Implicit Preferences Inferred from Choice*

PRELIMINARY DRAFT, COMMENTS WELCOME

Tom Cunningham† Jonathan de Quidt‡

October 22, 2016

Latest version available here
Appendix available here

Abstract

A longstanding distinction in psychology is between implicit and explicit preferences. Implicit preferences are ordinarily measured by observing non-choice data, such as response time. In this paper we introduce a method for inferring implicit preferences directly from choices. The necessary assumption is that implicit preferences toward an attribute (e.g. gender, race, sugar) have a stronger effect when the attribute is mixed with others, and so the decision becomes less "revealing" about one's preferences. We discuss reasons why preferences would have this property, advantages and disadvantages of this method relative to other measures of implicit preferences, and application to measuring implicit preferences in racial discrimination, self-control, and framing effects.

*We thank for comments, among others, Roland Benabou, Ed Glaeser, Sendhil Mullainathan, Antonio Rangel, Tomasz Strzalecki and seminar audiences at USC, Columbia, Caltech, Stanford, Facebook, Harvard, Princeton, & Santa Cruz.

†Facebook.

‡CESifo and Institute for International Economic Studies, Stockholm.

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

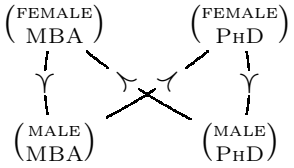
1 Introduction

In this paper we show how simple choices can, by themselves, reveal two separate sets of preferences. The idea is best illustrated with an intransitive cycle. Suppose you observe a recruiter's decisions between pairs of job applicants, each of whom is either male or female, and has either an MBA or a PhD. Suppose you observe both that:

1. the recruiter chooses a female candidate over a male candidate whenever the two candidates' qualifications are the same,

2. the recruiter chooses a male candidate over a female candidate whenever the two candidates' qualifications differ.

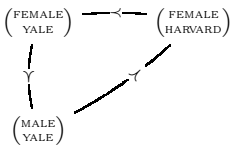
Graphically, using $A \succ B$ to represent the choice of A from $\{A, B\}$:



These choices are inconsistent with maximization of a utility function. Nevertheless they form an intuitive pattern, which we describe as a "figure 8," and seem to reveal the existence of two distinct attitudes towards female candidates: a positive preference revealed in the vertical choice sets (between candidates of opposite sexes but otherwise identical), and a negative preference revealed in the diagonal choice sets (between candidates who differ in another respect besides gender).

Our paper generalizes this observation, that choices can sometimes reveal two distinct sets of preferences. We study choice over bundles of binary attributes (male/female, black/white, aisle/window), and we rank choice sets according to how *revealing* they are about each attribute. For example, in the diagram above, we say that the diagonal choice sets are less revealing about preferences over gender, compared to the vertical choice sets. We define an implicit preference for an attribute as a preference that becomes stronger in less revealing choice sets: the figure-8 above reveals an implicit preference for male over female candidates.

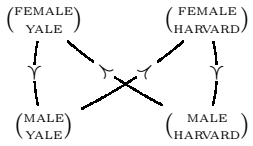
Right Triangle



Scissor

$$y\left(FY, \left\{ {FY \quad FH} \right\}\right) < y\left(FY, \left\{ {FY \quad \quad MH} \right\}\right)$$

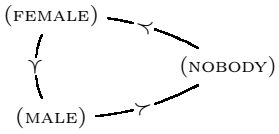
Figure 8



Parallel Scissors

$$y\left(FY, \left\{ {FY \quad FH} \right\}\right) > y\left(FY, \left\{ {FY \quad \quad MH} \right\}\right)$$
$$y\left(MY, \left\{ {MY \quad MH} \right\}\right) < y\left(MY, \left\{ {MY \quad \quad MH} \right\}\right)$$

Isosceles



Joint-Separate

$$y(\mathrm{F}, \{\mathrm{F}\}) > y(\mathrm{F}, \{\mathrm{F}, \mathrm{M}\})$$
$$y(\mathrm{M}, \{\mathrm{M}\}) < y(\mathrm{M}, \{\mathrm{F}, \mathrm{M}\})$$

Figure 1: Six different patterns of behaviors which can identify - under certain assumptions - an implicit preference for Male over Female. Each subfigure in the first column represents an intransitive cycles of choices. Each subfigure in the second column represents menu-dependent evaluations of outcomes, where $y(a, \{a, b\})$ represents the evaluation of outcome a when the decision-maker is evaluating both a and b. Each example in this column consists of evaluations which change depending on the set of outcomes being evaluated.

The formal results in this paper are mainly concerned with deriving sufficient conditions on choice data from which we can infer the existence of an implicit preference regarding some attribute, and giving counterexamples, i.e. examples where these conditions do not hold, and where such choices would not reveal implicit preferences.

The formal framework we develop additionally shows how implicit preferences can be revealed in data on continuous evaluations, such as willingness-to-pay, criminal sentencing, or exam grading. Suppose we observe a judge assigning sentences to defendants, and we find that (1) when a black and a white defendant are sentenced alongside each other, there is no difference in the sentence received; but (2) when two black defendants are sentenced, they both get relatively long sentences, and when two white defendants are sentenced they both get relatively short sentences. Under our model this behavior identifies an implicit preference in favor of white defendants.

We do not know of any prior theoretical papers which have identified this figure-8 pattern in choices, or which have shown how it can be used to identify implicit preferences; existing theories of menu-dependent preferences do not predict this pattern.[1] Nevertheless we think that the idea of implicit preferences being revealed by indirect choices taps into a commonsense understanding of decision-making, and most of our formal results correspond to natural intuitions. We discuss a few empirical papers we have found which can be interpreted as identifying implicit preferences.

Our introductory examples show how we can identify implicit discrimination - a topic of great recent interest.[2] But the possible applications are broad: in principle we can detect implicit preferences over any attribute, and there are many contexts in which we might expect them. Figure 2 shows a variety of figure-8 cycles in different domains. The choices indicated are our conjectures, to illustrate implicit preferences that we believe to be intuitive.

- **Consumption A.** Consider a person who chooses a diet soda over a full-sugar soda when they are of the same brand, but the full-sugar soda when they are of different brands. This reveals an explicit preference for diet soda, but an implicit preference for full-sugar soda.

- **Consumption B.** Consider a person who, when choosing which movie to watch, would always choose a documentary over a comedy when playing in the same cinema, but would always choose a comedy over a documentary if playing in different cinemas. This reveals an implicit preference for comedies over documentaries.

- **Framing.** Consider a person choosing between two lotteries, one of which emphasizes the probability of winning, the other the probability of losing. The person is indifferent when the lotteries share the same probability and payoff, but choose the positively-framed lottery when the payoffs differ. These choices would reveal an implicit preference for the positive description, but no explicit preference. More generally, we think that certain classical framing effects can be thought of as cases

[1] E.g. "salience" (**?**), "relative thinking" (Bushong et al. (2014)), "magnitude effects" (**?**), or "focusing" (**?**). To the best of our knowledge, Cunningham (2014) is the only existing paper with an explicit identification of a figure-8 intransitive cycle.

[2] Bertrand et al. (2005) discuss the economic importance of implicit 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 giving a way of measuring the relative ambiguity of choices sets. Mullainathan (2015) gives a recent overview of evidence of implicit discrimination. People often make a distinction between statistical and taste-based discrimination: both are compatible with being implicit.

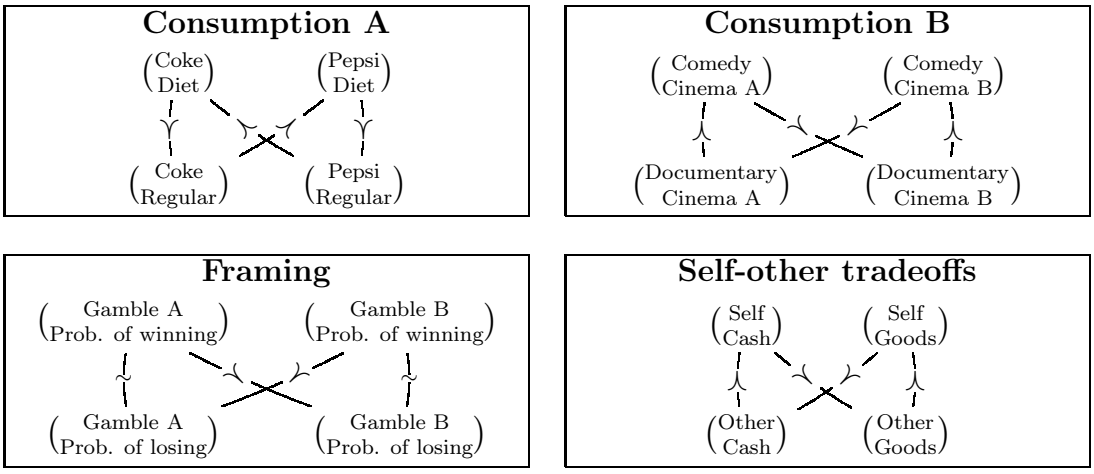


Figure 2: Figure 8 intransitivities applied to different domains.

where a decision-maker has an implicit preference, but no explicit preference, for some attribute.

- **Self-other tradeoffs.** Consider a person who would always choose to give cash or goods to charity, rather than keep it for themselves. But when the payoffs differ in kind (e.g. cash to me vs goods to charity), then they choose in favor of themselves. This reveals an explicit preference in favor of the charity, but an implicit preference in favor of themself.[3]

We think of the *implicitness* of preferences as analogous to other non-parametric properties of choice, such as complementarity or substitutability. Complementarity is defined by certain patterns in choice - for example the fact that someone chooses to drink their tea with milk, but their coffee without. In the same way we think that certain choices can be described as revealing implicit preferences, while remaining agnostic about the *reason* why preferences are implicit, just as we are typically agnostic about the underlying psychological reason why someone finds tea and milk to be complementary.

Why would preferences change when the choice set becomes more revealing? We discuss three interpretations. Consider our introductory example of gender discrimination. First, people could be *unaware* of having a preference over gender, and correct

[3]The experiments in Exley (2015) have a similar structure, although that paper introduces a risk/safety tradeoff, rather than a cash/goods tradeoff, and does not use a figure-8 identification. We discuss Exley (2015) in detail later in the paper.

for it only insofar as they can detect it in their own instincts, implying that gender would have a bigger effect on judgments in less revealing choice sets.[4] Second, people could be *aware* of their gender bias, but would like to conceal it from an observer, i.e. they wish to signal their preferences (the decision-maker could be their own observer, as in models of self-signaling). Again this implies that gender would have a larger effect in less revealing choices. Third, people could be constrained by what we call *ceteris paribus* rules, such as, for example, "never choose a man over an equally-qualified woman." Most of our theoretical work is agnostic about the underlying cause of implicit preferences, but we also discuss ways in which the interpretations can be distinguished - most simply, someone who signals their preferences will be constrained by precedents set by prior choices, while someone who learns their preferences (as in the unconscious influences model) will not.

Economic implications of implicit preferences. There are many influential theories of human behaviour in which motives are in some sense hidden: Freudian and subsequent psychoanalytic theory; recent claims in social psychology about unconscious processes;[5] claims in judgment and decision-making about unconscious influences;[6] evolutionary theories of self-deception in humans and other animals;[7] and economic theories of signaling in social behaviour.[8] There is also a case to be made from introspection: we often are unsure about, for example, whether we would have liked a wine equally much if it had cost \$10 instead of \$50; whether we would have liked an academic paper as much if it had been submitted under a different name; whether we would have treated a student the same way if they had been of a different gender or race. This ignorance represents a window through which implicit influences can enter. More narrowly, as economists we are interested domains where it is widely thought people have serious internal conflicts - in decisions concerning race, charitable giving, politics, and status goods. Our paper gives a rigorous foundation for estimating the strength of implicit

[4]Suppose you get a good feeling about the male candidate, and a bad feeling about the female candidate. If they have the same qualifications, then you can infer exactly why you have different feelings. If they have different qualifications, then the feeling may be attributed, in part, to the difference in qualifications. We formalize this theory in the Appendix using an application of the model in Cunningham (2014). In that model the conscious system must rely on pre-conscious systems for interpreting information, and therefore can be influenced by aspects of a stimulus that it regards as normatively irrelevant.

[5]For example, in the recent popular books "Blink", "Subliminal", "The Hidden Brain", "The Invisible Gorilla", and "Incognito".

[6]e.g. **?**.

[7]Von Hippel and Trivers (2011)

[8]Spence (1973), Hanson (2008)

influences in each of these domains.

Our theory has implications for applied industrial organization because it predicts that demand for a good can vary systematically with features of the choice set. Firms will bundle implicitly desired features or products along with other features, for example bundling pornographic pictures with journalism, when people have an implicit desire for pornography.[9]

Our measure of implicit preferences can be compared with the Implicit Association Test, which uses response time in a categorization task.[10] An important advantage of our measure is that it is based only on ordinary decision-making, so needs little additional interpretation to be used in interpreting economic outcomes, and can be computed directly from observational data.

Prior experiments on implicit preferences. A few prior experimental studies have relied on the intuition that we are attempting to formalize: Snyder et al. (1979) on implicit discrimination against the handicapped, Exley (2015) on implicit preferences over giving to charity, and Bohnet et al. (2015) on implicit gender discrimination. For each of these papers we show that, although they study implicit preferences in our sense, the statistical tests which they use to identify implicit discrimination are imperfect (i.e., they would identify implicit preferences where none exist), and we describe alternative appropriate tests. We also reanalyze an existing dataset from DeSante (2013) and find evidence for an implicit preference in favor of white over black welfare applicants.

Section 2 contains the main formal results, giving assumptions under which implicit preferences can be inferred from each of the patterns in behaviour illustrated in Figure 1. Section 3 discusses alternative ways of identifying implicit preferences; how to analyze different types of dataset; plausible foundations that generate implicit preferences; and relates our interpretations to existing literature. Section 4 discusses interpretation of data from four relevant empirical papers. Section 5 gives a brief overview of economic applications, and Section 6 concludes. Appendices contain proofs, statements of the three models that generate implicit preferences (*ceteris paribus*, signaling, and implicit knowledge), additional formal results and discussion.

[9] Chance and Norton (2009).

[10] Greenwald et al. (1998)

2 Model

We consider *outcomes* which are bundles of binary attributes (e.g., male/female, short/tall, day/night). In most of the paper we consider data on either *choice* between a pair of outcomes, or *evaluation* of both members a pair of outcomes (e.g. stating willingness to pay for each of a pair of goods). We derive techniques for detecting implicit preferences in the two types of dataset. Most of our results establish conditions under which the data are sufficient to establish the direction of an implicit preference, i.e. whether the implicit preference is positive or negative with respect to some attribute. The identification is entirely through observing violations of rationality - either by observing an intransitive cycle, or by observing that evaluation of a good is sensitive to the identity of the other outcome being evaluated (the "comparator" good).

If we impose the restriction that implicit preferences can exist over only one attribute (for example just over gender), then the task is relatively straight-forward: we can infer the direction of the implicit preference by observing just one 3-element intransitive cycle, or observing a pair of evaluations. The task becomes more complicated when implicit preferences could exist over multiple attributes, for example, someone could have implicit preferences over both gender and qualification. Much of our formal work shows how such effects can be disentangled.

For each result we have tried to present a minimal set of assumptions, although this comes at a cost of somewhat greater complexity in the framework. Our principal can be summarized as follows:

Results for choice data:

1. **Right-triangle cycle.** Observing an intransitive cycle among three outcomes, where one outcome is *between* the other two (defined below), reveals that an implicit preference exists for at least one of the attributes on which the polar outcomes differ. If sufficiently many right-triangle cycles are observed, implicit preferences over a single attribute can be inferred.

2. **Figure-8 cycle.** Observing a figure-8 intransitive cycle (as in the introduction) reveals an unambiguous implicit preference for one attribute.

Additional results for choice data:

3. **Isosceles cycle in a ternary space.** In some settings it is natural to consider attributes with three values (e.g. male/female/no gender). Under a minor extension to our definition of betweenness we can identify an unambiguous implicit preference from a single cycle with three elements (an *isosceles* cycle).

4. **Aggregation.** We give conditions under which aggregate choice data (i.e., between-subjects data) can establish the extent of implicit preferences in the population.

Results for evaluation data:

5. **Scissor effect.** Observing that evaluation of some outcome changes when its comparator changes, in a manner that satisfies betweenness, reveals a disjunction among a set of implicit preferences over all of the outcome's attributes.

6. **Parallel scissors.** Observing that the evaluations of a pair of outcomes which differ only in one attribute move in opposite directions when there are symmetric changes to each of their comparators reveals an unambiguous implicit preference over that attribute.

Additional results for evaluation data:

7. **Joint and separate evaluation.** Observing that evaluations of a pair of outcomes which differ only in one attribute move in opposite directions when moving from separate to joint evaluation reveals an unambiguous implicit preference over that attribute.

Theoretical foundations for implicit preferences. We outline in the paper, and formally present in an Appendix a few natural foundations which generate implicit preferences.

8. **Separable implicit preferences.** We first introduce a general model, called separable implicit preferences, in which all the binary attribute space results hold (i.e. all results above except number 3).

9. ***Ceteris paribus* rules.** We show that a decision-maker who maximizes utility, but is constrained to choose outcomes which dominate in one attribute, will exhibit separable implicit preferences.

10. **Signaling.** We show that a linear-Gaussian model of a decision-maker who wishes to signal his preferences to an observer will exhibit separable implicit preferences, so long as the observer does not have a strong prior over the decision-maker's preferences on any of the attributes.

11. **Implicit knowledge.** We show that a linear-Gaussian two-system decision-maker, with imperfect knowledge of their own preferences, will exhibit separable implicit preferences in choice, provided the outcomes in the choice set differ by no more than two attributes.

2.1 Setup

The space of outcomes is $X = \{-1, 1\}^n$, representing the combination of $n$different binary attributes. The set of choice sets is $\mathcal{A} = 2^X \backslash \emptyset$.

Our fundamental assumption is that the choice set affects the weight put on each of the attributes. We thus assume a menu-dependent utility function, which depends on $x \in X$ and $A \in \mathcal{A}$, with the following functional form:

$$u(x, A) = v(x) + \sum_{i=1}^{n} x_i \kappa_i \Phi_i(A).$$

The first term, $v(x)$, represents menu-independent preferences. The second term represents implicit preferences, with three components:

- x_i is the polarity of attribute i for the outcome x,
- κ_i is the strength of the implicit preference – if κ_i is positive then the decision-maker has a positive implicit preference for attribute i, and vice versa,
- $\Phi_i(A)$ represents how much this choice set activates implicit preferences over attribute i.

This utility function can be read as saying that if a decision-maker has positive implicit preferences with respect to attribute i, then choice sets that are more revealing about attribute i will cause the decision-maker to value outcomes with that attribute less, and value outcomes without that attribute more.

We will make a series of assumptions about how revealingness depends on the choice set.

Definition 1. For any $x, x', x'' \in X$, x' is **between** x and x'' if for all i, either $x'_i = x_i$ or $x'_i = x''_i$.

Assumption 1 (Betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then*

$$\forall i \in \{1, \ldots, n\},\ x_i \neq x'_i \implies \Phi_i(\{x, x'\}) \geq \Phi_i(\{x, x''\}).$$

Assumption 2 (Equivalence)**.** *For any $x, x', x'', x''' \in \mathcal{A}$, if, for all $i \in \{1, \ldots, n\}$, $|x_i - x'_i| = |x''_i - x'''_i|$ then,*

$$\forall i \in \{1, \ldots, n\},\ \Phi_i(\{x, x'\}) = \Phi_i(\{x'', x'''\}).$$

Assumption 3 (Strong betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then*

$$\begin{aligned} &\forall i \in \{1, \ldots, n\}, \\ &\quad x \neq x'_i \implies \Phi_i(\{x, x'\}) \geq \Phi_i(\{x, x''\}), \\ &\quad x_i \neq x'_i \implies \Phi_i(\{x, x'\}) \leq \Phi_i(\{x, x''\}). \end{aligned}$$

Assumption 4 (Uniqueness)**.** *For every $x \in X$, $\Phi_i(\{x\}) = \Phi_i(\{x, x\})$.*

2.2 Choice Results

Proposition 1 (Right Triangle Cycle)**.** *For any $x, x', x'' \in X$, if x' is between x and x'', and $x \succeq x' \succeq x'' \succeq x$, with at least one relation strict, then the decision-maker must have a negative implicit preference ($\kappa_i < 0$) for one of the attributes which x has, and x'' does not.*

Proof. Normalize the attribute space such that $\forall i,\ x_i = 1$. Suppose, for contradiction, that u has weakly positive implicit preferences for all the attributes on which x and x'' differ. By betweenness, $\{x, x''\}$ is less revealing than $\{x, x'\}$ about all the attributes on which x and x' differ. So, using the definition of implicit preferences, this implies,

$$u(x, \{x, x''\}) \geq u(x', \{x, x''\}).$$

The same logic applies for the comparison between x' and x'', yielding:

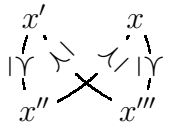
$$u(x', \{x, x''\}) \geq u(x'', \{x, x''\}).$$

But the observed choice between x and x'' implies that,

$$u(x, \{x, x''\}) \leq u(x'', \{x, x''\}).$$

If one of the three preferences is strict then one of these three inequalities is strict, yielding a contradiction. □

Proposition 2 (Figure 8 Cycle)**.** *For any $x, x', x'', x''' \in X$ (normalizing $x_i = 1, \forall i$), if (1) x' is between x and x'', (2) $x'''_i \neq x''_i \iff x_i \neq x'_i$, and (3) preferences are such that:*



with at least one preference strict, then $u(\cdot, \cdot)$ must have a negative implicit preference for at least one attribute on which x and x''' differ.

Proof. We start by noting that, if A and B are equally revealing about all attributes, then they must generate the same set of rankings of all elements.

$$(A =_i B, \forall i) \implies (u(x, A) \geq u(x', A) \iff u(x, B) \geq u(x', B))$$

Therefore the preferences invoked by the four pairs can be represented with just two different utility functions. We use u^V (vertical) to denote the preferences evoked by $\{x, x'''\}$ and $\{x', x''\}$, and u^D (diagonal) to denote the preferences evoked by $\{x, x''\}$ and $\{x', x'''\}$. Suppose, for contradiction, that there is no negative implicit preference for any of the attributes on which x and x''' differ. Then, because u^D is less revealing about the vertical attributes, u^D must be weakly more favorable to the North, in North-South comparisons, i.e.:

$$\begin{aligned} u^V(x) \geq u^V(x''') &\implies u^D(x) \geq u^D(x''') \\ u^V(x') \geq u^V(x'') &\implies u^D(x') \geq u^D(x''). \end{aligned}$$

But this yields:

$$u^D(x) \geq u^D(x''') \geq u^D(x') \geq u^D(x'') \geq u^D(x),$$

with one of these inequalities strict, which is a contradiction. □

2.3 Evaluation Results

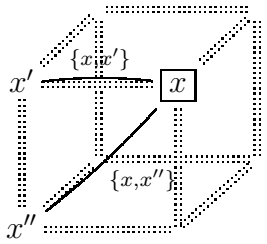
We now adapt the framework used for *choice* and apply it to *evaluation*. We mean by evaluation, for example, reports of willingness-to-pay, or some other continuous measure of value. We are interested in when more than one outcome is evaluated at once, because the comparisons affect implicit preferences. We thus define a menu-dependent value function $y(x, A)$, where x is an outcome, and A is the set of outcomes being evaluated, with $x \in A$. We furthermore assume that this function is identical to the menu-dependent utility function above, i.e. $y(x, A) = u(x, A)$.

Proposition 3 (Scissor)**.** *For any $x, x', x'' \in X$, with x' between x and x'' and*

$$y(x, \{x, x''\}) > y(x, \{x, x'\}),$$

then either (i) y has a positive implicit preference for some attribute i ($\kappa_i > 0$) on which x and x' disagree ($x_i \neq x'_i$); or (ii) y has a negative implicit preference ($\kappa_i < 0$) for some attribute i on which x and x' agree ($x_i = x'_i$).

Example 1. Consider the diagram below, representing two different evaluation sets, $\{x, x'\}$ and $\{x, x''\}$. If the evaluation of x increases with a change of comparator from x' to x'' this implies either a positive implicit preference for x's value on the horizontal dimension (for which revealingness has decreased) or a negative implicit preference for x's value on another dimension (for which revealingness has increased).



Proof. Consider the graphical case just above. Suppose, for the sake of contradiction, that y has a weakly negative implicit preference for every attribute on which x and x' disagree (for which $\{x, x''\} <_i \{x, x'\}$), and a weakly positive implicit preference for every attribute on which x and x' agree (for which $\{x, x''\} >_i \{x, x'\}$). Then, by the definition of implicit preferences, it must be the case that $y(x, \{x, x''\}) \leq y(x, \{x, x'\})$, contradicting the premise. □

Proposition 4 (Parallel Scissor)**.** *For some i, and $\underline{x}, \bar{x}, \underline{x}', \bar{x}', \underline{x}'', \bar{x}'' \in X$, with $\bar{x}_i = \underline{x}_i + 1$, $\bar{x}_j = \underline{x}_j$, $\forall j \neq i$, $\bar{x}'$ between $\bar{x}$ and $\bar{x}''$, and $|\bar{x} - \bar{x}'| = |\underline{x} - \underline{x}'|$, and $|\bar{x} - \bar{x}''| = |\underline{x} - \underline{x}''|$, if we observe*

$$\begin{aligned} y(\bar{x}, \{\bar{x}, \bar{x}'\}) &\geq y(\bar{x}, \{\bar{x}, \bar{x}''\}) \\ y(\underline{x}, \{\underline{x}, \underline{x}'\}) &\leq y(\underline{x}, \{\underline{x}, \underline{x}''\}), \end{aligned}$$

with one inequality strict, then $\lambda_i > 0$.

Proof. First note that, by equivalence, just two evaluation functions are invoked, denote them

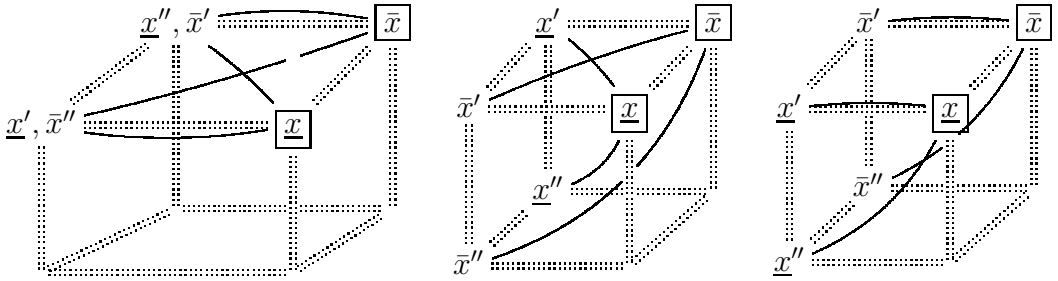
$$\begin{aligned} y^A(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}'\}) = y(\cdot, \{\underline{x}, \underline{x}'\}) \\ y^B(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}''\}) = y(\cdot, \{\underline{x}, \underline{x}''\}), \end{aligned}$$

from which we can rewrite the inequalities,

$$\begin{aligned} y^A(\bar{x}) &\geq y^B(\bar{x}) \\ y^A(\underline{x}) &\leq y^B(\underline{x}). \end{aligned}$$

Assume that $\lambda_i \leq 0$. But, by the monotonicity assumption, this implies B is weakly more favorable to $\bar{x}$ than $\underline{x}$, contradicting the two observed inequalities (assuming one is strict). □

This proposition yields a rich variety of tests for implicit preferences. These tests can be put into three categories, depending on whether $\bar{x}'$ and $\bar{x}''$ agree with $\bar{x}$ on the attribute of interest. Consider the following three diagrams, which illustrate the three types of parallel scissor effects, constructed around the outcomes $\bar{x}$ and $\underline{x}$ which differ on attribute i (e.g., gender). We normalize $\bar{x}_j = 1, \forall j$, and let $\bar{x}_i = 1$ (male) and $\underline{x}_i = 0$ (female). If the evaluations of $\bar{x}$ and $\underline{x}$ shift in opposite directions when their comparators undergo parallel transformations, this reveals the existence and direction of an implicit preference over attribute i.



Type 1 Type 2 Type 3

Type 1 $\bar{x}'$ agrees with $\bar{x}$ on attribute i, and $\bar{x}''$ disagrees. Thus the shift from $\bar{x}'$ to $\bar{x}''$ increases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$). Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when her comparator changes from female to male, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator changes from male to female.

Type 2 both $\bar{x}'$ and $\bar{x}''$ disagree with $\bar{x}$ on attribute i. Then the shift from $\bar{x}'$ to $\bar{x}''$ decreases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$). Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when her comparator becomes more similar on the non-gender dimensions, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator becomes more similar.

Type 3 both $\bar{x}'$ and $\bar{x}''$ agree with $\bar{x}$ on attribute i. Then the shift from $\bar{x}'$ to $\bar{x}''$ increases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$).Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when her comparator becomes more similar on the non-gender dimensions, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator becomes more similar. This third case may be the weakest method of detecting implicit preferences, because the change in revealingness might be expected to be small, given that there is no variation in the attribute of interest (attribute i) within either of the evaluation sets.

[note that joint-vs-separate is a subtype of Type 1]

Example 2. Consider a female and male candidate who are identical on all other attributes. If the female candidate's evaluation increases and the male candidate's evaluation decreases when evaluated jointly, we identify an implicit preference for male candidates.

3 Foundations of implicit preference

We discuss three formal models which would generate implicit preferences. Derivations are given in an Appendix.

3.1 Signaling

Suppose that you are concerned about the outward appearance of your preferences, for instance you might enjoy country music, but prefer your family not to know. A choice set that is more revealing about attribute i will tend to be one for which the observer's beliefs about your preferences over i are more sensitive to your choice. The more sensitive the observer's beliefs, the more the decision-maker will attempt to disguise their true motivations, therefore generating implicit preferences. We give a fully worked-out model in the Appendix: a decision-maker possesses, for each attribute, a coefficient representing their intrinsic preference, and a coefficient representing their concern about how other people perceive their intrinsic preferences. The sign of the second coefficient corresponds to the direction of the implicit preference that can be identified from choice.[11]

Some economists have argued that much social behavior is motivated by signaling concerns, for example that education is to signal ability (Spence (1973)), conspicuous consumption is used to signal status (Veblen (1899)), or generosity is used to signal altruism (Bénabou and Tirole (2006)).[12] If correct then demand for education, consumption and generosity should be lower in less revealing choice situations.

The signaling model can also be interpreted as self-signaling, as in Benabou and Tirole (2003) and Bodner and Prelec (2003), in which you distort your actions to persuade

[11]The model in the Appendix assumes that the observer has independent Gaussian priors over the intrinsic preferences. We assume that the observer's priors are mean-zero, and explain why betweenness can be violated when this is not true. We also assume a naive observer, i.e., the observer does not appreciate that the decision-maker has signaling motivation, but we believe that similar results would obtain in the equilibrium of a realistic model with a sophisticated observer.

[12]See Hanson (2008) for an expansive argument about the importance of signaling.

your own future self that you are generous, or clever, or hard-working. In these models, for the signal to be effective, the future self must be assumed to forget the present-self's motivations or circumstances.

3.2 Maximizing with *Ceteris Paribus* Rules

Implicit preferences could be generated by an ordinary decision-maker who is constrained by one or more rules, each of which requires that a certain attribute be preferred when all other attributes are equal. We call these *ceteris paribus* rules, and give a formal model of this type of decision-making in the Appendix. Each rule will manifest as an implicit preference, and therefore can be identified from behavior using the conditions we have derived.

This type of decision-making appears in a variety of real-world contexts: in a bureaucracy, rules are often explicitly written as ceteris paribus rules, e.g. "never appoint a male when there is an equally qualified female candidate."[13] Universities are often forbidden from discriminating on the basis of race (and are often thought to discriminate on attributes correlated with race). It also seems that many people take care to never *overtly* discriminate on the basis of race, sex, or political affiliation, but do allow those factors do influence their decisions when the comparison is less revealing. In individual decision-making we sometimes observe people following rules such as "you must always choose the diet version of a soda, when available."[14]

Viewed from the perspective of signaling these rules express an "innocent until proven guilty" philosophy, under which people are only penalized when their action incontrovertibly reveals a forbidden preference. This behavior is difficult to reconcile with the linear-Gaussian signaling model, in which the expression of implicit preferences varies continuously with revealingness.[15]

Finally, *ceteris paribus* decision-making is a special case of decision by "lexicographic semiorder", discussed in the Appendix.

[13]Or "fly economy class when it is available," or "if two bids are otherwise equivalent, choose the lowest bidder."

[14]It has commonly been observed that people adopt rigid "personal rules." For example: going to the gym at the same time every day; never making a withdrawal from your savings account; always forgoing dessert. Models which rationalize personal rules include Ainslie (1992), **?**, Bodner and Prelec (2003), **?**.

[15]Under the linear-Gaussian model, even when evaluating a man and woman side by side, who are otherwise equal, they would not receive the same evaluation: the intrinsic preferences and signaling preferences will be traded off, meaning any bias would be diminished, but not eliminated.

theory	typical evidence	typical findings
Freudian "deep psychology"	dreams, slips of the tongue, forgetting, jokes	sexual fixations
1970s social psychology[a]	influence of primes on judgment and decision-making	self-serving bias, social desirability bias
implicit motives[b]	Thematic Apperception Test (free response to a picture)	desire for power, achievement, emotional affiliation
implicit associations[c]	response time in an association task	discriminatory associations

Table 1: Some Theories of Subconscious/Implicit Motives

[a] In an influential paper **?** made the case that Subjects are sometimes (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response. **?** articulate some serious criticisms of this literature.

[b] See McClelland et al. The Achievement Motive, and Schultheiss and Brunstein (2010).

[c] Greenwald et al. (2009)

3.3 Implicit Knowledge

The idea that there are important subconscious influences on behavior did not become widespread until the 19th century (Ellenberger (1970)). Since then there have been many theories of such influences, and various techniques of identifying them, a few are summarized in Table 1. All of these techniques remain controversial. We consider our method to be an alternative means of identifying unconscious influences on behavior: a factor is unconscious if its influence judgment systematically differs with the revealingness of the situation.

For example, suppose we find that judgment of a drink's flavor is influenced by its color; judgment of a person's honesty is influenced by the clothes they wear; judgment of the value of a house is influenced by the glossiness of the brochure; or judgment of the severity of a crime is influenced by whether it was committed by a Republican or Democrat. Each influence could be conscious or unconscious: we can test for the consciousness of each of these influences by seeing if they vary as the revealingness is varied - e.g., by eliciting judgments side by side.

In an Appendix we state a model with two stages: you first get an "intuition" about the value of each outcome, and then you adjust each intuition, based on additional considerations, before making a final decision. Formally, two mental processes work sequentially, each forming an estimate of value, but each with access to private information. This implies that you have intuitions that are informative because they

incorporate knowledge to which you do not have conscious access. This model predicts systematic *comparison* effects in decision-making, because each new element in the choice set can reveal different information about the implicit knowledge. The model in this paper is a simplified version of that given in Cunningham (2014).

We show that the model meets our definition of implicit preferences in choice when the outcomes differ in at most two respects. If we discover a positive implicit preference for some attribute, e.g. for male job candidates, this implies one of two things: (1) that the decision-maker believes gender to be irrelevant, but has unconscious positive associations with men; (2) that the decision-maker does believe gender to be relevant, but has unconscious negative associations with men (and hence the difference in evaluation declines more in revealing choice sets). When the outcomes differ in more than 2 attributes then the techniques we use (triangle and figure-8s) are not appropriate for identifying implicit knowledge in this model. We discuss this further in the Appendix.[16]

We believe that this model can give a good account of framing effects: they are due to associations that are *normally* relevant, but irrelevant in the current context. This corresponds to a common description of biases, rarely formalized, as being byproducts of rational heuristics (**?**). We give examples and further discussion in section 6.3.

A final variation on the implicit knowledge model would be one with *motivated* bias: we sometimes talk of people deceiving themselves into making decisions by finding an excuse for their preferred outcome.[17]

[16]The model could also explain implicit race or sex bias under the assumptions that (1) people have learned associations with race or sex regarding which they have imperfect knowledge, and (2) people believe those associations to be are irrelevant for typical decisions. This explanation is similar to common descriptions of behavior in the implicit association test (IAT) - that people are unaware of their race-based instincts, and attempt to correct for them. However if this explanation is correct it remains a puzzle why people would remain unaware of their associations despite relatively frequent experience with making race-based and sex-based decisions.

[17]It would be possible to write down a model with an expert and a decision-maker, such that the expert's bias will be mixed into their advice, and predict that the expert's preferences will manifest as implicit preferences detectable in the decision-maker's behaviour. However it is much easier to achieve this pattern in decisions if the decision-maker is imperfectly informed about the expert's biases, otherwise the decision-maker could simply correct the advice to account for their bias. Thus there remains an element of this self-deception that is unexplained, because it seems that most people are aware of the direction of their own biases – e.g., in favor of their preferred political party, in favor of unhealthy foods, against physical exertion – yet those biases still seem to distort their judgments.

3.4 Distinguishing Between Interpretations

The interpretations given above cannot be distinguished on the basis of simple binary choice or evaluation, because they all fit the general model of implicit preferences. However we discuss a variety of ways to distinguish between them, with: (1) a change in incentives or observability of the choice, (2) variation in the identity of the preceding choice set, (3) variation in the order of preceding choice sets, or (4) choice from larger choice sets.

First, under the "signaling" interpretation the decision-maker will be sensitive to the implementation of their decision: the strength of implicit preferences should therefore be increasing in the probability of the decision being implemented (because this decreases the relative importance of the signaling motive), and decreasing in the probability of the decision being observed (which increases the relative importance of the signaling motive). Under an "implicit knowledge" interpretation neither change should affect the relative weight of implicit and explicit preferences.

Second, the models have different implications about the effects of preceding choice sets. Under implicit knowledge if some choice set is completely revealing about attribute i then the decision-maker will learn their preference over i, and so this should eliminate implicit preferences over i in subsequent choices. For example, if I am asked to choose between $\binom{\text{MALE}}{\text{MBA}}$ and $\binom{\text{FEMALE}}{\text{MBA}}$, this will reveal to me my implicit bias, and I should not exhibit any implicit preferences over gender in subsequent questions, for example in tradeoffs between MALE/FEMALE and OXFORD/CAMBRIDGE. This is not true in the signaling model.[18]

Third, the *ceteris paribus* model implies that choices will set precedents, and so constrain subsequent choices, leading to *order* effects that would not occur in the implicit knowledge model. Consider the following two sequences of three choice sets, which are identical except for the order of the first two sets:

$$\left(\left\{\binom{\text{FEMALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\},\left\{\binom{\text{FEMALE}}{\text{PHD}},\binom{\text{MALE}}{\text{MBA}}\right\},\left\{\binom{\text{MALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\}\right)$$
$$\left(\left\{\binom{\text{FEMALE}}{\text{PHD}},\binom{\text{MALE}}{\text{MBA}}\right\},\left\{\binom{\text{FEMALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\},\left\{\binom{\text{MALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\}\right)$$

A decision-maker with implicit knowledge will condition on the information learnt in the prior choice sets, but the order of those choice sets should not matter. In contrast

[18] This point courtesy of Luke Miner.

a *ceteris paribus* decision-maker who is not allowed to choose a male over a female, and who chooses the male candidate in the first choice set, will be forced to choose, in the third choice set, whichever candidate has the qualification which the male had in the first set. This follows from assuming that they are forbidden from making a choice which, when combined with prior choices, implies a violation of a *ceteris paribus* constraint through transitivity.[19]

Finally, the models differ in their predictions about choice from 3-element choice sets. Consider the following two 3-element choice sets, represented spatially for clarity:

$$\left\{ \begin{matrix} \binom{\text{FEMALE}}{\text{MBA}} & \\ \binom{\text{MALE}}{\text{MBA}} & \binom{\text{MALE}}{\text{PHD}} \end{matrix} \right\} \quad \left\{ \begin{matrix} & \binom{\text{FEMALE}}{\text{PHD}} \\ \binom{\text{MALE}}{\text{MBA}} & \binom{\text{MALE}}{\text{PHD}} \end{matrix} \right\}$$

A *ceteris-paribus* decision-maker with a rule not to choose a man over a similar woman, and a sufficiently strong implicit preference for men, would choose $\binom{\text{MALE}}{\text{PHD}}$ from the left-hand choice-set, and $\binom{\text{MALE}}{\text{MBA}}$ from the right-hand one, a violation of GARP. A decision-maker with implicit knowledge would never make such choices because both choice sets would be equally informative about her unknown preference parameters, and so would both evoke the same set of preferences.

4 Discussion

4.1 Comparison of Alternative Ways of Identifying Implicit Preferences

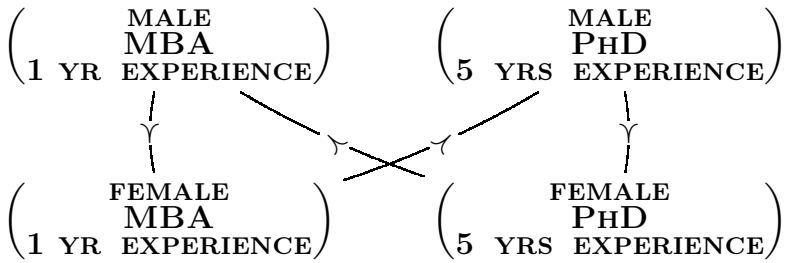
Our theoretical exposition takes as given that we know what the decision-maker would choose or what her evaluation would be in each choice or evaluation set. In practice, of course, choices and evaluations must be observed or elicited, opening up a number of interesting methodological issues.

[19]The two sequences are chosen so that, by the third step, the history is identical, but the order varies. A similar effect of precedents seems natural in the signaling model, but it is somewhat more difficult to model the desire for consistency. Interestingly, the order effects generated by *ceteris paribus* decision-making allow for strategic effects in agenda-setting: the decision-maker's final choice can be manipulated by gradually revealing alternatives, and eliciting intermediate choices.

4.1.1 Choices in a Binary Space

History Effects in Within-Subject Studies. If we do not know what a given subject will choose from each choice set, the natural approach is to measure it by presenting her with all relevant choices and recording what she does, i.e. collect within-subject data. However, there are reasons to expect significant history effects: one's decision is influenced by the prior choice set (in the "implicit knowledge" story), or by prior choices (under the other foundations). This makes within-subject data more complicated to interpret.[20] Under some assumptions, separating choices with decoy questions or time intervals (if she is forgetful), or making her feel un-observed by exploiting administrative data (if she has a signaling motive) could alleviate the problem.

Calibration. Even a decision-maker with substantial implicit preferences will not reveal them if we do not observe the right kind of choices. For example, if the decision-maker has a significant preference for MBAs over PhDs, then the choice-sets presented in the introduction might not detect any implicit preference over gender, even if one exists, because they will always choose the candidate with the MBA. This is essentially a calibration problem. Fortunately, by varying an additional attribute we may be able to bring the decision-maker closer to indifference:[21]



[20] Of course this challenge is not unique to our proposed approach, and is the reason why between-subject designs are more commonly used in economics and psychology experiments.

[21] A common way to deal with such calibration problems is by using "multiple price list" to find the indifference point between two bundles of goods, e.g. answering "what value of x would make you indifferent between $\binom{1 \text{ can spinach}}{3 \text{ cans corn}}$ and $\binom{x \text{ cans spinach}}{1 \text{ can corn}}$?" This is sometimes called "matching." A disadvantage is that the act of choosing an x could be psychologically different than making a binary choice, and so have less external validity when predicting choice behavior. Also note that here we are treating "1 year experience" and "5 years experience" as two poles of a binary attribute.

Heterogeneity in Between-Subject Studies. If we instead use between-subject data then we have the problem that between-subject heterogeneity of preferences could again make implicit preferences undetectable, no matter how well-calibrated is the choice set. To establish the existence of at least one decision-maker with intransitive preferences over outcomes a, b, c the aggregate choices must violate the triangle inequality: for a cycle of 3 elements the average choice probability must exceed $\frac{2}{3}$, (i.e., $P(a \succ b) + P(b \succ c) + P(c \succ a) > 2$).[22] This problem is alleviated when there is reason to believe that most of the population will have aligned preferences over a certain attribute: e.g., if most people would hire a woman over an equally qualified man.

Trembling-hand choice errors (where with some probability ϵ subjects mistakenly choose the less-preferred option) will push choice probabilities towards $\frac{1}{2}$, making it harder to reject transitivity but also implying that a rejection is robust to such errors.

Indifference Collecting data on indifferences can help. The standard form of the triangle inequality assumes all strict preferences, but an equivalent form can be derived for weak preferences. For any three elements i, j, k, we violate the condition if $P(i \succsim j) + P(j \succsim k) + P(k \succ i) > 2$.[23] For four elements i, j, k, l we violate the condition if $P(i \succsim j) + P(j \succsim k) + P(k \succsim l) + P(l \succ i) > 3$. The advantage of collecting data on indifference is that in many cases we expect people to be indifferent in direct comparisons (e.g. between equally qualified male and female candidates, or between equivalent gambles framed differently). For example, if all subjects are indifferent along the verticals of a figure-8, then *any* difference in choice proportions along the diagonals will violate the condition. There are two weaknesses however. First, we are not aware of a widely-accepted method for collecting indifference data in an incentive-compatible way. Second, for a given three elements there are three variants of the three-element condition (varying the identities of i, j and k), likewise there are four variants of the four-element condition, and one can construct examples which only violate some variants.[24]

[22]For intuition, note that if $\frac{2}{3}$ of subjects report $a \succ b$, $\frac{2}{3}$ report $b \succ c$ and $\frac{2}{3}$ report $c \succ a$ this could be rationalized by a subject pool in which $\frac{1}{3}$ of subjects have transitive preference $a \succ b \succ c$, $\frac{1}{3}$ $b \succ c \succ a$ and $\frac{1}{3}$ $c \succ a \succ b$). If the cycle has four elements the requirements are stronger: the average choice probability must be greater than $\frac{3}{4}$. To *statistically* establish cyclical preferences in a finite sample will tend to require higher fractions because of sampling variation. The problem of heterogeneity is reflected in the observation that, although there are many well documented and strong framing effects, there are few clear demonstrations of intransitive choices in the laboratory (Regenwetter et al. (2011)).

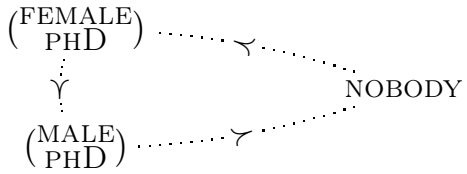
[23]Proof: $1 = P(i \succsim k) + P(k \succ i) \geq P(i \succsim j \succsim k) + P(k \succ i) = P(i \succsim j) + P(j \succsim k) - P(i \succsim j \cup j \succsim k) + P(k \succ i) \geq P(i \succsim j) + P(j \succsim k) + P(k \succ i) - 1$.

[24]For example, if all subjects have $a \sim b \sim c \succ a$ then we violate the condition with $i = a, j = b, k = c$ but not with $i = b, j = c, k = a$.

We suggest that researchers pre-specify which test they will run; or report all variants (possibly with a multiple-hypothesis correction).

4.1.2 Choices in a Ternary Space (Isosceles cycles).

Some choices do not naturally fit into a binary space. Suppose we observe a recruiter who would hire a female PhD over a male PhD, and hire a male PhD over hiring nobody, but would also hire nobody over hiring a female PhD, i.e. an intransitive cycle:



These choices seem to reveal an unambiguous implicit preference for male over female employees, yet do not have a natural analysis in a space composed only of binary attributes. We discuss in an Appendix how the binary model can be extended to license such an inference from cycles like the above, which we call *isosceles* cycles. An isosceles cycle is more parsimonious than a figure-8 cycle, having only three outcomes. In many cases we may also think of this method as more sensitive, in the sense that it induces greater variation in revealingness. For example a choice from $\{(^{\text{FEMALE}}_{\text{PHD}}), \text{NOBODY}\}$ seems intuitively less revealing about gender preferences than is a choice from $\{(^{\text{FEMALE}}_{\text{PHD}}), (^{\text{MALE}}_{\text{MBA}})\}$).

Nevertheless in this paper we principally concentrate on outcomes which can be represented in a binary space, for a number of reasons: (1) identifying a set of binary attributes in a set of outcomes often is less controversial; (2) for each isosceles cycle that identifies an implicit preference, a corresponding figure-8 cycle can often be constructed.[25] Finally, isosceles cycles could occur for reasons other than the existence of implicit preferences, if for example decision-makers are sensitive to the range of outcomes in a choice set, as in the theory of **?**. As we argue below, a figure-8 cycle is difficult to explain with existing theories of decision-making, and so is distinctive evidence of implicit preferences.

[25] In the example above, by replacing the "nobody" outcome with candidates who have MBAs.

4.1.3 Evaluation.

Using data from evaluation, instead of choice, will tend to be more sensitive to implicit preferences for three reasons.

Variation in Revealingness. Evaluations allow for greater variation in revealingness. This is because we can measure implicit preferences over an attribute using data on evaluations of choice sets which include only one realization of an attribute, for example by comparing evaluations among groups that are men-only, women-only, and mixed, whereas inference from choice can identify implicit gender preferences only from mixed sets.

Calibration. Second, with evaluation calibration problems largely disappear, i.e. the method can detect even very subtle implicit preferences, while, as noted above, choice data can only detect implicit preferences that are large enough to change the ranking of outcomes.

Power. Third, evaluation can be continuous, rather than discrete, tending to increase statistical power.

Disadvantages. A disadvantage of evaluation is that it may be less natural in domains where choice is more common, and therefore experimental findings would have lower external validity. Additionally, a choice is explicitly comparative, forcing subjects to consider every element of the choice set, while when forming an evaluation subjects do not have to consider every element of the evaluation set, yet will reveal their implicit preferences only if they do so.

Heterogeneity. Suppose we observe average evaluations over a population–as would occur in a between-subjects experiment–how does this affect our analysis? In particular, if we treat the average evaluations as those of a representative agent, and infer the implicit preferences of that agent, what can we then conclude about the population? If the direction of implicit preferences are not aligned within the population (i.e., if some people have a strictly positive implicit preference for attribute i, and others have a strictly negative one), then a representative agent may not exist, i.e., there may be no single set of implicit preferences which rationalize the average evaluations. However we conjecture that if implicit preferences are aligned then a representative agent will

exist, and thus the population's implicit preferences can be identified with the implicit preferences of that agent.

4.1.4 Sequential Evaluations

We often observe people making evaluations in a series: bidding on a series of paintings at auction, scoring a series of gymnastic performances. If we are willing to assume that the evaluation set consists of the current outcome under consideration plus the most recently considered outcome, then it is straightforward to apply our existing results for evaluation. We provide more details in Appendix A.3.

4.1.5 Other Issues

In the Appendix we discuss the relationship with other types of cycles: equilateral cycles, and cycles which indicate non-separable implicit preferences (section E.1), and extension to larger choice sets (section E.2).

4.2 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 divide alternative accounts into three classes, 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.[26] However 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. For example sensitivity to race would depend on the distribution of black and white elements. However in our model sensitivity to race will instead depend on the distribution of the *other* attributes - e.g., a decision-maker with implicit racial preferences would become more sensitive to race when the distribution of other attributes such as education becomes

[26]For example in **?** sensitivity is positively related to the range of values on an attribute, in Bushong et al. (2014) it is negatively related to the range, in **?** it is negatively related to the average, and in **?** it is - roughly - negatively related to the proportional range (range divided by the average).

more dispersed. None of the recent contingent-weighting models is consistent with a figure-8 intransitivity.[27]

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 - for example the difference in WTP for high-quality and low-quality goods tends to be higher in joint evaluation. Hsee et al. (1999) argue that this increased sensitivity is a general feature of joint evaluation, called "evaluability".[28] Again, this is a quite different principle to that used in implicit preferences. This mechanism could generate isosceles intransitivities and joint-separate differences in evaluation. However it could not generate a figure-8 cycle, by an analogous argument to footnote 27. See **?** for a Bayesian rationalization of increased sensitivity in joint evaluation.

Inference from the choice set. We have assumed that the attributes of one outcome are not informative about the value of other outcomes in the choice set. If they were informative then inference from the choice set could in principle rationalize *any* pattern in choice. The relevant question is what types of prior beliefs could generate the patterns we observe, and whether those beliefs seem realistic. Suppose we observe a cycle in choice among job candidates who vary in both race and in the school at which they

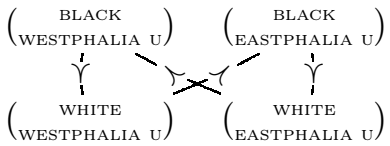
[27]Formally, suppose the utility function is separable in each attribute, in the sense that it can be written as,

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

where a_i^j is the ith attribute of the jth 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 **?**, but that model cannot generate intransitive cycles in binary choices with two attributes.

[28]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. **?** 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 **?**

studied:



These decisions could be rationalized by a decision-maker who (1) believes black candidates are better than white candidates, all else equal; but (2) believes that white candidates typically go to better schools, and therefore infers the quality of the school from the choice set. Thus in the diagonal choice sets they will prefer white candidates not because they are white, but because they went to the school that white people go to. In practice we believe that this alternative explanation of implicit preferences is not a realistic concern in most of our applications because (1) most examples we discuss use familiar attributes, so the scope for learning from the choice set seems small; and (2) the explanation requires that the *intrinsic* value of an attribute be opposite to its *informational* value (in this case, being white is a negative signal about the person, but it is a positive signal about the things which covary with being white).[29]

Inattention / Heuristics. Because much of our identification comes from comparing simple to complex choices (or direct to indirect choices), we may worry that inconsistencies are due to variation in complexity, as in models of inattention (**?**, **?**, **?**). It is intuitive that a decision-maker could become less sensitive to an attribute in a more complex choice situation, however we have not been able to find an inattention model in which an increase in complexity causes the polarity of an attribute to *reverse*, as necessary for the figure 8.[30]

[29]To explain a figure-8 with indifferences on the vertical comparisons, the intrinsic value of the vertical attribute must be zero and the informational value be non-zero, for example if race is believed to have no value in itself, but white students tend to go to better colleges.

There are cases where informational effects are certainly important: e.g., suppose one attribute is "Old Grouse" vs "Johnny Walker", and the other attribute is "labelled as Whisky of the Year" vs "no label." Naturally a decision-maker is indifferent about which bottle has the label, when the bottles are of the same brand, but strictly prefers the bottle with the label when they are of different brands. Our assumption is that attributes are informative about the *token*, not the *type* of an outcome.

[30]As was the case with inference, a figure-8 with indifferences could come from inattention if sensitivity to an attribute goes to zero in simple choices; though we are not aware of an inattention model with this feature.

4.3 Other Measures of Implicit Preference

We discuss a number of measures. An influential paper, Dana et al. (2007), reports a variety of experiments which show that pro-social choices are affected by "wiggle room." Each of their experiments falls under a different heading in the classification that follows.

Rationalization. Cherepanov et al. (2013) (CFS) propose a model of "rationalization" which is related to ours. Agents possess both a *true* preference relation and a set of *rationalizable* preference relations. A decision-maker will choose the item which is her favorite among those that would be chosen by at least one of the rationalizable preferences.[31]

The CFS model has a similar spirit to our model: their "rationalizable" preferences roughly corresponds to our "explicit" preference.[32] There are two important differences: (1) while CFS study choice among atomic elements, we study choice among bundles of attributes, making it easier to extrapolate behavior to new situations under our approach (for example detecting an implicit racial preference among one set of candidates has implications for choices among a completely different set). (2) We allow for choice to be a continuous mixture of implicit and explicit preferences, while in CFS the effects are binary: either a choice is rationalizable or not (the *ceteris paribus* model shares that feature with CFS).

Adding Noise (list elicitation and random response). Some experiments measure how preferences vary when noise is added to a decision. In the "list elicitation" method subjects are given a set of statements and record just the number of statements that they agree with.[33] In the "random response" method subjects are given one state-

[31]The principal working example is the following vignette: *"Dee decides to take time off from work to see a movie. However, prior to leaving the office she is informed that a colleague is in the local hospital and can accept visitors that afternoon. Dee reconsiders her decision to go to the movie and, instead, stays at work."* Dee's choices violate the weak axiom (WARP). Under the CFS model we can infer two facts: (1) Dee has the "true" preference order,

$$\text{MOVIE} \succ \text{WORK} \succ \text{VISIT SICK FRIEND},$$

but that (2) none of Dee's "rationalizable" preferences rank MOVIE above VISIT SICK FRIEND.

[32]In addition our *ceteris paribus* model, when there is a single ceteris paribus rule, obeys WWARP, the Weak Weak Axiom of Revealed Preference, the axiom which characterizes the CFS model (or, more generally, a lexicographic semiorder, discussed in Manzini and Mariotti (2012)).

[33]Miller (1984) the technique is also called "item count" or "randomized response." A post on Andrew Gelman's blog (Gelman, 2014) surveys some empirical work with these techniques and gives

ment, and then flip a coin with the instructions to mark "yes" if either (a) the coin lands heads (unobserved by the experimenter), or (b) they agree with the statement.

Under a signaling model these experiments could help identify implicit preferences - loosely reasoning that noise lowers the revealingness of a decision - so these techniques should reveal implicit preferences when compared with responses to the same questions asked separately.

A problem with both of these techniques is that, although adding noise reduces the incentive to distort, at the same time it increases the *ability* to distort, because the noise is private information to the decision-maker, allowing the decision-maker to misreport the noise. This means that adding noise has an ambiguous effect on reporting. This has been found in the data: for example, John et al. (2013) found that the random-response method did not increase the fraction of people who admitted to an embarrassing statement (in this case, admitting having cheated on an earlier test), in fact it decreased the number who admitted to it. John et al. conjecture that this was because some subjects answered "no" even when the coin landed heads-up (when they should have answered "yes", if following the instructions) due to a strong desire to signal that they did not cheat.[34] Thus either an increase *or* a decrease in reporting under these protocols can be interpreted as evidence for under-reporting in the ordinary protocol.

One solution to this problem is to add noise only after subjects make a decision, instead of letting subjects to add the noise themselves. This is used by Dana et al. (2007): they found that when decision-makers faced a chance of a donation decision not being implemented (and their decision was not observed by the beneficiary, only the implementation) then they tended to make more selfish decisions.

Verbal Explanation of Decisions. A series of papers has used verbal explanations of the decision-process as the dependent variable in a manipulation. Subjects are first asked to make a decision between two outcomes (bundles of attributes), and then asked what factors were most important in their decision. Papers in this literature typically report finding that (a) some attribute affects the decision without being described as important, while (b) another attribute that is *correlated* with the first attribute is

a pessimistic summary of their usefulness.

[34]The same logic holds for the item-count technique: when asked to sum the statements that they agree with, subjects have an increased ability to distort their answers. Gelman (2014) mentions some experiments that find this perverse effect.

described as important. For example Hodson et al. (2002) find that, in choice among black and white college applicants, subjects reported being uninfluenced by race, but when the white applicant had better grades then subjects were more likely to rate grades as an important factor.[35]

These studies are clearly related to the method advocated in this paper, but differ in using verbal judgments rather than choices. For instance, under a signaling interpretation the decision-maker is reporting the weights they put on attributes directly rather than weights being inferred by an observer.

Choice over Choice Sets. A variety of studies find situations in which subjects strictly prefer smaller choice sets (i.e., they will pay to avoid being given an additional alternative). In Dana et al. (2006) and Lazear et al. (2012) subjects have the choice whether to play a dictator game, or opt out of it at some cost, and many choose to opt out. Andreoni et al. (2011) similarly find that people are willing to pay to avoid a charity collector. These have a natural signaling interpretation: the decision-maker prefers to leave money on the table than to make a selfish choice that is observed by the recipient. In our signaling framework we identify concern for reputation via changes in choices or evaluations between more or less revealing situations, while here it is identified by willingness to pay to avoid a revealing situation.

Signalling and Crowding Out. Benabou and Tirole (2003) state a model in which providing an incentive for an action can change the signaling value of that action. In particular they predict a u-shaped effect: incentives decrease the signaling value when the action is rare (or unexpected), and increase the signaling value when the action is common (or expected). This occurs when the observer's priors regarding the actor's preferences are single-peaked - implying that an action is least informative about one's preferences when the observer puts a 50% chance on you performing the action (informativeness here means the difference in posterior means). They thus predict that providing an incentive for a pro-social act (e.g. giving blood) can crowd out the signaling incentive if the act is rarely performed (unexpected), because it causes the

[35]Interestingly Norton, Vandello & Darley (2004) use the same technique and find the opposite effect - a pro-black bias - perhaps because of difference in subject pools. Norton, Vandello and Darley (2004) find that, in a choice between candidates for a job in construction, when the female candidate had less education, then subjects were more likely to rate education as important. Norton (2010) found that, in a choice between magazines, when the magazine with swimsuit photos also had articles on sport, then subjects were more likely to rate sports-coverage as important.

act to become less diagnostic about one's pro-sociality.

Their results are related to the results from our signaling model: both show how changing the bundling of attributes can change the signaling value of a choice. They consider adding a feature with a known positive value, i.e. an incentive. Our model deals with adding features that have unknown values (with mean-zero expected value). We therefore consider their approach to be complementary.[36]

Choice of information. A variety of biases seem to be identified by choice to be strategically *ignorant*. A good example is reported in Dana et al. (2007)'s "hidden information" experiment. They find that subjects' choices are sensitive to the payoffs of their partner (a standard finding), but that, in addition, subjects prefer to remain ignorant about how their partner's payoff depends on the choice; and that when subjects are ignorant they tend to make the choice which maximizes their own payoff.[37]

Dana et al. refer to an "illusory preference for fairness." We might say that the possibility of not revealing the payoffs of the partner makes the decision under the treatment "less revealing," though the example does not fit neatly into our binary attribute framework. Their result is striking in particular because choosing to reveal should make the decision maker weakly better off (she is better able to trade off fairness and efficiency if she knows the payoffs), and strictly so unless she is very selfish. An interpretation which relates revealingness to the number of steps of reasoning required to determine if an action was selfish or not seems intuitively appropriate here.

Rabin (1995) proposes that people often treat moral considerations not as ends in themselves, but as constraints on maximizing self-regarding preferences. This motivation can be identified in information-seeking behavior: such people will choose to avoid information whenever that information will, in expectation, lead to decisions that lowers their selfish utility.[38]

[36]Bodner and Prelec (2003) also have a self-signaling model. Mijović-Prelec and Prelec (2010) has a useful discussion on the difference between self-deception and merely having biased beliefs.

[37]Subjects choose between allocations of money, denoted (self,other). Control subjects had to choose between a fair allocation $(5, 5)$ and an unfair allocation $(6, 1)$. Treatment subjects were given a choice between $(5, X)$ and $(6, Y)$. Pressing a button would reveal X and Y, which were either equal to 5 and 1, or 1 and 5 respectively. The generic pattern of choices was to choose $(5, 5)$ under the control, and $(not\, reveal, (6, Y))$ under the treatment, consistent with the uncertainty giving some "moral wiggle room."

[38]For example I might sincerely believe that the suffering of animals is not sufficient to become a vegetarian, but also avoid learning more for fear that I might revise upwards my estimate of suffering, and be forced to stop eating meat. This theory will only have empirical bite if the selfish payoff is nonlinear in beliefs (e.g., if my decision to eat meat is all-or-nothing). A more general treatment of this

Automatic Responses. Nosek et al. (2011) survey experimental measures of implicit social cognition. Most of those measures ask subjects to perform a classification task quickly, and test whether classification speed or accuracy is affected by semantic relationships among the stimuli used. Most famous is the Implicit Association Test, but there are many other variants.

5 Existing Data on Implicit Preferences

A small number of papers come close to measuring implicit preferences in the way we define. For the interested reader, Appendix D discusses the strengths and weaknesses of each in detail, we summarize our arguments briefly here.

Snyder et al. (1979) report an experiment which compares direct and indirect choices as a "general strategy for detecting motives that people wish to conceal." Their name for this general phenomenon is "attributional ambiguity," and their informal description comes very close to our basic analysis of revealingness and implicit preferences. Subjects chose between sitting in one of two booths, in each of which a movie was being shown. Subjects could see that each booth already contained one person: in one booth they were seated in a chair, in the other booth in a wheelchair. The treatments varied in whether the booths were showing the same, or different, movies. When the movies were the same, 75% (18/24) of subjects sat with the handicapped confederate, while when they were different only 33% (8/24) did so, suggesting an implicit preference against sitting with the handicapped individual: they write "avoidance of the handicapped ... masquerade[d] as a movie preference." However, in fact a rational decision-maker with strong preferences over movies and weak preferences over which confederate to sit with will exhibit the same pattern of choice. Instead we need to check for a figure-8 cycle, keeping in mind the appropriate triangle inequality. We find that the triangle condition is not violated, i.e. the choices observed can be rationalized by subjects with heterogeneous transitive preferences, and we provide an example.

In a design very similar to what we propose in this paper, Exley (2015) studies "excuse-driven risk preferences," finding that risk-preferences seem to change in a self-serving way when choosing between payoffs for self or charity. When the charity payoff is risky (and the self payoff riskless), subjects appear risk averse; but when the self payoff is risky (and the charity payoff riskless), then decision-makers become relatively

could identify, from choices over distributions of information (as in Kamenica and Gentzkow (2008)), a set of outcome-preferences separate from the preferences revealed in ordinary choice.

risk-loving. We show that Exley's subjects do exhibit implicit preferences in line with our definition: under a mild assumption her data reveal "two triangles" that identify an implicit preference for self-payoffs over charity-payoffs. Some subjects also exhibit a "figure-8" cycle revealing an additional implicit preference against risk.

DeSante (2013) finds racial bias in an experiment where subjects are asked to set welfare payments for applicants who vary in various attributes. In his experiment two applicants are evaluated at once, allowing us to test for implicit preferences. Reanalyzing the data we find evidence that his subjects have *implicit* biases: a negative implicit preference for black candidates, but also a negative implicit preference for candidates with high "work ethic."

Bohnet et al. (2015) study whether a decision-maker's choice between candidates for a task becomes more or less sensitive to certain attributes–gender and past performance–when the choice is either between an individual candidate and an unobserved "pool" alternative, or between two candidates and the pool (a paradigm closely related to joint and separate evaluation). They find that "disadvantaged gender" candidates are less likely to be selected when considered individually than when considered alongside an advantaged gender alternative. On the contrary, low ability candidates are more likely to be selected when considered individually than when the alternative is a high ability candidate.

While intuitively the variation in frequency of certain choices points to implicit preferences (as we have argued, considering multiple candidates increases revealingness with respect to their attributes), in fact it is not possible to infer implicit preferences from these data: heterogeneous transitive preferences can generate the patterns of choice observed. We do however show that Bohnet et al. (2015)'s subjects exhibit violations of WARP that point to implicit preferences, though are harder to assign to a given attribute. The most natural way to test for implicit preferences in their paradigm would be to collect *evaluation* data, and conduct our scissors tests.

6 Applications

In this section we discuss how certain anomalies in decision-making, across a variety of domains, can be interpreted as the expression of implicit preferences.

6.1 Implicit Discrimination

Since the mid 20th century it has become common, among philosophers and cultural theorists, to claim that our beliefs and preferences are subtly influenced by the culture we live in, in a way that is biased towards existing power structures. For example, that unspoken assumptions make it difficult to question existing class, sex, and race relations. Much intellectual work in Marxism, feminism, and race studies has tried to identify biases in different parts of everyday thought and culture. However the interpretation of the evidence, for example the analysis of texts, is notoriously disputable.

More recently an empirical case has been made for the implicitness of discrimination by comparing verbal reports of preference with actual behavior. This takes two forms: studies which find large differences in how people are treated, depending on their race or gender;[39] and studies which find differences in automatic associations.[40]

These approaches equate explicit preferences with stated preference, and implicit preference with revealed preference. Our claim is that we can identify *both* just from revealed preferences. Most closely related to our theory is Gaertner and Dovidio's (1986) work on "aversive racism" - they argue that most people in the US are no longer overtly racist, but their judgment and decisions reflect racial influences in hidden ways.

Our theory has a simple implication for experimental design: by varying revealingness we can determine the degree to which discrimination is implicit. Existing designs can be extended by asking subjects to consider two outcomes instead of one - either simultaneously or in sequence. This can also be applied in field experiments, as long as it is reasonable to believe that the subject will find the two outcomes to be salient comparisons - for example, sending two CVs in application for a job, or sending two testers to apply for an apartment or mortgage.[41] Put simply: between-subject studies and within-subject studies are expected to show different outcomes, and the difference will tell us about implicit preferences.

If a large part of discrimination is implicit, in our sense, this implies that it will be more pronounced in situations that are less revealing. In particular, we would expect discrimination to be stronger when cases are evaluated one-by-one, than when they are

[39] See Mullainathan (2015) for a selection of studies which find large effects of race discrimination.

[40] Most famously the "Implicit Association Test," which finds that most people perform significantly better at a task which asks them to associate white faces with positive words, and black faces with negative words, than the opposite combination.

[41] We have piloted an experiment in which subjects are shown two defendants, and asked to suggest appropriate sentences, varying the race and crime used. Preliminary results find little explicit racial discrimination, and significant implicit racial discrimination.

evaluated in groups. Consider two hiring policies: one in which job applications are evaluated as they arrive, and one in which applications accumulate and are evaluated in groups. We expect differences in treatment to decline under the second policy.[42] There are also interesting implications of providing, to a decision-maker, aggregated information about their own decisions, for example providing a judge with data on the average prison term they have sentenced defendants of different races to. If the implicit discrimination is due to implicit knowledge, this information will help the decision-maker to learn about their own biases and adjust for them. If it is due to signaling, it could have the opposite effect because the marginal effect of a sentence on an observer's beliefs could decrease.[43] Finally, the theory characterizes the subjective experience of people who are discriminated against; as put by Snyder et al. (1979): "the handicapped person may be repeatedly rebuffed in social encounters by people who give what may seem to them to be reasonable excuses."

6.2 Interpersonal Preferences.

Moral judgment is famously *opaque*: people find it easy to label actions as right or wrong, fair or unfair, but find it difficult to explain the reasoning behind their judgments. Much of moral philosophy proceeds by testing novel cases against intuition. These observations suggest that we have little direct introspection into our moral sense, and therefore that there could be large implicit effects. We make some suggestions of possible implicit influences, and discuss the relevant evidence that we are aware of.

Self-other tradeoffs. The most obvious implicit preference is a self-regarding bias: that people may put less weight on other peoples' payoffs, relative to their own, when the choice set becomes less revealing regarding that preference. This is a natural interpretation of the experiments in Exley (2015), who describes her results as "excuse driven." However we might also find the opposite implicit preference in some circumstances: Miller (1999) argues that contemporary American society exhibits a "norm of self-interest," which requires that people find a justification for their behavior on self-interested grounds: for example he claims that people are significantly more likely to contribute to charity when they are offered a trinket in exchange, because the exchange

[42]Our joint-separate result deals with groups of two. We discuss results for larger groups in the Appendix.

[43]This depends on the interpretation of the observer in the model - when judgments of n outcomes are aggregated, does the decision-maker care about the beliefs of n different observers?

gives them a selfish excuse to perform a generous act.

Inequality aversion. A large literature has studied aversion to inequality inside and outside the lab. We believe that these preferences may be importantly implicit: i.e., inequality may have a bigger effect on choice in less revealing contexts. An indication of this is found in an experiment by Bazerman et al. (1992) which asked subjects to rate the fairness of two different allocations of money:

$$\binom{\text{self=\$500}}{\text{neighbour=\$500}} \tag{1}$$

$$\binom{\text{self=\$600}}{\text{neighbour=\$800}} \tag{2}$$

They found that when the outcomes were presented separately then the subjects rated (1) more highly than (2), but when they were presented jointly the ranking reversed. A loose interpretation of these results is that people dislike getting less than their neighbor (as occurs in (2)), but that preference is implicit, and so its influence diminishes in joint evaluation.

Emotional/aesthetic aspects of a recipient. Patterns of giving to charity are famously difficult to reconcile with consequentialist preferences. We expect that peoples' implicit and explicit preferences regarding charity are quite different. As an illustration Kahneman and Ritov (1994) report that subjects rated a charity devoted to "skin cancer research" higher than one devoted to "saving Australian mammals," when the charities were evaluated jointly. However when the charities were evaluated separately the average rating was higher for the latter. Kahneman and Ritov (1994) report a series of similar findings.

Other influences. Schwitzgebel and Cushman (2012) report experimental results showing that judgments of moral responsibility are influenced by features which are often thought to be normatively irrelevant: whether the action is described as active or passive (action/omission); whether harm caused is a side-effect of aiming at a good outcome (the doctrine of double effect); and whether the outcome is under the decision-maker's control (moral luck). They additionally find that judgment is affected by the *order* of presentation: when asked about two situations, which vary only in one of these normatively-irrelevant features, respondents maintain consistency with their first

answer. We therefore interpret their findings as establishing implicit preferences for these features.

6.3 Framing Effects

A framing effect is usually thought of as an influence on choice by a normatively irrelevant feature of the choice context (**?**). Typical examples of framing effects are (1) the position of a reference point used in describing an outcome; (2) the position of an irrelevant anchor; (3) the designation of which alternative is the 'default' alternative; and (4) whether different aspects of an outcome are described separately or combined. However in each of these cases it is arguable whether the feature is indeed normatively irrelevant - the decision-maker may have preferences over that feature, or consider the feature informative.

An alternative definition - which does not require an assumption about which features are normatively relevant - can be given using our framework: a frame is an attribute over which there is an implicit preference, but no explicit preference. Any framing effect can therefore be described with an intransitive cycle. Some typical framing effects are represented in the following isosceles cycles.[44]

$$
\begin{array}{ccccccc}
z & \succ & x & \precsim & x' & \succ & z \\
\$1 & \succ & \binom{\text{gamble}}{\text{positive frame}} & \sim & \binom{\text{gamble}}{\text{negative frame}} & \succ & \$1 \\
\$1 & \succ & \binom{\text{10 good cards}}{\text{3 bad cards}} & \succ & (\text{10 good cards}) & \succ & \$1 \\
\$5 & \succ & \binom{\text{8oz ice-cream}}{\text{in 9oz cup}} & \succ & \binom{\text{7oz ice-cream}}{\text{in 5oz cup}} & \succ & \$5
\end{array}
$$

Our proposed definition does not fit all cases in the literature because sometimes a frame works at the level of the choice set, not at the level of an individual outcome. Consider the anchoring effect: it does not make much sense to ask a subject to separately state their WTP for two identical goods, one of which has been anchored at price p_1, another which has been anchored at price p_2 - here the anchor seems to affect the entire choice set, not an individual outcome.

[44] The effect of gamble frame is discussed in **?**. The choices with cards are reported in **?**, the choices with ice creams are discussed in **?**. Each could also be described in a binary space, though somewhat less naturally.

6.4 Implicit Preferences & Consumer Behavior

Consumer choice often involves choosing among *bundles* of attributes, and therefore revealingness will vary across consumption contexts. The methods used in this paper could be applied to consumption data, for example determining whether features of a house (bedrooms, hot tub, ocean view, central heating) have different implicit and explicit values.

Suppose consumers implicitly desire some product, in the sense that they have a positive implicit but a negative explicit preference it. Then the firm selling it will wish to make the purchase less revealing by bundling their product with other choices, for example bundling pornography with journalism, to make the purchase less revealing. Suppose instead that consumers implicitly *dislike* a product. Then the firm will wish to make the purchase *more* revealing by removing excuses to not buy the product.

Under the implicit knowledge model firms will also wish to bundle their product with attributes that the consumer knows to be valueless, but which evoke positive associations. Insofar as consumers are imperfectly aware of those associations they will attribute some of the positive feelings evoked to the true quality of the product.[45]

7 Conclusion

Many papers in behavioral economics propose modifying the classical utility function to accommodate observed choices - by adding a "taste" or an "aversion" regarding, for example, ambiguity, loss, gain, inequality, or relative consumption.

However we believe that in many cases behavior is not consistent with any single set of preferences - that instead people struggle with multiple different motivations. We also believe that the effects of these struggles can be detected in choice data - especially in intransitive choices.[46]

Of course any set of choices can be made consistent with a single set of preferences if one is willing to slice the space of outcomes thin enough. We mean that assuming an invariant utility function is often not the most parsimonious way of explaining observed

[45] This is elaborated on in Cunningham (2014).

[46] Another set of papers propose biases in beliefs - wedges between reality and perception - regarding, for example, self-assessments, exponential growth, or probabilities. We think of these cases in a similar way: that it is more fruitful to think of them not as arising from a single set of beliefs, but from an internal struggle between different sets of beliefs, and that the struggle can be identified in intransitive choices. And indeed many experiments use indirect methods to identify biases in belief, rather than just asking people to admit the bias directly.

choices. We think of this paper as a contribution towards formalizing, in a relatively nonparametric way, the choice effects of an internal struggle.[47] We suspect that many preferences which are strong in direct comparisons will become weak in indirect comparisons - for example preferences over equality of payoffs, preferences over ambiguity, and preferences over small risks. We also suspect that many preferences which are weak in direct comparison will become strong in indirect comparisons - for example preferences over race and sex, preference for relative status, and preferences over partisan political issues.

The basic intuition underlying our paper - that implicit attitudes are revealed in indirect comparisons - has been suggested before. However our discussion of existing work shows how difficult it can be to properly identify these effects, and we believe that our framework can serve as basis for much more systematic mapping of internal struggles between inconsistent preferences.

References

Ainslie, G. (1992). *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. New York: Cambridge University Press.

Andreoni, J., J. M. Rao, and H. Trachtman (2011, December). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. Working Paper 17648, National Bureau of Economic Research.

Bazerman, M. H., G. F. Loewenstein, and S. B. White (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 220–240.

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

Bénabou, 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.

[47] We think of **?**, **?**, and Cherepanov et al. (2013) as contributions to the same line of thought.

Bodner, R. and D. Prelec (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions 1*, 105–26.

Bohnet, I., M. H. Bazerman, and A. Van Geen (2015). When performance trumps gender bias: Joint versus separate evaluation. *Management Science, forthcoming*.

Bushong, B., M. Rabin, and J. Schwartzstein (2014). A model of relative thinking.

Chance, Z. and M. I. Norton (2009). I read playboy for the articles. *The Interplay of Truth and Deception: New Agendas in Theory and Research*, 136.

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

Cunningham (2014). Biases and implicit preferences. Technical report, Institute for International Economic Studies.

Dana, J., D. M. Cain, and R. M. Dawes (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2), 193 – 201.

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.

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

Ellenberger, H. F. (1970). The discovery ofthe unconscious. *New York, Basic Books*.

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies, forthcoming*.

Gelman, A. (2014).

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.

Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji (2009). Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of personality and social psychology 97*(1), 17.

Hanson, R. (2008). Hanson on signaling. *EconTalk*.

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.

John, L. K., G. Loewenstein, A. Acquisti, and J. Vosgerau (2013). Paradoxical effects of randomized response techniques.

Kahneman, D. and I. Ritov (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty 9*(1), 5–37.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1), 136–63.

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

Mijović-Prelec, D. and D. Prelec (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1538), 227–240.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Ph. D. thesis, George Washington University.

Mullainathan, S. (2015, January). Racial bias, even when we have good intentions. *The New York Times*.

Nosek, B. A., C. B. Hawkins, and R. S. Frazier (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences 15*(4), 152 – 159.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases. *Department of Economics, UCB*.

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

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.

Spence, M. (1973). Job market signaling. *The quarterly journal of Economics*, 355–374.

Veblen, T. (1899). *Theory of the Leisure Class*. Norwalk: Easton.

Von Hippel, W. and R. Trivers (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences 34*(01), 1–16.

A Appendix: Proofs and Further Discussion

A.1 Proofs

Proof of Proposition 1.

Proof. Normalize the attribute space such that $\forall i$, $x_i = 1$. Suppose, for contradiction, that u has weakly positive implicit preferences for all the attributes on which x and x'' differ. By betweenness, $\{x, x''\}$ is less revealing than $\{x, x'\}$ about all the attributes on which x and x' differ. So, using the definition of implicit preferences, this implies,

$$u(x, \{x, x''\}) \geq u(x', \{x, x''\}).$$

The same logic applies for the comparison between x' and x'', yielding:

$$u(x', \{x, x''\}) \geq u(x'', \{x, x''\}).$$

But the observed choice between x and x'' implies that,

$$u(x, \{x, x''\}) \leq u(x'', \{x, x''\}).$$

If one of the three preferences is strict then one of these three inequalities is strict, yielding a contradiction. □

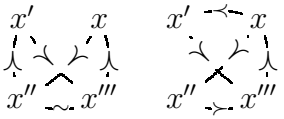
Proof of Proposition ??: Disjunctions

Proposition (13)**.** *To establish an unambiguous implicit preference from right-triangle-cycles of span m requires observing at least 2^{m-1} such cycles.*

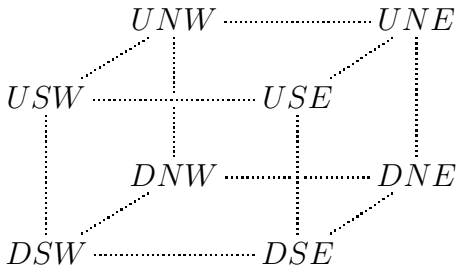
Proof. Assume that all cycles span the same m dimensions (if they span different dimensions, then more will be required). A cycle of span m establishes a disjunction with m terms, of the form $(\lambda_1 > 0) \vee \ldots \vee (\lambda_m < 0)$, which is equivalent to the negation of a conjunction of m terms, of the form $\neg((\lambda_1 \leq 0) \wedge \ldots \wedge (\lambda_m \geq 0))$. We can therefore establish $\lambda_1 > 0$ by collecting negations of disjunctions which rule out $\lambda_1 \leq 0$ along with every other permutation of the remaining attributes. There will be 2^{m-1} such permutation, therefore we must observe 2^{m-1} cycles. □

An unambiguous implicit preference can therefore be established by 2 cycles which span 2 dimensions, or by 4 cycles which span 3 dimensions, or 8 cycles which span 4 dimensions, etc.

As illustration, consider the following examples of pairs of cycles which establish implicit preferences in the 2-dimensional case. Both examples establish a positive implicit preference on the vertical dimension (Northwards). Here each dimension on the page represents a single attribute. Both cases contain a pair of right-triangle cycles, one of which establishes an implicit preference pointing either North or East, the other of which establishes an implicit preference pointing either North or West, and therefore the pair together imply the existence of an implicit preference pointing North.[48]



Moving to three dimensions, consider the following box labelled with initials for Up/Down, North/South, and East/West.



Consider the right-triangle cycle

$$UNW \succ DSE \succ DNE \succ UNW,$$

which satisfies betweenness (DNE is between UNW and DSE), and so implies the existence of at least one implicit preference, either a positive implicit preference for Up, for North, or for West (these being the three attributes belonging to the alternative which is preferred along the hypotenuse of the right-triangle cycle). This can be written either as a disjunction or the negation of a conjunction:

$$(\lambda_U > 0) \vee (\lambda_N > 0) \vee (\lambda_W > 0) \iff \neg((\lambda_U \leq 0) \wedge (\lambda_N \leq 0) \wedge (\lambda_W \leq 0)).$$

To establish the existence of a positive implicit preference for Up over Down, it would be sufficient to observe four right-triangle cycles of span 3, in which each of the Up

[48]Note that the first example has an indifference, and the second example has a non-monotonicity (i.e., one of the horizontal choices goes East, the other goes West). In 2 dimensions there are no examples which can establish an unambiguous implicit preference without one of these features.

elements is chosen over its diametrically opposite Down element, e.g. $UNW \succ DSE$, and in which the Down element is indirectly chosen over the Up element, e.g. $DSE \succ DNE \succ UNW$. Thus four cycles sufficient to establish an implicit preference for Up would be:[49]

$$\begin{aligned} UNW &\succ DSE \succ DNE \succ UNW \\ UNE &\succ DSW \succ DSE \succ UNE \\ USE &\succ DNW \succ DNE \succ USE \\ USW &\succ DNE \succ DSE \succ USW. \end{aligned}$$

Proof of Proposition 2.

Proof. We start by noting that, if A and B are equally revealing about all attributes, then they must generate the same set of rankings of all elements.

$$(A =_i B, \forall i) \implies (u(x, A) \geq u(x', A) \iff u(x, B) \geq u(x', B))$$

Therefore the preferences invoked by the four pairs can be represented with just two different utility functions. We use u^V (vertical) to denote the preferences evoked by $\{x, x'''\}$ and $\{x', x''\}$, and u^D (diagonal) to denote the preferences evoked by $\{x, x''\}$ and $\{x', x'''\}$. Suppose, for contradiction, that there is no negative implicit preference for any of the attributes on which x and x''' differ. Then, because u^D is less revealing about the vertical attributes, u^D must be weakly more favorable to the North, in North-South comparisons, i.e.:

$$\begin{aligned} u^V(x) \geq u^V(x''') &\implies u^D(x) \geq u^D(x''') \\ u^V(x') \geq u^V(x'') &\implies u^D(x') \geq u^D(x''). \end{aligned}$$

But this yields:

$$u^D(x) \geq u^D(x''') \geq u^D(x') \geq u^D(x'') \geq u^D(x),$$

with one of these inequalities strict, which is a contradiction. □

[49]Note that these cycles violate monotonicity: in the two direct North-South comparisons, North wins in one case, South wins in the other. It is true that in 2 dimensions, if using cycles only with strict preferences, then an unambiguous implicit preference can be inferred only if monotonicity is violated. It is not clear whether this is true for any number of dimensions.

A.2 Combining Scissor Effects

In proposition 3 we established that a scissor effect ($y(x|\{x, x'\}) \gtrless y(x|\{x, x''\})$) will establish a disjunction among implicit preferences over all n possible attributes. As discussed above, the implication can be expressed either as a disjunction of strict inequalities, or the negation of a conjunction of weak inequalities:

$$(\lambda_1 > 0) \vee \ldots \vee (\lambda_m < 0) \iff \neg((\lambda_1 \leq 0) \wedge \ldots \wedge (\lambda_m \geq 0)).$$

The latter formulation allows us to think of each scissor-effect as eliminating a set of cells in the space of all possible implicit preferences. For example, the proposition $\neg((\lambda_1 \leq 0) \wedge (\lambda_2 \leq 0))$ can be represented graphically, where "$\times$" eliminates a possible state:

		λ_1		
		-1	0	1
	-1	$\times$	$\times$	
λ_2	0	$\times$	$\times$	
	1			

Inference from multiple scissor-effects can be represented as combining these eliminations. This way of seeing things has some implications:

1. Given some x, x', x'', if we find that $y(x|\{x, x'\}) \neq y(x|\{x, x''\})$, this will rule out one of a pair of *opposite* points in λ-space, e.g. for two attributes it could establish $\neg((\lambda_1 \geq 0) \vee (\lambda_2 \geq 0))$ or $\neg((\lambda_1 \leq 0) \vee (\lambda_2 \leq 0))$.

2. As with choice, identifying a single unambiguous implicit preference will require at least 2^{n-1} scissor-effects.

3. Suppose that the decision-maker has non-zero implicit preferences with respect to every attribute (i.e., $\lambda_i \neq 0, \forall i$). It may then be possible to identify, using scissor effects, the full vector of implicit preferences, λ, by observing sufficiently many scissor effects. The minimum is $2^n - 1$, i.e. the number of pairs which include the true state as one element. It is also clear that we could never definitely establish that $\lambda_i = 0$ for any attribute i, because no set of observations could rule out both $\lambda_i > 0$ and $\lambda_i < 0$.

4. Testing all possible scissor effects (every combination of x, x', x'', with x' between x and x'') is not guaranteed to reveal the true λ, or even to unambiguously identify

any single implicit preference. Suppose we have just two attributes. Testing all pairs is guaranteed to reveal at least one implicit preference (if evaluation changes in every scissor), but you cannot know in advance which attribute it will work on. With 3 or more dimensions it is possible that you will not learn any unambiguous fact about any dimension's implicit preferences.

A.3 Sequential Evaluations

Suppose we observe data on evaluations where outcomes are considered sequentially, such as a judge sentencing a series of defendants, a critic reviewing a series of films, a referee calling a series of fouls. Denote the outcomes as $x^1, \ldots, x^T$, and the evaluations as $y^1, \ldots, y^T$. A simple way of analyzing this data is to treat the current and just-prior case as the comparison set, i.e. assume:

$$y^t = y(x^t, \{x^t, x^{t-1}\}). \tag{SEQ}$$

Given this assumption, and our results on evaluation, there are some simple tests for implicit preferences derived from Proposition 4. We illustrate with an example. Given data on sequential sentencing, an implicit bias against black defendants predicts the following three effects:

Type 1 The sentence y^t will be higher when x^{t-1} is black, compared to when they are white.

Type 2 If the defendant x^{t-1} has a different race to x^t, then the sentence y^t will be lower if x^t is black and the defendants x^t and x^{t-1} are more similar (compared to when they are less similar), and higher if x^t is white and x^t, x^{t-1} are more similar (compared to when they are less similar).

Type 3 If the defendant x^{t-1} has the same race as x^t, then the sentence y^t will be higher if x^t is black and the defendants x^t and x^{t-1} are more similar (compared to when they are less similar), and lower if x^t is white and x^t, x^{t-1} are more similar (compared to when they are less similar).

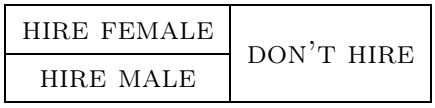
Two assumptions underlie (SEQ). First, that only the prior case is relevant as comparator. Second, that the evaluation of an outcome x will be affected in the same way by a comparator x', whether x' is evaluated at the same time as x, or prior to x. The

second assumption will hold in the implicit-knowledge model: in that model, System 2 conditions on signals received from System 1 which are independent of the choice set, and there is no reason to treat them differently based on the timing of those signals (unless System 2 is more likely to forget earlier signals). Turning to the signaling model, our earlier derivation found that $\frac{d\hat{\omega}_i}{dy}$ is independent of y'. This implies that when choosing an evaluation y the marginal signaling incentives do not depend on the evaluation given to the comparator, y' (though they do depend on the identity of the comparator, x'). Thus the assumption above, (SEQ), will hold in our signaling model. However the assumption will not hold in the *ceteris-paribus* model: suppose that, when evaluated separately, a man is given a salary of \$60K, and a woman is given a salary of \$50K, but they are both given \$55K salaries when evaluated jointly. Then, if the pair is evaluated sequentially, both will be given the same salary, determined by whoever is evaluated first.[50]

B A ternary attribute space

In this Appendix we extend the model to allow 3-valued attributes, allowing us to infer implicit preferences from an additional class of intransitivities - which we call "isosceles" cycles. As before we make assumptions about the relationship between the geometry of the choice set and the revealingness of choices. However, for this extension, we do not further justify the assumptions with deeper models, we only conjecture that there exists a relatively simple extensions of the existing models which would imply the assumptions made in this section.

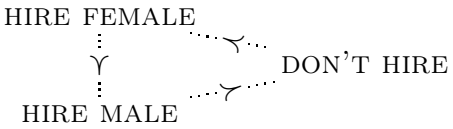
Suppose we observe a hiring decision with the 3 elements {HIRE FEMALE, HIRE MALE, DON'T HIRE}. It seems natural to think of these elements as having a resemblance relationship, which could be represented graphically as:



The horizontal dimension has a natural binary representation (whether or not to hire), but the vertical dimension has 3 distinct values, and we could describe the no-hire outcome as being "neutral" on the gender dimension. Our assumption on revealingness will be that, given two cases x and x' which differ only on attribute i, then if a third

[50]This uses the extension of the *ceteris paribus* model to evaluation, discussed later in this Appendix.

case z is neutral on that attribute (i.e., has the new 3rd value), then the choice set $\{x, x'\}$ is relatively more revealing about dimension i, than either $\{x, z\}$ or $\{x', z\}$, and vice versa for all other attributes. This allows us to conclude, from the following cycle, that the decision-maker has an implicit preference for male candidates:



Let $X = \{0, 1, \varnothing\}$.

Definition 2. For any $x, x', z \in X$, we say that z **is a vertex of** $\{x, x'\}$ if and only if there exists some j such that

$$\begin{aligned} z_j &= \varnothing \\ x_j &= 1 - x'_j \\ x_i = x'_i &\iff i \neq j. \end{aligned}$$

Assumption 5. *If z is a vertex of x, x', then*

$$\begin{aligned} \forall i,\ \{x, z\} &=_i \{x', z\} \\ \forall i\, x_i \neq x'_i,\ \{x, z\} &<_i \{x, x'\} \\ \forall i\, x_i = x'_i,\ \{x, z\} &>_i \{x, x'\} \end{aligned}$$

We finally extend the definition of implicit preferences to the 3-valued space, by assuming an ordering among features $(0, \varnothing, 1)$, and assuming that changes in implicit preferences will respect this ordering (i.e., the relatively higher value will come to be relatively preferred).

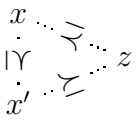
Definition 3. We say that $u(x, A)$ has **implicit preferences** $\lambda \in \{-1, 0, 1\}^n$, if, for any $x, x' \in X$, and $A, B \in \mathcal{A}$ (normalizing such that $\forall j, x_j \in \{\varnothing, 1\}, x'_j \in \{-1, \varnothing\}$) and for every i with $x'_i \neq x_i$,

$$\begin{aligned} A >_i B &\implies \lambda_i \geq 0 \\ A <_i B &\implies \lambda_i \leq 0, \end{aligned}$$

then

$$u(x, A) > u(x', A) \implies u(x, B) > u(x', B).$$

Proposition 5 (Isosceles Cycle)**.** *If z is a vertex of x, x', and we observe choices such that,*



with at least one preference strict, then $u(\cdot, \cdot)$ must have a negative implicit preference over the attribute j for which $x_j \neq x'_j$.

Proof. We know that $\{x, z\}$ and $\{x', z\}$ induce the same utility function because they are equally revealing on all dimensions, by assumption. So,

$$u(x', \{x, z\}) \geq u(z, \{x, z\}) \geq u(x, \{x, z\}).$$

While we also know that,

$$u(x', \{x, x'\}) \leq u(x, \{x, x'\}),$$

with one of these three inequalities strict. Normalize $x_i = 1, \forall i$. We know that the choice set $\{x, x'\}$ is more revealing about the attribute on which x and x' differ, compared to $\{x, z\}$, so if there was a weakly positive implicit preference for attribute j, then:

$$u(x', \{x, z\}) \leq u(x, \{x, z\}),$$

yielding a cycle of inequalities under $u(\cdot, \{x, z\})$, with one inequality strict, and therefore contradicting the assumption, and establishing a negative implicit preference for attribute j. □

C Foundations for Implicit Preference

In this Appendix we give three formal foundations for implicit preference: *ceteris paribus*, signaling, and implicit knowledge. We begin by defining *separable* implicit preferences, a subclass of menu-dependent utility functions which exhibit implicit preferences, and then show that the three parametric foundations all are members of this class.

C.1 Separable Implicit Preferences

Lemma 1. *If a menu-dependent utility function $u(x, A)$ can be written as,*

$$u(x, A) = v(x) - \sum_{i=1}^{n} (x_i - \frac{1}{2})\kappa_i \Phi_i(A) \tag{SEP}$$

with $\kappa_i \in \mathbb{R}$, and $\Phi_i : \mathcal{A} \to \mathbb{R}$, then, with respect to the orderings on A induced by Φ_i, $u(x, A)$ will (a) have relative implicit preferences, and (b) have absolute implicit preferences, and (c) satisfy monotonicity, with,

$$\lambda_i = \begin{cases} -1 & , \kappa_i < 0 \\ 0 & , \kappa_i = 0 \\ 1 & , \kappa_i > 0. \end{cases}$$

Proof. We begin with relative implicit preferences. Take some $x, x' \in X$, and $A, B \in \mathcal{A}$, normalizing $x_j = 1, \forall j$, such that for every i with $x'_i = 0$,

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

and

$$u(x, A) > u(x', A).$$

We reorder the attributes such that $x'_i = 0$ if $i \in \{m, \ldots, n\}$, for some m. Substituting in the utility functions we get:

$$\begin{aligned} v(x) - \sum_{i=1}^{n} \kappa_i (x_i - \frac{1}{2})\Phi_i(A) &> v(x') - \sum_{i=1}^{n} \kappa_i (x'_i - \frac{1}{2})\Phi_i(A) \\ v(x) - v(x') - \sum_{i=m}^{n} \kappa_i \Phi_i(A) &> 0. \end{aligned}$$

We wish to prove that $u(x, B) > u(x', B)$, and we can see that this will be true if:

$$\begin{aligned} \sum_{i=m}^{n} \kappa_i \Phi_i(B) &\leq \sum_{i=m}^{n} \kappa_i \Phi_i(A). \\ \sum_{i=m}^{n} \kappa_i \left[\Phi_i(B) - \Phi_i(A)\right] &\leq 0. \end{aligned}$$

By assumption, whenever $\Phi_i(A) \geq \Phi_i(B)$, $A \geq_i B$, so $\lambda_i \geq 0$ and so $\kappa_i \geq 0$, and so $\kappa_i \left[\Phi_i(B) - \Phi_i(A)\right] \leq 0$. The same conclusion holds when $\Phi_i(A) \leq \Phi_i(B)$. So the whole expression is negative proving the result.

We next show that $u(x, A)$ is an implicit evaluation function. Again consider some $x \in X$ (normalizing $x_j = 1, \forall j$) and $A, B \in \mathcal{A}$ such that

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

We wish to show that this implies,

$$u(x, A) \leq u(x, B).$$

We can express the difference:

$$u(x, A) - u(x, B) = -\frac{1}{2} \sum_{i=1}^{n} \kappa_i \left[\Phi_i(A) - \Phi_i(B)\right].$$

And, by assumption, whenever $\Phi_i(A) \geq \Phi_i(B)$, $A \geq_i B$, which implies $\lambda_i \geq 0$, and $\kappa_i \geq 0$. The converse holds when $\Phi_i(A) \leq \Phi_i(B)$, implying that every term in the sum will be positive, and the expression as a whole will be negative, proving the proposition.

Finally we wish to show monotonicity: if, for every $x, x' \in X$, $A, B \in \mathcal{X}$, with normalization $x_i = 1$, $\forall i$, and $x'_i = 0 \Leftrightarrow m \leq i \leq n$, and for all $i \in \{m, \ldots, n\}$:

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

then

$$y(x', A) < y(x', B) \implies y(x, A) < y(x, B).$$

The antecedent implies:

$$\begin{aligned} y(x', A) &< y(x', B) \\ -\sum_{i=1}^{n}(x'_i - \frac{1}{2})\kappa_i\Phi_i(A) &< -\sum_{i=1}^{n}(x'_i - \frac{1}{2})\kappa_i\Phi_i(B) \\ -\frac{1}{2}\sum_{i=1}^{m-1}\kappa_i\Phi_i(A) + \frac{1}{2}\sum_{i=m}^{n}\kappa_i\Phi_i(A) &< -\frac{1}{2}\sum_{i=1}^{m-1}\kappa_i\Phi_i(B) + \frac{1}{2}\sum_{i=m}^{n}\kappa_i\Phi_i(B) \\ \frac{1}{2}\sum_{i=1}^{m-1}\kappa_i[\Phi_i(A) - \Phi_i(B)] &> \frac{1}{2}\sum_{i=m}^{n}\kappa_i[\Phi_i(A) - \Phi_i(B)] \end{aligned}$$

When considering the corresponding expression for x, the LHS will be identical, and the RHS will have the opposite sign (because $x_i = 1, x'_i = 0$ for $i \in \{m, \ldots, n\}$), so the conclusion will hold if the RHS is positive, i.e. if:

$$\sum_{i=m}^{n}\kappa_i[\Phi_i(A) - \Phi_i(B)] \geq 0,$$

which will hold by our assumptions which imply that $\kappa_i \geq 0 \Leftrightarrow \Phi_i(A) \geq \Phi_i(B)$, and $\kappa_i \leq 0 \Leftrightarrow \Phi_i(A) \leq \Phi_i(B)$. □

Aggregation of the parallel construction test. Before moving on, separability allows us to prove a useful aggregation result for implicit preferences. Suppose we collect data on evaluations on a set of outcomes, and we wish to check for data which passes the "parallel construction" test (Proposition 4) - i.e. two outcomes which differ in one attribute, and which move in opposite directions when their comparators change in equivalent ways. If we have between-subjects data, and there are many attributes, then it is likely that we will not have enough observations to test a single parallel construction (a single pair $\bar{x}$ and $\underline{x}$), but we could check if the condition holds on *average*. The following proposition shows that, with separable preferences, if the parallel construction test is true on average, then it must hold in at least one individual case, and therefore it will establish an unambiguous implicit preference .

Proposition 6 (average evaluations)**.** *If $y(\cdot, \cdot)$ satisfies SEP, then for some set of outcomes $\{\bar{x}^k, \bar{x}'^k, \bar{x}''^k, \underline{x}^k, \underline{x}'^k, \underline{x}''^k\}_{k=1}^{l}$, with, for all $k \in \{1, \ldots, l\}$, $\bar{x}_i^k = 1 = 1 - \underline{x}^k$, and*

$\bar{x}^k_j = \underline{x}^k_j$ *for* $j \neq i$, *and* $\bar{x}'^k$ *between* $\bar{x}^k$ *and* $\bar{x}''^k$, *and*

$$\begin{aligned} \bar{x}^k = \bar{x}'^k &\Leftrightarrow \underline{x}_k = \underline{x}'_k, \forall i \\ \bar{x}^k = \bar{x}''^k &\Leftrightarrow \underline{x}_k = \underline{x}''_k, \forall i \end{aligned}$$

and if,

$$\begin{aligned} \sum_{k=1}^{l} y(\bar{x}^k, \{\bar{x}^k, \bar{x}'^k\}) - y(\bar{x}^k, \{\bar{x}^k, \bar{x}''^k\}) &> 0 \\ \sum_{k=1}^{l} y(\underline{x}^k, \{\underline{x}^k, \underline{x}'^k\}) - y(\underline{x}^k, \{\underline{x}^k, \underline{x}''^k\}) &< 0 \end{aligned}$$

then there exists a $k \in \{1, \ldots, l\}$ *with,*

$$\begin{aligned} y(\bar{x}^k, \{\bar{x}^k, \bar{x}'^k\}) - y(\bar{x}^k, \{\bar{x}^k, \bar{x}''^k\}) &> 0 \\ y(\underline{x}^k, \{\underline{x}^k, \underline{x}'^k\}) - y(\underline{x}^k, \{\underline{x}^k, \underline{x}''^k\}) &< 0. \end{aligned}$$

Proof. [COMING] □

C.2 Foundation for Implicit Preferences: *Ceteris Paribus* Rules

In this section we show that a decision-maker who maximizes an ordinary utility function, but is constrained by ceteris-paribus rules (e.g., "you may not choose a foreign over a domestic bidder, all else equal"), will exhibit implicit preferences in our sense.

We begin with the separable form above (equation SEP), and we let $\Phi_i(A)$ be an indicator function, equal to 1 whenever the choice set contains two elements which differ only on i. Then κ_i represents the penalty for choosing the disallowed option. For example, if the decision-maker incurs a penalty for choosing an element with $x_i = 1$ over an element which is otherwise identical but has $x_i = 0$, then κ_i will be a positive number. The rules will always be respected if we assume $|\kappa_i| \in \{0, \max_{x \in X}\{|v(x)|\}\}$. We therefore simply need to show that $\Phi_i(A)$ satisfies strong betweenness and equivalence.

Proposition 7. *The linear order induced by the function*

$$\Phi_i(\{x, x'\}) = \begin{cases} 1 & , (x_i \neq x'_i) \wedge (\forall j \neq i,\, x_j = x'_j) \\ 0 & , \textit{otherwise} \end{cases}$$

satisfies strong betweenness and equivalence.

Proof. First consider strong betweenness. Suppose that x' is between x and x''. Then for any i with $x_i \neq x'_i$, if $\Phi_i(\{x, x'\}) = 0$, we must also have $\Phi_i(\{x, x''\}) = 0$. And for any i with $x_i = x'_i$, $\Phi(\{x, x'\}) = \Phi(\{x, x''\}) = 0$.

Equivalence holds because the function Φ_i depends only on whether x and x' agree or disagree on each attribute, i.e. it is a function of the vector $|x - x'|$. □

There are two extensions of this model worth considering. First, to take into account the interaction of rules. Suppose you have two rules - one against choosing men over women (all else equal), and one against choosing white over black candidates (all else equal). As written, the model does not disallow the decision-maker from choosing a white man over a black woman. A fuller version of this model would disallow any choices which imply, by transitive extension, that at least one of the ceteris paribus rules is broken. We are not sure whether this extended version can be written as having separable implicit preferences (SEP).

Second, this basic idea could be adapted to apply to evaluation. Assume that the decision-maker incurs the penalty of κ_i if they evaluate an outcome x more highly than x', when x and x' differ only in x possessing attribute i. This would predict that, for example, an implicitly sexist decision-maker would pay an otherwise-identical man and a woman exactly the same salary when evaluated side by side, but a fixed discrepancy would appear when there was some difference in their qualifications, or when they were evaluated separately. Roughly speaking, we think this could be achieved by assuming a separable intrinsic utility function, $v(x) = \sum \omega_i x_i$, and letting $\kappa_i = \omega_i$ when there was a ceteris paribus rule on attribute i.

C.3 Foundation: Gaussian Signaling and Choice

In this section we present a linear-Gaussian signaling model of choice, and show that it can be described as having separable implicit preferences (SEP), with a comparison function $\Phi_i(\cdot)$ that satisfies strong betweenness and equivalence. In the following section we show that $\Phi_i(\cdot)$ differs slightly for evaluation, but against satisfies strong betweenness and equivalence.

The important assumptions will be (1) that the decision-maker's utility function is separable in each attribute, and regarding each belief of the observer about those

attribute-weights; (2) that the observer has mean-zero independent Gaussian priors over all those weights, regarding both attributes and beliefs.

We assume the decision-maker has the following utility function::

$$u(x, A) = \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i(x, A), \tag{3}$$

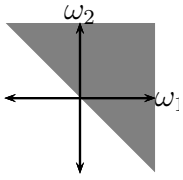
Here each ω_i represents the decision-maker's intrinsic preference for attribute i, $\hat{\omega}_i(x, A)$ represents the observer's mean posterior regarding ω_i given that they observed the choice of x from A, and κ_i represents the decision-maker's preference over the observer's beliefs. For example, $\omega_i > 0$ implies that the decision-maker likes attribute i, while $\kappa_i < 0$ implies that they wish for the observer to think that they don't like it.

The observed choice function will therefore be:

$$c(A) = \arg\max_{x \in A} \left\{ \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i(x, A) \right\}.$$

We assume that the observer has independent and symmetric priors over all the weights, $\omega_1, \ldots, \omega_n$ and $\kappa_1, \ldots, \kappa_n$.

Once a choice is observed, the posteriors will therefore be a truncated multivariate normal. For example, if we observe that $x \in c(\{x, x'\})$, with $x = \binom{1}{1}$ and $x' = \binom{0}{0}$, this implies that $\omega_1 + \omega_2 \geq 0$. Graphically, the posteriors will be contained in the shaded region:



Unfortunately the mean marginals of this distribution, $\hat{\omega}_i(x, A)$, do not have a simple expression. However we need only a qualitative result: that the *difference* in mean posteriors is smaller when the two alternatives differ among a strictly larger set of attributes. Define $\Delta_i^{\{x,x'\}}$ as the difference in posteriors regarding ω_i , between that generated by the choice of x, and that generated by the choice of x', i.e.:

$$\Delta_i^{\{x,x'\}} \equiv \left| E\left[\omega_i | \sum_{j=1}^{n} \omega_j (x_j - x'_j) + \sum_{i=1}^{n} \kappa_i \Delta_i^{\{x,x'\}} \geq 0\right] - E\left[\omega_i | \sum_{j=1}^{n} \omega_j (x_j - x'_j) + \sum_{i=1}^{n} \kappa_i \Delta_i^{\{x,x'\}} \leq 0\right] \right|.$$

We will assume that this implicit definition has a unique answer.

Proposition 8. *For binary choice sets, choice maximizes the following utility function,*

$$\bar{u}(x, \{x, x'\}) = \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i (x_i - \frac{1}{2}) \Delta_i^{\{x,x'\}} \tag{4}$$

which satisfies SEP.

Proof. We wish to show that, for any x, x',

$$u(x, \{x, x'\}) \geq u(x', \{x, x'\}) \iff \bar{u}(x, \{x, x'\}) \geq \bar{u}(x', \{x, x'\})$$

The left-hand side implies that,

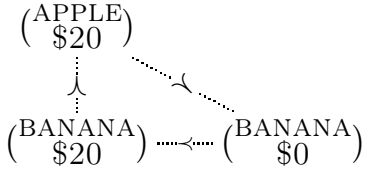
$$\begin{aligned}
\sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i(x, \{x, x'\}) &\geq \sum_{i=1}^{n} \omega_i (x'_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i(x', \{x, x'\}) \\
\sum_{i=1}^{n} \omega_i [x_i - x'_i] &\geq \sum_{i=1}^{n} \kappa_i \left[\hat{\omega}_i(x', \{x, x'\}) - \hat{\omega}_i(x, \{x, x'\})\right] \\
\sum_{i=1}^{n} \omega_i [x_i - x'_i] &\geq \sum_{i=1}^{n} \kappa_i (x'_i - x_i) \Delta_i^{\{x,x'\}} \\
\sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n} \kappa_i x_i \Delta_i^{\{x,x'\}} &\geq \sum_{i=1}^{n} \omega_i x'_i + \sum_{i=1}^{n} \kappa_i x'_i \Delta_i^{\{x,x'\}} \\
\bar{u}(x, \{x, x'\}) &\geq \bar{u}(x', \{x, x'\})
\end{aligned}$$

which completes the proof. □

We now turn to determining whether $\Delta_i^{\{x,x'\}}$ satisfies strong betweenness: we wish to show that the size of the effect on posteriors (Δ_i^A) will decrease when the attribute-wise distance between x and x' increases.

We are assuming that the observer has mean-zero priors over each of the intrinsic preferences, ω_i. To explain the intuition, consider choice among breakfasts: the choice set $\{\binom{\text{APPLE}}{\text{COFFEE}}, \binom{\text{BANANA}}{\text{COFFEE}}\}$, seems more informative about the relative preference for apple over banana than is the choice set $\{\binom{\text{APPLE}}{\text{COFFEE}}, \binom{\text{BANANA}}{\text{TEA}}\}$. However, consider the choice sets: $\left\{\binom{\text{APPLE}}{\$0}, \binom{\text{BANANA}}{\$0}\right\}$, and $\left\{\binom{\text{APPLE}}{\$20}, \binom{\text{BANANA}}{\$0}\right\}$. If you choose a banana from the first choice set, I will infer a mild preference for bananas; but if you choose a banana from the second choice set I will infer a very strong preference

for bananas over apples. This happens because we already have strong priors about the difference in value between \$0 and \$20, so seeing you choose a banana over \$20 causes an extreme revision of my beliefs about your fruit preferences. This implies a violation of strong betweenness: the second choice set could be *more* revealing about fruit preferences, despite the alternatives differing in a greater number of attributes. This could generate a right-triangle cycle such as the following:



by someone who (i) prefers bananas, (ii) wants people to believe that he prefers apples, and (iii) doesn't care much about money. Then he would choose $\binom{\text{APPLE}}{\$20}$ over $\binom{\text{BANANA}}{\$0}$ because of its strong signaling value.

We therefore add an assumption sufficient to guarantee strong betweenness: that the observer has mean-zero priors about all the weights:

$$\forall i,\ E[\omega_i] = 0 \tag{MZ}$$

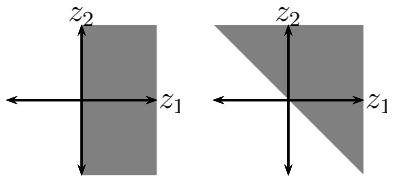
In practice this means that our inferences regarding implicit preferences are only valid when the attributes used do not having a high *valence*: if an observer has strong prior over preferences with respect to some attribute, then mixing that attribute with another can increase revealingness instead of decreasing it.

We first give a lemma to show the sufficiency of symmetric probability density functions (for a Gaussian distribution, a zero mean implies that the distribution is symmetric around zero).

Lemma 2. *For any two independent random variables, $z_1, z_2 \in \mathbb{R}$ with symmetric probability density functions,*

$$E[z_1 | z_1 > 0] \quad > \quad E[z_1 | z_1 + z_2 > 0].$$

Proof. When we observe that $z_1 > 0$, our posteriors will be truncated by a vertical line, as shown in the first panel below. When we observe $z_1 + z_2 > 0$, the posteriors will be truncated along a diagonal line, as shown in the second panel:



We can write the difference between the two expectations as follows, using $f(z_1)$ and $g(z_2)$ to represent probability density functions:

$$\begin{aligned}
E[z_1|z_1+z_2>0]-E[z_1|z_1>0] &= \frac{\int_{-\infty}^{\infty}\int_{-z_1}^{\infty} z_1 f(z_1)g(z_2)dz_2dz_1}{\int_{-\infty}^{\infty}\int_{-z_1}^{\infty} f(z_1)g(z_2)dz_2dz_1} - \frac{\int_0^{\infty} z_1 f(z_1)dz_1}{\int_0^{\infty} f(z_1)dz_1} \\
&= \frac{\int_{-\infty}^{0}\int_{-z_1}^{\infty} z_1 f(z_1)g(z_2)dz_2dz_1}{\int_{-\infty}^{0}\int_{-z_1}^{\infty} f(z_1)g(z_2)dz_2dz_1} - \frac{\int_0^{\infty}\int_{-\infty}^{-z_1} z_1 f(z_1)g(z_2)dz_2dz_1}{\int_0^{\infty}\int_{-\infty}^{-z_1} f(z_1)g(z_2)dz_2dz_1} \\
&= \frac{\int_{-\infty}^{0}\int_{-z_1}^{\infty} z_1 f(z_1)g(z_2)dz_2dz_1 - \int_0^{\infty}\int_{-\infty}^{-z_1} z_1 f(z_1)g(z_2)dz_2dz_1}{\int_{-\infty}^{0}\int_{-z_1}^{\infty} f(z_1)g(z_2)dz_2dz_1} \\
&< 0
\end{aligned}$$

The second-last step follows by symmetry of $f(\cdot)$ and $g(\cdot)$. □

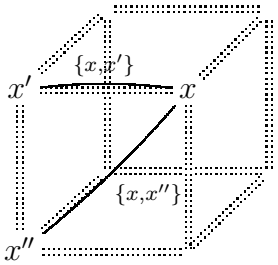
Lemma 3. *For any two independent random variables, $z_1, z_2 \in \mathbb{R}$ with symmetric probability density functions, and $\lambda > 1$,*

$$E[z_1|z_1+z_2>0] \;>\; E[z_1|z_1+\lambda z_2>0].$$

Proof. [coming soon] □

Proposition 9. *Ordering binary choice sets by Δ_i^A satisfies betweenness and equivalence.*

Proof. Let x' be an outcome between x and x'':



Let $x_i = 1$ for all i, and for convenience reorder the attributes such that $x'_i = 0$ if and only if $m < i \leq n$. We assume that x'' differs from x' in only one attribute, i.e. $x''_i = 0$ if and only if $i \in \{m, \ldots, n\}$. This is not without loss of generality, but the more general result can be derived by induction. Then we can write the following:

$$\begin{aligned}
\Delta_i^{\{x,x'\}} &= \left| E[\omega_i | \sum_{j=m+1}^{n} (\omega_j + \kappa_j \Delta_j^{\{x,x'\}}) \geq 0] - E[\omega_i | \sum_{j=m+1}^{n} (\omega_j + \kappa_j \Delta_j^{\{x,x'\}}) \leq 0] \right| \\
\Delta_i^{\{x,x''\}} &= |E[\omega_i | \sum_{j=m+1}^{n} (\omega_j + \kappa_j \Delta_j^{\{x,x''\}}) + \omega_m + \kappa_m \Delta_m^{\{x,x''\}} \geq 0] \\
&\quad - E[\omega_i | \sum_{j=l+1}^{n} (\omega_j + \kappa_j \Delta_j^{\{x,x''\}}) + \omega_m + \kappa_m \Delta_m^{\{x,x''\}} \leq 0]|
\end{aligned}$$

We wish to show that, for $i \in \{m+1, \ldots, n\}$, the second choice set is weakly less revealing, i.e. $\Delta_i^{\{x,x''\}} \leq \Delta_i^{\{x,x'\}}$.

As in Lemma 2, we are dealing with the sum of a set of symmetric independent random variables ($\omega_m, \ldots, \omega_n$ and $\kappa_m, \ldots, \kappa_n$).

First we show that it cannot be the case that $\Delta_i^{\{x,x''\}} > \Delta_i^{\{x,x'\}}$ for all $i \in \{m+1, \ldots, n\}$. That would lead to a contradiction, because by Lemmas 2 and 3, the posteriors must shift closer to zero under $\{x, x''\}$, and so Δ_i must decrease.

Second, suppose that one Δ_i increased.

==

This would be straightforward in a naive signaling model - where the observer believed that $\kappa_i = 0$. In that case the second choice set ($\{x, x''\}$) would result from adding ω_m, a mean-zero symmetric random variable, and so Δ_i would fall, according to Lemma 2.

Suppose the opposite, that $\Delta_i^{\{x,x''\}} > \Delta_i^{\{x,x'\}}$ for some $i \in \{m+1, \ldots, n\}$.

However we need to be more careful because for $j \in \{m+1, \ldots, n\}$ the Δ_j terms will *decrease* when moving from the first choice set to the second. This decreases the variance of the $\kappa_j \Delta_j$ terms.

From Lemma 2 we know that the difference in .

From these equations, and Lemma 2, we can see that strong betweenness will hold for Δ_i^A, because:[51]

[51]Note that the sum of mean-zero Gaussians is itself a mean-zero Gaussian distribution, so Lemma 2 can be used with the sum.

- For the variables on which x and x' disagree ($m < i \leq n$), $\{x, x''\}$ is weakly less revealing ($\Delta_i^{\{x,x''\}} \leq \Delta_i^{\{x,x'\}}$).
- For the attributes on which x' and x'' disagree ($l < i \leq m$), $\{x, x''\}$ is weakly more revealing ($\Delta_i^{\{x,x''\}} \geq \Delta_i^{\{x,x'\}} = 0$).
- For the attributes on which x and x'' agree ($i \leq l$), $\{x, x''\}$ is equally revealing ($\Delta_i^{\{x,x''\}} = \Delta_i^{\{x,x'\}} = 0$).

Finally, $\Delta_i^{\{x,x'\}}$ satisfies equivalence, because it depends just on the absolute values of each difference. □

C.4 Foundation: Gaussian Signaling and Evaluation

The same basic signaling model can also be applied to evaluation, but Φ differs somewhat.

Consider the evaluation of an outcome, y, as a cost paid to receive the outcome, and assume that it directly enters the utility function:

$$U(x, y, \hat{w}) = \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i - y,$$

and we define $y(x, A)$ as the value of y which the DM would state under the following elicitation procedure (for example, the BDM):

$$y(x, A) = \arg\max_{y \in \mathbb{R}} \left\{ \int_{-\infty}^{y} \left[\sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) - \bar{y} \right] f(\bar{y}) d\bar{y} + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i \right\}$$

The solution to this problem is:

$$y(x, A) = \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \frac{d\hat{\omega}_i}{dy} \tag{5}$$

From the observer's point of view, they wish to infer the weights $\omega_1, \ldots, \omega_n$ from y and y'. Assume that the κ_i. This is a simple signal extraction model: they will .

From the logic of Gaussian signal extraction, we get

$$\hat{\omega}_i = \begin{cases} \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{y + y'}{2} & , i \in S \\ \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \frac{y + y'}{2} & , i \in D \end{cases}$$

where $i \in S \iff x_i = x'_i$, and $i \in D$ otherwise. This .

We can therefore derive the sensitivity of $\hat{\omega}_i$ to y:

$$\frac{d\hat{\omega}_i}{dy} = \begin{cases} \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{1}{2} & , i \in S \\ \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \frac{1}{2} & , i \in D \end{cases}$$

Proposition 10. *Evaluation according to 5 can be written as SEP.*

Proof. We know that $\frac{d\hat{\omega}_i}{dy}$ has the same sign as $(x_i - \frac{1}{2})$, which implies:

$$\frac{d\hat{\omega}_i}{dy} = 2(x_i - \frac{1}{2}) \left| \frac{d\hat{\omega}_i}{dy} \right|,$$

allowing us to write the evaluation function in the form SEP:

$$y(x, A) = \sum_{i=1}^{n} \omega_i (x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i 2(x_i - \frac{1}{2}) \left| \frac{d\hat{\omega}_i}{dy} \right|.$$

We will now make the assumption that, for any x, x' we consider, the observer has more uncertainty about common factors ($i \in S$) than about differential factors ($i \in D$). □

Assumption 6 (Dominance of Uncertainty about Common Factors)**.** $\sum_{j \in S} \sigma_j^2 > \sum_{j \in D} \sigma_j^2$.

We regard this assumption as merely technical, mainly necessary because of the stylized way in which we have treated attributes, where the values of each realization are opposites. In an alternative version of the model, where the two realizations of the attributes are treated as independent dummy variables, with independent weights, this assumption is not necessary. Results for this version of the model are available on request.

Proposition 11. *Ranking binary choice sets by* $\left| \frac{d\hat{\omega}_i}{dy} \right|$ *satisfies strong betweenness and equivalence.*

Proof. Given three outcomes, $x, x', x'' \in X$, where x' is between x and x'', we wish to show that, for every attribute i on which x and x' disagree,

$$\left| \frac{dE[\omega_i | \binom{x}{x'}, \binom{y}{y'}]}{dy} \right| \geq \left| \frac{dE[\omega_i | \binom{x}{x''}, \binom{y}{y''}]}{dy} \right|$$

and the reverse for every attribute on which x and x' agree. This will hold by Assumption 6. □

C.5 Foundation: Implicit Knowledge

We now present a model of implicit knowledge and show that it will generate implicit preferences - a more extensive version of this model is given in **?**. A decision-maker with implicit knowledge will satisfy betweenness, but not strong betweenness, meaning that our techniques for identifying implicit preferences will work in choice but not evaluation. This is because the interactions among different aspects of implicit knowledge are complex, and become difficult to disentangle.

The basic assumption in the model is that we possess knowledge that we do not have direct conscious access to. The knowledge is accessible only indirectly, through its contribution to inferences that are formed pre-consciously. Thus we can say things like, "I like this car, but I don't know why." Interesting patterns of choice will arise if we assume that, in addition, there is some information which is available to the conscious system and not to the pre-conscious system. The final output of the two systems will therefore reflect an imperfect aggregation of the two sets of information, and we will show that the decisions produced by such a setup will exhibit implicit preferences in our sense.

Implicit preferences in choice can reflect one of two different effects. First, suppose there is some attribute that we consciously wish to ignore, then it will tend to have a bigger effect in less revealing comparisons, because in those comparisons it will become harder to infer the contribution of that attribute. For example suppose that I regard weather as irrelevant in judging the quality of an apartment, but I know that weather may influence my intuitions about quality: then weather would have a bigger influence when comparisons are less revealing, i.e. it would manifest as an implicit preference in choice over apartments. Second, suppose there is some attribute that we do not wish to ignore (i.e., we trust our intuition regarding this attribute), then it will tend to have a smaller effect in less revealing comparisons, because in those situations it becomes harder for us to infer the contribution of that attribute. For example suppose we trust our instincts about the color of a car, but not about other aspects, then our instincts will have a larger effect when comparing two cars which differ only in their color, than when the cars differ in other respects, i.e. it would manifest as an implicit preference in choice over cars.

Formally, the evaluation of an outcome depends on, besides x, two vectors of random variables, $\omega \in \mathbb{R}^n$, and $\pi \in \{0,1\}^n$, all of which are independent, with the functional form:

$$u(x, \pi, \omega) = \sum_{i=1}^{n} (x_i - \frac{1}{2}) \pi_i \omega_i.$$

There are two agents in the model, which we call System 1 and System 2, and which operate sequentially. System 1 first calculates the expected utility of each x, using ω, but without access to π:

$$v = E[u|x, \omega] = \sum_{i=1}^{n} (x_i - \frac{1}{2}) E[\pi_i] \omega_i.$$

We describe the variables $\omega_1, \ldots, \omega_n$ as *associations*, and assume that they are constant across outcomes (i.e., x and x' share the same associations). System 2 receives a vector of evaluations from System 1 (evaluations $\mathbf{v}$ of cases $\mathbf{x}$), and forms its own evaluations, adding its own knowledge of π, calculating:

$$y = E[u|\mathbf{x}, \pi, \mathbf{v}] = \sum_{i=1}^{n} (x_i - \frac{1}{2}) \pi_i E[\omega_i|\mathbf{x}, \mathbf{v}].$$

We also assume that System 2 has Gaussian priors over the weights ($\omega_i \sim N(0, \sigma_i^2)$), and we will normalize $E[\pi_i]$ to equal 1 for all i.

Proposition 12. *The decisions and evaluations of an implicit-knowledge decision-maker can both be represented with a separable implicit preference utility function,*

$$y(x, A) = \sum_{i=1}^{n} (x_i - \frac{1}{2}) \omega_i (\pi_i - \frac{1}{2}) \Gamma_i(A)$$

where

$$\Gamma_i(A) = \begin{cases} \frac{1}{\pi_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} \pi_j & , i \in S \\ \frac{1}{\pi_i - \frac{1}{2}} \frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \sum_{j \in D} \pi_j & , i \in D \end{cases}$$

and $i \in S \iff x_i = x_i'$, *and* $i \in D \iff x_i \neq x_i'$.

Proof. First consider what happens when evaluating a single outcome, x: the logic of Gaussian signal extraction implies that inferring $\omega_1, \ldots, \omega_n$ from v simply involves splitting up the sum v into its parts, weighted by their respective variances (σ_i^2). When we have two outcomes (x, x') and two evaluations (v, v'), we can use the same principle

but now we are inferring the common attributes ($i \in S$) from the average ($\frac{v+v'}{2}$), and the distinctive attributes ($i \in D$) from the difference ($\frac{v-v'}{2}$):

$$E[\omega_i|x, x', v, v'] = \begin{cases} \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2 (x_i - \frac{1}{2})^2}{\sum_{j \in S} \sigma_j^2 (x_j - \frac{1}{2})^2} \frac{v+v'}{2} & , i \in S \\ \frac{1}{x_i - \frac{1}{2}} \frac{\sigma_i^2 (x_i - \frac{1}{2})^2}{\sum_{j \in D} \sigma_j^2 (x_j - \frac{1}{2})^2} \frac{v-v'}{2} & , i \in D \end{cases}$$

Substituting this into the expression for y we get:

$$\begin{aligned} y(x, A) &= \sum_{i \in S} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{v+v'}{2}\right) + \sum_{i \in D} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \frac{v-v'}{2}\right) \\ &= \sum_{i \in S} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} \omega_i (x_i - \frac{1}{2})\right) + \sum_{i \in D} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \sum_{j \in D} \omega_i (x_i - \frac{1}{2})\right) \\ &= \sum_{i \in S} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} \omega_j (x_j - \frac{1}{2})\right) + \sum_{i \in D} \pi_i \left(\frac{\sigma_i^2}{\sum_{j \in D} \sigma_j^2} \sum_{j \in D} \omega_j (x_j - \frac{1}{2})\right) \\ &= \left(\sum_{i \in S} \omega_i (x_i - \frac{1}{2})\right) \left(\frac{\sum_{i \in S} \pi_i \sigma_i^2}{\sum_{i \in S} \sigma_i^2}\right) + \left(\sum_{i \in D} \omega_i (x_i - \frac{1}{2})\right) \left(\frac{\sum_{i \in D} \pi_i \sigma_i^2}{\sum_{i \in D} \sigma_i^2}\right) \end{aligned}$$

We can write the last expression as,

$$y(x, A) = \left(\sum_{i \in S} \omega_i (x_i - \frac{1}{2})\right) \bar{\pi}_S + \left(\sum_{i \in D} \omega_i (x_i - \frac{1}{2})\right) \bar{\pi}_D \tag{6}$$

where $\bar{\pi}_S$ and $\bar{\pi}_D$ represent weighted averages of π_i, among the S and D attributes, respectively. The proposition follows by substituting $\frac{\pi_i - \frac{1}{2}}{\pi_i - \frac{1}{2}}$ into both expressions. □

The expression 6 has a simple interpretation: we infer the associations which feed into the intuitive valuations (v and v'), by separately identifying the common attributes from the average valuation ($\frac{v+v'}{2}$), and identifying the distinctive attributes from the difference in evaluations ($\frac{v-v'}{2}$). Thus the marginal effect of an association on a final valuation will depend on whether it is a common or distinctive attribute:

$$\frac{dy}{d\omega_i} = \begin{cases} \bar{\pi}_S & , i \in S \\ \bar{\pi}_D & , i \in D \end{cases}$$

Ideally System 2 wishes to weight each association ω_i according to its corresponding factor π_i. Because it does not know the associations it cannot do that, and it instead

uses the *average* correction factors $\bar{\pi}_S$ and $\bar{\pi}_D$. It will be able to perfectly identify π_i when the two outcomes differ only in attribute i (because then $\bar{\pi}_D = \pi_i$). But when the outcomes differ in more than one attribute we will tend to over-react to attributes we wish to ignore ($\kappa_i = 0$), and under-react to associations we wish to follow ($\kappa_i = 1$).

Proposition 13. *The function $\Gamma_i(A)$ satisfies betweenness and equivalence among choice sets which differ in two or fewer attributes (i.e., for any $A = \{x, x'\}$ such that $\sum_{i=1}^{n} |x_i - x'_i| \leq 2$).*

Proof. We wish to show that, for every attribute on which x and x' differ, Γ_i will be smaller when the comparison comes to be more different (i.e., in $\{x, x''\}$ compared to $\{x, x'\}$). This means showing that,

$$\frac{1}{\pi_i - \frac{1}{2}}\pi_i \geq \frac{1}{\pi_i - \frac{1}{2}}\frac{\sigma_i^2\pi_i + \sigma_j^2\pi_j}{\sigma_i^2 + \sigma_j^2}.$$

It can be seen that this inequality will hold both for $\pi_i = 0$ and for $\pi_i = 1$, whatever the value of π_j. Equivalence follows directly. □

Betweenness may be violated when outcomes differ on more than two attributes. This is because, when bundling some attribute i with additional attributes, the direction of the effect on $\bar{\pi}_D$ is ambiguous: it will increase if the new attribute is relevant ($\pi_i = 1$), but it will decrease if the new attribute is irrelevant ($\pi_i = 0$). For example suppose that someone is trying to ignore race: then when outcomes differ on both race and education this will increase the effect of race, but increasing the number of differences by adding variation in age has an ambiguous effect. For this reason the model does not generate implicit preferences, in our sense, when outcomes differ in more than 2 attributes.

For a similar reason strong betweenness does not hold, and therefore our propositions about identification of implicit preference from evaluation-data will not apply to this model. This is because changes to the bundling of attributes have an ambiguous effect on $\bar{\pi}_C$, and so an ambiguous effect on evaluations.

These qualifications imply that, under the implicit knowledge model, implicit preferences can be identified endogenously only when choice involves variation in 2 attributes. However the model still remains useful in other sense .

The model can also be applied to data on evaluations, but it requires ancillary assumptions. Suppose that we have data on sentencing. Then if we assume that people are trying to ignore race ($\pi_i = 0$), but follow their instincts for other attributes

($\pi_i = 1$), then this will generate an array of testable implications: (1) evaluations of both black and white defendants should increase when the comparator is white; and (2) evaluations of black defendants should decrease, and evaluations of white defendants should increase, when the comparator becomes same-race.

D Existing Data on Implicit Preferences

D.1 Snyder et al. (1979) on discrimination

Snyder, Kleck, Strenta, and Mentzer (1979) (SKSM) report an experiment which compares direct and indirect choices as a "general strategy for detecting motives that people wish to conceal." Their name for this general phenomenon is "attributional ambiguity," and their description, although not formalized, comes very close to our basic analysis of revealingness and implicit preferences. Subjects were invited to choose between sitting in one of two booths, in each of which a movie was being shown. Each booth already contained another person, who was either seated or was sitting in a wheelchair. The treatments varied in whether the booths were labelled to show either the same, or different, movies. The paper found that when the movies were the same, 75% (18/24) of subjects sat with the handicapped confederate, while when they were different only 33% (8/24) chose to sit with the handicapped confederate, intuitively pointing to an implicit preference against sitting with the handicapped individual: they write "avoidance of the handicapped ... masquerade[d] as a movie preference."

However, this is not an appropriate test for implicit preferences, and in fact a rational decision-maker with strong preferences over movies and weak preferences over which confederate to sit with will exhibit the same pattern of choice. Instead we need to check for a figure-8 cycle, keeping in mind the appropriate triangle inequality (Regenwetter et al. (2011)) since the data are between-subjects. We find that the condition is not satisfied, i.e. the choices observed can be rationalized by conventional transitive preferences, and we provide an example. Collecting data on indifferences, collecting within-subjects data, or collecting data on evaluation may increase the ability to detect implicit preferences in this paradigm.

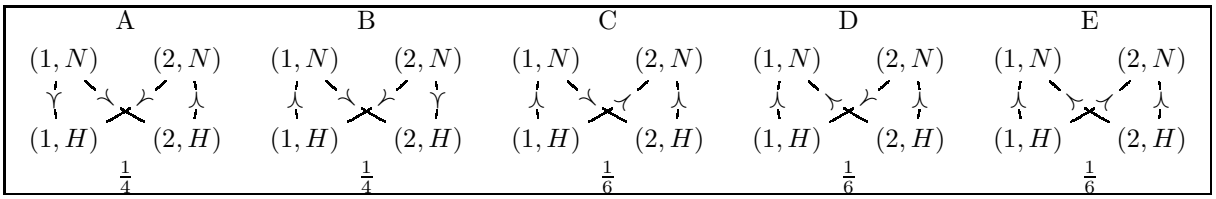
Subjects were invited to choose between sitting in one of two booths, in each of which a movie was being shown. Each booth already contained another person, who

was either seated or was sitting in a wheelchair. The treatments varied in whether the booths were labelled to show either the same, or different, movies (the identities of the movies were cross-randomized against the assignment of the wheelchair). The paper found that when the movies were the same, 75% (18/24) of subjects sat with the handicapped confederate, while when they were different only 33% (8/24) chose to sit with the handicapped confederate. SKSM test for a significant difference between the fractions, and conclude that "avoidance of the handicapped ... masquerade[d] as a movie preference."

However a simple difference in choice proportions does not establish that these preferences are inconsistent: a difference in proportions can occur if subjects have sufficiently strong preferences over which movie to see. Denoting movies by $i \in \{1, 2\}$ and confederates by $j \in \{N, H\}$ (Non-handicapped, Handicapped), SKSM test the null hypothesis that $Pr((1, H) \succ (2, N)) + Pr((2, H) \succ (1, N)) = Pr((1, H) \succ (1, N)) + Pr((2, H) \succ (2, N))$. Even when subjects have context-independent, transitive preferences this condition will be violated in all but knife-edge cases. For example, the condition is violated (a strict inequality $>$) if all subjects have the preferences $(1, N) \succ (1, H) \succ (2, N) \succ (2, H)$, i.e. they prefer to see a given movie with a non-handicapped person, but strongly prefer movie 1 to movie 2.

Instead, we need to check whether people exhibit a figure-8 cycle. As discussed earlier, for aggregate data on a 4-element choice cycle to be irreconcilable with heterogeneous, transitive preferences the average rate of preference must exceed $\frac{3}{4}$ (Regenwetter et al. (2011)). SKSM find only an average rate of preference along their cycle of 71% ((75% + (1 − 33%))/2). In other words, the data that they observe could be generated by a mixture of agents each of who has a transitive choice function.[52]

[52]For example the following 5 types would generate the observed choice fractions (denoting movies by $i \in \{1, 2\}$ and confederates by $j \in \{N, H\}$ (Non-handicapped, Handicapped)):



This hypothetical distribution of preferences would imply that 75% of subjects choose to sit with the handicapped person when the movies are the same (B, C, D, E for movie 1, and A, C, D, E for movie 2), and 33% sit with the handicapped person when the movies are different (C and E when the handicapped person is viewing movie 1, and D and E when they are viewing movie 2). If we assume that preferences are separable in the movie type, then transitivity of underlying preferences requires that the average rate of preference along the cycle exceed $\frac{2}{3}$, meaning the observed data does violate transitivity, although the difference is unlikely to be statistically significant. This discussion underlines

Collecting data on data on indifferences, collecting within-subjects data, or collecting data on evaluation may increase the ability to detect implicit preferences.

D.2 Exley (2015) on self-serving biases

Exley (2015) experimentally studies "excuse-driven risk preferences," and finds that risk-preferences seem to *change* in a self-serving way. Subjects choose between payoffs to charity and payoffs to themselves: when the payoff to charity has some risk, then decision-makers are risk averse; but when the payoff to themselves has risk, then decision-makers become relatively risk-loving. Her experimental design is the closest that we are aware of to the approach we propose in this paper, and we show that her data do indeed reveal implicit preferences: under a mild assumption her data reveal "two triangles" that identify an implicit preference for self-payoffs over charity-payoffs. Some subjects also exhibit a "figure-8" cycle that reveals an additional implicit preference over risk: some subjects are implicitly risk-averse. Exley shows that excuse-driven behavior correlates with selfishness as well as the propensity to "wiggle" in a moral wiggle-room task (Dana et al. (2007)), suggesting that all three behaviors capture a common feature of preferences. This correlation is consistent with the intuition that "wiggle room" behavior can also be thought of under the banner of implicit preferences.

Subjects make five types of choices. In Exley's notation, she first elicits (using a choice list), the X that makes each individual subject indifferent between $(10, 0)$ and $(0, X)$, where $(10, 0)$ denotes \$10 for self, and \$0 for charity. Thereafter, she elicits four types of certainty equivalent. $Y^S(P^S)$ is the certain amount for self, expressed as a percentage of \$10, that makes the subject indifferent to a self-lottery P^S that pays \$10 to self with probability P^S, nothing otherwise. $Y^C(P^C)$ is the certain amount for charity, as a percentage of $\$X$, that yields indifference to a charity-lottery paying $\$X$ to charity with probability P^C. $Y^S(P^C)$ is the analogous self-dollar valuation of the charity-lottery, and $Y^C(P^S)$ the charity-dollar valuation of the self-lottery. Hence a higher value for $Y^i(P^j)$ implies a higher value assigned to that j-lottery when measured in i-dollars.

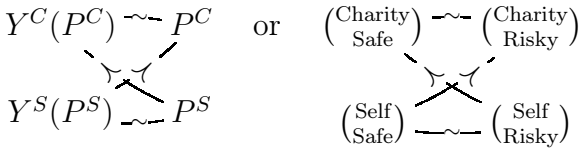
The basic choice behavior of interest, presented both graphically and in regressions, can be summed up by Exley's Prediction 3 (Excuse-driven preferences). It can be written as follows:

the care that needs to be taken in testing for implicit preferences.

$$Y^C(P^S) > Y^S(P^S) = Y^C(P^C) > Y^S(P^C)$$

In words, a self-lottery is assigned a higher valuation in charity-dollars than in self-dollars, while a charity-lottery is assigned a lower valuation in self-dollars than in charity dollars. Yet the self-dollar valuation of the self-lottery is equal to the charity-dollar valuation of the charity-lottery. This does appear intuitively consistent with an implicit preference for self, manifesting as context-specific risk preferences. Exley regresses $Y^i(P^j)$ on dummies for $\mathcal{I}[j = C]$ ("charity"), $\mathcal{I}[i \neq j]$ ("tradeoff"), and $\mathcal{I}[i = Charity] * \mathcal{I}[i \neq j]$ ("charity*tradeoff"), which allows her to check whether the average valuations satisfy the three inequalities in Prediction 3.

Perhaps surprisingly, the *observed choices alone* as described are consistent with transitivity. In order to establish intransitive cycles one must make an additional assumption, although one we think is reasonable. We translate the observed choices of an individual who is consistent with Prediction 3 into our graphical representation below:[53]



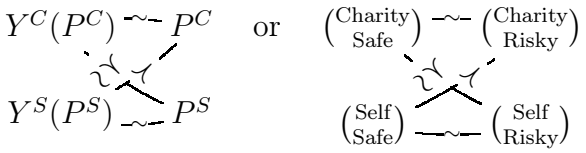
These choices represented are not intransitive, in our terminology they are consistent with a purely explicit preference for self, choosing the self-favoring option when available. Adding an assumption about preferences in the vertical choice sets allows us to establish two triangles. For example, adding $Y^S(P^S) \precsim Y^C(P^C)$ or $P^S \precsim P^C$ would be sufficient to establish a positive implicit preference for self. While subjects did face a choice between $(10, 0)$ and $(0, X)$, they did not face a choice between $(10 * Y^S(P^S), 0)$ and $(0, X * Y^C(P^C))$, and $Y^S(P^S) = Y^C(P^C)$ does not necessarily imply $Y^S(P^S) \sim Y^C(P^C)$ (e.g. it might be that $(9, 0) \succ (0, 0.9X)$). To establish inconsistency, Exley assumes that charity-dollar amounts can be translated into self-dollar amounts at exchange rate $10/X$, implying that $Y^C(P^C) = Y^S(P^S) \Leftrightarrow Y^C(P^C) \sim Y^S(P^S)$ and thus yielding two triangles that together reveal an implicit preference for self (triangle 1 is $Y^C(P^C) \sim P^C \succ Y^C(P^S) \sim Y^C(P^C)$ and triangle 2 is $Y^S(P^S) \sim P^S \succ Y^C(P^C) \sim Y^S(P^S)$).

We do believe this is a reasonable assumption, because a) Exley detects strong

[53]The horizontal indifferences follow from the fact that the horizontals are directly elicited via the choice list. The diagonals follow from the fact that, for example, $Y^C(P^S) > Y^C(P^C)$, and therefore we know by the structure of the choice list that the subject also chose P^S over $Y^C(P^C)$ when presented with the choice (because preferences are monotone in self-dollars and charity-dollars).

effects even for probabilities P^i close to one, where it is reasonable to assume that approximate indifference between self and charity safe payoffs is preserved, and b) as Exley points out, by construction $P^C \sim P^S$ is implied by any theory of risk preferences that obeys the Independence axiom.[54] We raise the point because it determines what is, under our framework, the optimal approach to detecting implicit preferences in Exley's data. *With* the assumption, one need not run any regression. By exploiting the within-subjects nature of the data one can directly identify the individuals who exhibit the relevant cycles.[55]Meanwhile, pooled regressions (which compare average valuations) do not exploit the power of within-subject data and risk missing the presence of implicit preferences amidst the sampling variation. They also require stronger restrictions on the underlying heterogeneity of implicit preferences. *Without* the assumption either analysis risks mistaking transitive underlying preferences for implicit preferences. We note that an advantage of the pooled regressions is the ability to correlate the magnitude of deviations from risk-neutrality in the diagonal choice sets, with the tendency to exhibit "wiggle-room" behavior, and with selfishness (high value of X). A similar exercise would also be possible from a binary classification of subjects by their implicit preferences as identified in the within-subjects data.

Finally, we do also find some data on subject-level intransitive choices, reported in Exley's footnote 29: for $P = 0.95$, 42 percent of subjects made intransitive choices summarized by $Y^S(P^S) > Y^S(P^C)$ and $Y^C(P^S) \leq Y^C(P^C)$. Using $P^S \sim Y^S(P^S)$, $Y^C(P^C) \sim P^C$, and the structure of the choice set (which tells us, for example, that the subject chose $Y^S(P^S)$ over P^C when given the choice) we can translate the choices into our representation, revealing an intransitive figure-8 cycle:



These choices identify an implicit negative preference for *risk*: 42% of subjects are systematically more risk averse in the less revealing (diagonal) choice sets. An

[54]Additionally, valuations are very close to risk neutral when valuing self-lotteries in self-dollars and charity-lotteries in charity-dollars, and furthermore when the linear translation is applied she finds that $Y^S(P^S) \approx Y^C(P^C)$ for all probabilities, i.e. linearity seems to do a good job of predicting behavior.

[55]Although Exley does not explicitly report how many individuals are cyclical, she tells us that 78% of choices exhibit $Y^i(P^j) \neq Y^j(P^j)$. Assuming that all of these differences go in the predicted direction, this implies that at least 78% of subjects revealed at least one of the diagonal preferences (at least one triangle), and that at least 56% of people exhibited both diagonals at least once for a given P (two triangles).

explanation could be that riskiness makes people uneasy, but they find it difficult to rationalize that uneasiness, unless the riskiness is bundled some other other attribute, as it is in the diagonal choice sets.

Taking the evidence as a whole, it seems that Exley's subjects do exhibit implicit preferences over Self-Charity, and we also find evidence for implicit preferences over Risky-Safe.

D.3 DeSante (2013) on racial discrimination

DeSante (2013) finds racial bias in an experiment where subjects are asked to set welfare payments for applicants who vary in various attributes. In his experiment two applicants are evaluated at once, allowing us to test for implicit preferences. Reanalyzing the data we find evidence for a negative implicit preference for black candidates, and additionally a negative implicit preference for candidates with high "work ethic".

In DeSante (2013) 1,000 subjects were asked to provide recommendations for state welfare payments to two hypothetical recipients - i.e. evaluation data on pairs of outcomes. Subjects were told they had $1,500 which they should allocate between two applicants presented side-by-side, with any remainder to be "added to state funds." Applicants differ in race (Black, White) signaled by name, and "work ethic" (Poor, Excellent).[56] The design is not ideal because the budget constraint creates a trade-off between the two applicants, whereas true joint evaluation does not impose choice-set dependence. Nevertheless, it is the closest example to our proposed method of which we are aware. Since the paper does not perform our specific comparisons of interest, we obtained and reanalyzed the experimental data.

We focus on the half of the subjects who were shown applicants with both attributes.[57] For simplicity we ignore "background" attributes which are held constant, which is equivalent to assuming there are no implicit preferences over these attributes. There are four types of evaluation sets: $\{WE, BP\}$, $\{WE, WP\}$, $\{BE, WP\}$ and

[56] Latoya and Keisha for black applicants, Laurie and Emily for whites. Because only two names are used for each race, and in fact only one name for each race appears in the mixed-race evaluation sets it is difficult to conclusively separate implicit preferences over names from implicit preferences over race.

[57] The other half did not have work ethic information, so in our framework can be thought of as having only one attribute, race. This can be analyzed as a variant of joint/separate evaluation in our framework (the evaluation sets are $\{W, W\}, \{B, B\}, \{W, B\}$ where the "separate" sets contain two applicants of the same race) and actually yield stronger evidence consistent with an implicit bias (Blacks applicants are are allocated $557 on average when alongside another black applicant, and $600 on average alongside a white applicant, while white applicants receive $583 when alongside a white and $556 when alongside a black applicant), but we focus on the two-attribute treatment.

$\{BE, BP\}$. There exist two parallel scissors:

$$(1) \quad y(WE|BP) - y(WE|WP) \text{ and } y(BE|WP) - y(BE|BP)$$
$$(2) \quad y(WP|BE) - y(WP|WE) \text{ and } y(BP|WE) - y(BP|BE)$$

For (1) we find $y(WE|BP) < y(WE|WP)$, inconsistent with a weakly positive implicit preference for both Black and Excellent, and $y(BE|WP) < y(BE|BP)$, inconsistent with a weakly positive preference for both White and Excellent. Hence the data in (1) imply a negative implicit preference for Excellent, but are inconclusive about race.[58]

For (2) we find $y(WP|BE) < y(WP|WE)$, inconsistent with a weakly positive implicit preference for both Black and Poor and $y(BP|WE) > y(BP|BE)$, inconsistent with a weakly positive implicit preference for both Black and Excellent, and therefore implying a negative implicit preference for Black applicants.[59]

D.4 Bohnet et al. (2015) on gender preferences

A recent experimental paper by Bohnet et al. (2015) can be interpreted as studying implicit preferences. They study whether a decision-maker's choice between candidates for a task becomes more or less sensitive to certain attributes–gender and past performance–when the choice is either between an individual candidate and an unobserved "pool" alternative, or between two candidates and the pool (a paradigm closely related to joint and separate evaluation, although it elicits choices, not evaluations). They find that "disadvantaged gender" candidates are less likely to be selected when considered individually than when considered alongside an advantaged gender alternative. On the contrary, low ability candidates are more likely to be selected when considered

[58]Note that this test is *not* a valid parallel scissors in general for detecting an implicit preference over work ethic: the target in both scissors is Excellent, while parallel scissors would require one Poor and one Excellent target (see Proposition 4). We are nevertheless able to make a conclusive statement about work ethic preferences because we have only two attributes, and hence two scissors can be sufficient (see Appendix A.2).

[59]For simplicity, as in our discussion of Snyder et al. (1979) we have treated the sample averages as though they were population averages. In fact, of the four scissors only $y(WE|BP) - y(WE|WP) < 0$ is statistically significant ($p = 0.052$). There are various approaches one could take for performing joint hypothesis tests, a non-trivial issue due to the compound directional hypotheses under consideration. For example, a standard F- or Chi-square test would test the joint null that all scissors are equal to zero, hence a rejection provides evidence for some choice-set dependence but not immediately about its form.

individually than when the alternative is a high ability candidate.

While intuitively the variation in frequency of certain choices points to implicit preferences (as we have argued, considering multiple candidates increases revealingness with respect to their attributes), in fact we show that it is not possible to infer implicit preferences from these data: regular, transitive preferences will generate the patterns of choice that Bohnet et al. (2015)'s tests interpret as varying sensitivity. A simple example illustrates the basic point: imagine a decision-maker whose preferences over quality are $High \succ Low \succ Pool$. Then they are "more sensitive" to quality in the choice set $\{High, Low, Pool\}$ (High is always chosen and Low is never chosen), than in $\{High, Pool\}$ and $\{Low, Pool\}$ (High and Low are equally likely to be chosen in each case). We do however show that Bohnet et al. (2015)'s subjects exhibit violations of WARP that point to implicit preferences, though we believe are harder to interpret. Instead, the most natural way to test for implicit preferences in their paradigm would be to collect *evaluation* data, and conduct our scissors tests.

In the experiment, lab subjects choose between candidates to perform a task, and are rewarded according to the chosen candidate's performance. Subjects observe the candidates' gender and prior performance (above/below average). They can also always choose instead to take a random draw from the pool of other candidates. Subjects face a single choice from one of two kinds of choice set: "separate" (a choice between one candidate or the pool), and "joint" (a choice between a male candidate and a female candidate, each with different level of prior performance, and the pool). Finally, Bohnet et al. (2015) also vary the type of task (math/verbal), to test whether gender bias is task-specific, with the maintained hypothesis that males are the "advantaged" gender in math, and females in verbal. To pool the data, they recode gender as "advantaged/disadvantaged." The paper is interested in whether subjects' choices reveal a gender influence that varies between treatments, with the hypothesis that the influence of gender decreases (and the influence of quality increases) in the joint treatment.

To aid comparison, we need to clarify terminology. Bohnet et al. (2015) refer to the two treatments as *separate evaluation* and *joint evaluation* respectively. In our terminology they are eliciting choices, not evaluations. While we will retain their use of "joint" and "separate," formally we are studying choices from either binary or ternary choice sets.

Bohnet et al. (2015)'s main analysis studies choice proportions, comparing the relative likelihood that the advantaged/disadvantaged candidate (averaging over quality) or high/low candidate (averaging over gender) is chosen, between joint and separate.

They find that advantaged candidates are more likely to be chosen from "separate" choice sets, but approximately equally likely in joint, and they find the opposite pattern for high performance candidates. This is interpreted as showing that choices are more sensitive to gender, and less sensitive to quality, in the separate treatment. Next, they run a Probit regression estimating the likelihood that a candidate is chosen as a function of gender and quality. In the separate treatment the coefficient on gender is large and significant and the coefficient on quality is small and not significant, and the reverse is found in the joint treatment. This is again interpreted as evidence for changing sensitivity between modes.

It is clear that the aggregate sensitivity of choice likelihood varies between joint and separate treatments, a fact that is surely of interest to policymakers. Interestingly, however, this analysis is *not* sufficient to conclude that subjects' underlying gender and quality preferences are choice-set dependent (or, specifically, implicit). To see this, observe that the test for changing sensitivity to quality has the following null (where gender is A/D, quality is H/L and Pool is "P"):

$$\begin{aligned}
&\frac{Pr(AH = c\{AH, DL, P\}) + Pr(DH = c\{DH, AL, P\})}{2} \\
-&\frac{Pr(AL = c\{AL, DH, P\}) + Pr(DL = c\{DL, AH, P\})}{2} \\
=&\frac{Pr(AH = c\{AH, P\}) + Pr(DH = c\{DH, P\})}{2} \\
-&\frac{Pr(AL = c\{AL, P\}) + Pr(DL = c\{DL, P\})}{2}
\end{aligned}$$

which simplifies to $Pr(AH \succ DL \succ P) + Pr(DH \succ AL \succ P) = Pr(DL \succ AH \succ P) + Pr(AL \succ DH \succ P)$. This condition will be violated–and hence a large sample test will reject with probability one–in all but knife-edge cases, a simple example would be where all subjects have $DH \succ AH \succ DL \succ AL \succ P$. Hence a rejection of the null is fully consistent with context-independent (transitive) underlying preferences.[60] An equivalent argument shows that the test for changing sensitivity to gender is also

[60] One can also argue that Bohnet et al.'s finding of higher sensitivity to quality in Joint (i.e. a strict inequality $>$) should be *expected.* Inspection of the condition reveals that for the opposite inequality ($<$) to hold, a relatively large fraction of subjects must exhibit strong preferences over gender such that a Low candidate of one gender is preferred to a High candidate of the other. Of course, knowing whether gender preferences are strong or weak is of interest, but is quite distinct from the hypothesis of context-dependent preferences.

Table 2: Heterogeneity and choice proportions

	Gender			Quality		
	A	D	Gap	H	L	Gap
Separate	75%	50%	25%	62.5%	62.5%	0%
Joint	37.5%	37.5%	0%	50%	25%	25%

Note: choice proportions constructed from example in the text.

invalid.

One can construct populations of purely transitive individuals that replicate the pattern of Bohnet et al.'s results for both gender and quality. In Example 3 we show that conventional context-*independent* preferences can generate aggregate choice proportions which under the approach in Bohnet et al. would be interpreted as evidence for context-*dependent* preference.

Example 3. Suppose subjects had the following transitive preferences, in equal proportions: (1) $AH \succ AL \succ Pool \succ DH \succ DL$; (2) $DH \succ AH \succ DL \succ AL \succ Pool$; (3) $DH \succ DL \succ AH \succ AL \succ Pool$; (4) $Pool \succ DH \succ DL \succ AH \succ AL$. Averaging over the six possible choice sets in Bohnet et al. (2015)[61], we would observe an apparent gender advantage in separate that disappears in joint, and an apparent quality advantage in joint that disappears in separate (see Table 2). In other words, aggregate sensitivity to gender and quality varies, while the sensitivity of the underlying preferences does not. The assumed individual preferences actually tend to favor *disadvantaged* candidates: types 3 and 4 always rank D above A, while type 2 prefers D to A in 3 out of 4 pairwise comparisons.

In the regression analysis, the maintained assumption is that subjects' preferences over candidates can be characterized by a representative utility function, perturbed by independent (Gaussian or extreme-value distributed) shocks. It therefore cannot capture the correlation between preferences generated, for example, by the types in our above examples, and will tend to incorrectly identify instability in preferences if the underlying individual utilities are stable but correlated.

Turning to our own model, because we have defined it only over binary choice sets (with the exception of the discussion in B), and binary attributes (which the "Pool" is not), it is difficult to prescribe the optimal test for implicit preferences Bohnet et al.'s

[61] Specifically, $\{AH, Pool\}, \{DH, Pool\}, \{AL, Pool\}, \{DL, Pool\}, \{AH, DL, Pool\}, \{AL, DH, Pool\}$.

Table 3: WARP violations in Bohnet et al. (2015) data

	Joint choice	Separate choice	p-value
Math	$Pr(FH = c\{FH, ML, P\}) = 57\%$	$Pr(FH = c\{FH, P\}) = 44\%$	0.14
Math	$Pr(P = c\{FH, ML, P\}) = 40\%$	$Pr(P = c\{ML, P\}) = 35\%$	0.33
Math	$Pr(P = c\{MH, FL, P\}) = 42\%$	$Pr(P = c\{MH, P\}) = 34\%$	0.28
Verbal	$Pr(P = c\{FH, ML, P\}) = 38\%$	$Pr(P = c\{FH, P\}) = 19\%$	0.08*

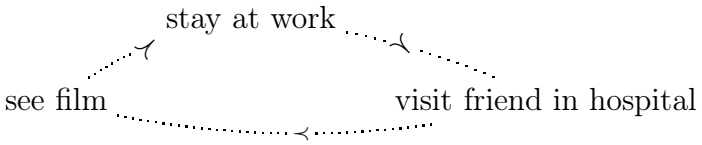
Note: p-values from one-sided test of proportions.

data. However, there is one natural test for violations of rationality: one can check for violations of the Weak Axiom of Revealed Preferences (WARP). WARP requires that subjects do not become more likely to choose a given candidate when the choice set increases in size. Under the maintained assumption that deviations from rationality are generated only by implicit preferences, detecting such a violation confirms their existence (but does not, we believe, identify over which attribute). We find four WARP violations, of which one is marginally significant, detailed in Table 3.

E Additional Notes on Interpretation

E.1 Other types of cycle

We briefly mention some cycles that are more difficult to fit into our existing framework. Consider the following example, adapted from Cherepanov et al. (2013):



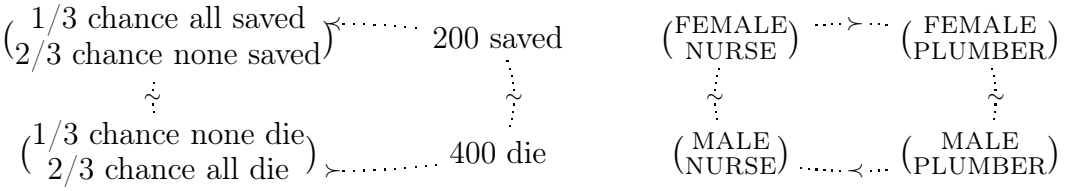
Here there seems to be some sort of implicit preference, but each alternative is idiosyncratic - i.e. it is not obvious that the cycle can be described as due to any one pair of outcomes sharing an attribute that the remaining outcome lacks (as occurs in a right-triangle cycle). We could call this type of cycle an *equilateral* cycle.

An intuitive way to rationalize this cycle in the signaling model would be if the value of staying at work tended to be highly uncertain to an observer, due to you having private information about how urgent your work is. In our signaling model the revealingness of a choice declines as the variance of prior beliefs over the value of each

alternative increases: i.e., if I have very broad priors over the value of attribute i, then when I observe you choosing i over j, I will not update much about the value of j. Thus one interpretation of why this hypothetical cycle seems compelling is that we feel the decision-maker to be motivated to visit their friend by, in part, signaling considerations, and so when making a choice against staying at work the signaling incentive is relatively weaker, i.e. people will not infer much about your dedication to your friend from this choice.

Broadening this observation, outcomes with values is strongly unobservable will tend to amplify implicit preferences. In everyday life we have certain dimensions of choice which are deeply unobservable - one can always say "I'm tired," "I'm sick," or "I have a lot of work to do," - and when available these choices allow a person to express their implicit preferences with a lower reputational cost. When designing an experiment it may be wise to include attributes that are deeply unobservable in this sense, to maximize power in detecting implicit preferences.

One other type of cycle is worth noting: cases where the *separability* of implicit preferences is violated. The following two examples show intransitive cycles which appear to be driven by some kind of implicit preference, but which violate separability. On the left we show choices from the "Asian disease" problem described in **?**, represented in a space with two attributes (safe/risky and gain/loss), and supplemented with the assumption that people would be indifferent between outcomes which differ only in whether described as gains or losses. On the right we show choices from a decision-maker who implicitly prefers nurses to be female, and implicitly prefers plumbers to be male.



In both of the cases discussed we could design an experiment to isolate a single implicit preference by holding one of the attributes fixed: e.g., test for implicit preferences for women just among nurses (either by adding a third attribute, or by using an isosceles cycle).

E.2 Larger choice sets.

It is natural to ask how implicit preferences will be revealed in choice and evaluation of sets larger than two elements. However in attempting to answer these questions we find that the predictions of the different foundational models diverge. Thus the *general* concept of an implicit preference, which is agnostic among different foundations, is limited to binary choice. More precisely, we find that revealingness is most naturally interpreted as a property of a *choice* in the signaling and ceteris-paribus models, while it is most naturally intepreted as a property of the *choice set* in the implicit knowledge model. When considering only binary choice sets we can consider it as a property of the choice set because, in the signaling and ceteris-paribus models, both choices have the same revealingness (i.e., the same differential effect on the observer's beliefs).

Here we make a few observations in lieu of a systematic discussion of choice and evaluation of larger sets of outcomes.

1. In the implicit knowledge model adding an extra outcome increases revealingness with respect to all attributes because the extra outcome causes the decision-maker to learn more about his own preferences (both for choice and evaluation).

2. In the signaling model with evaluation, adding an alternative can *decrease* the revealingness with respect to a particular dimension. Consider a decision-maker giving sentences to two defendants, with attributes $\binom{\text{BLACK}}{\text{ASSAULT}}$ and $\binom{\text{WHITE}}{\text{BURGLARY}}$. Their sentences will be restrained by signaling considerations. Now consider a set of three defendants:

$$\binom{\text{BLACK}}{\text{BURGLARY}} \qquad \binom{\text{BLACK}}{\text{ASSAULT}}$$

$$\binom{\text{WHITE}}{\text{BURGLARY}}$$

   The decision-maker's racial preference is exactly identified by the difference between the sentences given to $\binom{\text{BLACK}}{\text{BURGLARY}}$ and $\binom{\text{WHITE}}{\text{BURGLARY}}$, therefore the sentence given to $\binom{\text{BLACK}}{\text{ASSAULT}}$ will have no effect on the observer's beliefs about racial preferences. Thus, in the signaling model, adding an outcome can reduce the revealingness of an evaluation.

3. We discuss above, in Section 3.4, how choice from 3-element sets can distinguish between the ceteris-paribus and implicit-knowledge models.

4. An interesting implication of the implicit-knowledge model is that discrimination

should be a U-shaped function of the composition of a group. Consider a decision-maker who must assign salaries to a set of employees, each of whom differs in some idiosyncratic attribute, as well as varying in gender. The model predicts that gender bias in salaries will be decreasing and then increasing in the fraction of male employees, because the decision-maker learns the most about her bias when the fraction is 50-50.[62] It is not clear whether this will hold in the signaling model: on the one hand, a 50-50 composition will reveal, to the observer, the most information about the decision-maker's gender bias; but on the other hand, it is more costly to implement equal salaries when there is a 50-50 composition.

[62]Suppose that male number i has a value of $\omega_i + \omega_M$, and that female i has a value of $\omega_i - \omega_M$. If there is only a single female, then it is not possible to separate the contribution of ω_M from her idiosyncratic value ω_i. As we increase the share of females we learn more about ω_M.