Implicit Preferences Inferred from Choice[*]

PRELIMINARY DRAFT, COMMENTS WELCOME

Tom Cunningham[†] Jonathan de Quidt[‡]

February 28, 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, Sendhil Mullainathan, Antonio Rangel, and seminar audiences at USC, Columbia, Caltech, Stanford, Facebook, and the BABEEW.

[†] 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.

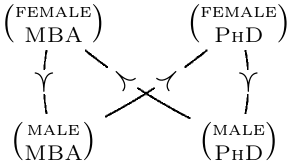
1 Introduction

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 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 who are 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.

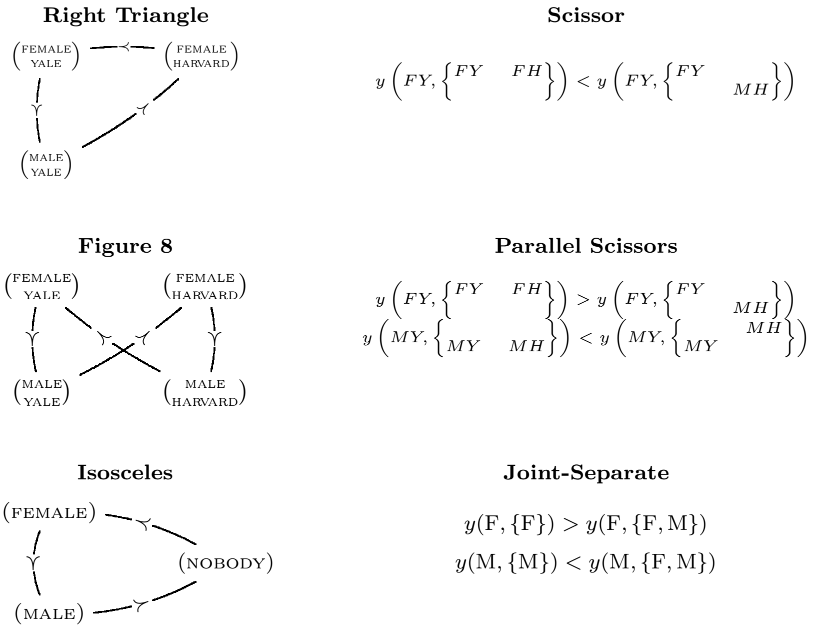


Figure 1: Six different patterns of behaviors which can identify - under certain assumptions - an implicit preference for Male over Female. Each subfigure in the first column represents an intransitive cycles of choices. Each subfigure in the second column represents menu-dependent evaluations of outcomes, 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.

We look for systematic differences in the preferences expressed in less revealing choice sets, and we define those preferences as *implicit* preferences (*explicit* preferences are those used in more revealing choice sets). Given the choices above we would infer a positive explicit preference, but a negative implicit preference, for female candidates. 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. situations in which these behaviors would not reveal implicit preferences.

The formal framework we develop additionally shows how implicit preferences can be revealed in data on evaluations. For example, 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 the 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.** 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.

- **Self-other tradeoffs.** Consider a person who would always choose to give an object of value to charity, whether it is cash or goods. But when the payoffs are different (cash to one, goods to the other), 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]

[1] E.g. "salience" (Bordalo et al. (2012)), "relative thinking" (Bushong et al. (2014)), "magnitude effects" (Cunningham (2012)), or "focusing" (Kőszegi and Szeidl (2011)). 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.

[3] The experiments in Exley (2015) have a similar structure, although that paper introduces a

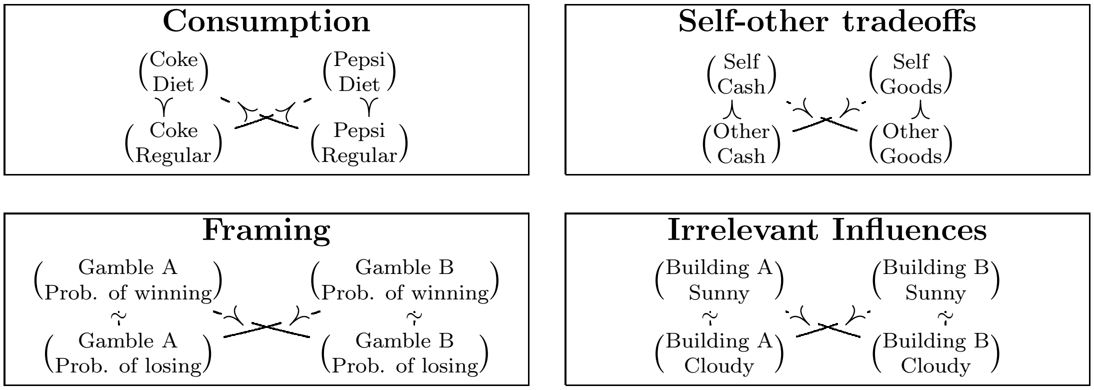


Figure 2: Figure 8 intransitivities applied to different domains.

- **Framing.** Consider a person choosing between two lotteries, one of which is described in terms of the probability of winning, the other in terms of the probability of losing. Suppose that the person is indifferent between two lotteries when they share the same objective payoffs, but when the lottery payoffs differ, they choose the one that is described with the emphasis on winning. These choices would reveal an implicit preference for the positive description, but no explicit preference. More generally, we think that many classical framing effects can be thought of as cases where a decision-maker has no explicit preference, but an implicit preference, for some attribute.[4]

- **Irrelevant Influences on Decision-Making.** Consider a person viewing two apartments, one on a sunny day and the other on a cloudy day. They are indifferent when the apartments are in the same building (i.e., identical), but when the apartments are in different buildings they tend to prefer the apartment viewed on the sunny day. This reveals an implicit preference for apartments viewed in good weather, but no explicit preference. [5]

Why would preferences change when the choice set becomes more revealing? We

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.

[4] Framing effects are usually defined as choice being influenced by a normatively irrelevant attribute, where "normative irrelevance" is imposed by assumption. Experiments rarely test whether framing effects survive in side-by-side evaluation, it is usually taken for granted that they do not. One exception is Mazar et al. (2013).

[5] A number of papers find that certain economic decisions are significantly influenced by the weather - Hirshleifer (2001), Simonsohn (2010), and Busse et al. (2013).

discuss three interpretations. Consider our introductory example of gender discrimination. First, people could be *unaware* of having a preference over gender, and correct 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.[6] 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;[7] claims in judgment and decision-making about unconscious influences;[8] evolutionary theories of self-deception in humans and other animals;[9] and economic theories of signaling in social behaviour.[10] 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 leaves opens a door for implicit influences. More narrowly, as economists we are interested domains where it is widely thought people have serious internal conflicts

[6] 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.

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

[8] e.g. Kahneman (2011).

[9] Von Hippel and Trivers (2011)

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

- in decisions concerning race, charitable giving, politics, and status goods. Our paper gives a rigorous foundation for estimating the strength of implicit influences in all 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, to make the purchase less revealing.[11]

Our measure of implicit preferences can be compared with the Implicit Association Test, which uses response time in a categorization task.[12] 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 models that generate implicit preferences (*ceteris paribus*, signaling, and implicit knowledge), additional formal results and discussion.

[11] Chance and Norton (2009).

[12] 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 (e.g. stating willingness to pay) of both members a pair of outcomes. We derive parallel 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 an outcome changes when the identity of the other outcome being evaluated changes (the "comparator").

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 either a single 3-element intransitive cycle in choices, or a single comparator-effect on evaluation. The task becomes more complicated when implicit preferences could exist over multiple attributes, for example, over both gender and qualification. Much of the 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.

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 ex-

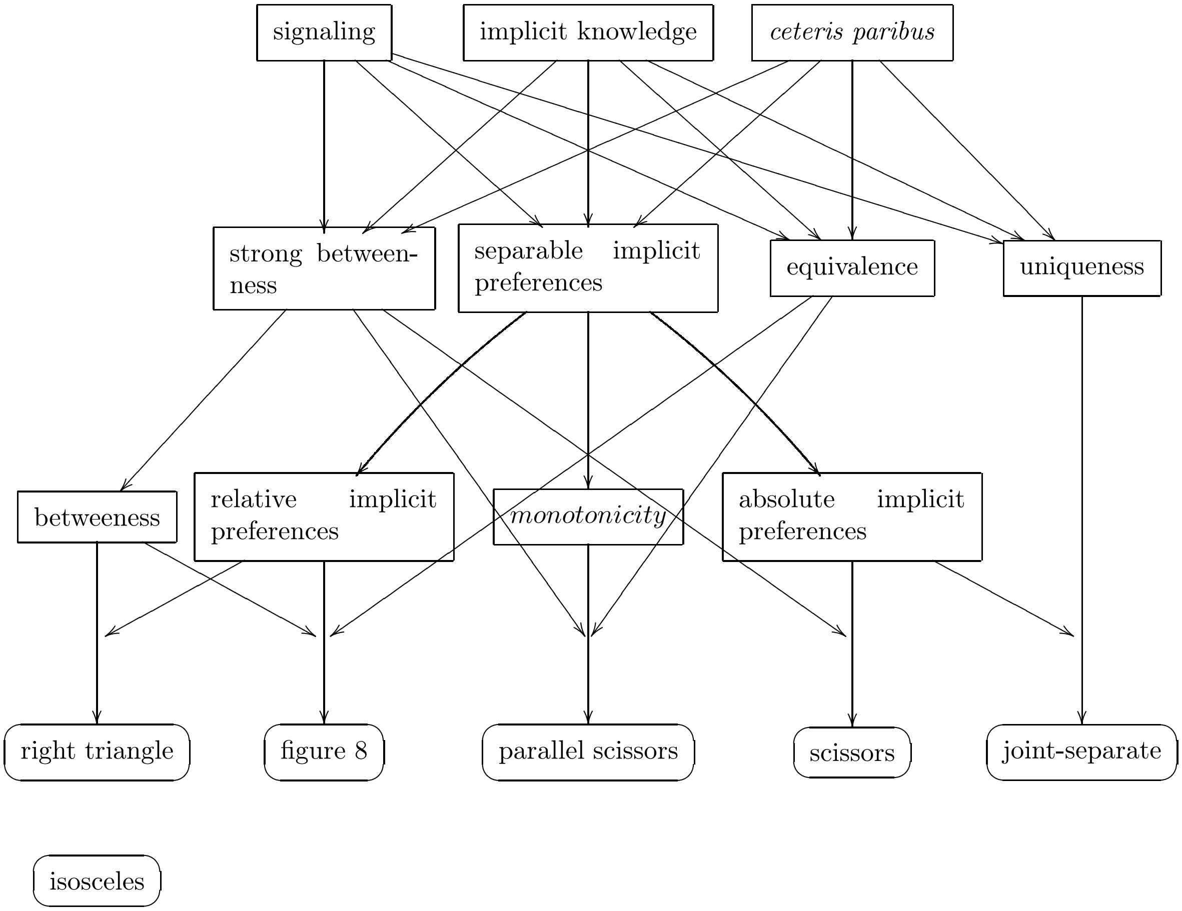


Figure 3: The relation between assumptions and results in this paper. The first row shows three parameterized models of decision-making. The last row shows five types of behavior (either choice or evaluation behavior) which have implications about implicit preferences. When arrows join before arriving at a box, it means that the gathering assumptions are jointly sufficient.

tension 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 constrained by what we term *ceteris paribus* rules 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.

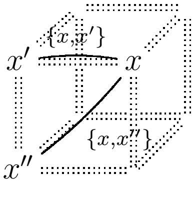
11. **Implicit knowledge.** We show that a linear-Gaussian two-system decision-maker, with imperfect knowledge of their own preferences, will exhibit separable implicit preferences in choice, provided the outcomes in the choice set differ by no more than two attributes.

2.1 Choice in a Binary Space

The space of outcomes is defined by n binary attributes, i.e. $X = \{0,1\}^n$. In many cases we will without loss of generality consider outcomes with $x_i = 1$, $\forall i$. We consider only choice sets with two elements for the majority of this paper, so the set of choice sets is $\mathcal{A} = \{\{x,y\} : x,y \in X, x \neq y\}$. A menu-dependent utility function is a function $u : X, \mathcal{A} \to \mathbb{R}$.

We assume that choice sets can be ranked in terms of *revealingness* regarding each attribute. Formally we assume that there exists a set of simple orders among choice sets, denoted $\geq_i$, where $A \geq_i B$ means that choice-set A is weakly more revealing than choice-set B with respect to attribute i.[13] The semantic interpretation of revealingness differs between the different foundations of implicit preferences. However they all satisfy the following assumption: that a choice set is less revealing about some attribute when more other attributes are bundled with it - in other words, when it becomes more diluted.

To state this clearly we first define an outcome x' as being *between* x and x'' if it is a convex combination: in the following diagram x' is between x and x''. [14] The betweenness assumption implies that the choice set $\{x, x'\}$ is relatively more revealing about the horizontal dimensions than the choice set $\{x, x''\}$.



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

[13] The symbols $>_i$, $=_i$, and $\neq_i$ are defined in the usual way relative to $\geq_i$.

[14] The diagram is drawn in 3 dimensions, but can represent an arbitrary number of attributes bundled into three groups: the attributes on which x and x' disagree plotted on the horizontal axis, those on which x' and x'' disagree plotted in the vertical axis, and those on which all three elements agree plotted in the remaining axis.

or $x'_i = x''_i$.

We now make the assumption that a strict increase in the dimension-wise distance between two elements will lower the revealingness about the attributes on which they already differ.

Assumption 1 (Betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then $\{x, x'\} \geq_i \{x, x''\}$ for all i such that $x_i \neq x'_i$.*

We now define implicit preferences: roughly speaking, a decision-maker has a positive implicit preference over an attribute if they become more likely to choose outcomes with that attribute when revealingness with respect to that attribute decreases. As an example:

Example 1. Consider a decision-maker with a positive implicit preference for whites, and consider a white and a black job candidate who are equal in every other respect. If the white candidate is preferred in one context, then they will also be preferred when revealingness with respect to race decreases.

Definition 2. We say that a menu-dependent utility function $u(x, A)$ has **relative implicit preferences** $\lambda \in \{-1, 0, 1\}^n$ with respect to a set of orderings on $\mathcal{A}$, $\{>_i\}_{i=1}^n$, if, for any $x, x' \in X$ (normalizing $x_j = 1, \forall j$) and $A, B \in \mathcal{A}$, such that for every i with $x'_i = 0$,

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

then

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

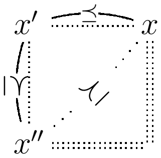
Assumption 2. *$u(x, A)$ has relative implicit preferences.*

The definition says that if the choice set B is less revealing about the attributes of x which are implicitly preferred, and is more revealing about the attributes of x which are implicitly dispreferred (relative to x'), then under B the ranking of x relative to x' can only improve.

The vector λ summarizes the implicit preferences: if $\lambda_i = +1$, then we say that u has a positive implicit preference for attribute i, if $\lambda_i = -1$ a negative implicit preference, and if $\lambda_i = 0$ no implicit preference. Our definition assumes separability of implicit preferences: for example, if the attributes are male/female and white/black, then we allow for an implicit preference for men, and for white candidates, but not one which applies just to white men. However we make no assumption about the separability of $u(x, A)$ in x_i and x_j. Specifically, conditional on the choice set we allow an arbitrary ranking of outcomes, but changes to that ranking must obey the vector of implicit preferences when revealingness changes.[15]

This definition of implicit preferences, along with betweenness, is sufficient to make basic inferences from certain intransitive choices. We will use $\succeq$ as a shorthand to denote choice from a binary choice set, i.e. $x \succeq x'$ if and only if $x \in c(\{x, x'\})$, which in turn is true if and only if $u(x, \{x, x'\}) \geq u(x', \{x, x'\})$.

We first show that a 3-element intransitive cycle which satisfies betweenness establishes a disjunction among implicit preferences. In the following diagram the observed choices reveal that the decision-maker must have a negative implicit preference for one of the attributes which x and x'' disagree on, because the relative preference for x over x'' declines in the less revealing comparison (the hypotenuse of the triangle).



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 u must have a negative implicit preference for one of the attributes on which x and x'' differ (normalizing $x_i = 1$, $\forall i$).*

Observing a single right-triangle cycle only establishes a disjunction among implicit preferences. If we are willing to assume that there exists an implicit preference on at most one, given attribute, a single right-triangle cycle is sufficient to identify it.

[15] Consider a set of candidates who are White (W) or Black (B) and Male (M) or Female (F), and a decision-maker with a positive implicit preference for males and none over race. For any given choice set A we allow the decision-maker's preferences to be non-separable in race and gender, for example: $u(WM, A) = u(BF, A) > u(BM, A) = u(WF, A)$. However, if a choice set B is less revealing with respect to gender, the decision-maker's preferences must shift in favor of males, i.e. $u(WM, B) \geq u(BF, B)$ and $u(BM, B) \geq u(WF, B)$.

Example 2. Suppose we observe the following preferences over films:

$$\begin{tikzcd}
 & \binom{\text{DOCUMENTARY}}{\text{WEDNESDAY NIGHT}} \arrow[d, dash, "\curlyvee"'] \\
\binom{\text{COMEDY}}{\text{TUESDAY NIGHT}} \arrow[ur, dash, dashed, "\prec"] \arrow[r, dash, "\sim"] & \binom{\text{COMEDY}}{\text{WEDNESDAY NIGHT}}
\end{tikzcd}$$

Under the assumption that implicit preferences exist only over film-genres this is sufficient to identify a positive implicit preference for the comedy over the documentary. However if there also could exist implicit preferences over days of the week, this cycle would not unambiguously identify the direction of any implicit preference.

We can infer unambiguous implicit preferences by combining multiple observations of choice cycles. Define the **span** m of a right-triangle-cycle as the number of dimensions on which the outcomes that lie on the hypotenuse differ (i.e., $m = \sum_{i=1}^{n} 1\{x_i \neq x_i''\}$).

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

Importantly, note that when outcomes differ in at most two attributes (such as our Male/Female-MBA/PhD example), only two right-triangles are needed. For example, consider the following:

$$\begin{tikzcd}
x' \arrow[r, dotted, dash] \arrow[r, dashed, dash, bend left, "\prec"] \arrow[d, dotted, dash] \arrow[d, dashed, dash, bend right, "\curlyvee"'] \arrow[dr, dash, "\succ"] & x \arrow[d, dotted, dash] \arrow[d, dash, bend right, "\prec"'] \\
x'' \arrow[r, dotted, dash] \arrow[r, dashed, dash, bend right, "\succ"'] \arrow[ur, dash, "\prec"'] & x'''
\end{tikzcd}$$

However note that in this case, to identify an implicit preference, we must observe either (i) two preferences on the horizontal dimension which go in different directions (a *non-monotonicity*); or (ii) two indifferences along the horizontal dimension. Further discussion and examples of combining multiple cycles can be found in the Appendix.

There is a more parsimonious way of inferring implicit preferences: from a figure-8 cycle of intransitivities. This requires an additional assumption: that revealingness depends only on the set of dimensions which differ between the outcomes considered. For example, we assume that the two choice sets, $\left\{\binom{\text{MALE}}{\text{MBA}}, \binom{\text{FEMALE}}{\text{PHD}}\right\}$ and $\left\{\binom{\text{FEMALE}}{\text{MBA}}, \binom{\text{MALE}}{\text{PHD}}\right\}$ are equally revealing about each of the attributes. In the Appendix we show that this assumption will hold in all the underlying models of implicit preferences that we consider.[16]

[16] If equivalence did not hold then the figure-8 shown could occur without any North-South implicit preferences. Suppose that $\{x, x''\}$ was more revealing about East-West preferences than $\{x', x'''\}$. A positive implicit preference for East could then cause $x'' \succ x$, while $x''' \succ x'$.

Assumption 3 (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 for all $i \in \{1, \ldots, n\}$, $\{x, x'\} =_i \{x'', x'''\}$.*

Example 3. Consider the diagram in Proposition 3. The following pairs of choice sets are equally revealing about both attributes: $\{\{x, x'\}, \{x'', x'''\}\}$, $\{\{x', x''\}, \{x, x'''\}\}$, and $\{\{x', x'''\}, \{x, x''\}\}$.

Proposition 3 (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:*

$$\begin{tikzcd}
x' \arrow[d, phantom, sloped, "\succeq"] \arrow[dr, phantom, sloped, "\succeq"] & x \arrow[d, phantom, sloped, "\succeq"] \arrow[dl, phantom, sloped, "\preceq"] \\
x'' & x'''
\end{tikzcd}$$

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.

A figure 8 does not require, unlike the pair of right-triangles, a non-monotonicity or indifference with respect to one attribute.

2.2 Evaluation in a Binary Space

We now turn to data on evaluations, applicable to, for example, bids in an auction, statements of willingness to pay, assignment of scores in judging sports, etc. $\mathcal{A}$ now represents the set of *evaluation sets*: pairs of outcomes to which the decision-maker simultaneously assigns evaluations. A menu-dependent evaluation function is a function $y : X, \mathcal{A} \to \mathbb{R}$.[17]

The main results in this section parallel those in the section on choice. We first slightly strengthen the betweenness assumption by assuming that, for a common attribute (which is irrelevant in studying choice), the revealingness weakly decreases when the total number of common attributes increases. Strong betweenness holds in all of our foundational models, and follows from the logic of signal extraction: if we think of the evaluation of each outcome in the evaluation set as informative about the value associated with the common attributes, then reducing the correlation of those evaluations will increase the accuracy of our inference about the remaining attributes.

[17] We emphasize the necessity of evaluation *sets*. It is not possible to extract implicit preferences solely from evaluations made in isolation.

Assumption 4 (Strong betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then $\{x, x'\} \geq_i \{x, x''\}$ for all i such that $x_i \neq x'_i$, and $\{x, x'\} \leq_i \{x, x''\}$ for all i such that $x_i = x'_i$.*

Second, we appropriately modify the definition of implicit preferences. Previously, in less revealing situations, preferences would switch in favor of the implicitly favored outcome, all else equal. We now assume that the effect is not just marginal but absolute: in less revealing situations the evaluations given to the implicitly favored outcome will increase and the evaluation of the disfavored outcome will decrease.

Formally, if a choice set B is less revealing about the attributes of x which are implicitly preferred, and more revealing about the attributes of x which are implicitly dispreferred, then the evaluation of x must increase when changing from A to B.

Definition 3. We say that $y(x, A)$ has **absolute implicit preferences** $\lambda \in \{-1, 0, 1\}^n$ with respect to a set of orderings on $\mathcal{A}$, $\{>_i\}_{i=1}^n$ if, for any $x \in X$ (normalizing $x_j = 1, \forall j$) and $A, B \in \mathcal{A}$ such that:

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

then

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

Assumption 5. *$y(x, A)$ has absolute implicit preferences.*

From this it follows that, if we observe the evaluation of x increase when the comparator shifts strictly away from x,[18] then there must exist either a negative implicit preference for one of x's attributes which the original comparator agreed on (for which revealingness has increased), or a positive implicit preference for one of x's attributes which the original comparator disagreed on (for which revealingness has decreased). We call this a "scissor effect."

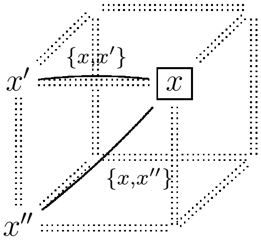
Proposition 4 (Scissor Effect)**.** *For any $x, x', x'' \in X$, with x' between x and x'' and*

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

[18] Where "strictly away" is used in the betweenness sense, that the old comparator is between x and the new comparator.

then (normalizing $x_i = 1\ \forall i$) either (i) y has a positive implicit preference for some attribute i ($\lambda_i > 0$) on which x and x' disagree ($x_i \neq x'_i$); or (ii) y has a negative implicit preference ($\lambda_i < 0$) for some attribute i on which x and x' agree ($x_i = x'_i$).

Example 4. Consider the diagram below. 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. □

Example 5. Suppose we observe the following pattern of willingness-to-pay for films:

$$y\Bigg(\begin{pmatrix}\text{COMEDY}\\ \text{WEDNESDAY NIGHT}\end{pmatrix}, \Bigg\{\begin{pmatrix}\text{COMEDY}\\ \text{WEDNESDAY NIGHT}\end{pmatrix}, \begin{pmatrix}\text{COMEDY}\\ \text{TUESDAY NIGHT}\end{pmatrix}\Bigg\}\Bigg) > y\Bigg(\begin{pmatrix}\text{COMEDY}\\ \text{WEDNESDAY NIGHT}\end{pmatrix}, \Bigg\{\Bigg(\begin{matrix}\text{COME}\\ \text{WEDNESDA}\end{matrix}$$

Under the maintained assumption that implicit preferences can exist only over film genre, this is sufficient to identify a positive implicit preference for comedies (aka a negative implicit preference for documentaries).

We discuss in Appendix A.2 how the disjunctions derived from multiple scissor effects can be combined to infer unambiguous implicit preferences. As before, a single scissor will be sufficient to identify a unique implicit preference if and only if we are willing to assume that there are no implicit preferences over other attributes.

With the addition of some minor assumptions we can infer implicit preferences from a much smaller dataset. Suppose we observe two different scissor effects composed of outcomes that are identical but flipped with respect to attribute i, meaning that the second scissors is composed of outcomes identical to the first, but for having opposite values of attribute i, compared to the corresponding outcomes in the first scissors. This essentially enables us to focus on the influence of attribute i, "controlling for" the influence of the remaining attributes $j \neq i$. If the two scissors cause opposite shifts in evaluation then we identify an unambiguous implicit preference over attribute i. We term this a *parallel scissor effect.*

The parallel scissor effect relies on two additional assumptions. First, the equivalence assumption, described above, so that revealingness is comparable between the evaluation sets.[19] Second, we assume that implicit preferences are *monotonic*, in the following sense: suppose switching evaluation sets from A to B raises the evaluation of x', then if B is less revealing about the distinctive attributes of x which are implicitly preferred, and B is more revealing about the distinctive attributes of x which are implicitly dispreferred (relative to x' and B), then switching from A to B must also raise evaluation of x.

Assumption 6 (Monotonicity). *For any $x, x' \in X$ and $A, B \in \mathcal{A}$ (normalizing $x_i = 1, \forall i$) if, for all j with $x'_j = 0$,*

$$\begin{aligned} A \geq_j B \ &\Leftrightarrow \ \lambda_j \geq 0 \\ A \leq_j B \ &\Leftrightarrow \ \lambda_j \leq 0, \end{aligned}$$

then,

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

Example 6. Consider a decision-maker with a positive implicit preference for males, and a male and female candidate who are identical in other respects. Monotonicity implies that, if B is less revealing about gender than A, and if the evaluation of the female increases when switching from evaluation set A to B, then the evaluation of the male must also increase when switching from A to B.

[19] The equivalence assumption is stronger when applied to evaluation than when applied to choice. Briefly - equivalence could be violated in a signaling model if there is differential uncertainty about the weights on each of a pair of attributes - e.g. if your evaluation of $\binom{\text{BLACK}}{\text{PHD}}$ and of $\binom{\text{WHITE}}{\text{PHD}}$ could be differentially revealing about your PhD-preference, if an observer is more certain of your white-preference than your black-preference. This issue does not seem to be important in choice, where an observer only gets information about the difference between the two realizations of an attribute (i.e., the black-white difference).

Monotonicity is not guaranteed by our basic definition of implicit preferences, because switching from A to B can change revealingness about the entire range of attributes. Monotonicity imposes that the effect of these other attributes on evaluation cannot overwhelm the effect of gender (and note that monotonicity will automatically hold if $y(\cdot,\cdot)$ is separable in the attributes of the outcome).

Proposition 5 (Parallel Scissor Effects)**.** *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

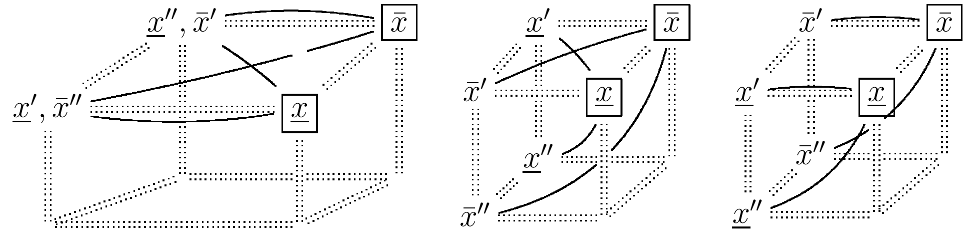
$$\begin{aligned} y^A(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}'\}) = y(\cdot, \{\underline{x}, \underline{x}'\}) \\ y^B(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}''\}) = y(\cdot, \{\underline{x}, \underline{x}''\}), \end{aligned}$$

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 surprisingly 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.

2.2.1 Joint and Separate Evaluation

With a minor additional assumption, the same logic will also allow us to infer implicit preferences from comparison of *joint* and *separate evaluation* of two outcomes $\bar{x}$ and $\underline{x}$, where joint evaluation considers evaluation set $\{\bar{x}, \underline{x}\}$, while separate evaluation considers $\{\bar{x}\}$ or $\{\underline{x}\}$.

Assumption 7 (Uniqueness)**.** *For every $x \in X$, $\{x\} =_i \{x, x\}$.*

This implies that that $y(x|\{x\}) = y(x|\{x, x\})$.[20] Then, if we find that $\underline{x}$ and $\bar{x}$ move in opposite directions when evaluated jointly relative to when evaluated separately, this reveals an implicit preference with respect to attribute i.

Example 7. 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.

2.2.2 Testing Evaluation Data

Given these results, how should one analyze a dataset on evaluations? Suppose there are 2 attributes, implying 4 outcomes and 16 conditional evaluations of the form $y(x|\{x, x'\})$.[21] Then, for each attribute, we can run 10 separate tests for implicit preferences.[22] Each test could identify either a positive implicit preference, a negative implicit preference, or an ambiguous result. A test of the theory as a whole can be performed by checking that the data never identifies, for the same attribute, both positive and negative implicit preferences.

3 Foundations of implicit preference

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

[20] Note that x is trivially between x and any x'.

[21] If there are n attributes then there are 2^n possible outcomes, and so 2^{2n} potential observations of the form $y(x|x')$. For each attribute there will be 2^{n-1} pairs of outcomes, $\bar{x}$ and $\underline{x}$, and then a variety of $\bar{x}'$ and $\bar{x}''$. Our calculations include evaluation sets with the same element repeated twice (technically a multiset).

[22] There are two possible choices for $\bar{x}$. If $\bar{x}' = \bar{x}$ then there are 3 choices for $\bar{x}''$. There are two other choices for $\bar{x}'$, and for each $\bar{x}''$ is unique (it is the exact opposite of $\bar{x}$). Thus there are ten tests in total.

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.[23]

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)).[24] 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 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

[23] 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.

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

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.”[25] 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.”[26]

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.[27]

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

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

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

[26] 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), Bénabou and Tirole (2004), Bodner and Prelec (2003), Brocas and Carrillo (2008).

[27] 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[28]	influence of primes on judgment and decision-making	self-serving bias, social desirability bias
implicit motives[29]	Thematic Apperception Test (free response to a picture)	desire for power, achievement, emotional affiliation
implicit associations[30]	response time in an association task	discriminatory associations

Table 1: Some Theories of Subconscious/Implicit Motives

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.[31]

[31] The model could also explain implicit race or sex bias under the assumptions that (1) people

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 (Tversky and Kahneman (1981)). 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.[32]

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

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.

[32] 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.

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 $\binom{\textsc{male}}{\text{MBA}}$ and $\binom{\textsc{female}}{\text{MBA}}$, 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.[33]

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:

$$\left(\left\{\begin{pmatrix}\textsc{female}\\ \text{MBA}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \textsc{PhD}\end{pmatrix}\right\},\left\{\begin{pmatrix}\textsc{female}\\ \textsc{PhD}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \text{MBA}\end{pmatrix}\right\},\left\{\begin{pmatrix}\textsc{male}\\ \text{MBA}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \textsc{PhD}\end{pmatrix}\right\}\right)$$
$$\left(\left\{\begin{pmatrix}\textsc{female}\\ \textsc{PhD}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \text{MBA}\end{pmatrix}\right\},\left\{\begin{pmatrix}\textsc{female}\\ \text{MBA}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \textsc{PhD}\end{pmatrix}\right\},\left\{\begin{pmatrix}\textsc{male}\\ \text{MBA}\end{pmatrix},\begin{pmatrix}\textsc{male}\\ \textsc{PhD}\end{pmatrix}\right\}\right)$$

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 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.[34]

Finally, the models differ in their predictions about choice from 3-element choice

[33] This point courtesy of Luke Miner.

[34] 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.

sets. Consider the following two 3-element choice sets, represented spatially for clarity:

$$\left\{\begin{array}{cc} \binom{\textsc{female}}{\text{MBA}} & \\ \binom{\textsc{male}}{\text{MBA}} & \binom{\textsc{male}}{\textsc{PhD}} \end{array}\right\} \left\{\begin{array}{cc} & \binom{\textsc{female}}{\textsc{PhD}} \\ \binom{\textsc{male}}{\text{MBA}} & \binom{\textsc{male}}{\textsc{PhD}} \end{array}\right\}$$

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 $\binom{\textsc{male}}{\textsc{PhD}}$ from the left-hand choice-set, and $\binom{\textsc{male}}{\text{MBA}}$ 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.

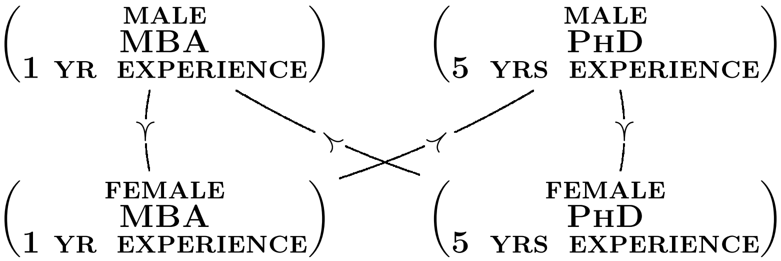
4.1.1 Choices in a Binary Space

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.[35] Under some assumptions, separating choices with decoy questions or time

[35] 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.

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:[36]



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., $P(a \succ b) + P(b \succ c) + P(c \succ a) > 2$).[37] This problem is alleviated when there is rea-

[36] 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.

[37] 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,

son 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 $P(i \succsim j) + P(j \succsim k) + P(k \succ i) > 2$.[38] 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.[39] 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,

there are few clear demonstrations of intransitive choices in the laboratory (Regenwetter et al. (2011)).

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

[39] 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 $i = b, j = c, k = a$.

but would also hire nobody over hiring a female PhD, i.e. an intransitive cycle:

$$\begin{tikzcd}
\binom{\textsc{Female}}{\textsc{PhD}} \arrow[dd, dotted, no head, "\curlyvee" description] \arrow[rrd, dotted, no head, "\prec" description] & & \\
& & \textsc{Nobody} \\
\binom{\textsc{Male}}{\textsc{PhD}} \arrow[rru, dotted, no head, "\succ" description] & &
\end{tikzcd}$$

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 $\{\binom{\textsc{Female}}{\textsc{PhD}}, \textsc{nobody}\}$ seems intuitively less revealing about gender preferences than is a choice from $\{\binom{\textsc{Female}}{\textsc{PhD}}, \binom{\textsc{Male}}{\textsc{MBA}}\}$).

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.[40] 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 Hsee and Zhang (2010). 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.

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

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

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.[41] 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 more dispersed. None of the recent contingent-weighting models is consistent with a figure-8 intransitivity.[42]

[41] For example in Kőszegi and Szeidl (2011) 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 Cunningham (2012) it is negatively related to the average, and in Bordalo et al. (2012) it is - roughly - negatively related to the proportional range (range divided by the average).

[42] 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, 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".[43] 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 42. See Cunningham (2012) 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 studied:

$$\begin{tikzcd}
\begin{pmatrix}\textsc{black}\\ \textsc{westphalia u}\end{pmatrix} \arrow[d, dash, dashed, sloped, "\succ" description] \arrow[dr, dash, dashed, sloped, "\prec" description] & \begin{pmatrix}\textsc{black}\\ \textsc{eastphalia u}\end{pmatrix} \arrow[d, dash, dashed, sloped, "\succ" description] \arrow[dl, dash, dashed, sloped, "\prec" description] \\
\begin{pmatrix}\textsc{white}\\ \textsc{westphalia u}\end{pmatrix} & \begin{pmatrix}\textsc{white}\\ \textsc{eastphalia u}\end{pmatrix}
\end{tikzcd}$$

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

because they have the same marginal distributions, and that utility function prefers Male to Female, all else equal. But this contradicts the choice observed in the vertical choice sets (where Female is chosen over Male). Separability holds for all the models discussed above except Bordalo et al. (2012), but that model cannot generate intransitive cycles in binary choices with two attributes.

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

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).[44]

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 (Sims (2003), Caplin and Martin (2011), Woodford (2012)). 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.[45]

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.[46]

[44] 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.

[45] 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.

[46] 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

The CFS model has a similar spirit to our model: their “rationalizable” preferences roughly corresponds to our “explicit” preference.[47] 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.[48] In the “random response” method subjects are given one statement, 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),

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.

[47] 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)).

[48] 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 a pessimistic summary of their usefulness.

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.[49] 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 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.[50]

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.

[49] 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.

[50] 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.

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 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.[51]

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

[51] 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.

are ignorant they tend to make the choice which maximizes their own payoff.[52]

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.[53]

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.

[52] 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."

[53] 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 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.

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 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;[54] and studies which find differences in automatic associations.[55]

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

[55] Most famously the "Implicit Association Test," which finds that most people perform significantly

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.[56] 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 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.[57] 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.[58] Finally, the theory characterizes the subjective experience of people who are discriminated against; as put by Snyder et al. (1979): "the handicapped

better at a task which asks them to associate white faces with positive words, and black faces with negative words, than the opposite combination.

[56] 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.

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

[58] 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?

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 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 (Tversky and Kahneman (1981)). 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.[59]

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

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.

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

[59] The effect of gamble frame is discussed in Levin et al. (1987). The choices with cards are reported in List (2002), the choices with ice creams are discussed in Hsee and Zhang (2010). Each could also be described in a binary space, though somewhat less naturally.

attribute some of the positive feelings evoked to the true quality of the product.[60]

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.[61]

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. What we mean is that assuming an invariant utility function is often not the most parsimonious way of explaining observed choices. We think of this paper as a contribution towards formalizing, in a relatively nonparametric way, the choice effects of an internal struggle.[62] 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.

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

[61] 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.

[62] We think of Rubinstein (1988), Hsee (1996), and Cherepanov et al. (2013) as contributions to the same line of thought.

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 (2004). Willpower and personal rules. *Journal of Political Economy 112*(4), 848–886.

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.

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*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience and consumer choice. Technical report, National Bureau of Economic Research.

Brocas, I. and J. Carrillo (2008). The brain as a hierarchical organization. *The American Economic Review 98*(4), 1312–1346.

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

Busse, M. R., D. G. Pope, J. C. Pope, and J. Silva-Risso (2013). The overinfluence of weather fluctuations on convertible and 4-wheel drive purchases. University of Chicago Working Paper.

Caplin, A. and D. Martin (2011). A testable theory of imperfect perception. Working paper 17163, National Bureau of Economic Research.

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.

Cunningham, T. (2012). Comparisons and choice. *Working Paper*.

Cunningham, T. E. (2013). Biases and implicit knowledge. Working Paper.

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*.

Hirshleifer, D. (2001). Investor psychology and asset pricing. *The Journal of Finance 56*(4), 1533–1597.

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. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes 67*(3), 247–257.

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

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. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

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

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.

Kőszegi, B. and A. Szeidl (2011). A model of focusing in economic choice. Working Paper.

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.

Levin, I., R. Johnson, and M. Davis (1987). How information frame influences risky decisions: Between-subjects and within-subject comparisons* 1. *Journal of economic psychology 8*(1), 43–54.

List, J. (2002). Preference reversals of a different kind: The 'more is less' phenomenon. *The American Economic Review 92*(5), 1636–1643.

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

Mazar, N., B. Koszegi, and D. Ariely (2013). True context dependent preferences? the causes of market dependent valuations. *Journal of Behavioral Decision Making 27*(3), 200–208.

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*.

Newell, B. R. and D. R. Shanks (forthcoming). Unconscious influences on decision making: a critical review. *Behavioral and Brain Sciences*.

Nisbett, R. E. and T. D. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review 84*(3), 231–259.

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.

Rubinstein, A. (1988). Similarity and decision-making under risk (is there a utility theory resolution to the allais paradox?). *Journal of Economic Theory 46*(1), 145–153.

Simonsohn, U. (2010). Weather to go to college. *The Economic Journal 120*(543), 270–280.

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

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

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

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.

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.

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