice and evaluation data.

Assumptions 1–3 define our workhorse model, "Separable Implicit Preferences."

**Definition 5** (Separable Implicit Preferences)**.** *A* Separable Implicit Preferences *decision maker satisfies Assumptions 1, 2, and 3.*

Assumptions 1–3 are sufficient for all of our analysis of choice data and for a number of identification results in evaluation data. But for some evaluation results we need to be able to rank opacities for attributes that change status, from shared to non-shared or vice versa, and Dilution has nothing to say in these cases.[15] Our final assumption allows us to do this. It assumes that one attribute, $k \in \{1, \ldots, n\}$, is "special" in the sense that opacity is greater for attributes that have the same status as $k$. If $k$ is shared, then opacity is greater for shared attributes, if $k$ is non-shared, opacity is greater for non-shared attributes.

**Assumption 4** (Dominance of attribute $k$)**.** *For any* $i \in \{1, \ldots, n\} \setminus k, \boldsymbol{\delta}, \boldsymbol{\delta}' \in \Delta$*,*

$$\underbrace{\delta_i = \delta_k \wedge \delta'_i \neq \delta'_k}_{\substack{\text{same status as } k \text{ in } \delta \\ \text{diff status from } k \text{ in } \delta'}} \quad \Longrightarrow \quad \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\text{more opaque} \\ \text{under } \delta \text{ than } \delta'}}$$

[15] Dilution is sufficient for analysis of choice data because implicit preferences on shared attributes have no influence on choice.

We think it is most natural to think of $k$ as a shared attribute, capturing what is "held constant" across comparisons, implying that opacity is higher for shared attributes. In section 4 we provide conditions under which each foundation implies Assumption 4. For the foundations based on signal extraction (signaling and implicit associations) the intuition is that there is high uncertainty about the value of attribute $k$, such that little can be learned about attributes that share status with $k$.[16]

## 2.2 Choice and Evaluation as Datasets

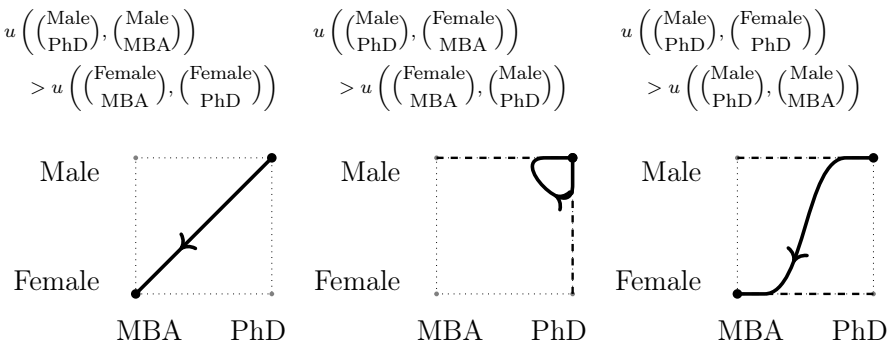
We have defined a dataset as a set of inequalities between comparative utilities. We now show how data from choice and evaluation can be represented in a dataset.

For binary choice we treat each bundle as the other bundle's comparator: a strict revealed preference for $\boldsymbol{x}$ over $\boldsymbol{z}$ gives a strict inequality, $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{z}, \boldsymbol{x})$, while indifference gives a pair of weak inequalities, $u(\boldsymbol{x}, \boldsymbol{z}) \geq u(\boldsymbol{z}, \boldsymbol{x})$ and $u(\boldsymbol{z}, \boldsymbol{x}) \geq u(\boldsymbol{x}, \boldsymbol{z})$. Thus a choice cycle $\boldsymbol{x}^a \succ \boldsymbol{x}^b \succ \boldsymbol{x}^c \succeq \boldsymbol{x}^a$ will correspond to a dataset with $m = 4, \bar{m} = 2$: $u(\boldsymbol{x}^a, \boldsymbol{x}^b) > u(\boldsymbol{x}^b, \boldsymbol{x}^a)$; $u(\boldsymbol{x}^b, \boldsymbol{x}^c) > u(\boldsymbol{x}^c, \boldsymbol{x}^b)$; $u(\boldsymbol{x}^c, \boldsymbol{x}^a) \geq u(\boldsymbol{x}^a, \boldsymbol{x}^c)$; and $u(\boldsymbol{x}^a, \boldsymbol{x}^c) \geq u(\boldsymbol{x}^c, \boldsymbol{x}^a)$.

Theorem 1 is also applicable to data on continuous *evaluations* of bundles when each evaluated bundle has a comparator, for example when two bundles are evaluated simultaneously. We assume that evaluations are a strictly increasing function of utility: $y(\boldsymbol{x}, \boldsymbol{z}) = f(u(\boldsymbol{x}, \boldsymbol{z})), f' > 0$. To apply our theorem we construct a set of inequalities sufficient to represent the ordinal relationship of the evaluations: we first rank each evaluation (breaking ties arbitrarily), and enter an inequality into the dataset for each pair of consecutive evaluations. When two evaluations are equal we use two opposing weak inequalities. For example suppose we observe the following joint evaluations of willingness to pay: $y(\boldsymbol{x}^a, \boldsymbol{x}^b) = \$310$, $y(\boldsymbol{x}^b, \boldsymbol{x}^a) = \$200$, $y(\boldsymbol{x}^a, \boldsymbol{x}^c) = \$200$, $y(\boldsymbol{x}^c, \boldsymbol{x}^a) = \$150$. Then we would construct a dataset with $m = 4, \bar{m} = 2$: $u(\boldsymbol{x}^a, \boldsymbol{x}^b) > u(\boldsymbol{x}^b, \boldsymbol{x}^a)$, $u(\boldsymbol{x}^b, \boldsymbol{x}^a) \geq u(\boldsymbol{x}^a, \boldsymbol{x}^c)$, $u(\boldsymbol{x}^a, \boldsymbol{x}^c) \geq u(\boldsymbol{x}^b, \boldsymbol{x}^a)$, $u(\boldsymbol{x}^a, \boldsymbol{x}^c) > u(\boldsymbol{x}^c, \boldsymbol{x}^a)$.

**Graphical representation of inequalities** We will frequently use diagrams to visualize sets of inequalities. An arrow from $\boldsymbol{x}$ to $\boldsymbol{x}'$ shows the sign of the inequality, with the entry and exit angles pointing towards each bundle's comparator, $\boldsymbol{z}$ and $\boldsymbol{z}'$. We use dashed lines to show the comparisons: they run from $\boldsymbol{x}$ to $\boldsymbol{z}$ and from $\boldsymbol{x}'$ to $\boldsymbol{z}'$. Figure 2 gives three examples.

[16] Consider a judge assigning sentences to two defendants simultaneously. Intuitively, the long sentence assigned to two Black defendants could be explained by their race, but also by prevailing sentencing rules, the leniency of the judge, the time of day, and so on. These shared attributes cannot explain why sentences differ between otherwise similar Black and white defendants. Hence, it is intuitive that opacity about race is greater when race is shared than when it is non-shared. Our second empirical application has this structure.




Solid lines with arrows show the sign of the inequality between two target bundles. Dashed lines point to each target's comparator (not visible in the case of choice, since each target is the other's comparator).

Figure 2: Examples of graphical representations of inequalities.

# 3 Canonical Examples

We now define a set of important classes of dataset, from both choice and evaluation, picked to encompass those that are most useful for applications. We derive the implications of each as a corollary of Theorem 1. The proofs are mechanical so we consign them to web appendix B.1. All definitions are stated in terms of strict inequalities, however all results will go through provided at least one inequality in each cyclical selection is strict.

For choice, we introduce the *right triangle*, the shortest choice cycle (three choices) that provides unambiguous restrictions on $\boldsymbol{\kappa}$. It provides a disjunction over the implicit preferences on all attributes that vary in the cycle. Second, we discuss the *figure 8*, which is the shortest choice cycle (four choices) that can unambiguously identify a single implicit preference. Third, we show how pairs of *parallel right triangles* (five or six choices) can provide further restrictions relative to the single triangle, including the possibility of unambiguous identification.

Turning to evaluation we introduce the *convex scissor*, the smallest cyclical selection (two joint evaluations) that can provide unambiguous restrictions on *some elements of* $\boldsymbol{\kappa}$. It implies the existence of at least one implicit preference, but only restricts the sign of a subset. Next, we show how pairs of *parallel convex scissors* (three or four joint evaluations) can refine identification relative to the single scissor, including the possibility of unambiguous identification. We also show how Assumption 4 (Dominance of attribute $k$) further refines identification.

Finally, we present two examples that reveal the presence of an implicit preference but no further restrictions, and three that falsify the theory.

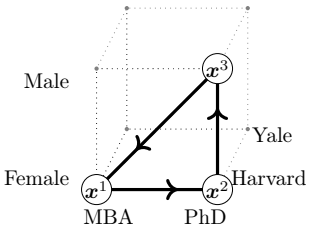
For each corollary we provide worked examples with variation in two or three attributes. We state each example's implications for $\boldsymbol{\kappa}$ and in natural language. E.g., $\kappa_1 = 1$ means we learn $\kappa_1$ is positive, $\kappa_1 \neq 0$ means we learn there is an implicit preference for attribute 1 but not its sign, and so on. In natural language, we always state preferences relative to the $+1$ pole of the attribute. +Male means an implicit preference favoring men (relative to women), −Male means an implicit preference favoring women (i.e., against men), and ±Male means we learn there is an implicit gender preference not its sign.

**Choice examples.** We begin with choice data. We assume throughout that the decision maker satisfies Separable Implicit Preferences (Assumptions 1–3).

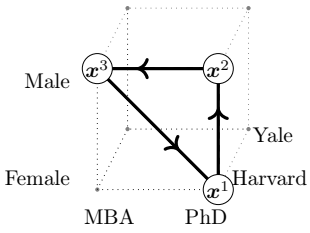
**Definition 6** (Right triangle)**.** *A right triangle is a choice cycle over three distinct bundles, ordered* $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$*, with corresponding differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \boldsymbol{\delta}^3$*, in which the differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2$ *are orthogonal, i.e.* $\boldsymbol{\delta}^1$ *and* $\boldsymbol{\delta}^2$ *differ on distinct sets of attributes.*

**Corollary 3.** *A right triangle implies at least one nonzero implicit preference favoring* $\boldsymbol{x}^3$*'s realization of an attribute on which it differs from* $\boldsymbol{x}^1$*:*

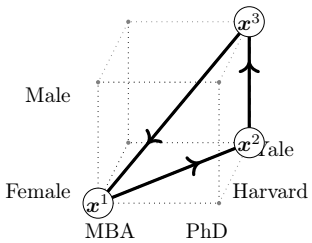
$$\bigvee_{i: x_i^3 \neq x_i^1} (x_i^3 \kappa_i = 1).$$




$(\kappa_1 = 1) \vee (\kappa_2 = 1)$
+PhD ∨ +Male.




$(\kappa_1 = -1) \vee (\kappa_2 = 1)$
−PhD ∨ +Male.




$(\kappa_1 = 1) \vee (\kappa_2 = 1) \vee (\kappa_3 = 1)$
+PhD ∨ +Male ∨ +Yale.

A single right triangle cannot unambiguously identify an implicit preference without further restrictions on $\boldsymbol{\kappa}$, because $\boldsymbol{x}^3$ and $\boldsymbol{x}^1$ must differ on at least two attributes.

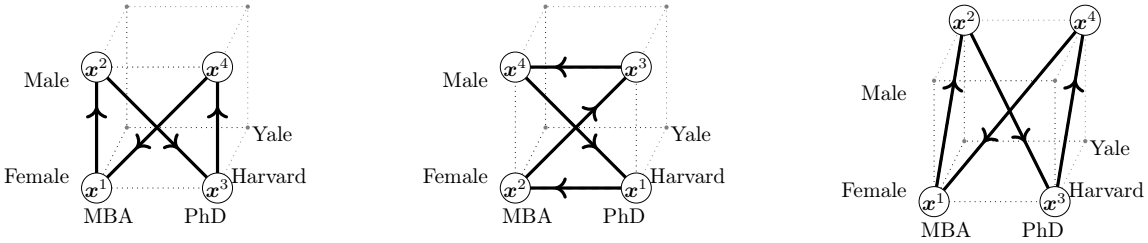
For intuition note that the first right triangle can be thought of as containing two reversals of preference: (1) a female PhD is chosen over a male PhD, but the gender preference is reversed when the man has an MBA (which dilutes the gender attribute); (2) a female MBA is chosen over a female PhD, but the qualification preference is reversed when the PhD holder is male (which dilutes the qualification attribute). The cycle thus implies the presence of at least one implicit preference but we cannot distinguish between one favoring men, one favoring PhDs, or both.

**Definition 7** (Figure 8)**.** *A figure 8 is a choice cycle over four distinct bundles, ordered*

$\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^4 \succ \boldsymbol{x}^1$, *with corresponding differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \boldsymbol{\delta}^3, \boldsymbol{\delta}^4$. *It must satisfy two conditions: (1) there are only two sets of differences* $\boldsymbol{\delta}^1 = \boldsymbol{\delta}^3$ *and* $\boldsymbol{\delta}^2 = \boldsymbol{\delta}^4$*; and (2) the even-numbered comparisons differ on a superset of attributes:* $\boldsymbol{\delta}^2 > \boldsymbol{\delta}^1$.

**Corollary 4.** *A figure 8 implies at least one nonzero implicit preference, favoring* $\boldsymbol{x}^4$*'s realization of an attribute on which it differs from* $\boldsymbol{x}^3$*:*

$$\bigvee_{i: x_i^3 \neq x_i^4} (x_i^4 \kappa_i = 1).$$




$(\kappa_2 = 1)$ +Male.

$(\kappa_1 = -1)$ −PhD.

$(\kappa_2 = 1) \vee (\kappa_3 = 1)$ +Male ∨ +Yale.

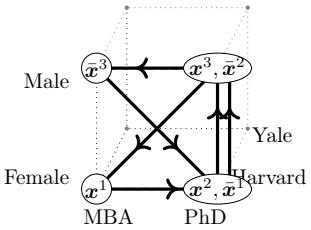
Since $\boldsymbol{x}^3$ and $\boldsymbol{x}^4$ may differ on only one attribute, a figure 8 cycle can unambiguously identify an implicit preference.

The figure 8 can be thought of as containing two preference reversals. In the leading example, a female candidate is chosen over a man with the same qualification, but is rejected whenever the qualifications differ (which dilutes the gender attribute). One reversal favors male MBAs, the other favors male PhDs. Under our assumptions, an implicit preference on the qualification dimension cannot generate both, so there must be an implicit preference favoring men.

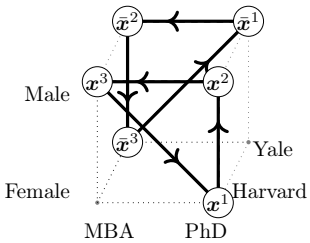
**Definition 8** (Parallel right triangles)**.** *A pair of parallel right triangles is a cyclical selection consisting of two right triangles* $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$ *and* $\bar{\boldsymbol{x}}^1 \succ \bar{\boldsymbol{x}}^2 \succ \bar{\boldsymbol{x}}^3 \succ \bar{\boldsymbol{x}}^1$*, satisfing two conditions: (1) identical* signed *differences on* $\{\boldsymbol{x}^2, \boldsymbol{x}^3\}$ *and* $\{\bar{\boldsymbol{x}}^1, \bar{\boldsymbol{x}}^2\}$ *(that is,* $\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$*, implying* $\boldsymbol{\delta}^2 = \bar{\boldsymbol{\delta}}^1$*); and (2) opposing* signed *differences on* $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$ *and* $\{\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3\}$ *(that is,* $\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$*, implying* $\boldsymbol{\delta}^1 = \bar{\boldsymbol{\delta}}^2$*).*

**Corollary 5.** *A pair of parallel right triangles implies at least one implicit preference, favoring* $\boldsymbol{x}^3$*'s realization of an attribute on which it differs from* $\boldsymbol{x}^2$*:*

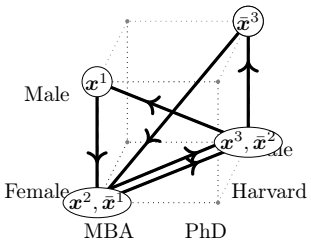
$$\bigvee_{i: x_i^3 \neq x_i^2} (x_i^3 \kappa_i = 1).$$




$(\kappa_2 = 1)$
+Male.




$(\kappa_1 = -1)$
−PhD.




$(\kappa_1 = 1) \vee (\kappa_3 = 1)$
+PhD ∨ +Yale.

Because $\boldsymbol{x}^3$ and $\boldsymbol{x}^2$ can differ on only one attribute, a pair of parallel right triangles can unambiguously identify a single implicit preference. They achieve this by ruling out part of each triangle's individual disjunctions. In particular, they eliminate all attributes that vary in $\boldsymbol{\delta}^1$ and $\bar{\boldsymbol{\delta}}^2$ (the attributes over which the triangles have opposing preferences), leaving only the attributes that vary in $\boldsymbol{\delta}^2$ and $\bar{\boldsymbol{\delta}}^1$ (where the triangles agree). For intuition, observe that the first example is constructed from two of the individual right triangles given above. They agree on the gender attribute, and disagree on the qualification attribute, allowing us to isolate the implicit preference on gender.

**Examples without direct comparisons.** All of our examples that unambiguously identify a single implicit preference include "direct" comparisons in which a single attribute varies. It is natural to ask whether this is essential. The answer is no: we can unambiguously identify a single implicit preference without ever observing a direct comparison, if we observe more than one intransitive cycle.[17]

**Evaluation examples.** We now turn to evaluation data. In each case we first state the result under the assumption of Separable Implicit Preferences (Assumptions 1–3). Then we show how adding Assumption 4 can refine identification.

**Definition 9** (Convex scissor)**.** *A convex scissor is a pair of evaluations of a single bundle $\boldsymbol{x}$ with two different comparators: $y^1 = y(\boldsymbol{x}, \boldsymbol{z}^1), y^2 = y(\boldsymbol{x}, \boldsymbol{z}^2)$ (differences $\boldsymbol{\delta}^1$ and $\boldsymbol{\delta}^2$). Two conditions must be satisfied: (1) the evaluations are not equal ($y^1 \neq y^2$), and (2) the second comparison differs on a superset of attributes ($\boldsymbol{\delta}^2 > \boldsymbol{\delta}^1$).*

**Corollary 6.** *A convex scissor implies at least one nonzero implicit preference:*

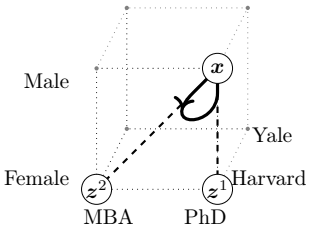
$y^2 > y^1$ *(i) favoring $\boldsymbol{x}$'s realization of an attribute that it does not share with $\boldsymbol{z}^1$,*
*(ii) disfavoring $\boldsymbol{x}$'s realization of an attribute that it shares with $\boldsymbol{z}^2$, or*
*(iii) with unrestricted sign on any other attribute.*

[17] E.g., our third figure-8 example above has no direct choices. It reveals $(\kappa_2 = 1) \vee (\kappa_3 = 1)$. If we observed a second figure-8 revealing $(\kappa_2 = 1) \vee (\kappa_3 = -1)$, we could conclude that $\kappa_2 = 1$ for sure.

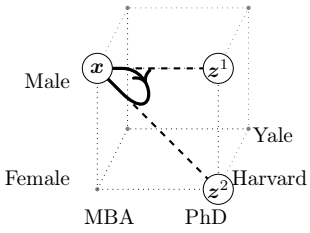
$y^2 < y^1$ *(implies the reverse of* $y^2 > y^1$*)*

*Defining* $\Upsilon = sgn(y^2 - y^1) \in \{-1, 1\}$*, we can write:*

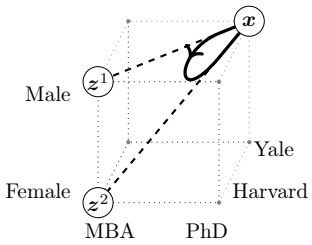
$$\bigvee_{i:x_i \neq z_i^1} (x_i \kappa_i \Upsilon = 1) \vee \bigvee_{i:x_i = z_i^2} (x_i \kappa_i \Upsilon = -1) \vee \bigvee_{i:z_i^1 \neq z_i^2} (\kappa_i \neq 0).$$




$y^2 > y^1$
$(\kappa_1 \neq 0) \vee (\kappa_2 = 1) \vee (\kappa_3 = 1)$
$\pm$PhD $\vee+$Male $\vee+$Yale.




$y^2 < y^1$
$(\kappa_1 = 1) \vee (\kappa_2 \neq 0) \vee (\kappa_3 = -1)$
$+$PhD $\vee\pm$Male $\vee-$Yale.




$y^2 < y^1$
$(\kappa_1 = -1) \vee (\kappa_2 \neq 0) \vee (\kappa_3 = -1)$
$-$PhD $\vee\pm$Male $\vee-$Yale.

Intuitively, the shift of comparison from $\boldsymbol{z}^1$ to $\boldsymbol{z}^2$ changes opacity for every attribute. Those that are shared in both comparisons become less dilute (as the set of shared attributes shrinks), so evaluation becomes less sensitive to implicit preferences on these attributes. Those that are non-shared in both comparisons become *more* dilute (as the set of non-shared attributes grows) so their implicit preferences have more influence. But those that were shared in $\{\boldsymbol{x}, \boldsymbol{z}^1\}$ but are not shared in $\{\boldsymbol{x}, \boldsymbol{z}^2\}$ are not restricted by Assumption 3 (Dilution) so we cannot sign their implicit preferences. Assumption 4 (Dominance of attribute $k$) resolves the ambiguity.

**Corollary 7** (Convex scissor with Dominance of attribute $k$)**.** *When Assumption 4 holds, a convex scissor implies:*

$$\bigvee_{i:x_i \neq z_i^1} (x_i \kappa_i \Upsilon = 1) \vee \bigvee_{i:x_i = z_i^2} (x_i \kappa_i \Upsilon = -1) \vee \bigvee_{i:z_i^1 \neq z_i^2} (x_i \kappa_i \Upsilon = -\Theta),$$

*where* $\Theta$ *equals* 1 *when* $k$ *is shared (opacity is higher for shared attributes), and* $-1$ *when* $k$ *is non-shared (opacity is higher for non-shared attributes).*

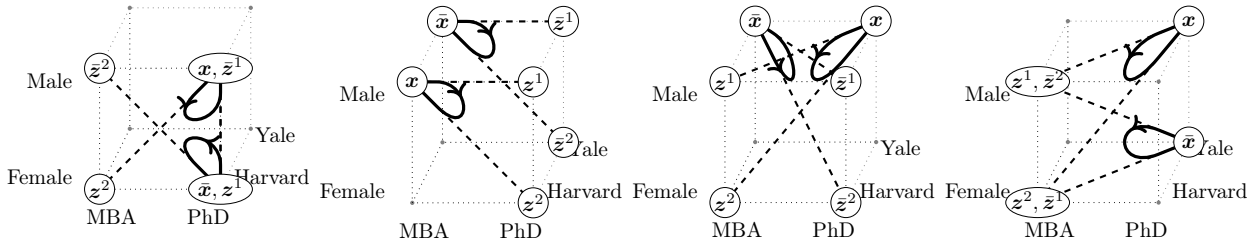
**Definition 10** (Parallel convex scissors)**.** *A pair of parallel convex scissors is a dataset consisting of two convex scissors,* $y(\boldsymbol{x}, \boldsymbol{z}^1) \neq y(\boldsymbol{x}, \boldsymbol{z}^2)$ *and* $y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^1) \neq y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^2)$*,* $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$*. We label the differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \bar{\boldsymbol{\delta}}^1, \bar{\boldsymbol{\delta}}^2$*, the evaluation values* $y^1, y^2, \bar{y}^1, \bar{y}^2$*, and the signs of changes in evaluation values* $\Upsilon = sgn(y^2 - y^1)$ *and* $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$*.*

*Two conditions must be satisfied: (1) identical or opposing* signed *differences on* $\{\boldsymbol{x}, \boldsymbol{z}^1\}$ *and* $\{\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^1\}$ *(i.e., either* $\boldsymbol{x} - \boldsymbol{z}^1 = \bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^1$ *or* $\boldsymbol{x} - \boldsymbol{z}^1 = -(\bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^1)$*); and (2) identical* absolute

*differences on* $\{\boldsymbol{x}, \boldsymbol{z}^2\}$ *and* $\{\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^2\}$ *(i.e.,* $\boldsymbol{\delta}^2 = \bar{\boldsymbol{\delta}}^2$*).*[18]

**Corollary 8.** *A pair of parallel convex scissors imply at least one nonzero implicit preference. There are many cases, which depend on the relationships between* $\boldsymbol{x}, \bar{\boldsymbol{x}}, \Upsilon$*, and* $\bar{\Upsilon}$*. The cases are summarized in the following disjunction:*

$$\bigvee_{i:x_i \neq z_i^1} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) = 2\right) \vee \bigvee_{i:x_i = z_i^2} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) = -2\right) \vee \bigvee_{i:z_i^1 \neq z_i^2} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) \neq 0\right).$$




| $y^2 > y^1,\ \bar{y}^2 < \bar{y}^1$ | $y^2 > y^1,\ \bar{y}^2 > \bar{y}^1$ | $y^2 < y^1,\ \bar{y}^2 > \bar{y}^1$ | $y^2 < y^1,\ \bar{y}^2 > \bar{y}^1$ |
|---|---|---|---|
| $(\kappa_2 > 0)$ | $(\kappa_1 > 0) \vee (\kappa_2 \neq 0)$ | $(\kappa_1 < 0)$ | $(\kappa_2 \neq 0)$ |
| +Male. | +PhD $\vee\pm$Male. | −PhD | $\pm$Male. |

Just like the pair of parallel right triangles, parallel convex scissors can provide a refinement on the implications of their constituent scissors. This occurs when there are attributes with $x_i\Upsilon = -\bar{x}_i\bar{\Upsilon}$, in which the terms associated with those attributes equal zero and drop out of the disjunction. Intuitively, the observed behavior cannot be explained by those attributes if evaluation moves in contradictory directions in the two scissors.

Often we can eliminate all but one attribute, in which case we may unambiguously identify that attribute's implicit preference. The first example shows that Assumption 4 is not required to accomplish this. In contrast, the fourth example illustrates a case where we identify which attribute must have an implicit preference, but Assumption 4 is needed to learn its sign.

**Corollary 9** (Parallel convex scissors with Dominance of attribute $k$)**.** *When Assumption 4 holds, a pair of parallel convex scissors implies:*

$$\bigvee_{i:x_i \neq z_i^1} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) = 2\right) \vee \bigvee_{i:x_i = z_i^2} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) = -2\right) \vee \bigvee_{i:z_i^1 \neq z_i^2} \left(\kappa_i(x_i\Upsilon + \bar{x}_i\bar{\Upsilon}) = -2\Theta\right),$$

*where* $\Theta$ *equals* 1 *when* $k$ *is shared (opacity is higher for shared attributes), and* $-1$ *when* $k$

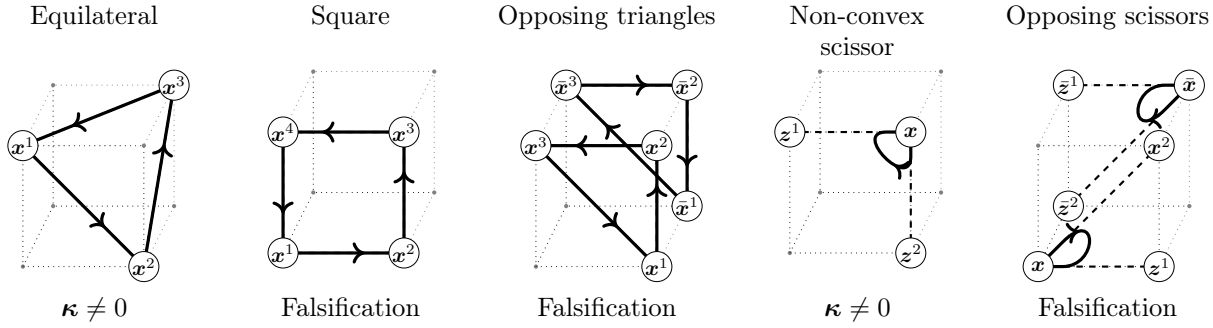
[18]We also assume that the only information derived from the evaluations is the ranking of $y^1, y^2$ and the ranking of $\bar{y}^1, \bar{y}^2$, i.e. we either do not have, or do not exploit, the ranking of evaluations between the scissors. In principle such information could be used in combination with functional form assumptions to extract additional information, but we do not model this for sake of brevity.

*is non-shared (opacity is higher for non-shared attributes).*

Suppose we assume that the dominating attribute $k$ is shared, so opacity is higher for shared attributes ($\Theta = 1$). In our second parallel convex scissors example, this implies $(\kappa_1 = 1) \vee (\kappa_2 = 1)$, an implicit preference favoring PhDs, or men. In our third example it implies $\kappa_2 = 1$, an implicit preference favoring men.

**Examples without direct comparisons.** As for choice, it is possible to unambiguously identify an implicit preference without ever observing a "direct" comparison (in which only one attribute varies). See e.g. our third and fourth examples of parallel convex scissors.

**Other examples** Figure 3 gives five examples that either fail to identify any implicit preference, or cannot be rationalized by any $\boldsymbol{\kappa}$, falsifying the model. The falsifications have in common that there exists a cyclical selectin with equal numbers of wins and losses, for every $i, \boldsymbol{\delta}$, meaning that their score vector equals zero. Thus, wins trivially opacity dominate losses, and vice versa, so no implicit preference can rationalize the dataset (Corollary 2).




An **equilateral triangle** is a choice cycle over three bundles, where Dilution does not rank its comparisons. Its wins neither opacity dominate its losses, nor vice versa. We cannot rule out any $\boldsymbol{\kappa}$.

A **square** cycle is a choice cycle over four bundles, in which each choice is twinned with another with the same $\boldsymbol{\delta}$ but opposing preference. Thus its score is a vector of zeros, so cannot be rationalized.

A pair of **opposing triangles** is two right triangles with opposing signed differences on each choice. Just like the square cycle, its score vector equals zero, so cannot be rationalized.

A **non-convex scissor** is one in which Dilution does not rank its comparisons, so under Separable Implicit Preferences we cannot rule out any $\boldsymbol{\kappa}$.

A pair of **opposing scissors** is one in which $\boldsymbol{x} = -\bar{\boldsymbol{x}}$ but both scissors move in the same direction. Its score vector equals zero, so cannot be rationalized.

Figure 3: Examples that do not identify an implicit preference, or falsify the model

# 4 Foundations

We now provide models of three types of decision maker: one constrained by rules (*ceteris paribus*), one concerned for their reputation (*signaling*), and one rationally influenced by

unconscious *implicit associations*.

To keep the discussion concise, for each foundation we provide the setup of the model and then state the main result that the foundation satisfies Separable Implicit Preferences (Assumptions 1–3). At the end, we provide conditions under which each foundation also satisfies "Dominance of attribute $k$" (Assumption 4). In the text we focus on intuition. Derivations and proofs are provided in the web appendix.

We can explain briefly why each model satisfies our core intuition: that the influence of implicit preferences increases as comparisons become more dilute. In the *ceteris paribus* model, the decision maker is constrained by rules that apply to certain comparisons (e.g., a male MBA versus a female MBA) but turn off as additional differences are introduced (a male MBA versus a female PhD). As a result, they can express their implicit preferences more strongly as the set of differences between bundles grows. In the signaling model, the more that an attribute is mixed with others, the less an observer can infer about the decision maker's preference for that attribute, so they feel freer to express their true preferences. In the implicit associations foundation, the more an attribute is mixed with others, the less the decision maker can infer about potential unconscious influences on their preferences, so the more inclined they are to go with their gut instincts.

It will be useful to define the set of shared attributes for comparison $|\boldsymbol{x} - \boldsymbol{z}|$:

$$S^{|\boldsymbol{x}-\boldsymbol{z}|} = \{i : |x_i - z_i| = 0\}.$$

Non-shared attributes are those not in $S$. We suppress the superscript unless needed.

## 4.1 *Ceteris Paribus* Decision Maker

Suppose our hiring manager normally chooses whichever candidate they prefer, except when comparing a male candidate to an otherwise identical female candidate, in which case they are required to hire the woman. We state a general model of *ceteris paribus* decision makers who are constrained by rules that apply "all else equal," but otherwise maximize menu-independent utility. Rules can be interpreted as internal to the decision maker (e.g. a moral obligation) or external (e.g. a bureaucratic rule).[19]

We will also allow multiple rules which can compound or counteract one another, in which case "all else equal" is taken to mean when all *non-rule-governed* attributes are equal. Suppose a manager is instructed to both (1) prefer female candidates all else equal, (2) prefer Black candidates all else equal. We will assume that the rules combine such that they must

[19]For example, job advertisements at the Norwegian School of Economics (NHH) routinely include the sentence "In the event of equivalent qualifications, female applicants will be given preference."

choose a Black woman over a white man (otherwise equal), but when choosing between a white woman and a Black man the decision is governed by whichever rule has more force.

**Definition 11.** *A **ceteris paribus utility function** has the form:*

$$u^{CP}(\boldsymbol{x}, \boldsymbol{z}) = \underbrace{g(\boldsymbol{x})}_{\substack{\text{menu-independent} \\ \text{utility}}} + \sum_{i \notin S} x_i \underbrace{\lambda_i}_{\substack{\text{bonus or} \\ \text{penalty}}} \underbrace{\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\}}_{\substack{=1 \text{ iff all non-rule-governed} \\ \text{attributes are shared}}},$$

*for some* $g : \{-1, 1\}^n \to \mathbb{R}$*, and* $\boldsymbol{\lambda} \in \mathbb{R}^n$*.*

When $\lambda_i \neq 0$ we say attribute $i$ is governed by a rule. Thus the bonus/penalty $\lambda_i$ is applied to a bundle if and only if (a) attribute $i$ is non-shared ($i \notin S$); and (b) every attribute that is not governed by a rule ($\lambda_j = 0$) is shared ($j \in S$).

Applied to choice, $\lambda_i$ is a bonus/penalty for choosing one bundle over another. Rules could demand hiring a Black candidate, booking the cheapest flight, or ordering a low-calorie meal. If $\lambda = \infty$ the rule is inviolable. Applied to evaluation, $\lambda_i$ is a bonus/penalty applied to reported values. For example, someone might give women lower scores except when they are compared to an otherwise-identical man.

**Proposition 1.** $u^{CP}(\boldsymbol{x}, \boldsymbol{z})$ *satisfies Separable Implicit Preferences.*

For interpretation, consider our hiring manager example. When the candidates differ only on gender, opacity is low (the rule is applied) decreasing the utility of the male candidate. When the candidates differ in other attributes as well, opacity increases (the rule is turned off), increasing the utility of the male candidate increases. Hence, the rule manifests as an implicit preference favoring men.

## 4.2 Signaling Decision Maker

Suppose the decision maker holds intrinsic values over attributes, but also has reputational preferences. They care about the beliefs that some other person—perhaps their own future self—holds over those intrinsic values. We represent their intrinsic values as:

$$g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i$$

where $g(\boldsymbol{x})$ is assumed to be common knowledge, while $w_i$ terms ("weights") are the decision maker's private information. We assume the observer holds mean-zero, independent Normal priors over the weights, and forms posteriors based on the decision maker's actions.

The observer's information differs between choice and evaluation, so we will describe separate signaling models for each type of behavior. We assume throughout that the bundles $\boldsymbol{x}$ and $\boldsymbol{z}$ are chosen by Nature and are common knowledge (i.e., we abstract from strategic choice over choice sets).

### 4.2.1 Choice

When the decision maker chooses $\boldsymbol{x}$ over $\boldsymbol{z}$ the observer will update their beliefs $\hat{w}_i$ about the decision maker's weights on attributes where $\boldsymbol{x}$ and $\boldsymbol{z}$ differ. The core intuition is that when the bundles differ on a superset of attributes, the observer updates less about each $w_i$, so the decision maker's signaling incentives weaken.

We make two strong assumptions, which amount to the observer expecting the decision maker to be indifferent *ex ante*. First, the observer's priors over all intrinsic values have identical mean, which we normalize to zero: $g(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{X}$.[20] Second, we assume the observer is *naïve*, meaning they are not aware of the decision maker's reputational motives (otherwise they would expect a particular bundle to be chosen). These may be quite strong assumptions, but our conclusions should extend to small deviations. We discuss their relevance to applications in Section 5.

We define a comparative utility function $u^{SC}(\boldsymbol{x}, \boldsymbol{z})$, interpreted as the utility of choosing $\boldsymbol{x}$ when the observer knows the choice set was $\{\boldsymbol{x}, \boldsymbol{z}\}$. We assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct, so there is at least one non-shared attribute. We also assume that all preferences are expressed strictly.[21]

**Definition 12.** *A **signaling-choice utility function** has the form:*

$$\underbrace{u^{SC}(\boldsymbol{x}, \boldsymbol{z})}_{\substack{\text{utility of} \\ \text{choosing } \boldsymbol{x} \\ \text{from } \{\boldsymbol{x}, \boldsymbol{z}\}}} = \underbrace{\sum_{i=1}^{n} x_i w_i}_{\substack{\text{intrinsic} \\ \text{value}}} + \sum_{i=1}^{n} \underbrace{\lambda_i}_{\substack{\text{reputational} \\ \text{preference} \\ \text{for attribute } i}} \underbrace{\cdot E\left[ w_i \middle| \sum_{i=1}^{n} x_i w_i > \sum_{i=1}^{n} z_i w_i \right]}_{\substack{\text{observer's naïve posteriors} \\ \text{over weights when } \boldsymbol{x} \text{ is chosen}}},$$

*for some $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$ (observer's priors over weights).*

[20] If the observer had reason to believe the decision maker prefers one bundle over another, more dilute comparisons can sometimes be *more* informative about an attribute rather than less, violating assumption 3. For example, choosing a male PhD over a female MBA is plausibly less informative about gender preferences than choosing a male PhD over a female PhD. But choosing a male PhD over a female Nobel prize winner is clearly *more* informative, in the sense of posteriors being farther apart.

[21] It is possible to show that a decision maker would choose to express indifference, with its consequent reputational effects, only if they received equal utility from expressing indifference and expressing either of the two strict preferences, i.e. $u^{SC}(\boldsymbol{x}, \boldsymbol{z}) = u^{SC}(\boldsymbol{z}, \boldsymbol{x})$. Thus the function we derive for the 2-action world correctly predicts behavior in a 3-action world, so the model can be applied to data containing indifferences. A derivation is available on request.

$\lambda_i$ captures the decision maker's utility of shifting the observer's posterior over weight $w_i$. Because of the separable setup, the observer will only update about the weights on non-shared attributes.

**Proposition 2.** *$u^{SC}(\boldsymbol{x}, \boldsymbol{z})$ satisfies Separable Implicit Preferences.*

Consider a hiring manager that prefers men but wants the observer to believe they prefer women. When candidates differ on few attributes, the observer infers a lot about their gender preferences from their choice. As additional attributes vary, the observer updates less about gender, lowering the reputational cost of hiring a man. The implicit preference has the opposite sign to its associated signaling motive $\lambda_i$: a motive to signal a preference for women manifests as an implicit preference favoring men.

### 4.2.2 Evaluation

In evaluation we assume the decision maker reports their utility of two bundles, $\boldsymbol{x}$ and $\boldsymbol{z}$, with a quadratic cost of inaccuracy. An observer then makes inferences about the decision maker's weights $w_i$. Unlike the choice setting, we do not need to assume that the observer has constant priors over the intrinsic values, nor that they are naïve.

We will define a signaling evaluation function, $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$, show that it corresponds to an equilibrium strategy in a signaling game, and finally that it satisfies Separable Implicit Preferences.

**Definition 13.** *A **signaling evaluation utility function** is:*

$$u^{SE}(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} x_i \lambda_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\} \sum_{j \in S} \sigma_j^2 + \mathbf{1}\{i \notin S\} \sum_{j \notin S} \sigma_j^2}$$

*for some $g : \{-1, 1\}^n \to \mathbb{R}$, $\boldsymbol{\lambda} \in \mathbb{R}^n$, $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$.*[22]

**Lemma 1.** *Reporting the value of $y^x = u^{SE}(\boldsymbol{x}, \boldsymbol{z}), y^z = u^{SE}(\boldsymbol{z}, \boldsymbol{x})$, is an optimal strategy in a pure-strategy Perfect Bayes Equilibrium of a signaling game in which:*

1. *Player 1 first chooses $y^x$ and $y^z$ to maximize*

$$U^1 = \underbrace{-\frac{1}{2}\Big(y^x - g(\boldsymbol{x}) - \sum_{i=1}^{n} w_i x_i\Big)^2 - \frac{1}{2}\Big(y^z - g(\boldsymbol{z}) - \sum_{i=1}^{n} w_i z_i\Big)^2}_{\text{quadratic loss from inaccuracy}} + \underbrace{\sum_{i=1}^{n} \lambda_i \hat{w}_i(y^x, y^z)}_{\text{reputational gain}}.$$

[22] The assumption of mean-zero priors on $\boldsymbol{w}$ is without loss of generality, as $g(\boldsymbol{x})$ is unrestricted.

2. *Player 2 observes* $y^x, y^z$ *and chooses* $\hat{\boldsymbol{w}}$ *to maximize* $U^2 = -E\left[\sum_{i=1}^{n}(\hat{w}_i - w_i)^2 \middle| y^x, y^z\right]$, *with* $g(\cdot)$ *and* $\boldsymbol{\lambda}$ *common knowledge, and priors* $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$.

$\lambda_j$ captures the decision maker's utility of shifting the observer's posteriors over $w_j$, while $\sigma_j^2$ is the variance of the observer's prior on $w_j$. The final term in $u^{SE}$ captures how the decision maker adjusts her evaluations to influence the observer's beliefs. The adjustment to attribute $i$ is proportional to the observer's uncertainty about $w_i$ ($\sigma_i^2$), and inversely proportional to the total uncertainty about the weights on attributes with the same status (shared or non-shared) as $i$.[23]

**Proposition 3.** $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ *satisfies Separable Implicit Preferences.*

The intuition behind how signaling motives manifest as implicit preferences in evaluation is very similar to the choice example, with the exception that the observer now updates about both shared and non-shared attributes, because they observe distinct signals about both bundles' values rather than just their ranking.

## 4.3 Implicit Associations Decision Maker

Finally we describe a decision maker made up of two agents, each with private information relevant to the value of a bundle:[24]

$$\underbrace{f(\boldsymbol{x})}_{\substack{\text{true value}\\ \text{of bundle } \boldsymbol{x}}} = \underbrace{g(\boldsymbol{x})}_{\substack{\text{known}\\ \text{by both}}} + \sum_{i=1}^{n} \underbrace{x_i}_{\substack{\text{known}\\ \text{by both}}} \cdot \underbrace{\lambda_i}_{\substack{\text{known}\\ \text{by first}\\ \text{agent}}} \cdot \underbrace{\pi_i}_{\substack{\text{known}\\ \text{by second}\\ \text{agent}}} .$$

The first agent can be thought of as the pre-conscious brain, drawing on knowledge of "associations" ($\boldsymbol{\lambda} \in \mathbb{R}^n$) between each attribute and true value, and the second agent can be thought of as the conscious brain, which has access to "adjustments" ($\boldsymbol{\pi} \in \mathbb{R}^n_+$), high-level contextual information used to adjust the value of each association.

Sequencing is as follows. The first agent calculates expected values for bundles $\boldsymbol{x}$ and $\boldsymbol{z}$ ($E[f(\boldsymbol{x})|\boldsymbol{\lambda}]$ and $E[f(\boldsymbol{z})|\boldsymbol{\lambda}]$). The second agent then makes decisions taking into account the first agent's two estimates, plus its own private information ($\boldsymbol{\pi}$), but without access to the underlying associations ($\boldsymbol{\lambda}$). The theory predicts that the second agent's estimate of

[23] We stated at the beginning that we did not need to assume naïveté, and solved the model assuming full sophistication. In fact, the quadratic loss function means that player 1's best response is independent of the observer's prior beliefs about $\boldsymbol{\lambda}$ and so our solution will to hold if the observer has incorrect priors, including full naïveté (believing $\boldsymbol{\lambda} = \boldsymbol{0}$).

[24] This model is based on Cunningham (2016), which discusses more generally conditions under which sequential aggregation of information will be efficient.

$\boldsymbol{x}$'s value will be affected by a comparator $\boldsymbol{z}$ insofar as the comparison is informative about associations, $\boldsymbol{\lambda}$.

The core intuition is that associations are generally informative (otherwise the second agent would ignore the first agent's estimates). However, the second agent has access to contextual information that leads her to adjust the first agent's estimates. Her ability to apply these adjustments depends on the degree to which she can separately distinguish the influence of each association. For example, the decision maker might have an association with gender, $\lambda_i \neq 0$, e.g. an unconscious positive attitude toward men. However, the conscious brain has reason to believe that in the current setting any such association is normatively irrelevant ($\pi_i = 0$). Such a decision maker will exhibit a pro-male bias that decreases as the comparison becomes less opaque.

**Definition 14.** *An **implicit associations utility function** has the form:*

$$u^{IA}(\boldsymbol{x}, \boldsymbol{z}) = E\big[f(\boldsymbol{x})|\boldsymbol{\pi}, E[f(\boldsymbol{x})|\boldsymbol{\lambda}], E[f(\boldsymbol{z})|\boldsymbol{\lambda}]\big],$$

*with*

$$\begin{aligned} &\pi_i \in \mathbb{R}_+ \;\&\; E[\pi_i] = 1 && \textit{(1st agent's priors)} \\ &\boldsymbol{\lambda} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2)) && \textit{(2nd agent's priors)} \\ &\boldsymbol{\pi} \perp\!\!\!\perp \boldsymbol{\lambda} && \textit{(independence of priors).} \end{aligned}$$

$u^{IA}$ represents the second agent's best guess at the true value $f(\boldsymbol{x})$.

In this model the sensitivity of utility to attribute $i$ will be proportional to a weighted average of the adjustments (the $\pi$s) on all attributes with the same status (shared or non-shared) as $i$. So, in general a dilution of attribute $i$ can increase or decrease opacity, depending on whether the dilution increases or decreases that average. This is inconsistent with Assumption 3. The foundation satisfies Assumption 3 in two special cases: (i) when there are exactly two attributes ($n = 2$); or (ii) when an unexpected realization of information occurs for at most one attribute (at most one $i$ has either $\lambda_i \neq 0$ or $\pi_i \neq 1$). We adopt the second assumption for the remainder of the section, which implies that there can be an implicit preference for at most one attribute.

**Proposition 4.** *$u^{IA}(\boldsymbol{x}, \boldsymbol{z})$ satisfies Separable Implicit Preferences if at most one attribute has a non-zero implicit association and/or non-unitary adjustment factor: $\sum_{i=1}^{n} \mathbf{1}\{\lambda_i \neq 0 \text{ or } \pi_i \neq 1\} \leq 1$.*

For intuition of how implicit associations can be interpreted as implicit preferences consider the hiring manager that has a positive association with male candidates ($\lambda_i > 0$), but believes gender is normatively irrelevant ($\pi_i = 0$). When the candidates differ only on gender, opacity

is low, as the second agent can directly detect and override the influence of $\lambda_i$. As gender is diluted, the influence of gender is mixed with other possible associations. The second agent believes some of those associations might contain valuable information, so they do not ignore them entirely. Thus $\lambda_i$ influences their decision, increasing the utility of the man and manifesting as an implicit preference favoring men.[25]

### 4.4 Dominance of attribute $k$

Assumption 4 permits ranking of opacity when an attribute changes status, from shared to non-shared or vice-versa. That is useful for analysis of evaluation data. It says that for some special attribute $k$, opacity about $i$ increases when $i$ has the same status as $k$. Our final proposition provides sufficient conditions for this assumption to hold in all foundations.

**Proposition 5.** *The* ceteris paribus *decision maker of Proposition 1 satisfies Assumption 4 if $k$ is not governed by a rule. The* signaling evaluation *decision maker of Proposition 3 satisfies Assumption 4 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$. The* implicit associations *decision maker of Proposition 4 satisfies Assumption 4 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$.*

## 5 Guidance for applications

We expect that the canonical examples introduced in Section 3 will be especially useful for applications. We now discuss some other practical guidance for applications of our theory.

**Multivalued attributes** Some attributes might take multiple values. For example, we might observe job candidates with three different qualifications (MBA/PhD/JD) instead of two. Since our theory is based on binary attributes, the data need to be transformed in order to apply it. The appropriate transformation depends on the setting. Attribute values could be grouped together, or the dataset partitioned to focus on parts of the attribute space. Our analysis of Exley (2016)'s data needs to address the fact that one attribute (the probability of winning a lottery prize) is multivalued. We construct binary attribute spaces around each probability, and analyze them separately.

[25]The sign of the implied implicit preference depends on $\lambda_i(1 - \pi_i)$. If $\pi_i > 1$, the second agent wants to *amplify* their implicit associations (in a sense, they think the first agent is too conservative). This generates an implicit preference with the opposite sign to $\lambda_i$. In our example, this increases the value of men when the candidates differ only on gender, and weakens as gender is mixed with other attributes, generating an implicit preference favoring women. Note also that there is only an implicit preference if both $\lambda_i \neq 0$ *and* $\pi_i \neq 1$: the first agent must have a nonzero implicit association and the second agent must want to adjust it.

***Ambivalence* in choice data** In choice data an important consideration arises that we refer to as *ambivalence*. Choice sets should be constructed such that participants are *expected to be close to indifferent*. There are two reasons for this, a statistical one and a theoretical one. The statistical (or "calibration") motive is that it is difficult to observe intransitivities, even when they exist, if the ordering of the explicit values $v(.)$ is strong enough. The theoretical motive is that our signaling-choice foundation relies on the observer having equal priors over the utility of both bundles.

When there are multiple non-ambivalent attributes in the dataset, one solution is to group them together so that their combination plausibly satisfies ambivalence. For instance, while a hiring manager is unlikely to be close to indifferent between a candidate with a BA and one with a PhD, they might plausibly be so between a BA with work experience, versus a PhD without.

Our analysis of Exley (2016)'s data faces this issue. The basic attributes that vary in her experiment (Recipient, Prize, and Probability of winning) are unlikely to satisfy ambivalence: all else equal, we would expect payments to self to be preferred to payments to charity, and higher prizes or probabilities to lower. We therefore construct two new attributes (Attitude and Risk) by grouping payments to self with lower prizes than payments to charity, and sure payoffs with lower prizes than risky ones, in order to restore Ambivalence.

**Within-subjects data** The theory assumes we observe the revealed preferences of a single decision-maker, that is, we observe within-subjects data. A concern in such datasets is order effects: participants' later decisions may be influenced by their earlier ones. The usual experimental technique to minimize order effects is to spread decisions over time, intersperse them with "filler" tasks or questions, or in other ways make their earlier decisions less salient or harder to remember. This appears to have been successful in Exley (2016)'s experiments, in which many participants reveal within-subject inconsistencies.[26]

If order effects are a serious concern, the standard response is to collect between-subjects data in which each participant makes only one or a small number of choices. This has different implications for analysis of choice and evaluation data.

**Between-subjects choice data** Establishing the presence of intransitivities in between-subjects data is challenging, because intransitivity is difficult to distinguish from underlying heterogeneity in preferences (similar to the Condorcet paradox in pairwise voting). A strong solution is to impose homogeneity restrictions on preferences. Alternatively, one can test for

[26]A related concern is experimenter demand effects: participants may guess what the experimenter is looking for from the sequence of decisions they observe. Recent work that directly manipulates such beliefs finds mostly modest effects (de Quidt et al., 2018; Mummolo and Peterson, 2018).

violations of the Triangle inequality (see Regenwetter et al. (2011) for extensive discussion). To establish the presence of at least one intransitive decision-maker in choice over $a, b, c$, we would need to observe $Pr(a \succ b) + Pr(b \succ c) + Pr(c \succ a) > 2$, i.e. the average choice probability must strictly exceed $2/3$. For four-element cycles the threshold increases to $3/4$. It may be challenging to find a setting with sufficiently strong intransitive preferences to satisfy such conditions (Müller-Trede et al., 2015).[27]

**Between-subjects evaluation data** Our tools for evaluation data carry over well to between-subjects data, provided we are willing to impose some restrictions on heterogeneity and functional form. Our application to DeSante (2013) is an example of such an analysis. Suppose we observe $t = 1, \ldots, T$ iid sampled individuals' evaluations of $\boldsymbol{x}$ with comparator $\boldsymbol{z}$. We allow for heterogeneity in $v(.)$ and $\boldsymbol{\kappa}$, with population averages $\overline{v(\boldsymbol{x})}$ and $\overline{\boldsymbol{\kappa}}$, while assuming $\boldsymbol{\theta}$ is common and determined by the structure of the comparison set.[28] We also assume evaluations are affine in utility: $y(\boldsymbol{x}, \boldsymbol{z}) = a + b \times u(\boldsymbol{x}, \boldsymbol{z})$. Normalizing $a = 0, b = 1$, the average evaluation is:

$$\frac{1}{T}\sum_{t=1}^{T}\left[v_t(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \kappa_{i,t} \theta_i(|\boldsymbol{x} - \boldsymbol{z}|)\right] \xrightarrow[T\to\infty]{} \overline{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i sgn(\overline{\kappa}_i)|\overline{\kappa}_i|\theta_i(|\boldsymbol{x} - \boldsymbol{z}|).$$

This is the comparative utility function of a representative agent who has implicit preferences $\kappa_i^{rep} = sgn(\bar{\kappa}_i)$ and opacity function $\theta_i^{rep} = |\overline{\kappa}_i|\theta_i$. It satisfies Assumption 1. Thus our usual tools can identify $sgn(\bar{\kappa}_i)$, the sign of the *average* implicit preference in the population. If $sgn(\bar{\kappa}_i)$ is positive, we learn that at least some part of the population has positive $\kappa_i$. If we also assume that implicit preferences are aligned in the population (have weakly the same sign), we learn that sign.

# 6 Applications

## 6.1 Implicit Risk and Social Preferences (Exley, 2016)

Exley (2016) studies "the use of risk as an excuse not to give." She conducts two experiments in which participants make a sequence of choices between lotteries or sure payments, where the beneficiaries can be either themselves, or charity.[29] She uses the choice data to construct

[27] As an example, the choice proportions in Snyder et al. (1979)'s experiment do not satisfy the criterion and could be explained by heterogeneous, transitive preferences.

[28] We could allow for heterogeneity of the form $\theta_{i,t} = \alpha_{i,t}\theta_i$, then we would identify $sgn(E[\kappa_i \alpha_i])$.

[29] In her second experiment the other beneficiary is another participant in the study, we use "charity" throughout for brevity.

certainty equivalents, such that each lottery to self, and each lottery to charity, is valued in terms of money to self, and in terms of money to charity, and tests for variation in these certainty equivalents as the trade-off between self and charity varies. The canonical pattern is one in which participants tolerate more risk when the risk favors them (high certainty equivalents), and less risk when the risk favors charity (low certainty equivalents), relative to when there is no trade-off between payoffs to self or to charity, suggesting *implicit selfishness*.

Reanalyzing Exley's dataset using our methods for choice data, we confirm this interpretation: 51 percent of participants reveal an implicit selfish preference. Our approach also yields new insights in the form of *implicit risk preferences*. 30 percent of participants become more risk averse when opacity about risk increases, while 15 percent become more risk tolerant.

### 6.1.1 Data

We need to do a little work to place Exley's data in a binary attribute framework. Web appendix B.3 provides a detailed description of the data structure, how it can reveal preferences on a binary attribute space, and how we can use the same assumptions as Exley's analysis to impute certain choices that are not directly observed in the data. Here we provide a brief summary.

Exley's dataset consists of an initial *normalization* choice (to figure out at roughly what exchange rate the participant is indifferent between money to self and to charity), followed by a sequence of choice lists. Each subsequent choice is between a safe payoff or a simple lottery paying a single prize with probability $P$. There are four types of choice list, across which the recipient of the lottery and the safe payoff vary: either both are to self, both are to charity, or one is to self and the other to charity. She repeats the exercise for a total of seven values of $P$: $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We do not observe choices between lotteries with different prizes or probabilities, so perform separate analysis for each value of $P$. Thus, each participant has seven separate opportunities to reveal their implicit preferences.

We show in the Appendix how, for each $P$, we can represent the empirical content of the data as four binary choices over two binary attributes, which we label Social $\in$ {Generous, Selfish} and Risk $\in$ {Safe, Risky}. Figure 4 shows the attribute space.

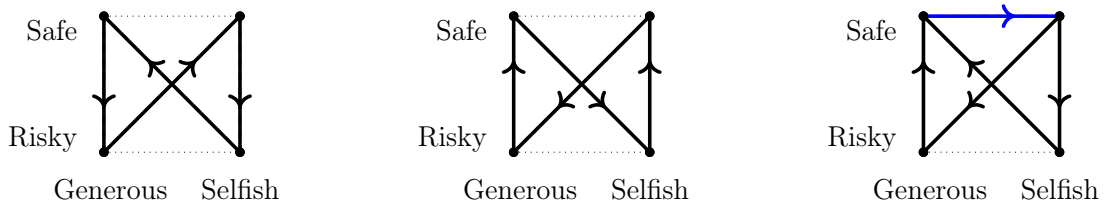
Not every choice set on this binary space is observed in the data. Specifically, participants do not make direct choices on the horizontal edges, (between two sure payments or between two lotteries). The horizontal choices are needed for analysis of implicit social preferences. In her analysis, Exley uses a *linearity in payoffs* assumption to compare choices involving money to self to those involving money to charity. Given how we construct the binary attribute space, that same assumption allows us to impute the choice

(Generous, Safe) $\succ$ (Selfish, Safe), marked in blue on the diagram. We do not observe the choices we would need to impute the opposing preferences on the horizontal choice sets. See the Appendix for details.

Exley excludes from most of her analysis participants whose initial normalization choices were censored or inconsistent, since their later choice lists cannot be properly calibrated. We do the same. We pool the data from both of her experiments, giving us 86 participants.

### 6.1.2 Observable cycles in the data

There are three possible patterns of choice that unambiguously identify an implicit preference. Figures 4a and 4b show figure-8 cycles that reveal implicit preferences over the Risk attribute. Figure 4c shows a pair of parallel right triangles that jointly reveal an implicit preference for Selfish.[30] Since we do not observe the opposing horizontal choice, we cannot detect implicit Generous preferences, nor can we detect inconsistencies in implicit preferences over this attribute.




(a) Implicit pro-Risky (b) Implicit pro-Safe (c) Implicit pro-Selfish

Choices marked in black are observable in the data. The blue horizontal choice is imputed.

Figure 4: Exley (2016) data structure

Table 1 presents the empirical frequencies of each of the cycles shown in Figure 4, averaged over the 86 participants and 7 values of $P$ for a total of 602 observations. Overall, participants exhibit one of the cycles of interest 33 percent of time, but at different frequencies. Only 5 percent of choices exhibit pro-Risky cycles, 10 percent are pro-Safe, while 18 percent are pro-Selfish.

### 6.1.3 Classifying individuals by implicit preference type

We begin by analyzing implicit risk preferences. We classify participants into one of four categories, by counting their number of pro-Risky and pro-Safe cycles across the seven values

[30] Note that identification of implicit Selfish preferences depends on the imputed horizontal choice. Without it, the observed preferences could be consistent with entirely *explicit* selfish preferences. Specifically, (Selfish, Safe) $\succ$ (Selfish, Risky) $\succ$ (Generous, Risky) $\succ$ (Generous, Safe), which always ranks Selfish bundles above Generous ones. Exley's analysis depends on her linearity assumption in a similar way.

| Cycle | | | Frequency | s.e. | 95% CI |
|---|---|---|---|---|---|
| Implicit pro-Risky | (a) |  | .048 | (.01) | [.032, .073] |
| Implicit pro-Safe | (b) |  | .103 | (.018) | [.073, .144] |
| Implicit pro-Selfish | (c) |  | .181 | (.024) | [.139, .233] |

This table shows the frequency of each of the cycles in Figure 4. Standard errors clustered at the participant level. Statistical tests: $p(a = b) = .009$, $p(a = c) < .001$, $p(b = c) = .019$, $p(a = b = c) < .001$, $p(a + b + c = .25) < .001$.

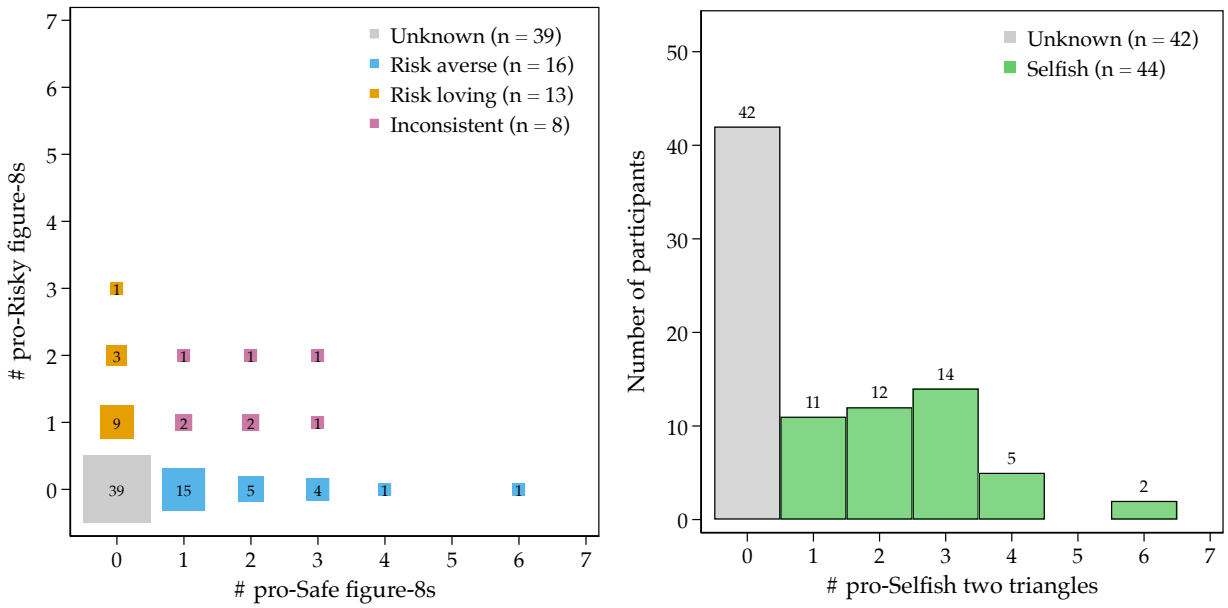
Table 1: Frequencies of different cycles in Exley (2016) data

of $P$. The classifications are: *Unknown* (no type (a) or (b) cycles); *Implicit pro-Risky* (at least one (a) cycle, no (b) cycles); *Implicit pro-Safe* (at least one (b) cycle, no (a) cycles); and *Inconsistent* (at least one of each). Figure 5a presents the findings, plotting the joint distribution of participant-level cycle counts. Of the 86 participants, 39 do not reveal any implicit risk preferences, 26 are implicitly pro-Safe, 13 implicitly pro-Risky, and 8 are inconsistent.

Implicit risk attitudes are prevalent in the sample, and tend to be implicitly risk-averse. This could have important implications for real-world decisions. For example, an implicitly risk-averse decision-maker might make more risk-averse choices when choosing between pension plans with different attributes (where opacity about risk is high) than she would when choosing between different variants of the same plan (where opacity is lower). That could have substantial implications for wealth at retirement.

However, relatively few participants exhibit more than one of these cycles. Among the consistent participants, 11 participants exhibit two or more pro-Safe cycles, while 4 exhibit two or more pro-Risky cycles. This suggests that implicit risk attitudes may be weak at the individual level. We assess the overall strength of the findings in Section 6.1.4 below.

Turning to implicit Social preferences, we classify participants according to their number of pro-Selfish cycles. They can be either *Unknown* (no (c) cycles), or *Implicit pro-Selfish* (at least one (c) cycle). Figure 5b shows that just over half of the participants (44 in total) are classified as implicitly pro-Selfish. Of those, 75% exhibit this pattern twice or more. Thus implicit selfishness is more widespread, and expressed more frequently, than either pro-Safe or pro-Risky implicit preferences. However, as noted above, we cannot assess the extent of inconsistency in this preference.




(a) Implicit risk preferences

(b) Implicit social preferences

Panel (a) classifies participants according to their number of implicit pro-Risky and pro-Safe cycles. Cell size and numeric labels indicate the number of participants in each cell, colors indicate the type classification. Panel (b) classifies participants according to their number of implicit pro-Selfish cycles (the data structure means that we do not observe pro-Generous cycles).

Figure 5: Type classifications in Exley (2016) data

### 6.1.4 Statistical analysis

Our analysis so far assumes behavior is deterministic, giving rise to discrete type classifications. In reality some of the heterogeneity we observe is likely a result of errors, or noise in the data. Here we examine how systematic are the patterns we observe.

First we verify that the data are inconsistent with purely random behavior. Table 1 shows that the frequency of each type of cycle is heterogeneous, and a joint test strongly rejects equality of frequencies across types of cycle ($p < .001$). Additionally, we reject equality of each pair of frequencies. We observe a strong systematic tendency toward pro-Selfish cycles, and a strong tendency toward pro-Safe relative to pro-Risky cycles.[31]

Next, we examine whether these general tendencies are consistent with homogeneous implicit preferences, plus noise. We do this using permutation tests, under which we specify a precise null hypothesis, and then compare the distribution of behavior that we observe to what we would expect under that null. Web Appendix B.3.5 provides full details. We

[31] We also simulate a large dataset where the switching point in each choice list is uniformly random. In the simulated dataset, each type of cycle occurs with frequency .083 (well outside the 95% CIs for type (a) and (c) cycles), and the frequency of at least one cycle is .25 (in the data the rate is 0.33, with 95% CI [.278, .386]). Thus the observed behavior differs substantially from this random choice benchmark.

conclude that while noise is an important component of the observed behavior, there is evidence of both systematic and heterogeneous implicit preferences in the sample.[32]

In sum, like Exley, we find substantial evidence of implicit selfishness. Our analysis demonstrates that our general-purpose method can pick this up when applied to experimental choice data, and can extract new findings (implicit risk preferences) from data collected for another purpose.

## 6.2 Implicit Racial Discrimination (DeSante, 2013)

DeSante (2013) conducted an experiment on a US representative sample, in which participants were asked to recommend state welfare payments for hypothetical applicants. The paper asks whether people reward hard work in a "color-blind manner," i.e. whether the relationship between the applicant's reported "work ethic" and the funds allocated to them is the same for Black and White applicants. We will show how the experimental data can also be analyzed through the lens of our model to test for implicit preferences along race lines. Specifically, we will test whether participants tend to award more money to applicants of one race, and less to the other, when the comparison is more opaque about race.

Participants were presented with two hypothetical application forms (constructed from real applications) side-by-side. They were asked to allocate up to a total of $1,500 to the two applicants, with the remainder going to to "offset the deficit." We therefore interpret the decision as joint evaluation.[33]

The key attribute of interest is the applicant's Race $\in$ {Black, White}, signaled by their name (Latoya and Keisha for Black applicants, Laurie and Emily for Whites). Crucially, some participants evaluate two applicants of the same Race, while others evaluate one from each Race.[34] Second, in some conditions there is also an assessment of each applicant's Work Ethic $\in$ {Good, Bad}.[35] When reported, this attribute always varies within the comparison

[32]Test 2 asks whether, conditional on the fact that some cycles are more frequent than others, there is systematic heterogeneity in behavior across individuals. The null hypothesis is that all participants have the same likelihood of exhibiting each distinct type of cycle. We strongly reject this hypothesis ($p < .001$). Test 3 investigates whether the *sign* of implicit preferences is homogeneous, but individuals vary in how strongly or frequently they reveal them. The null hypothesis that the probability of a cycle is heterogeneous, but, conditional on exhibiting a cycle, everyone has the same probability of each type of cycle. We find evidence against this hypothesis too, albeit somewhat weaker ($p = .065$).

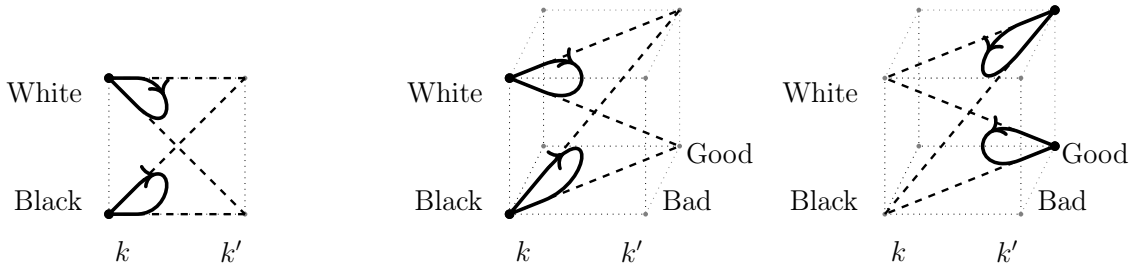
[33]The total budget constraint introduces a slight complication, since when it binds, a participant that wants to assign a high value to one applicant is constrained to give less to the other. This could make it harder to detect implicit preferences, as it is expected to particularly constrain allocations when the comparison set contains two of the most implicitly-preferred applicants. In the data 31 percent of participants allocate the whole $1,500 to the two applicants.

[34]Simonsohn (2016) points out that names might signal something additional to race, e.g. socioeconomic status. So we might be observing implicit preferences over SES instead of race.

[35]The language in the experiment is "Excellent/Poor", we use "Good/Bad" for compactness. Race is hidden

set. Third, there are some less salient additional characteristics (e.g. the ages of the applicants' children), which are randomized independently of race and work ethic. These are not observed in the data, so we will treat them as a third "background" attribute Kids $\in \{k, k'\}$ which always differs within the comparison set, with no implicit preference attached to it.

Figure 6 represents the data structure graphically. We observe some evaluations over bundles with two attributes (Panel (a)), and some with three (Panel (b)). Each applicant is evaluated alongside a Black comparator and a White comparator, who are otherwise identical to each other. For example, candidate (Black, Bad, $k$) is evaluated alongside (Black, Good, $k'$) and (White, Good, $k'$). From these we can construct six convex scissors and three pairs of parallel convex scissors (see Section 3).




(a) Work ethic concealed (b) Work ethic revealed

Labels $k$ and $k'$ indicate the background attribute Kids which varies independently of the other attributes and over which we assume there are no implicit preferences. On each diagram we mark the observed comparisons, and inequalities in evaluation which would reveal a positive implicit preference for White applicants.

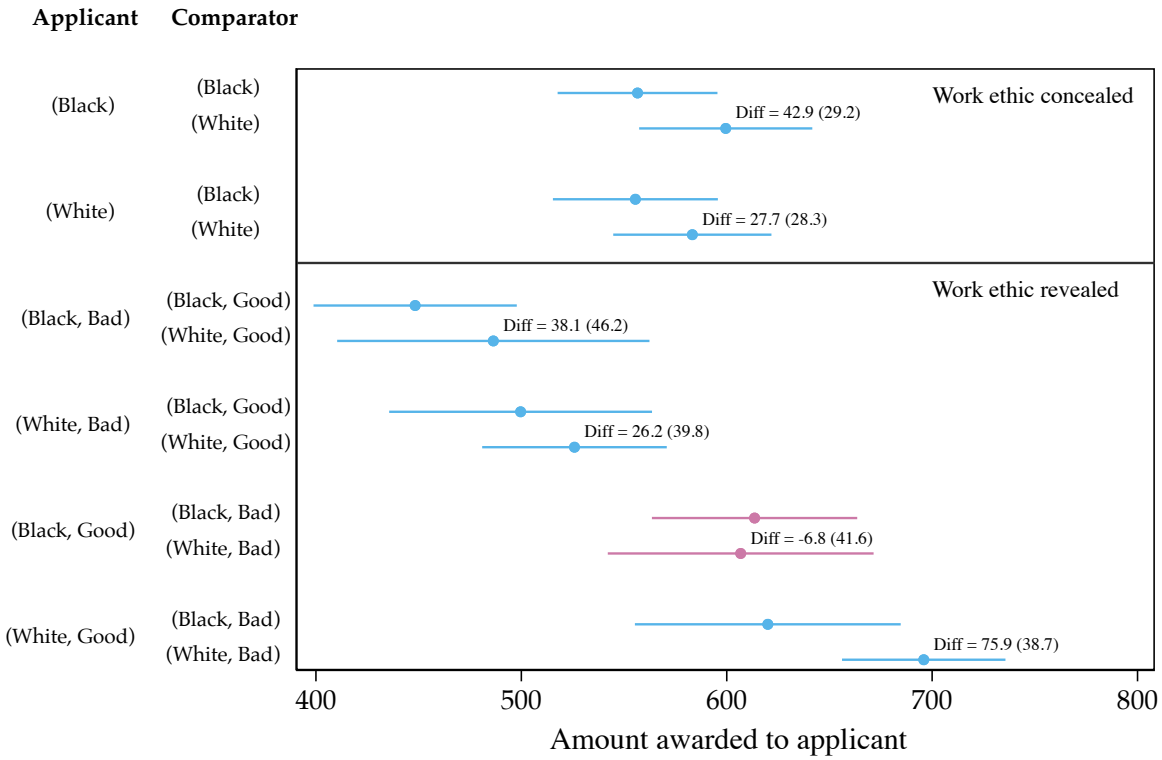
Figure 6: DeSante (2013) data structure

Our workhorse mode, Separable Implicit Preferences, does not rank opacity between these comparisons, because race switches from shared to non-shared. It can thus tell us if there is an implicit racial preference, but not its sign (see Corollary 8). We thus make use of Assumption 4. Specifically, we assume that one of the many attributes that are shared between comparisons (e.g., the fact that all applicants are female), satisfies the assumption, such that opacity is higher for shared attributes than non-shared. This tells us that we learn more about attitudes toward Black applicants relative to Whites when there is one Black and one White applicant, than when both are Black or White.

Given this setup, an implicit pro-White preference will manifest as (1) Higher evaluations of Black applicants when compared to White comparators, than when compared to Black comparators, and (2) Higher evaluations of White applicants when compared to White comparators than when compared to Black comparators. In other words, evaluations should always increase when the comparator switches from Black to White.

in DeSante's experimental conditions 1–3. These conditions have no variation in opacity, so we drop them.

The experiment uses a between-subjects design, that is, each participant reports exactly one pair of evaluations, corresponding to one of the comparison sets in Figure 6. We therefore cannot identify implicit preferences at the individual level. Instead we will compare average evaluations across different comparison sets, and interpret these averages as revealing the preferences of a representative agent, as explained in Section 5.




Each pair of points corresponds to one "scissor." Positive values of "Diff" (blue) are consistent with pro-White implicit preferences, negative values (red) are inconsistent. $N = 753$ participants and 1,506 evaluations. Standard errors clustered at participant level. The average difference across conditions is 43.0 (s.e. 13.5). Joint test for no influence of any comparator, $F(6, 752) = 1.91, p = 0.0765$.

Figure 7: Reanalysis of DeSante (2013) data

Figure 7 presents the results. We group evaluations in pairs that correspond to six convex scissor tests in three pairs of parallel convex scissors. We find positive differences in five out of six scissors, meaning that the general pattern is as predicted by implicit pro-White preferences. Only one of these individual differences (the fourth, with (White, Good) as target) is statistically significant, but the average of the differences across conditions is \$43, and highly significantly different from zero ($p < 0.01$). An F-test of the null of no difference in any Scissor has a p-value of 0.08. Overall, we find moderate evidence of implicit pro-White preferences in this dataset.[36]

[36]One could also use these data to test for implicit preferences over Work Ethic (when this attribute is

# 7 Related Theories

Our identification of implicit preferences relies on inconsistencies in choice and in evaluation. However inconsistencies could occur for other reasons. In this section we discuss three alternatives, and argue that each is unlikely or unable to produce the specific patterns in choice and evaluation that we associate with implicit preferences.

**Contingent weighting.** Models of contingent weighting in multi-attribute choice, like our theory, assume that preferences depend on the choice set.[37] However most existing theories rely on a very different intuition: they assume that the sensitivity to a given attribute depends on the observed distribution over that attribute. In contrast, our model assumes that sensitivity depends on what other attributes share status with the attribute of interest. None of the recently published contingent-weighting models is consistent with a figure-8 intransitivity.[38]

A similar point applies to the literature on comparing joint and separate evaluation of outcomes: Hsee et al. (1999) give many examples. Most of these studies find that people are more sensitive to an attribute when presented jointly (two bundles simultaneously) than separately (one at a time). They argue that this increased sensitivity is a general feature of joint evaluation, called "evaluability."[39] Again, this is a quite different principle to that

included). Opacity about Work Ethic is higher in the comparisons where Work Ethic co-varies with Race, so implicit pro-Good preferences would manifest as higher evaluations of Good candidates in these comparisons. Figure 7 does not suggest any systematic pro-Good or pro-Bad patterns.

[37]For example in Kőszegi and Szeidl (2012) sensitivity is positively related to the range of values on an attribute, in Bushong et al. (2020) it is negatively related to the range, in Cunningham (2013) it is negatively related to the average, and in Bordalo et al. (2013) it is (roughly) negatively related to the proportional range (range divided by the average).

[38] Formally, suppose the utility function is entirely separable in each attribute, in the sense that it can be written as,

$$u(x, A) = \sum_i u_i(x_i, \{a_i^j\}_{j=1}^m),$$

where $a_i^j$ is the $i$th attribute of the $j$th element of the choice set, $A$, then a figure-8 intransitivity could never occur because - using the gender example - the marginal distribution of the gender attribute remains the same in all four choice sets, thus the difference in attribute-utility ($u_i$) between "Male" and "Female" must remain the same. The two diagonal choice-sets must evoke the same utility function, because they have the same marginal distributions, and that utility function prefers Male to Female, all else equal. But this contradicts the choice observed in the vertical choice sets (where Female is chosen over Male). Separability holds for all the models discussed above except Bordalo et al. (2013), but that model cannot generate intransitive cycles in binary choices with two attributes.

[39]For example subjects were found to state a higher WTP for a dictionary with 10,000 entries when it was evaluated alone, than when it was evaluated alongside a dictionary with 20,000 entries and a torn cover. Kahneman and Frederick (2005) discuss a similar phenomenon: that subjects are generally more sensitive to changes in within-subjects experiments than in between-subjects experiments. The theory is further developed in Hsee and Zhang (2010).

used in this paper. Hsee's mechanism could not generate a figure-8 cycle, by an analogous argument to footnote 38. See Cunningham (2013) for a Bayesian rationalization of increased sensitivity in joint evaluation.

**Inattention/Heuristics.** Our identification comes from comparing where more or fewer attributes vary. If the former are more complex than the latter, we might worry that inconsistencies are due to complexity variation, as in models of inattention (Sims (2003), Caplin and Martin (2014), Woodford (2012)). It is intuitive that a decision-maker could become less sensitive to an attribute in a more complex choice situation, however it would be unusual for an increase in complexity to causes the preference for an attribute to *reverse*, as necessary for the figure 8 choice pattern.[40] An exception is Cubitt et al. (2018), in whose model the decision maker puts less weight on each attribute when more attributes vary, *but* treats money separately from other attributes. That model cannot generate strict cycles over non-monetary attributes.

**Inference from the choice set.** We assume that the attribute values of one bundle are uninformative about the value of other bundles. If not, in principle, any pattern of choice could be rationalized. The relevant question is what types of prior beliefs could generate the patterns we observe, and whether those beliefs seem realistic. Consider our leading example of gender bias in hiring. These decisions could be rationalized by a hiring manager who (1) prefers women to men, all else equal; but (2) believes that men have better qualifications. Thus in the diagonal choice sets they prefer men, not because they are men, but because they have the qualification that men have. In applications with familiar attributes this seems unlikely to be important, as the scope for learning from the choice set seems small. Moreover, the explanation requires that the *intrinsic* value of an attribute be opposite to its *informational* value (in this case, being male is a negative signal about the person, but a positive signal about things that covary with maleness).[41]

# 8 Conclusion

Our paper formalizes an assumption that is latent in a number of empirical papers: that people maintain two layers of preference for a given attribute, such that one preference–the

[40] A figure-8 with indifferences could come from inattention if sensitivity to an attribute goes to zero in complex choices, though we are not aware of an inattention model with this feature.

[41] This inference-based explanation can more easily rationalize a figure-8 which has indifferences on the vertical comparisons: e.g. if the decision-maker was indifferent between men and women, all else equal, but would choose men on the diagonal based beliefs about their qualifications.

*implicit* preference–becomes stronger when the comparison between outcomes becomes more indirect.

By formalizing this assumption we are able to give precise guidance for inferring a person's implicit preferences from their decisions in a way that is applicable to many existing empirical datasets. We provide two empirical applications, but the number of domains in which the approach can be applied is large.

A natural formal extension would be to apply the representation theorem to other definitions of comparison and opacity. For example, we assumed that comparisons depend only on which attributes are shared or non-shared, an alternative would preserve the direction of differences ($\delta(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x} - \boldsymbol{z}$). This could give conditions for identification of implicit preferences in additional models, such as a sophisticated signaling model.

A natural empirical extension would be to run fresh experiments designed to systematically map out the existence, strength, consistency, and out-of-sample predictiveness of implicit preferences across a range of different attributes.

# References


Andreoni, J. and B. D. Bernheim (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton University Press.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2022). Explicit and implicit belief-based gender discrimination: A hiring experiment. *mimeo*.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Benabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies 70*(3), 489–520.

Benabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review*, 94–98.

Bertrand, M. and E. Duflo (2017). Field Experiments on Discrimination. In *Handbook of Field Experiments*, pp. 309–393. Elsevier.

Bodner, R. and D. Prelec (2003). Self-signaling and diagnostic utility in everyday decision making. In I. Brocas and J. D. Carrillo (Eds.), *The Psychology of Economic Decisions Volume One: Rationality and Well-Being*. Oxford: Oxford University Press.

Bohnet, I., A. van Geen, and M. Bazerman (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science 62*(5), 1225–1234.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019). Innacurate statistical discrimination. *SSRN Electronic Journal*.

Bohren, J. A., P. Hull, and A. Imas (2022). Systemic discrimination: Theory and measurement. *mimeo*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*(5), 803–843.

Border, K. C. (2013). Alternative linear inequalities, version 2020.10.15::09.50, accessed 2022-04-24. *https://kcborder.caltech.edu/Notes/Alternative.pdf*.

Bursztyn, L., G. Egorov, I. Haaland, A. Rao, and C. Roth (2022). Justifying dissent. *mimeo*.

Bushong, B., M. Rabin, and J. Schwartzstein (2020). A model of relative thinking. *The Review of Economic Studies 88*(1), 162–191.

Caplin, A. and D. Martin (2014). A testable theory of imperfect perception. *The Economic Journal 125*(582), 184–202.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics 134*(3), 1163–1224.

Caruso, E. M., D. A. Rahnev, and M. R. Banaji (2009). Using conjoint analysis to detect discrimination: Revealing covert preferences from overt choices. *Social Cognition 27*(1), 128–137.

Chance, Z. and M. I. Norton (2009). "I Read Playboy for the Articles": Justifying and Rationalizing Questionable Preferences. In M. S. McGlone and M. L. Knapp (Eds.), *The Interplay of Truth and Deception: New Agendas in Theory and Research*, Chapter 9. Routledge.

Cherepanov, V., T. Feddersen, and A. Sandroni (2013). Rationalization. *Theoretical Economics 8*(3), 775–800.

Corno, L., E. L. Ferrara, and J. Burns (2018). Interaction, stereotypes and performance. evidence from south africa. *IFS Working Paper W19/03*.

Cubitt, R., R. McDonald, and D. Read (2018). Time matters less when outcomes differ: Unimodal vs. cross-modal comparisons in intertemporal choice. *Management Science 64*(2), 873–887.

Cunningham, T. (2013). Comparisons and choice. *Stockholm University mimeo*.

Cunningham, T. (2016). Hierarchical aggregation of information and decision-making. *Stockholm University mimeo*.

Dana, J., R. A. Weber, and J. X. Kuang (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory 33*(1), 67–80.

de Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and bounding experimenter demand. *American Economic Review 108*(11), 3266–3302.

DeSante, C. D. (2013). Working twice as hard to get half as far: Race, work ethic, and america's deserving poor. *American Journal of Political Science 57*(2), 342–356.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies 83*(2), 587–628.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *The Quarterly Journal of Economics 132*(3), 1219–1260.

Greenwald, A. G., M. R. Banaji, and B. A. Nosek (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology 108*(4), 553–561.

Greenwald, A. G. and L. H. Krieger (2006). Implicit bias: Scientific foundations. *California Law Review 94*(4), 945.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology 74*(6), 1464–1480.

Hodson, G., J. F. Dovidio, and S. L. Gaertner (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin 28*(4), 460–471.

Hsee, C. K., G. F. Loewenstein, S. Blount, and M. H. Bazerman (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin 125*(5), 576–590.

Hsee, C. K. and J. Zhang (2010). General evaluability theory. *Perspectives on Psychological Science 5*(4), 343–355.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D. and S. Frederick (2005). A model of heuristic judgment. *The Cambridge handbook of thinking and reasoning*, 267–294.

Kőszegi, B. and A. Szeidl (2012). A model of focusing in economic choice. *The Quarterly Journal of Economics 128*(1), 53–104.

Manzini, P. and M. Mariotti (2007). Sequentially rationalizable choice. *American Economic Review 97*(5), 1824–1839.

Manzini, P. and M. Mariotti (2012). Choice by lexicographic semiorders. *Theoretical Economics 7*(1), 1–23.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012). Revealed attention. *American Economic Review 102*(5), 2183–2205.

Müller-Trede, J., S. Sher, and C. R. M. McKenzie (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision 2*(4), 280–305.

Mummolo, J. and E. Peterson (2018). Demand effects in survey experiments: An empirical assessment. *American Political Science Review 113*(2), 517–529.

Norton, M. I., J. A. Vandello, and J. M. Darley (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology 87*(6), 817–831.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology 105*(2), 171–192.

Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–61.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012). Spontaneous giving and calculated greed. *Nature 489*(7416), 427–430.

Regenwetter, M., J. Dana, and C. P. Davis-Stober (2011). Transitivity of preferences. *Psychological Review 118*(1), 42.

Ridout, S. (2021). Choosing for the right reasons. *Working paper*.

Simonsohn, U. (2016). DataColada[51] Greg vs. Jamal: Why Didn't Bertrand and Mullainathan (2004) Replicate? *https://datacolada.org/51*, *accessed 2022-05-01*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics 50*(3), 665–690.

Snyder, M. L., R. E. Kleck, A. Strenta, and S. J. Mentzer (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of personality and social psychology 37*(12), 2297.

Uhlmann, E. and G. L. Cohen (2005). Constructed criteria: redefining merit to justify discrimination. *Psychological Science 16*(6), 474–80.

Woodford, M. (2012). Inattentive valuation and reference-dependent choice. *Unpublished Manuscript, Columbia University*.

# A Appendix

## A.1 Proof of Theorem 1

We first express the rationalizability problem as a matrix of inequalities. Each inequality in the dataset can written as:

$$v(\boldsymbol{x}^j) + \sum \boldsymbol{x}_i^j \kappa_i \theta_i(\delta(\boldsymbol{x}^j, \boldsymbol{z}^j)) \begin{array}{c} > \\ \geq \end{array} v(\boldsymbol{x}'^j) + \sum \boldsymbol{x}_i'^j \kappa_i \theta_i(\delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j)).$$

The unobserved functions, $v(\cdot)$ and $\theta_i(\cdot)$, can be written as vectors $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|}$ and $\boldsymbol{\theta} \in \mathbb{R}^{n|\Delta|}$, with elements $v_x = v(\boldsymbol{x})$, and $\theta_{i,\delta} = \theta_i(\delta)$. We will express rationalizability of the dataset with two matrix inequalities: $[\,\hat{P}\ \hat{X}\,]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] \gg 0$ representing the $\bar{m}$ strict inequalities, and $[\,\bar{P}\ \bar{X}\,]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] \geq 0$ representing the $m - \bar{m}$ weak inequalities. The matrix $P = \left[\begin{smallmatrix}\hat{P}\\ \bar{P}\end{smallmatrix}\right] \in \mathbb{Z}^{m \times |\mathcal{X}|}$ holds the coefficients on $\boldsymbol{v}$, with entries:

$$P_{\underbrace{j}_{\substack{\text{row}\\ j\in 1,\ldots,m}},\underbrace{\boldsymbol{x}}_{\substack{\text{column}\\ x\in\mathcal{X}}}} = \underbrace{\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\}}_{\text{LHS of inequality}} - \underbrace{\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}}_{\text{RHS of inequality}}.$$

The matrix $X = \left[\begin{smallmatrix}\hat{X}\\ \bar{X}\end{smallmatrix}\right] \in \mathbb{Z}^{m \times n|\mathcal{X}|}$ holds the coefficients on $\boldsymbol{\theta}$, with entries:

$$X_{\underbrace{j}_{\substack{\text{row}\\ j\in 1,\ldots,m}},\underbrace{i\delta}_{\substack{\text{column}\\ i\in 1,\ldots,n\\ \delta\in\Delta}}} = x_i^j \kappa_i \underbrace{\mathbb{1}\{\delta = \delta(\boldsymbol{x}^j, \boldsymbol{z}^j)\}}_{\substack{=1 \text{ if LHS of inequality } j\\ \text{has comparison } \delta}} - x_i'^j \kappa_i \underbrace{\mathbb{1}\{\delta = \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j)\}}_{\substack{=1 \text{ if RHS of inequality } j\\ \text{has comparison } \delta}}.$$

Finally matrix $Q \in \mathbb{Z}^{n|\Delta|^2 \times n|\Delta|}$ holds coefficients on $\boldsymbol{\theta}$ which encode Assumption 1. Each row represents a combination of an attribute $k$ and two comparisons $\bar{\delta}, \bar{\delta}'$, and has non-zero entries only if $\bar{\delta} \sqsupseteq_k \bar{\delta}'$:

$$Q_{\underbrace{k\bar{\delta}\bar{\delta}'}_{\substack{\text{row}\\ k\in\{1,\ldots,n\}\\ \bar{\delta},\bar{\delta}'\in\Delta}},\underbrace{i\delta}_{\substack{\text{column}\\ i\in\{1,\ldots,n\}\\ \delta\in\Delta}}} = \mathbb{1}\{\underbrace{(i=k)}_{\substack{\text{if column}\\ \text{corresponds to } k}} \wedge \underbrace{(\bar{\delta} \sqsupseteq_i \bar{\delta}')}_{\substack{\text{if opacity}\\ \text{dominance}}}\}(\underbrace{\mathbb{1}\{\delta = \bar{\delta}\}}_{\substack{\text{if column}\\ \text{corresponds to } \bar{\delta}}} - \underbrace{\mathbb{1}\{\delta = \bar{\delta}'\}}_{\substack{\text{if column}\\ \text{corresponds to } \bar{\delta}'}}$$

Rationalizability can then be expressed in the following condition:

**Condition 1.** *There exists a real-valued vector* $\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right]$ *satisfying*

$$\begin{bmatrix}\hat{P} & \hat{X}\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{bmatrix} \gg \mathbf{0} \quad \textit{(all positive)}$$
$$\begin{bmatrix}\bar{P} & \bar{X}\\ 0 & Q\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{bmatrix} \geq \mathbf{0} \quad \textit{(all non-negative).}$$

Applying Motzkin's Rational Transposition Theorem (Border (2013)) to this set of linear inequalities implies that Condition 1 will be true if and only if another condition, Condition 2, is false. Intuitively Condition 2 expresses that a non-negative combination of the rows of the matrix $\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]$ can be summed to make a row of zeroes.

**Condition 2.** *There exist integer-valued vectors* $\hat{\boldsymbol{p}} \in \mathbb{Z}^{\bar{m}}$, $\bar{\boldsymbol{p}} \in \mathbb{Z}^{m-\bar{m}}$, $\boldsymbol{q} \in \mathbb{Z}^{n|\Delta|}$ *(with* $\boldsymbol{p} \equiv \left[\begin{smallmatrix}\hat{p}\\ \bar{p}\end{smallmatrix}\right]$*), satisfying:*

$$\hat{\boldsymbol{p}}^T\begin{bmatrix}\hat{P} & \hat{X}\end{bmatrix} + \bar{\boldsymbol{p}}^T\begin{bmatrix}\bar{P} & \bar{X}\end{bmatrix} + \boldsymbol{q}^T\begin{bmatrix}\mathbf{0} & Q\end{bmatrix} = \begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix}\begin{bmatrix}P & X\\ \mathbf{0} & Q\end{bmatrix} = \mathbf{0}^T,$$

$$\hat{\boldsymbol{p}} > 0 \quad \textit{(all non-negative, at least one positive)}$$
$$\bar{\boldsymbol{p}} \geq 0 \quad \textit{(all non-negative)}$$
$$\boldsymbol{q} \geq 0 \quad \textit{(all non-negative).}$$

Loosely speaking, implicit preferences $\boldsymbol{\kappa}$ can rationalize a dataset if and only if there is no combination of inequalities on $\boldsymbol{v}$ and $\boldsymbol{\theta}$ (i.e. rows in $\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]$), which exactly cancel, because that would lead to a contradiction.[42] It remains to show that Condition 2 is equivalent to the condition given in the theorem:

**Condition 3.** *There exists a cyclical selection* $\boldsymbol{s} \in \mathbb{N}^m$ *in which, for every* $\kappa_i = 1$*, losses opacity dominate wins, and for every* $\kappa_i = -1$*, wins opacity dominate losses.*

**Proof that condition 3 implies condition 2.** We will construct the two Motzkin vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, from the selection and the matching:

$$\forall j \in \{1, \ldots, m\}, \qquad p_j = s_j$$
$$\forall i \in \{1, \ldots, n\}, \delta, \delta' \in \Delta, \quad q_{i\delta\delta'} = M_{i,\delta,\delta'}$$

We can verify that $\hat{\boldsymbol{p}} > 0$, $\bar{\boldsymbol{p}} \geq 0$, from the definition of a cyclical selection, and $\boldsymbol{q} \geq 0$ from the definition of a matching.

[42] Condition 1 implies $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] > \mathbf{0}$, condition 2 implies $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] = \mathbf{0}$.

For any element of the vector $\boldsymbol{p}^T P \in \mathbb{Z}^{|\mathcal{X}|}$, we can write:

$$\sum_{j=1}^{m} p_j P_{j,\boldsymbol{x}} = \sum_{j=1}^{m} s_j P_{j,\boldsymbol{x}} = \underbrace{\sum_{j:\boldsymbol{x}^j=\boldsymbol{x}} s_j}_{\substack{\boldsymbol{x} \text{ appears} \\ \text{on LHS of} \\ \text{inequality } j}} - \underbrace{\sum_{j:\boldsymbol{x}'^j=\boldsymbol{x}} s_j}_{\substack{\boldsymbol{x} \text{ appears} \\ \text{on RHS of} \\ \text{inequality } j}} = 0.$$

Where the last step follows from the definition of a cyclical selection: each bundle $\boldsymbol{x}$ must appear equally often on the left-hand and right-hand side. Thus $\boldsymbol{p}^T P = \mathbf{0}$.

We next show that $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} = \mathbf{0}$. An element of the vector $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} \in \mathbb{Z}^{n|\Delta|}$, indexed by $(i\delta)$, can be expressed as:

$$\underbrace{\sum_{j=1}^{m} p_j X_{j,i\delta}}_{\text{elements of } X \text{ selected by } \boldsymbol{p}} + \underbrace{\sum_{k=1}^{n} \sum_{\bar{\delta}\in\Delta} \sum_{\bar{\delta}'\in\Delta} q_{k\bar{\delta}\bar{\delta}'} Q_{k\bar{\delta}\bar{\delta}',i\delta}}_{\text{elements of Q selected by } \boldsymbol{q}}.$$

Using the definitions of $X$ and $Q$ we can write this as:

$$\underbrace{\sum_{j:\delta(x^j,z^j)=\delta} p_j x_i^j \kappa_i - \sum_{j:\delta(x'^j,z'^j)=\delta} p_j x_i'^j \kappa_i}_{\text{inequalities on } \boldsymbol{\theta} \text{ from selection}} + \underbrace{\sum_{\bar{\delta}':\delta\sqsupseteq_i\bar{\delta}'} q_{i,\delta,\bar{\delta}'}}_{\substack{\text{where } \delta \\ \text{dominates}}} - \underbrace{\sum_{\bar{\delta}:\bar{\delta}\sqsupseteq_i\delta} q_{i,\bar{\delta},\delta}}_{\substack{\text{where } \delta \\ \text{is dominated}}} \tag{2}$$

Given $\boldsymbol{p} = \boldsymbol{s}$ the first two terms will be equal to the score for that combination of $i$ and $\delta$:

$$\sum_{j:\delta(x^j,z^j)=\delta} s_j x_i^j \kappa_i - \sum_{j:\delta(x'^j,z'^j)=\delta} s_j x_i'^j \kappa_i = \kappa_i c_{i,\delta}.$$

We can then take the last two terms of (2), using $q_{i\delta\delta'} = M_{i,\delta,\delta'}$:

$$\begin{aligned} \sum_{\bar{\delta}':\delta\sqsupseteq_i\bar{\delta}'} M_{i,\delta,\bar{\delta}'} - \sum_{\bar{\delta}:\bar{\delta}\sqsupseteq_i\delta} M_{i,\bar{\delta},\delta} &= \begin{cases} -c_{i,\delta} & , \kappa_i = 1 \quad \text{(when losses dominate wins)} \\ c_{i,\delta} & , \kappa_i = -1 \quad \text{(when wins dominate losses)} \end{cases} \\ &= -\kappa_i c_{i,\delta} \end{aligned}$$

Combined with the prior step we thus have $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} P & X \\ 0 & Q \end{bmatrix} = \mathbf{0}$, establishing Condition 2.

**Proof that condition 2 implies condition 3.** We construct our selection vector $\boldsymbol{s}$ and a set of matrices $M_i$ for $i = \{1, \ldots, n\}$ as:

$$\forall j \in \{1, \ldots, m\}, \qquad s_j = p_j$$
$$\forall i \in \{1, \ldots, n\}, \delta, \delta' \in \Delta, \quad M_{i,\delta,\delta'} = \begin{cases} q_{i\delta\delta'} & , \delta \sqsupseteq_i \delta' \\ 0 & , \text{otherwise.} \end{cases}$$

We can verify that $s_j \geq 0$ and $M_{i,\delta,\delta'} \geq 0$ because $\bar{\boldsymbol{p}}, \boldsymbol{q} \geq 0$, and that $s_j > 0$ for at least one $j \leq \bar{m}$ because $\hat{\boldsymbol{p}} > 0$. To confirm that $\boldsymbol{s}$ is a cyclical selection we need to show that

$$\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} = \sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\},$$

which follows because $\boldsymbol{p}'P = \boldsymbol{0}$, and:

$$\begin{aligned}(\boldsymbol{p}'P)_{\boldsymbol{x}} &= \sum_{j=1}^{m} p_i P_{j,\boldsymbol{x}} \\ &= \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} - \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}.\end{aligned}$$

We finally verify that for each $i$ with $\kappa_i = 1$ the losses opacity dominate the wins, meaning the matrix $M_i$ satisfies the conditions of Definition 4 (the same argument will show that, when $\kappa_i = -1$, the wins opacity dominate the losses).

1. Matches obey dominance: $\forall \delta, \delta' \in \Delta,\ (M_{i,\delta,\delta'} > 0) \implies (\delta \sqsupseteq_i \delta')$. This immediately follows from our construction of $M$ from $\boldsymbol{q}$ above.
2. All scores are accounted for, i.e. for every $\delta \in \Delta$:

$$\begin{aligned}\underbrace{\sum_{\bar{\delta}' \in \Delta} M_{i,\delta,\bar{\delta}'}}_{\delta \text{ dominating}} - \underbrace{\sum_{\bar{\delta} \in \Delta} M_{i,\bar{\delta},\delta}}_{\delta \text{ dominated}} &= \sum_{\bar{\delta}' : \delta \sqsupseteq_i \bar{\delta}'} q_{i,\delta,\bar{\delta}'} - \sum_{\bar{\delta} : \bar{\delta} \sqsupseteq_i \delta} q_{i,\bar{\delta},\delta} && \text{(from construction of } M\text{)} \\ &= (\boldsymbol{q}^T Q)_i && \text{(from definition of } Q\text{)} \\ &= -(\boldsymbol{p}^T X)_i && \text{(from condition 2)} \\ &= -\sum_{j : \delta(\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} p_j x_i^j + \sum_{j : \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} p_j x_i'^j && \text{(from definition of } X \text{ with } \kappa_i = 1\text{)} \\ &= -\sum_{j : \delta(\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} s_j x_i^j + \sum_{j : \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} s_j x_i'^j && \text{(from construction of } \boldsymbol{s}\text{)} \\ &= -c_{i,\delta} && \text{(from definition of score)}\end{aligned}$$

# B Web appendix

## B.1 Derivations for Section 3

All of our examples can be analyzed by partitioning the set of attributes into three disjoint and collectively exhaustive "groups," $A, B, C$, each containing at least one attribute, where all attributes within a group are perfectly correlated, so we can represent them using three grouped attributes, $\boldsymbol{x} = (x_A, x_B, x_C)$.[43]

Since attributes are perfectly correlated within dimensions, they will have identical differences in a given comparison (e.g. we have $\delta_i = \delta_j, \forall i, j \in A$). Identical differences do not translate into identical opacities (the opacity function can vary in magnitude across attributes) but all opacity dominance relationships will be shared, so for example $\theta_i(\boldsymbol{\delta}) \geq \theta_i(\boldsymbol{\delta}') \Leftrightarrow \theta_j(\boldsymbol{\delta}) \geq \theta_j(\boldsymbol{\delta}'), \forall i, j \in A$. Therefore, rather than laboriously write out conditions and matrices for all $n$ attributes, we can conduct all our analysis using the three grouped attributes $x_A, x_B, x_C$, where $x_A \kappa_A \theta_A(\boldsymbol{\delta}) := \sum_{i \in A} x_i \kappa_i \theta_i(\boldsymbol{\delta})$. Implications that we derive on a grouped attribute will imply a disjunction over all attributes within the group (essentially, because we do not know which attribute(s) within a group are responsible for the observed behavior), namely:

$$(x_A \kappa_A = 1) \Rightarrow \left( \bigvee_{i \in A} x_i \kappa_i = 1 \right).$$

**Applying the Theorem** The proof of Theorem 1 shows how to represent a dataset and opacity dominance relationship in terms of an $X$ and $Q$ matrix, and use them to evaluate whether a given $\boldsymbol{\kappa}$ can be consistent with the data. For each example, we construct its $X$ matrix by counting up wins and losses for each attribute and comparison $\boldsymbol{\delta}$. Here we make use of Assumption 2, which implies $\delta(\boldsymbol{x}, \boldsymbol{z}) = |\boldsymbol{x} - \boldsymbol{z}|$. We construct the $Q$ matrix by assembling all opacity relationships that can be derived for the set of comparisons we observe. We include rows corresponding to each Dilution relationship (Assumption 3) and rows corresponding to the Dominance of attribute $k$ assumption (Assumption 4).

We can use a number of shortcuts to simplify the analysis. First, every cyclical selection must put the same weight on each inequality in a given cycle (sequence of inequalities that starts and ends with the same target bundle $\boldsymbol{x}$, such as an intransitive choice cycle or a scissor). As a result, for our analysis we can sum the rows of $X$ that correspond to a single cycle, collapsing it down to a single row. Each entry in the row equals the sum of wins minus

[43] So, $A \cup B \cup C = 1, \ldots, n$; $A \neq \emptyset, B \neq \emptyset, C \neq \emptyset$; and $A \cap B = A \cap C = B \cap C = \emptyset$. For example, if $A = \{1, 2, 3\}$ we might have $x_A = -1 \Leftrightarrow (x_1, x_2, x_3) = (-1, 1, -1)$ and $x_A = 1 \Leftrightarrow (x_1, x_2, x_3) = (1, -1, 1)$.

losses (i.e., the score $c_{i,\delta}$), multiplied by $\kappa_i$.

Second, we can ignore columns in $X$ and $Q$ where all entries in $X$ are zero since they will not restrict any $\boldsymbol{\kappa}$.[44]

Third, we eliminate rows from $Q$ that do not restrict any observed cyclical selection (that is, there is no cyclical selection in $X$ with nonzero entries in both of the columns that are restricted by this row of $Q$). We label these simplified matrices $X^*$ and $Q^*$. Figure 8 presents the matrix representation of each example.

Applying the theorem requires asking the question: for a candidate $\boldsymbol{\kappa}$ vector, do there exist vectors $\boldsymbol{p}, \boldsymbol{q}$ such that $\left[\begin{smallmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{smallmatrix}\right] \left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right] = \mathbf{0}$? Our approach will be to write out the terms of $\left[\begin{smallmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{smallmatrix}\right] \left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right]$ and ask for what values of $\boldsymbol{\kappa}$ at least one term must be nonzero, meaning that $\boldsymbol{\kappa}$ is not ruled out by Theorem 1.

**Right triangle** Let $A = \{i : x_i^1 \neq x_i^2\}$, $B = \{i : x_i^2 \neq x_i^3\}$, $C = \{i : x_i^1 = x_i^3\}$. So $A$ is the set of attributes that vary in $\boldsymbol{\delta}^1$, $B$ is the set that vary in $\boldsymbol{\delta}^2$, $A \cup B$ is the set that vary in $\boldsymbol{\delta}^3$ (the "diagonal"), and $C$ is the set that do not vary in any $\boldsymbol{\delta}$. Orthogonality of $\boldsymbol{\delta}^1$ and $\boldsymbol{\delta}^2$ implies $A, B, C$ are disjoint and collectively exhaustive. Collapsing the set of attributes down to these three groups, we have:

$$\boldsymbol{\delta}^1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\delta}^2 = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \text{ and } \boldsymbol{\delta}^3 = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}.$$

The choice inequalities are $u(\boldsymbol{x}^1, \boldsymbol{x}^2) > u(\boldsymbol{x}^2, \boldsymbol{x}^1)$, $u(\boldsymbol{x}^2, \boldsymbol{x}^3) > u(\boldsymbol{x}^3, \boldsymbol{x}^2)$, and $u(\boldsymbol{x}^3, \boldsymbol{x}^1) > u(\boldsymbol{x}^1, \boldsymbol{x}^3)$. To construct the $X$ matrix we need to count the wins and losses for each $i, \boldsymbol{\delta}$ pair. Beginning with the first inequality, for each attribute $i$, the left-hand side gives us a win if $x_i^1 = 1$ and a loss otherwise. The right-hand side gives us a loss if $x_i^2 = 1$ and a win otherwise. From the conditions defining the right triangle, we know that $x_A^1 = -x_A^2 = -x_A^3$ while $x_B^1 = x_B^2 = -x_B^3$ and $x_C^1 = x_C^2 = x_C^3$. We work through each inequality in turn.

Inequality 1 gives us two wins for $A, \boldsymbol{\delta}^1$ if $x_A^1 = 1$ and two losses if $x_A^1 = -1$. Thus the entry in column $A, \boldsymbol{\delta}^1$ equals $2\kappa_A x_A^1$, which in turn equals $-2\kappa_A x_A^3$ by definition of $x^1$ and $x^3$. All other attributes do not vary and so have zero wins and losses for comparison $\boldsymbol{\delta}^1$.

Inequality 2 gives us two wins for $B, \boldsymbol{\delta}^2$ if $x_B^2 = 1$ and two losses if $x_B^2 = -1$. Thus the entry in column $B, \boldsymbol{\delta}^2$ equals $2\kappa_B x_B^2$, which in turn equals $-2\kappa_B x_B^3$ by definition of $x^2$ and $x^3$. All other attributes do not vary and so have zero wins and losses for comparison $\boldsymbol{\delta}^2$.

Inequality 3 gives us two wins for $A, \boldsymbol{\delta}^3$ if $x_A^3 = 1$ and two losses if $x_A^3 = -1$. Thus the entry in column $A, \boldsymbol{\delta}^3$ equals $2\kappa_A x_A^3$. Inequality 3 gives us two wins for $B, \boldsymbol{\delta}^3$ if $x_B^3 = 1$ and two losses if $x_B^3 = -1$. Thus the entry in column $B, \boldsymbol{\delta}^3$ equals $2\kappa_B x_B^3$. All other attributes

[44]Note that if all columns equal zero, the model is falsified, see Corollary 2.

$$X^* = \begin{array}{c} \\ \text{Right triangle 1} \\ \text{Right triangle 2} \\ \text{Figure 8} \\ \text{Convex scissor 1} \\ \text{Convex scissor 2} \end{array} \begin{array}{c} \begin{array}{cccccccccc} A,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & A,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & A,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \end{array} \\ \begin{bmatrix} -2\kappa_A x_A^3 & 0 & 2\kappa_A x_A^3 & 0 & -2\kappa_B x_B^3 & 2\kappa_B x_B^3 & 0 & 0 & 0 \\ -2\kappa_A \bar{x}_A^3 & 0 & 2\kappa_A \bar{x}_A^3 & 0 & -2\kappa_B \bar{x}_B^3 & 2\kappa_B \bar{x}_B^3 & 0 & 0 & 0 \\ -4\kappa_A x_A^4 & 0 & 4\kappa_A x_A^4 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\kappa_A x_A \Upsilon & 0 & \kappa_A x_A \Upsilon & -\kappa_B x_B \Upsilon & 0 & \kappa_B x_B \Upsilon & -\kappa_C x_C \Upsilon & 0 & \kappa_C x_C \Upsilon \\ -\kappa_A \bar{x}_A \bar{\Upsilon} & 0 & \kappa_A \bar{x}_A \bar{\Upsilon} & -\kappa_B \bar{x}_B \bar{\Upsilon} & 0 & \kappa_B \bar{x}_B \bar{\Upsilon} & -\kappa_C \bar{x}_C \bar{\Upsilon} & 0 & \kappa_C \bar{x}_C \bar{\Upsilon} \end{bmatrix} \end{array}$$

$$Q^* = \begin{array}{c} \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \sqsupseteq_A \left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] \\ \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \sqsupseteq_B \left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] \\ \left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] \sqsupseteq_C \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \\ \text{Dominance} \end{array} \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & \Theta & 0 & -\Theta & 0 & 0 & 0 \end{bmatrix}$$

$$X^* = \begin{array}{c} \\ \text{Equilateral triangle} \\ \text{Non-convex scissor} \\ \text{Falsification} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & 1,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & 1,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & 1,\left[\begin{smallmatrix}2\\0\\2\end{smallmatrix}\right] & 2,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & 2,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & 2,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & 2,\left[\begin{smallmatrix}0\\2\\2\end{smallmatrix}\right] & 3,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & 3,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & 3,\left[\begin{smallmatrix}0\\2\\2\end{smallmatrix}\right] & 3,\left[\begin{smallmatrix}2\\0\\2\end{smallmatrix}\right] \end{array} \\ \begin{bmatrix} 0 & 0 & -2\kappa_1 & 2\kappa_1 & 0 & 0 & 2\kappa_2 & -2\kappa_2 & 0 & 0 & -2\kappa_3 & 2\kappa_3 \\ -\kappa_1 & \kappa_1 & 0 & 0 & -\kappa_2 & \kappa_2 & 0 & 0 & \kappa_3 & -\kappa_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$Q^* = \text{empty}$$

**Top panel** corresponds to the main Corollaries. (1) Columns are labeled by attribute group ($i \in \{A, B, C\}$), and $\boldsymbol{\delta} \in \{0, 2\}^3$. (2) Rows correspond to $\kappa_i$ multiplied by the sum of wins and losses for that example (i.e., equal to $\kappa_i c_{i,\boldsymbol{\delta}}$). (3) $Q^*$ includes only rows that restrict at least one row of $X^*$. (4) For scissors, $\Upsilon \in \{-1, 1\}$ equals the sign of the evaluation change: $\Upsilon = sgn(y^2 - y^1)$. (5) $\Theta \in \{-1, 0, 1\}$ captures the sign of the Dominance of attribute $k$ assumption (Assumption 4). $\Theta = 0$ if the assumption does not apply, $\Theta = 1$ if opacity is higher for shared attributes ($k$ is shared), $\Theta = -1$ if opacity is higher for non-shared ($k$ is non-shared).
**Bottom panel** shows the additional examples given in Figure 3. Attributes $1, 2, 3$ correspond to horizontal, vertical, depth. Falsification gives an example of a cyclical selection where $c_{i,\boldsymbol{\delta}} = 0, \forall i, \boldsymbol{\delta}$. $Q^*$ is empty as $\sqsupseteq$ does not restrict any row of $X^*$.

Figure 8: Matrix representation of corollaries and examples from Section 3

do not vary and so have zero wins and losses for comparison $\boldsymbol{\delta}^3$.

Collapsing these entries to a single row we obtain "Right triangle 1" in Figure 8.

With a single cycle (Right triangle 1) we can set $\boldsymbol{p} = 1$ without loss of generality, obtaining (after ignoring columns that equal zero):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} = \begin{bmatrix} -2\kappa_A x_A^3 - q_1 & 2\kappa_A x_A^3 + q_1 & -2\kappa_B x_B^3 - q_2 & 2\kappa_B x_B^3 + q_2 \end{bmatrix},$$

where $q_1$ is the coefficient on the first row of $Q$ and $q_2$ is the coefficient on the second. There exist $q_1, q_2 \geq 0$ such that this vector equals 0 if and only if: $(\kappa_A x_A^3 \leq 0) \wedge (\kappa_B x_B^3 \leq 0)$. Hence, the data can be rationalized if and only if:

$$\left(\kappa_A x_A^3 = 1\right) \vee \left(\kappa_B x_B^3 = 1\right) \Leftrightarrow \bigvee_{\{i : x_i^3 \neq x_i^1\}} (\kappa_i x_i^3 = 1),$$

where the last part follows from the definitions of $A, B, \boldsymbol{x}^1, \boldsymbol{x}^3$.

**Figure 8** Let $A = \{i : x_i^1 \neq x_i^2\}$, $B = \{i : x_i^1 \neq x_i^3\}$, $C = \{i : x_i^1 = x_i^4\}$. So $A$ is the set of attributes that vary in the odd-numbered comparisons, $B$ is the set of additional attributes that varies in the even-numbered but not the odd-numbered comparisons, $A \cup B$ the set that vary in the even-numbered comparisons, and $C$ the set that are shared in all comparisons. By construction, $A, B, C$ are disjoint and collectively exhaustive. We also have:

$$\boldsymbol{\delta}^1 = \boldsymbol{\delta}^3 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \text{ and } \boldsymbol{\delta}^2 = \boldsymbol{\delta}^4 = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}.$$

As for the right triangle, we populate the matrix $X$ by calculating wins and losses for each $i, \boldsymbol{\delta}$ combination. Unlike the right triangle, all comparisons are concentrated on just two $\boldsymbol{\delta}$'s. Following the same proof strategy as for the right triangle, we set $\boldsymbol{p} = 1$ without loss of generality. We obtain (after eliminating zeros):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} = \begin{bmatrix} -4\kappa_A x_A^4 - q_1 & 4\kappa_A x_A^4 + q_1 \end{bmatrix},$$

where $q_1$ is the coefficient on the first row of $Q$. By the same argument as for the right triangle, the data can be rationalized if and only if:

$$\left(\kappa_A x_A^4 = 1\right) \Leftrightarrow \bigvee_{\{i : x_i^3 \neq x_i^4\}} (\kappa_i x_i^4 = 1),$$

where the last part follows from the definitions of $A, \boldsymbol{x}^3, \boldsymbol{x}^4$.

**Parallel right triangles** Let:

$$A = \{i : x_i^1 \neq x_i^2\} = \{i : \bar{x}_i^2 \neq \bar{x}_i^3\}$$
$$B = \{i : x_i^2 \neq x_i^3\} = \{i : \bar{x}_i^1 \neq \bar{x}_i^2\}$$
$$C = \{i : x_i^1 = x_i^3\} = \{i : \bar{x}_i^1 = \bar{x}_i^3\}.$$

In words, $A$ is the set of attributes that are not shared in $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$ and not shared in $\{\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3\}$, $B$ is the set of attributes that are not shared in $\{\boldsymbol{x}^2, \boldsymbol{x}^3\}$ and not shared in $\{\bar{\boldsymbol{x}}^1, \bar{\boldsymbol{x}}^2\}$, and $C$ is the set of attributes that do not vary within any comparison.[45] By construction, $A, B$, and $C$ are disjoint and collectively exhaustive.

We populate the second triangle's row in $X$ by calculating the wins and losses for each $i, \bar{\boldsymbol{\delta}}$ combination. As for right triangle 1 we exploit the definitions of the triangle and the sets $A, B, C$ to express them in terms of $\bar{\boldsymbol{x}}^3$.

When the dataset consists of a pair of parallel right triangles, a cyclical selection consists of $p_1 \geq 0$ copies of the first and $p_2 \geq 0$ copies of the second, giving us (ignoring zero elements):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} = \begin{bmatrix} -W_A & W_A & -W_B & W_B \end{bmatrix}$$
$$W_A = 2\kappa_A(p_1 x_A^3 + p_2 \bar{x}_A^3) + q_1 = 2\kappa_A(p_1 - p_2)x_A^3 + q_1$$
$$W_2 = 2\kappa_B(p_1 x_B^3 + p_2 \bar{x}_B^3) + q_2 = 2\kappa_B(p_1 + p_2)x_B^3 + q_2,$$

where $q_1$ is the coefficient on the first row of $Q$ and $q_2$ is the coefficient on the second. The second steps use the fact that $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.[46] Thus for a given $p_1, p_2$, the data can be rationalized if and only if:

$$\left(\kappa_A(p_1 - p_2)x_A^3 > 0\right) \vee \left(\kappa_B(p_1 + p_2)x_B^3 > 0\right).$$

When $p_1 = p_2$ (i.e. the selection contains an equal number of each cycle), the disjunction collapses to $(\kappa_B(p_1+p_2)x_B^3 = 1)$, so this condition must hold for the data to be rationalizable. Once this condition holds, the data can be rationalized for all $p_1, p_2$, so no further restrictions

[45] Note that while $C$ attributes do not vary within any comparisons, they might differ between the two triangles, that is it could be that $x_C \neq \bar{x}_C$. See e.g. the second example given with Corollary 5 where all candidates in triangle 1 have a Harvard degree while all candidates in triangle 2 have a Yale degree.

[46] The definition of the parallel right triangle, condition (1) ($\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$) allows us to pin down the values of the non-shared attributes in these comparisons (set $B$): $(\boldsymbol{x}_B^2 = \bar{\boldsymbol{x}}_B^1) \wedge (\boldsymbol{x}_B^3 = \bar{\boldsymbol{x}}_B^2)$ (to see this note that if $\boldsymbol{x}_B^2 - \boldsymbol{x}_B^3 = 2$, it must be that $\boldsymbol{x}_B^2 = 1$ and $\boldsymbol{x}_B^3 = -1$). Similarly, condition (2) ($\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$) allows us to pin down the values of the non-shared attributes in these comparisons (set $A$): $(\boldsymbol{x}_A^1 = -\bar{\boldsymbol{x}}_A^2) \wedge (\boldsymbol{x}_A^2 = -\bar{\boldsymbol{x}}_A^3)$. Finally, the definitions of $A, B$, and $C$ imply $x_A^3 = x_A^2 = -x_A^1, x_B^3 = -x_B^2 = -x_B^1, \bar{x}_A^3 = -\bar{x}_A^2 = -\bar{x}_A^1$ and $\bar{x}_B^3 = \bar{x}_B^2 = -\bar{x}_B^1$. Substitution yields $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.

are obtained by considering other $\boldsymbol{p}$s. Finally, using the definition of set $B$ we obtain the result, that a pair of parallel right triangles implies:

$$\bigvee_{i:x_i^3 \neq x_i^2} (x_i^3 \kappa_i = 1).$$

**Convex scissor without and with Dominance of attribute $k$.** Let $A = \{i : x_i \neq z_i^1\}$, $B = \{i : z_i^1 \neq z_i^2\}$, $C = \{i : x_i = z_i^2\}$. So $A$ is the set of attributes that vary in the first comparison, $B$ is the set of additional attributes that varies in the second comparison but not the first, $A \cup B$ the full set that vary in the second comparison, and $C$ the set that do not vary within either comparison. By construction, $A, B, C$ are disjoint and collectively exhaustive.

We construct the scissor's row in the $X$ matrix by counting losses and wins in the scissor's single inequality. If $y^2 > y^1$ we have $u(\boldsymbol{x}, \boldsymbol{z}^2) > u(\boldsymbol{x}, \boldsymbol{z}^1)$. The left-hand side corresponds to $\boldsymbol{\delta}^2 = \left[\begin{smallmatrix} 2 \\ 2 \\ 0 \end{smallmatrix}\right]$, giving us a win in column $i$, $\left[\begin{smallmatrix} 2 \\ 2 \\ 0 \end{smallmatrix}\right]$ if $x_i = 1$ and a loss otherwise. The right-hand side corresponds to $\boldsymbol{\delta}^1 = \left[\begin{smallmatrix} 2 \\ 0 \\ 0 \end{smallmatrix}\right]$, giving us a loss in column $i$, $\left[\begin{smallmatrix} 2 \\ 0 \\ 0 \end{smallmatrix}\right]$ if $x_i = 1$ and a win otherwise. If $y^2 < y^1$ then the left- and right-hand sides of the inequalities are switched. Thus, defining $\Upsilon = sgn(y^2 - y^1)$, we enter $\kappa_i x_i \Upsilon$ in the columns associated with $\boldsymbol{\delta}^2$, and $-\kappa_i x_i \Upsilon$ in the columns associated with $\boldsymbol{\delta}^1$. Thus, we obtain (setting $p_1 = 1$ and ignoring zeros):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} = \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C \end{bmatrix}$$
$$W_A = \kappa_A x_A \Upsilon + q_1$$
$$W_B = \kappa_B x_B \Upsilon - \Theta q_4$$
$$W_C = \kappa_C x_C \Upsilon - q_3.$$

where $q_1$ and $q_3$ are the coefficients on the first and third rows of $Q$, which correspond to the main Opacity assumption (Assumption 1), while $q_4$ is the coefficient on $Q$'s fourth row which captures the Dominance of attribute $k$ assumption (Assumption 4). $\Theta$ embeds the assumption: $\Theta = 0$ if the assumption does not apply, $\Theta = 1$ if opacity is higher for shared attributes ($k$ is shared), $\Theta = -1$ if opacity is higher for non-shared attributes ($k$ is non-shared).

When Assumption 4 does not hold, $\Theta = 0$. The data can be rationalized if and only if there is no $\boldsymbol{q} \geq 0$ such that $W_A = W_B = W_C = 0$, i.e. if and only if:

$$(\kappa_A x_A \Upsilon = 1) \vee (\kappa_B x_B \Upsilon \neq 0) \vee (\kappa_C x_C \Upsilon = -1).$$

Expanding this expression using the definitions of $A, B, C$, and $\Upsilon$ gives the result.

When Assumption 4 holds, $\Theta \neq 0$, so the data can be rationalized if and only if:

$$(\kappa_A x_A \Upsilon = 1) \vee (\kappa_B x_B \Upsilon = -\Theta) \vee (\kappa_C x_C \Upsilon = -1)\,.$$

Expanding this expression using the definitions of $A, B, C, \Upsilon$ and $\Theta$ gives the result.

**Parallel convex scissors without and with Dominance of attribute $k$.** The conditions (1) and (2) imply $\boldsymbol{\delta}^1 = \bar{\boldsymbol{\delta}}^1$ and $\boldsymbol{\delta}^2 = \bar{\boldsymbol{\delta}}^2$. Let:

$$\begin{aligned} A &= \{i : x_i \neq z_i^1\} = \{i : \bar{x}_i \neq \bar{z}_i^1\} \\ B &= \{i : z_i^1 \neq z_i^2\} = \{i : \bar{z}_i^1 \neq \bar{z}_i^2\} \\ C &= \{i : x_i = z_i^2\} = \{i : \bar{x}_i = \bar{z}_i^2\}. \end{aligned}$$

So $A$ is the set of attributes that vary in each scissor's first comparison, $B$ is the set of additional attributes that varies in the second comparisons but not the first (which is nonempty since the second comparisons differ on a superset of attributes), $A \cup B$ the full set that vary in the second comparisons, and $C$ the set that do not vary within any comparison. By construction, $A, B, C$ are disjoint and collectively exhaustive. Since the values of $\boldsymbol{x}, \bar{\boldsymbol{x}}, sgn(y^1 - y^2)$ and $sgn(\bar{y}^1 - \bar{y}^2)$ are unrestricted, there are many possible combinations of parallel convex scissor.

When the dataset consists of a pair of parallel convex scissors, a cyclical selection consists of $p_1 \geq 0$ copies of the first and $p_2 \geq 0$ copies of the second, giving us (ignoring zero elements):

$$\begin{aligned} \begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} &= \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C \end{bmatrix} \\ W_A &= \kappa_A (p_1 x_A \Upsilon + p_2 \bar{x}_A \bar{\Upsilon}) + q_1 \\ W_2 &= \kappa_B (p_1 x_B \Upsilon + p_2 \bar{x}_B \bar{\Upsilon}) - \Theta q_4 \\ W_3 &= \kappa_C (p_1 x_C \Upsilon + p_2 \bar{x}_C \bar{\Upsilon}) - q_3, \end{aligned}$$

where $q_1$ and $q_3$ are the coefficients on the first and third rows of $Q$, which capture the Dilution assumption (Assumption 3), while $q_4$ is the coefficient on $Q$'s fourth row which captures the Dominance of attribute $k$ assumption 4. $\Upsilon = sgn(y^2 - y^1)$ and $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$ capture the direction in which each evaluation changes when the comparator changes. $\Theta$ embeds Assumption $k$ as above.

By a similar argument to the parallel right triangles, the strongest restrictions on $\boldsymbol{\kappa}$ will be obtained when $p_1 = p_2$. This maximizes the number of terms in the disjunction that

become zero and drop out, and by so doing, reveals the set of restrictions that must hold in every selection. In other words, we can without loss of generality consider only the selection consisting of exactly one copy of each scissor ($p_1 = p_2 = 1$).

When Assumption 4 does not hold ($\Theta = 0$) the data can be rationalized if and only if:

$$\left(\kappa_A(x_A\Upsilon + \bar{x}_A\bar{\Upsilon}) = 2\right) \vee \left(\kappa_B(x_B\Upsilon + \bar{x}_B\bar{\Upsilon}) \neq 0\right) \vee \left(\kappa_C(x_C\Upsilon + \bar{x}_C\bar{\Upsilon}) = -2\right).$$

When Assumption 4 holds ($\Theta \neq 0$) the data can be rationalized if and only if:

$$\left(\kappa_A(x_A\Upsilon + \bar{x}_A\bar{\Upsilon}) = 2\right) \vee \left(\kappa_B(x_B\Upsilon + \bar{x}_B\bar{\Upsilon}) = -2\Theta\right) \vee \left(\kappa_C(x_C\Upsilon + \bar{x}_C\bar{\Upsilon}) = -2\right).$$

Expanding the expressions using the definitions of $A, B, C, \Upsilon, \bar{\Upsilon}$ and $\Theta$ gives the results.

Note that in each case, the term corresponding to $i \in \{A, B, C\}$ is eliminated if:

$$x_i\Upsilon = -\bar{x}_i\bar{\Upsilon},$$

that is, if either (i) the second scissor has an opposite realization of $x_i$ but evaluation moves in the same direction, or (ii) the second scissor has an identical realization of $x_i$, but evaluation moves in the opposite direction.

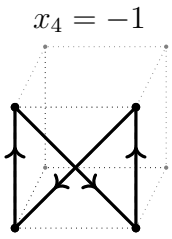
Note that parallel scissors can eliminate the term involving attribute group $B$, so it is possible to unambiguously identify an implicit preference without imposing Assumption 4.

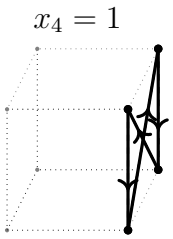
### B.1.1 Corollary 2 is sufficient but not necessary for falsification

Corollary 2 says that to falsify the model it is sufficient but not necessary that the dataset $D$ contains a cyclical selection where losses opacity dominate wins and wins opacity dominate losses. Here we provide an example of a falsification where the condition does not hold.

Consider a Separable Implicit Preferences decision maker (that is, $\sqsupseteq_i$ satisfies Assumptions 1, 2, and 3) choosing between bundles with $n = 4$ attributes. On the left side of Figure 9 we draw two figure-8 cycles, in three dimensions, holding the fourth fixed. The first figure-8 has $x_4 = -1$ and the second figure-8 has $x_4 = 1$. On the right of the diagram we show the matrix representation of the dataset (simplified as in the derivations above). Note that only attribute 2 has nonzero columns in $X^*$.

The first cycle rules out all $\boldsymbol{\kappa}$s with $\kappa_2 \neq 1$, while the second rules out all $\boldsymbol{\kappa}$s with $\kappa_2 \neq -1$, so there exists no comparative utility function $u$ that can rationalize the data. However, no single cyclical selection (weighted combination of rows of $X^*$) can rule out both.




$$X^* = \begin{array}{c} \\ \text{1st Figure-8} \\ \text{2nd Figure-8} \end{array} \begin{array}{c} 2, \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad 2, \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad 2, \begin{bmatrix} 0 \\ 2 \\ 2 \\ 0 \end{bmatrix} \\ \begin{bmatrix} -4\kappa_2 & 4\kappa_2 & 0 \\ 4\kappa_2 & 0 & -4\kappa_2 \end{bmatrix} \end{array}$$

$$Q^* = \begin{array}{c} \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \sqsupseteq_2 \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 2 \\ 2 \\ 0 \end{bmatrix} \sqsupseteq_2 \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \end{array} \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Figure 9: Corollary 2 is sufficient but not necessary for falsification

## B.2 Proofs for Section 4 (Foundations)

In proving some of these results we make use of an additional lemma that we call "Sums and Differences," which we state and prove first. Recall also the definition of the set of shared attributes: $S^{|\boldsymbol{x}-\boldsymbol{z}|} = \{i : |x_i - z_i| = 0\}$.

**Lemma 2** (Sums and Differences)**.** *Suppose we observe two linear combinations of $n$ independent Normal variables ("weights"), with $+1$ or $-1$ coefficients ("attributes"):*

$$\underbrace{\begin{bmatrix} \bar{y}^x \\ \bar{y}^z \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} x_1 & \dots & x_n \\ z_1 & \dots & z_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}}_{\boldsymbol{w}}$$

$$x_i, z_i \in \{-1, 1\}, \boldsymbol{w} = N(0, diag(\sigma_1^2, \dots, \sigma_n^2)),$$

*The Bayesian posterior for unobserved weight $w_i$, given observed $\boldsymbol{y}$ will be:*

$$E[w_i|\boldsymbol{y}] = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2} & , i \in S^{|\boldsymbol{x}-\boldsymbol{z}|} \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2} & , i \notin S^{|\boldsymbol{x}-\boldsymbol{z}|} \end{cases}.$$

In words, the posterior for the weight on each shared attribute depends only on the sum $\bar{y}^x + \bar{y}^z$, and the posterior for the weight on each non-shared attribute depends only on the difference $\bar{y}^x - \bar{y}^z$.

**Proof of Lemma 2** First we assume there exists at least one shared and one non-shared attribute (in other words, $\boldsymbol{x} \neq \boldsymbol{z}$ and $\boldsymbol{x} \neq -\boldsymbol{z}$). Given two multivariate Normals, $\boldsymbol{a}$ and $\boldsymbol{b}$, with covariance matrix: $Var\left[\begin{smallmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{smallmatrix}\right] = \begin{bmatrix} \Sigma_a & \Sigma_{a,b} \\ \Sigma'_{a,b} & \Sigma_b \end{bmatrix}$ we can write the conditional expectation:

$E[\boldsymbol{a}|\boldsymbol{b}] = E[\boldsymbol{a}] + \Sigma_{a,b}\Sigma_b^{-1}(\boldsymbol{b} - E[\boldsymbol{b}])$. In our case this implies:

$$E[\boldsymbol{w}|\boldsymbol{y}] = \Sigma_{w,y}\Sigma_y^{-1}\boldsymbol{y} \tag{3}$$

with components as follows:

$$\Sigma_y = X\Sigma_w X' = \begin{bmatrix} \sum_i x_i^2\sigma_i^2 & \sum_i x_i z_i \sigma_i^2 \\ \sum_i x_i z_i \sigma_i^2 & \sum_i z_i^2 \sigma_i^2 \end{bmatrix} = \begin{bmatrix} \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 & \sum_{i\in S}\sigma_i^2 - \sum_{i\notin S}\sigma_i^2 \\ \sum_{i\in S}\sigma_i^2 - \sum_{i\notin S}\sigma_i^2 & \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 \end{bmatrix}$$

$$\Sigma_y^{-1}\boldsymbol{y} = \frac{1}{4\sum_{i\in S}\sigma_i^2 \sum_{i\notin S}\sigma_i^2} \begin{bmatrix} \left(\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^x + \left(-\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^z \\ \left(-\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^x + \left(\sum_{i\in S} + \sigma_i^2 \sum_{i\notin S}\sigma_i^2\right)\bar{y}^z \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} + \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \\ \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} - \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \end{bmatrix}$$

$$\Sigma_{w,y} = \Sigma_w X' = \begin{bmatrix} x_1\sigma_1^2 & z_1\sigma_1^2 \\ \vdots & \vdots \\ x_n\sigma_n^2 & z_n\sigma_n^2 \end{bmatrix}$$

Thus, given (3), we obtain:

$$E[w_i|\boldsymbol{y}] = \frac{1}{4}\begin{bmatrix} \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} + \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \\ \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} - \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \end{bmatrix} \begin{bmatrix} x_i\sigma_i^2 \\ z_i\sigma_i^2 \end{bmatrix} = \begin{cases} x_i\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\frac{\bar{y}^x+\bar{y}^z}{2} & , i \in S \\ x_i\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\frac{\bar{y}^x-\bar{y}^z}{2} & , i \notin S \end{cases}$$

Where the last step uses $x_i + z_i = 2x_i\mathbf{1}\{i \in S\}$ and $x_i - z_i = 2x_i\mathbf{1}\{i \notin S\}$.

Next we show that the same formula applies to the two special cases we initially ruled out, $\boldsymbol{x} = \boldsymbol{z}$ and $\boldsymbol{x} = -\boldsymbol{z}$. We cannot use equation (3) because $X$ does not have full rank so $\Sigma_y$ is not invertible. If all attributes are shared ($\boldsymbol{x} = \boldsymbol{z}$) we have a Normal updating problem with a single observable, $\bar{y}^x = \bar{y}^z$, and each weight is updated in proportion to its share of the total variance. So, $E[w_i|\boldsymbol{y}] = x_i\frac{\sigma_i^2}{\sum_{j=1}^n\sigma_j^2}\bar{y}^x = x_i\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\frac{\bar{y}^x+\bar{y}^z}{2}$, as in the statement of the Lemma. If instead all attributes are non-shared ($\boldsymbol{x} = -\boldsymbol{z}$) then $\bar{y}^x = -\bar{y}^z$ and we have $E[w_i|\boldsymbol{y}] = x_i\frac{\sigma_i^2}{\sum_{j=1}^n\sigma_j^2}\bar{y}^x = x_i\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\frac{\bar{y}^x-\bar{y}^z}{2}$, once again corresponding to the statement of the Lemma. □

**Proof strategy for Propositions.** Our strategy will be to show that the utility function defined in each foundation can be expressed in a form satisfying Assumption 1, where $\theta_i()$ depends only on $|\boldsymbol{x} - \boldsymbol{z}|$ (Assumption 2) and is increasing as $|\boldsymbol{x} - \boldsymbol{z}|$ becomes more dilute (Assumption 3). Specifically, we will show that in each foundation we can express the utility

function as a special case of (1), where $\theta_i(.)$ takes form that depends on $i$'s status (shared or non-shared):

$$\theta_i(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & ,i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & ,i \notin S. \end{cases}$$

To show that Assumption 3 is satisfied we must show that $\theta_i()$ is weakly increasing as $i$ becomes more dilute. By definition of dilution, $|\boldsymbol{x}'-\boldsymbol{z}'| \sqsupseteq_i |\boldsymbol{x}-\boldsymbol{z}|$ when (a) $i$ does not change status (either $(i \in S^{|\boldsymbol{x}'-\boldsymbol{z}'|}) \wedge (i \in S^{|\boldsymbol{x}-\boldsymbol{z}|})$ or $(i \notin S^{|\boldsymbol{x}'-\boldsymbol{z}'|}) \wedge (i \notin S^{|\boldsymbol{x}-\boldsymbol{z}|})$); and (b) the set of attributes that share status with $i$ grows ($\{j : |\boldsymbol{x}'_j - \boldsymbol{z}'_j| = |\boldsymbol{x}'_i - \boldsymbol{z}'_i|\} \supseteq \{j : |\boldsymbol{x}_j - \boldsymbol{z}_j| = |\boldsymbol{x}_i - \boldsymbol{z}_i|\}$).

Part (a) implies that we can study the properties of $\theta_i^S$ and $\theta_i^N$ separately, since $i$ does not change status in a given dilution. Part (b) implies we need to show that $\theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|)$ weakly increases as the set of shared attributes grows (in a superset sense), and that $\theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|)$ weakly increases as the set of non-shared attributes grows.

**Proof of Proposition 1** First, note that $u^{CP}(\boldsymbol{x},\boldsymbol{z})$ can be rearranged to satisfy Assumption 1 (using the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$):

$$u^{CP}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(\boldsymbol{x},\boldsymbol{z})$$

$$\theta_i(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) = |\lambda_i| & ,i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) = |\lambda_i|(1 - \mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\}) & ,i \notin S. \end{cases}$$

It is easy to see that $\theta_i$ depends only on $(\boldsymbol{x},\boldsymbol{z})$ through $|\boldsymbol{x}-\boldsymbol{z}|$. $\theta_i^S$ is weakly increasing as the set of shared attributes grows since $\theta_i^S$ is a constant. We need to show that $\theta_i^N$ is weakly increasing as the set of non-shared attributes grows. Let $|\boldsymbol{x}'-\boldsymbol{z}'| \sqsupseteq_i |\boldsymbol{x}-\boldsymbol{z}|$. Consider the set of attributes that are shared under $|\boldsymbol{x}-\boldsymbol{z}|$ and become non-shared under $|\boldsymbol{x}'-\boldsymbol{z}'|$, i.e. $D = \{j : (j \in S^{|\boldsymbol{x}-\boldsymbol{z}|}) \wedge (j \notin S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\}$. If all of them are governed by a rule ($\forall j \in D, \lambda_j \neq 0$) then the rule-applying function is unaffected, so $\theta_i^N(|\boldsymbol{x}'-\boldsymbol{z}'|) = \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|)$. If one or more is not governed by a rule ($\exists j \in D : \lambda_j = 0$), then $\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\} = 0$, so $\theta_i^N(|\boldsymbol{x}'-\boldsymbol{z}'|) = |\lambda_i| \geq \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|)$. □

**Proof of Proposition 2** We begin by deriving an explicit solution for the observer's posterior.

**Lemma 3.** *Suppose a naïve observer sees the decision maker choose $\boldsymbol{x}$ from $\{\boldsymbol{x}, \boldsymbol{z}\}$, $\boldsymbol{x} \neq \boldsymbol{z}$. Their posterior over weight $w_i$ can be written as:*

$$E\left[w_i \middle| \sum_{i=1}^{n} x_i w_i > \sum_{i=1}^{n} z_i w_i\right] = \mathbf{1}\{i \notin S\} \frac{x_i \sigma_i^2}{\sqrt{\sum_{j \notin S} \sigma_j^2}} \frac{\phi(0)}{1 - \Phi(0)},$$

*where $\phi$ and $\Phi$ are the standard Normal density and cumulative density functions.*

**Proof of Lemma 3** The expectation of a Normally-distributed variable, $b$, conditioning on another Normal variable, $a$, exceeding some threshold $\bar{a}$ can be written as:

$$E[b|a > \bar{a}] = \mu_b + \frac{\text{Cov}(a,b)}{\sqrt{Var(a)}} \frac{\phi(\frac{\bar{a} - \mu_a}{\sqrt{Var(a)}})}{1 - \Phi(\frac{\bar{a} - \mu_a}{\sqrt{Var(a)}})}.$$

In our model each $w_i$ is Normally distributed, implying the difference in intrinsic utility between $\boldsymbol{x}$ and $\boldsymbol{z}$ will also be Normal, and so given $\boldsymbol{x}$ is chosen over $\boldsymbol{z}$ we have:

$$\begin{aligned} E\left[w_i \middle| \sum_{j=1}^{n} w_j(x_j - z_j) > 0\right] &= E[w_i] + \frac{Cov(w_i, \sum_{j=1}^{n} w_j(x_j - z_j))}{\sqrt{Var(\sum_{j=1}^{n} w_j(x_j - z_j))}} \frac{\phi(0)}{1 - \Phi(0)} \\ &= \frac{(x_i - z_i)\sigma_i^2}{\sqrt{\sum_{j=1}^{n} (x_j - z_j)^2 \sigma_j^2}} \frac{\phi(0)}{1 - \Phi(0)} \\ &= \mathbf{1}\{i \notin S\} \frac{x_i \sigma_i^2}{\sqrt{\sum_{j \notin S} \sigma_j^2}} \frac{\phi(0)}{1 - \Phi(0)}, \end{aligned}$$

since $(x_i - z_i) = 2x_i \times \mathbf{1}\{i \notin S\}$ and $(x_i - z_i)^2 = 4 \times \mathbf{1}\{i \notin S\}$. □

Three things are worth noting. First, the observer divides attribution for the choice among the weights $w_i$ on non-shared attributes, attributing more to those with larger variance $\sigma_i^2$. Second, the magnitude of the belief change on a given non-shared attribute $i$ is decreasing as the set of non-shared attributes grows, i.e. as the comparison becomes more dilute with respect to $i$. Third, they do not update at all about weights on shared attributes, since choice is uninformative about those weights. Thus there is no reputational effect for shared attributes.

Using Lemma 3, plus the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$, we can rearrange $u^{SC}$ to satisfy

Assumption 1:

$$u^{SC}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{\sum_{i=1}^{n} x_i \left( w_i + \lambda_i \sigma_i \frac{\phi(0)}{1-\Phi(0)} \right)}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(\boldsymbol{x},\boldsymbol{z})$$

$$\theta_i(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|\sigma_i \frac{\phi(0)}{1-\Phi(0)} & , i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|\sigma_i \left(1 - \frac{\sigma_i}{\sqrt{\sum_{j\notin S}\sigma_j^2}}\right) \frac{\phi(0)}{1-\Phi(0)} & , i \notin S. \end{cases}$$

It is easy to see that $\theta_i$ depends only on $(\boldsymbol{x},\boldsymbol{z})$ through $|\boldsymbol{x}-\boldsymbol{z}|$. We need to show that $\theta^S$ and $\theta^N$ are weakly increasing as the sets of shared and non-shared attributes grow respectively. $\theta^S$ is a constant. It is easy to see that $\theta^N$ increases as we add additional non-shared attributes. □

**Proof of Lemma 1** (reporting $u^{SE}$ is an optimal strategy in the signaling-evaluation game). We first define the *residual* evaluations $\bar{y}^x, \bar{y}^z$, after subtracting components which are common knowledge. For a bundle $\boldsymbol{x}$ define $\bar{y}^x$ as:

$$\bar{y}^x = y^x - g(\boldsymbol{x}) - \sum_{i=1}^{n} x_i \lambda_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\}\sum_{j\in S}\sigma_j^2 + \mathbf{1}\{i \notin S\}\sum_{j\notin S}\sigma_j^2} = \sum_{i=1}^{n} x_i w_i$$

Next, we show that player 1's strategy $y^x = u^{SE}(\boldsymbol{x},\boldsymbol{z})$, $y^z = u^{SE}(\boldsymbol{z},\boldsymbol{x})$ is optimal assuming that player 2's strategy is:

$$\hat{w}_i(y^x, y^z) = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2} & , i \notin S \end{cases}$$

Taking first-order conditions of $U^1$ with respect to $y^x$ and $y^z$ gives us optimal values of $y^x$ and $y^z$:

$$\begin{aligned} y^x(\boldsymbol{x},\boldsymbol{z}) &= g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} \lambda_i \frac{\partial \hat{w}_i(y^x, y^z)}{\partial y^x} \\ &= g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} x_i \lambda_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\}\sum_{j\in S}\sigma_j^2 + \mathbf{1}\{i \notin S\}\sum_{j\notin S}\sigma_j^2} \\ y^z(\boldsymbol{z},\boldsymbol{x}) &= g(\boldsymbol{z}) + \sum_{i=1}^{n} z_i w_i + \sum_{i=1}^{n} z_i \lambda_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\}\sum_{j\in S}\sigma_j^2 + \mathbf{1}\{i \notin S\}\sum_{j\notin S}\sigma_j^2}. \end{aligned}$$

Hence $y^x(\boldsymbol{x},\boldsymbol{z}) = u^{SE}(\boldsymbol{x},\boldsymbol{z})$ and $y^z(\boldsymbol{z},\boldsymbol{x}) = u^{SE}(\boldsymbol{z},\boldsymbol{x})$ as stated in the proposition.

Next we show that player 2's strategy is optimal, given player 1's. Taking first order conditions of $U^2$, and using Lemma 2, we obtain the desired result:

$$\hat{w}_i(y^x, y^z) = E[w_i|y^x, y^z] = E[w_i|\bar{y}^x, \bar{y}^z] = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2} & , i \notin S. \end{cases}$$

□

**Proof of Proposition 3** We can rearrange $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ in a form that satisfies Assumption 1 (using the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$):

$$u^{SE}(\boldsymbol{x}, \boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_{i=1}^{n}(w_i + \lambda_i)x_i}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(\boldsymbol{x}, \boldsymbol{z})$$

$$\theta_i(\boldsymbol{x}, \boldsymbol{z}) = \begin{cases} \theta_i^S(|\boldsymbol{x} - \boldsymbol{z}|) & = |\lambda_i| \left(1 - \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\right) & , i \in S \\ \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|) & = |\lambda_i| \left(1 - \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\right) & , i \notin S. \end{cases}$$

It is easy to see that $\theta_i$ depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x} - \boldsymbol{z}|$. It is easily seen also be seen that $\theta^S$ and $\theta^N$ are weakly increasing as we add additional shared and non-shared attributes respectively.

**Proof of Proposition 4** First, we show that utility takes a simple form:

**Lemma 4.** *An implicit associations utility function can be written.*

$$u^{IA}(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i \bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|)$$

$$\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|) = \begin{cases} \frac{\sum_{j\in S}\pi_j\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & , i \in S \\ \frac{\sum_{j\notin S}\pi_j\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & , i \notin S \end{cases}$$

**Proof of Lemma 4** Given agent 1's prior on $\boldsymbol{\pi}$, we have:

$$\hat{f}(\boldsymbol{x}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i E[\pi_i] = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i.$$

Next, we define the residual value $\bar{\hat{f}}(\boldsymbol{x})$ by subtracting the common-knowledge $g(\boldsymbol{x})$. We obtain $\bar{\hat{f}}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \lambda_i$. The second agent's posteriors for each $\lambda_i$ can then

be derived using Lemma 2:

$$E[\lambda_i|\hat{f}(\boldsymbol{x}),\hat{f}(\boldsymbol{z})] = E[\lambda_i|\bar{\hat{f}}(\boldsymbol{x}),\bar{\hat{f}}(\boldsymbol{z})] = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \frac{\bar{\hat{f}}(\boldsymbol{x})+\bar{\hat{f}}(\boldsymbol{z})}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \frac{\bar{\hat{f}}(\boldsymbol{x})-\bar{\hat{f}}(\boldsymbol{z})}{2} & , i \notin S \end{cases}$$
$$= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \frac{\sum_{i=1}^n (x_i+z_i)\lambda_i}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \frac{\sum_{i=1}^n (x_i-z_i)\lambda_i}{2} & , i \notin S \end{cases}$$
$$= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \sum_{j\in S} x_j\lambda_j & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \sum_{j\notin S} x_j\lambda_j & , i \notin S \end{cases}$$

where the final step uses $x_i - z_i = 2x_i\mathbf{1}\{i \notin S\}$ and $x_i + z_i = 2x_i\mathbf{1}\{i \in S\}$.

The second agent's overall evaluation of bundle $\boldsymbol{x}$ will thus be equal to:

$$\begin{aligned} E[f(\boldsymbol{x})|\boldsymbol{\pi},\hat{f}(\boldsymbol{x}),\hat{f}(\boldsymbol{z})] &= g(\boldsymbol{x}) + \sum_{i=1}^n x_i\pi_i E[\lambda_i|\hat{f}(\boldsymbol{x}),\hat{f}(\boldsymbol{z})] \\ &= g(\boldsymbol{x}) + \sum_{i=1}^n x_i^2\pi_i\sigma_i^2 \left( \frac{\mathbf{1}\{i\in S\}}{\sum_{j\in S}\sigma_j^2} \sum_{j\in S} x_j\lambda_j + \frac{\mathbf{1}\{i\notin S\}}{\sum_{j\notin S}\sigma_j^2} \sum_{j\notin S} x_j\lambda_j \right) \\ &= g(\boldsymbol{x}) + \frac{\sum_{i\in S}\pi_i\sigma_i^2}{\sum_{j\in S}\sigma_j^2} \sum_{j\in S} x_j\lambda_j + \frac{\sum_{i\notin S}\pi_i\sigma_i^2}{\sum_{j\notin S}\sigma_j^2} \sum_{j\notin S} x_j\lambda_j \\ &= g(\boldsymbol{x}) + \sum_{i\in S} x_i\lambda_i \frac{\sum_{j\in S}\pi_j\sigma_j^2}{\sum_{j\in S}\sigma_j^2} + \sum_{i\notin S} x_i\lambda_i \frac{\sum_{j\notin S}\pi_j\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} = u^{IA}(\boldsymbol{x},\boldsymbol{z}), \end{aligned}$$

where we use the convention ${}^{\mathbf{1}\{i\in S\}}/_{\sum_{j\in S}\sigma_j^2} = 0$ if there are no shared attributes, and equivalently for the non-shared (this saves us from explicitly writing out the special cases of all shared or all non-shared attributes). The third step uses $x_i^2 = 1$ and the fourth step uses a switch of index labels. □

The Proposition states that at most one attribute has either $\lambda_i \neq 0$ or $\pi_i \neq 1$. Assign index $t$ to this attribute. Our next goal is to show that the functional form derived in Lemma 4 satisfies Assumption 1.

First, observe that $\bar{\pi}_i(|\boldsymbol{x}-\boldsymbol{z}|)$ can be written as;

$$\begin{aligned}
\bar{\pi}_i(|\boldsymbol{x}-\boldsymbol{z}|) &= \begin{cases} 1-\frac{\sum_{j\in S}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & ,i\in S \\ 1-\frac{\sum_{j\notin S}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & ,i\notin S \end{cases} \\
&= \begin{cases} 1-\frac{(1-\pi_i)\sigma_i^2}{\sum_{j\in S}\sigma_j^2}-\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & ,i\in S \\ 1-\frac{(1-\pi_i)\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}-\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & ,i\notin S \end{cases} \\
&= \begin{cases} \pi_i+(1-\pi_i)\left(1-\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\right)-\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & ,i\in S \\ \pi_i+(1-\pi_i)\left(1-\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\right)-\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & ,i\notin S. \end{cases}
\end{aligned}$$

Substituting into the functional form derived in Lemma 4, plus the fact that $\lambda_i(1-\pi_i) = sgn(\lambda_i(1-\pi_i))|\lambda_i(1-\pi_i)|$, we obtain:

$$u^{IA}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x})+\sum_{i=1}^n x_i\lambda_i\pi_i}_{v(\boldsymbol{x})} + \sum_{i=1}^n x_i \underbrace{sgn(\lambda_i(1-\pi_i))}_{\kappa_i}\theta_i(\boldsymbol{x},\boldsymbol{z}) - B$$

$$\theta_i(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)|\left(1-\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\right) & ,i\in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)|\left(1-\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\right) & ,i\notin S, \end{cases}$$

where:

$$\begin{aligned}
B &= \sum_{i=1}^n x_i\lambda_i\left[\mathbf{1}\{i\in S\}\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2}+\mathbf{1}\{i\notin S\}\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2}\right] \\
&= x_t\lambda_t\left[\mathbf{1}\{t\in S\}\frac{\sum_{(j\in S)\wedge(j\neq t)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2}+\mathbf{1}\{t\notin S\}\frac{\sum_{(j\notin S)\wedge(j\neq t)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2}\right] \\
&\quad+\sum_{i\neq t}x_i\lambda_i\left[\mathbf{1}\{i\in S\}\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2}+\mathbf{1}\{i\notin S\}\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2}\right] \\
&= 0,
\end{aligned}$$

where the final step uses the facts that $\forall j\neq t,\ \pi_j = 1$ (so the first term equals zero) and $\lambda_j = 0$ (so the second term equals zero).

Thus we can see that $u^{IA}$ satisfies Assumption 1. To complete the proof, note that (1) $\theta_i$ depends only on $(\boldsymbol{x},\boldsymbol{z})$ through $|\boldsymbol{x}-\boldsymbol{z}|$, and (2) $\theta^S$ and $\theta^N$ are weakly increasing as the sets of shared and non-shared attributes grow, respectively.

**Proof of Proposition 5** We need to show that opacity for attribute $i$ is weakly greater when an attribute shares status with $k$ than when it does not. For each foundation we derive an expression for the change in $\theta_i$ when $i$ changes status from non-shared to shared, while no other attribute changes status. We show that this expression is weakly positive if $k \in S$, and weakly negative if $k \notin S$. (Note that the assumption holds trivially for $i = k$, since $i$ always shares status with itself).

**Ceteris paribus.** No implicit preference for $k$ means $\lambda_k = 0$.

$$\theta_i^S - \theta_i^N = |\lambda_i| \mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\}.$$

$\theta_i^S - \theta_i^N \geq 0$, so if $k$ is shared, shared attributes have weakly higher opacity. If $k$ is non-shared, $\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\} = 0$ (since $\lambda_k = 0$ and $k \notin S$), so $\theta_i^S - \theta_i^N = 0$ (opacity is the same for shared and non-shared). Thus, opacity for attribute $i$ is weakly greater when $i$ shares status with $k$.

**Signaling evaluation.** Let $\sigma_k^2 \geq \sum_{i \neq t} \sigma_i^2$. Then we have:

$$\theta_i^S - \theta_i^N = |\lambda_i| \sigma_i^2 \left( \frac{\sum_{j \in S} \sigma_j^2 - \sum_{j \notin S} \sigma_j^2}{\sum_{j \in S} \sigma_j^2 \sum_{j \notin S} \sigma_j^2} \right),$$

which is weakly positive if $k \in S$, weakly negative if $k \notin S$.

**Implicit associations.** Let $\sigma_k^2 \geq \sum_{i \neq t} \sigma_i^2$. Assumption 4 holds trivially for all attributes $j$ with $\lambda_j = 0$ or $\pi_j = 1$, since $\theta_j = 0$, so we only need to check if it holds for attribute $t$, the attribute that may have nonzero $\lambda$ or non-unity $\pi$. If $t$ coincides with $k$, Assumption 4 holds trivially. Suppose not, i.e. $t \neq k$ (and hence $\lambda_k = 0$ and $\pi_k = 1$). We have:

$$\theta_t^S - \theta_t^N = |\lambda_t (1 - \pi_t)| \sigma_t^2 \left( \frac{\sum_{j \in S} \sigma_j^2 - \sum_{j \notin S} \sigma_j^2}{\sum_{j \in S} \sigma_j^2 \sum_{j \notin S} \sigma_j^2} \right),$$

which is weakly positive if $k \in S$, weakly negative if $k \notin S$.

## B.3 Data appendix for analysis of Exley (2016)

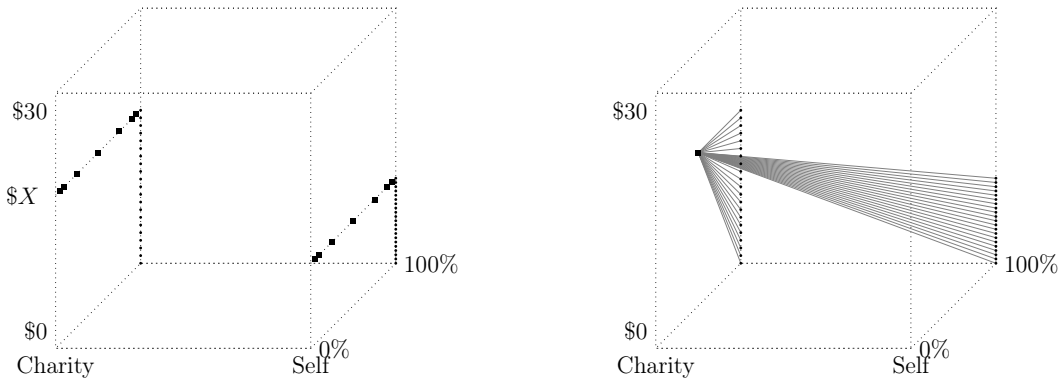
Exley (2016)'s experiment proceeds in three steps:

1. Normalization choice. For each participant she elicits using a choice list the smallest sure payment $\$X \in \{0, 2, \ldots, 30\}$ to charity (or to another participant – we refer to both as "charity") that is chosen over \$10 for self.

2. Using $X$, she constructs a sequence of participant-specific simple lotteries. These pay out with probability $P \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. Self lotteries, denoted by $P^S$, pay \$10 to self. Charity lotteries, denoted by $P^C$, pay \$$X$ to charity.

3. She elicits, using choice lists, preferences between each lottery and 21 different sure payoffs to self or to charity. We index these by $t = 0, ..., 20$. The sure payments are $Y_t^S = (0, 0.50, ..., 10)$ for self lotteries and $Y_t^C = (0, {}^X/_{20}, ..., X)$ for charity lotteries.

Thus, a bundle in this experiment is characterized by three basic attributes: a Recipient (Self or Charity), a Prize, and a Probability. Figure 10a shows graphically the full set of bundles that appear in the choice lists. Figure 10b shows every choice set a participant faces for one charity lottery.




(a) All lotteries and safe payments (b) Choice sets for charity lottery with $P = 0.5$

Square markers correspond to lotteries paying charity (left face, prize \$$X$) and self (right face, prize \$10) used in the experiment. Circular markers correspond to sure payoffs on the choice lists. Solid lines in panel (b) correspond are all choice sets containing the .

Figure 10: Exley (2016) data structure

We do not observe all possible choices over the bundles marked in Figure 10a. Specifically, we do not observe (a) choices between sure payoffs to Self and Charity (except for the normalization choices); and (b) choices between Self and Charity lotteries.

### B.3.1 Defining a utility benchmark

Exley's null hypothesis *standard risk preferences*, assumes two properties of utility. We will make use of the same assumptions to guide our representation of the choice data in a space of two binary attributes: Social $\in$ {Selfish, Generous} and Risk $\in$ {Safe, Risky} (Section B.3.2), and to impute some choices that are not observed in the data (Section B.3.3).

Exley's first assumption is *linearity in payoffs*, meaning that preferences over sure payoffs are preserved under linear rescaling. So, if the participant is indifferent between \$$y$ for

Charity and $\$y'$ for Self, she is also indifferent between $\$yr$ for Charity and $\$y'r$ for Self, for $r \geq 0$.

This assumption plays an important role in Exley's analysis. Her tests involve comparing certainty equivalents of Self and Charity lotteries, measured in terms of sure payments to Self and Charity. To establish inconsistencies in these certainty equivalents, she compares these certainty equivalents. To say that the participant values a given lottery more in dollars to Self than in dollars to Charity, she needs to be able to rank certainty equivalents measured in these units.

The second assumption is that preferences over bundles are preserved under linear rescaling of probabilities, so we refer to it as *separability in probabilities*. If the participant is indifferent between $\$y$ for Charity and $\$y'$ for Self, she is also indifferent between $\$y$ for Charity with probability $p$ and $\$y'$ for Self with probability $p$, for $p \in [0, 1]$.

Let us define a binary variable $c \in \{0, 1\}$ equal to one if the Recipient is Charity, and denote the Prize by $y$ and Probability by $p$. The assumptions imply a utility function of the following form (normalizing $v(0) = 0$, $\pi(0) = 0$, and $\pi(1) = 1$):

$$v(c, y, p) = \pi(p)v\left(\frac{y}{1 + \lambda c}\right)$$

Linearity in payoffs is captured by $\lambda$. The participant is indifferent between a prize of $y$ to Self and $(1 + \lambda)y$ to Charity. Separability in probabilities is captured via the probability weighting function $\pi(p)$. Preferences between two same-probability lotteries do not depend on $p$ (note that since all lotteries have exactly one non-zero prize, the assumption does not require *linearity* in probabilities).

Finally, if we add Constant Relative Risk Aversion (CRRA): $v(y) = y^\alpha$ we obtain utility function (4). We will use this to motivate construction of a binary attribute space in which "ambivalence" holds.

$$v(c, y, p) = \pi(p)\left(\frac{y}{1 + \lambda c}\right)^\alpha. \tag{4}$$

### B.3.2 Constructing a binary attribute space with "ambivalence"

We describe here how using these assumptions we can describe the environment in terms of two binary attributes: Social $\in$ {Selfish, Generous} and Risk $\in$ {Safe, Risky}. Our approach amounts to selecting in Figure 10a, which can be described by the binary attributes Social and Risk, over which we observe preferences. Figure 11 panels (a) and (b) provide an example.

We need to transform the data for two reasons. First, note that all three attributes are

"monotone": all else equal, we would expect the participant to prefer Self over Charity, and prefer larger Prizes or Probabilities to smaller. Therefore, choice sets that vary on only one of these dimensions at a time cannot satisfy Ambivalence (see Section 5): we cannot expect the participant to be close to indifferent. Second, note that Prize and Probability are multivalued, and so do not immediately fit into a binary attribute representation. To construct a binary attribute space, we do the following.

First, we analyze preferences within a set of choice lists defined by a given lottery probability $P$. We cannot make comparisons across different values of $P$, because we would not expect ambivalence to be satisfied and because in any case such choices are not observed. Thus, we will construct a separate binary attribute space for each value of $P$. Such a space contains two probability values: lotteries with probability $P$, and sure payoffs with probability 1.

Second, we divide up the Prize dimension, so that Self prizes are different to Charity prizes, and sure prizes are different to risky ones, in such a way that ambivalence plausibly holds. We are guided by the assumptions described in section B.3.1. In essence we ensure that an observer who believed the participant maximizes (4) would expect them to be close to indifferent.

Consider the self lottery $(0, 10, P)$ that pays \$10 to Self with probability $P$. Equation (4) implies the following utilities are equal:

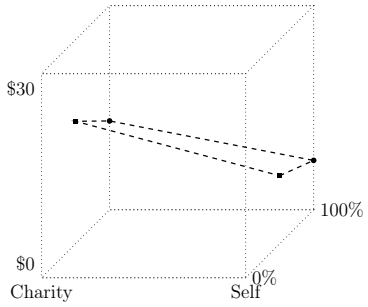
$$v\underbrace{(0,10,P)}_{\text{Self lottery}} = v\underbrace{(1,(1+\lambda)10,P)}_{\text{Charity lottery}} = v\underbrace{\left(0,\pi(P)^{\frac{1}{\alpha}}10,1\right)}_{\text{Self sure payoff}} = v\underbrace{\left(1,(1+\lambda)\pi(P)^{\frac{1}{\alpha}}10,1\right)}_{\text{Charity sure payoff}} \quad (5)$$

Our approach will be to focus on choices defined by two participant-specific scaling parameters, $L$ and $R(P)$, such that Charity prizes are an $L$-multiple of self prizes, and sure prizes are an $R(P)$-multiple of risky prizes. So, our binary attribute space consists of: (1) the Self lottery paying \$10 with probability $P$, (2) the Charity lottery paying \$$10L$ with probability $P$, (3) the Self sure payment of \$$10R(P)$, and (4) the Charity sure payment of \$$10LR(P)$. Ambivalence holds if $L \approx 1 + \lambda$ and $R(P) \approx \pi(P)^{\frac{1}{\alpha}}$.

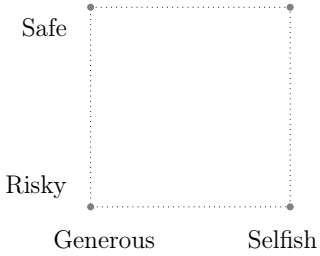
We calibrate $L$ and $R(P)$ using the participant's own revealed preferences. $L$ is set using the initial normalization choice in the experiment: $L = {}^{X}/_{10}$ (which is also the rate at which Exley compares self and charity payoffs). Recall that $X$ is the smallest payment to charity that was chosen over \$10 to self, from which we infer ${}^{X}/_{10} > 1 + \lambda > {}^{X-2}/_{10}$. Linearity in payoffs therefore implies that the participant can be expected to to have a slight preference for a payoff $LY$ to charity over $Y$ to self, but is also close to indifferent.

We consider two possible values for $R(P)$ set using the participant's own choices between

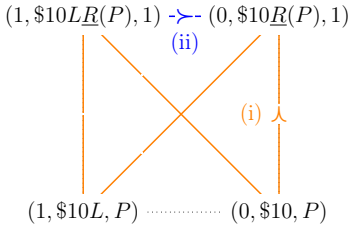
the self lottery and self sure payoffs. The first is based on the largest self sure payment that the participant rejected, which we denote by $\underline{Y}(P^S)$ and set $\underline{R}(P) = \underline{Y}(P^S)/10$. The second is based on the smallest self sure payoff that they accepted, which we denote by $\overline{Y}(P^S)$. This gives us $\overline{R}(P) = \overline{Y}(P^S)/10$. Since $\underline{R}(P)$ and $\overline{R}(P)$ are close to one another, we assume that the choices based on these parameters are informative about the same binary attribute space, depicted in Figure 11b. Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Risky) is chosen over (Selfish, Safe) (Figure 11c). Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Safe) is chosen over (Selfish, Risky) (Figure 11d). We assume that both preferences are close enough to indifferent that ambivalence holds.



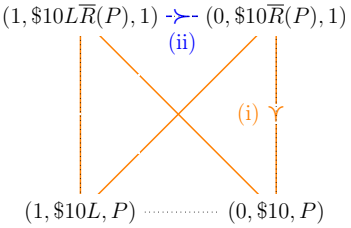

(a) Bundles selected for analysis




(b) Binary attribute representation




(c) Calibration based on $\underline{R}(P)$




(d) Calibration based on $\overline{R}(P)$

Panels (a) and (b) shows four bundles selected from the set in Figure 10a and in our binary-attribute representation. Panels (c) and (d) display these bundles in terms of $(Recipient, Prize, Probability)$ where $Recipient = 1$ denotes Charity and $Recipient = 0$ denotes Self. Choice sets marked in orange are observed in the data. Choices labeled (i) follow from the calibration of $\underline{R}(P)$ and $\overline{R}(P)$. Choices labeled (ii) are not directly observed in the data, but are imputed from the calibration of $L$ plus *linearity in payoffs*.

Figure 11: Binary attribute representation of Exley (2016)'s choice data

Note that Exley's analysis uses the midpoints between just-rejected and just-accepted payoffs to impute certainty equivalents (i.e. points of indifference) of different lotteries, which are the outcomes in her regression analyses. Our analysis uses the observed choices only, so is expressed in terms of strict preferences.

### B.3.3 Imputing non-observed choices

Figures 11c and 11d include preferences on the upper horizontal choice set (labeled (ii)). This choice set is not observed in the data. But *linearity in payoffs* implies that $\$LY$ to charity is preferred to $\$Y$ to self, for all $Y \geq 0$. We use this to impute the ranking of the two safe payoffs in our binary attribute space. In essence, we are assuming that choices in this direct comparison are governed by (4).

Our calibration of the binary attribute space is constrained by the lotteries that we observe, whose prizes Exley also calibrated from $X$, that is, charity lotteries pay $X = 10L$ and self lotteries pay 10. Thus we cannot examine payoffs that vary in other proportions, and therefore cannot observe or impute a choice set where (Selfish, Safe) $\succ$ (Charity, Safe).

### B.3.4 Separable representation

Now that our binary attribute space is defined, we show that the preferences defined by (4) have a separable representation over its attributes. This is helpful because our signaling-choice foundation assumes the observer believes preferences are a sum of mean-zero *weights*.

Consider the four bundles associated with lottery winning probability $P$, defined by our calibration parameters $L$ and $R(P)$. Using (4) we have:

$$\begin{array}{lll} v\left(Selfish, Risky\right) & = v(0, 10, P) & = \pi(P)\left(10\right)^{\alpha} \\ v\left(Generous, Risky\right) & = v(1, 10L, P) & = \pi(P)\left(\frac{10L}{1+\lambda}\right)^{\alpha} \\ v\left(Selfish, Safe\right) & = v(0, 10R(P), 1) & = \left(10R(P)\right)^{\alpha} \\ v\left(Generous, Safe\right) & = v(0, 10LR(P), 1) & = \left(\frac{10LR(P)}{1+\lambda}\right)^{\alpha} \end{array}$$

The ranking of bundles is invariant to any increasing transformation of $u$. Consider the transformation

$$\hat{v}(x) := 2\ln\left(\frac{v(x)^{\frac{1}{\alpha}}}{10}\right) - \ln\left(R(P)\pi(P)^{\frac{1}{\alpha}}\right) + \ln\left(\frac{1+\lambda}{L}\right)$$

Applying this, we obtain:

$$\begin{aligned}
\hat{v}\left(Selfish, Risky\right) &= \ln\left(\tfrac{1+\lambda}{L}\right) + \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
\hat{v}\left(Generous, Risky\right) &= -\ln\left(\tfrac{1+\lambda}{L}\right) + \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
\hat{v}\left(Selfish, Safe\right) &= \ln\left(\tfrac{1+\lambda}{L}\right) - \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
\hat{v}\left(Generous, Safe\right) &= -\ln\left(\tfrac{1+\lambda}{L}\right) - \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right)
\end{aligned}$$

Let $Social \in \{Generous, Selfish\}$ be $x_1 \in \{-1, 1\}$ and $Risk \in \{Risky, Safe\}$ be $x_2 \in \{-1, 1\}$. Then we have:

$$\hat{v}(x) = \ln\left(\frac{1+\lambda}{L}\right) x_1 - \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) x_2 \tag{6}$$

Thus, any choices over the four bundles can represented by a separable utility function in $x_1$ and $x_2$.

The Signaling foundation requires the observer to have mean-zero Gaussian priors over the weights. Zero mean requires $L$ and $R(P)$ to be appropriately calibrated to make the participant indifferent in expectation, i.e. such that $E \ln\left(\frac{1+\lambda}{L}\right) = 0$ and $E \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) = 0$. Mean-zero plus Normality requires ${}^{1+\lambda}/_{L} \sim \text{Lognormal}(0, \sigma_1^2)$ and ${}^{\pi(P)^{\frac{1}{\alpha}}}/_{R(P)} \sim \text{Lognormal}(0, \sigma_2^2)$.

### B.3.5 Permutation tests

We perform two simple permutation tests that ask whether our data are consistent with different assumptions about noise in behavior.

The starting point is the experimental dataset. An observation is $C_{iP}$ where $i \in (1, ..., 86)$ indexes participants and $P \in (.05, .1, .25, .5, .75, .9, .95)$ indexes lottery probabilities. $C \in \{0, 1, 2, 3\}$ records what type of cycle was observed for that participant-probability: 0 for no cycle, 1 for pro-Risky, 2 for pro-Safe, 3 for pro-Selfish.

**Null hypothesis 2: homogeneity.** The null hypothesis of our first permutation test is that, separately for each value of $P$, the likelihood of a given cycle $C' \in \{0, 1, 2, 3\}$ is the same for all participants. E.g., it could be that when $P = 0.5$, all participants have a 5% chance of a pro-Risky cycles, a 10% chance of a pro-Safe, cycle, and an 18% chance of a pro-Selfish cycle. Our permutation test thus asks whether permuting indices $i$ within each probability $P$ reproduces the same distribution over $C_{iP}$. Intuitively, this test asks whether

variation in cycling behavior could be explained by a single representative agent.

**Null hypothesis 3: homogeneity conditional on a cycle.** The null hypothesis of our second permutation test is that, separately for each value of $P$, *conditional on a cycle being observed* the likelihood of a given cycle is the same for all participants. E.g., it could be that when for all participants, $P = 0.5$, 15% of *cycles* are pro-Risky, 30% are pro-Safe, and 55% are pro-Selfish, but some participants can cycle more than others. Our permutation test therefore permutes indices $i$ within each probability $P$ *conditional on* $C_{iP} \neq 0$. Intuitively, this test asks whether variation in cycling behavior could be explained by heterogeneity in the *likelihood* of cycling, but otherwise homogeneity in implicit preferences.

The basic testing approach is as follows.

1. We represent each participant in the **sample** according to their number of cycles of each type (pro-Risky, pro-Safe, pro-Selfish). We then compute the fraction of participants exhibiting each possible combination. We call these the **sample proportions**. For example, 20 percent of participants have no cycles $(0, 0, 0)$. (Thus the dataset is represented as a distribution over the simplex $\{c \in \{0, ..., 7\}^3 : c_1 + c_2 + c_3 \leq 7\}$, since at most one cycle can be observed per $P$).
2. We duplicate the experimental dataset 10000 times, creating a **population** of 860,000 decision-makers that holds constant the frequency of each observed choice. We then randomly permute rows of this dataset according to our null hypothesis to generated a simulated population distribution of behavior under the null. We compute the fraction of the population exhibiting each possible combination of cycles, and call these the **population proportions**.
3. We compute the sum of squared differences between sample and population proportions, this is our **sample statistic** of interest. A small value of this statistic implies the sample distribution is similar to the population distribution.
4. Returning to the sample dataset with 86 participants, we generate 10,000 **simulated samples**, by permuting rows according to the null assumptions. For each, we compute the fraction of the simulated sample exhibiting each combination of cycles, and call these the **simulated proportions**. We compute the sum of squared differences between the simulated proportions and the population proportions, to obtain a 10,000 draws of the **simulated statistic**.
5. The p-value of the test is simply the fraction of simulated statistics that are larger than the sample statistic. A small p-value indicates that the sample statistic tends to be larger than we would expect it to be under the null hypothesis.

We present our findings in Figure 12. We strongly reject Null hypothesis 2 ($p < .001$). The main contributor to this rejection is substantial excess mass at $(0, 0, 0)$ in the sample relative to that expected under the null: 20 percent of participants have no cycles at all, whereas under the null only around 6 percent of participants should exhibit zero cycles across all seven probabilities. We also find evidence against Null hypothesis 3 ($p = .065$), albeit weaker.

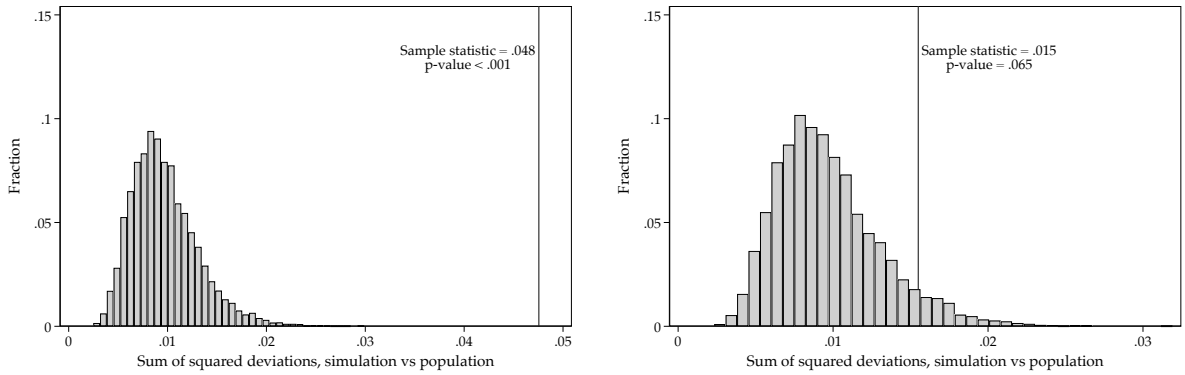



Figure 12: Permutation tests on Exley (2016) data

Left panel corresponds to Null hypothesis 2, right panel to Null hypothesis 3.