8 Appendix: Proofs and Further Discussion

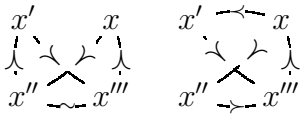
8.1 Proposition 2: Disjunctions

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

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

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

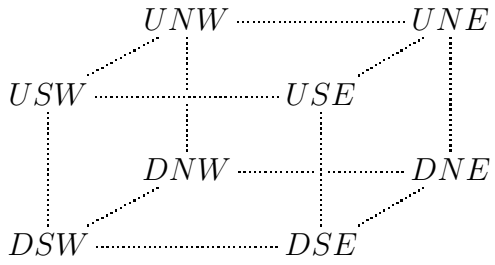
As illustration, consider the following examples of pairs of cycles which establish implicit preferences in the 2-dimensional case. Both examples establish a positive implicit preference on the vertical dimension (Northwards). Here each dimension on the page represents a single attribute. Both cases contain a pair of right-triangle cycles, one of which establishes an implicit preference pointing either North or East, the other of which establishes an implicit preference pointing either North or West, and therefore the pair together imply the existence of an implicit preference pointing North.[72]



Moving to three dimensions, consider the following box labelled with initials for

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

Up/Down, North/South, and East/West.



Consider the right-triangle cycle

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

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

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

To establish the existence of a positive implicit preference for Up over Down, it would be sufficient to observe four right-triangle cycles of span 3, in which each of the Up elements is chosen over its diametrically opposite Down element, e.g. $UNW \succ DSE$, and in which the Down element is indirectly chosen over the Up element, e.g. $DSE \succ DNE \succ UNW$. Thus four cycles sufficient to establish an implicit preference for Up would be:[73]

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

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

8.2 Additional Proofs

Proof of Proposition 1.

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

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

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

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

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

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

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

Proof of Proposition 3.

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

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

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

comparisons, i.e.:

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

But this yields:

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

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

8.3 Combining Scissor Effects

In proposition 4 we established that a scissor effect ($y(x|\{x, x'\}) \gtrless y(x|\{x, x''\})$) will establish a disjunction among implicit preferences over all n possible attributes. As discussed above, the implication can be expressed either as a disjunction of strict inequalities, or the negation of a conjunction of weak inequalities:

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

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

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

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

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

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

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

4. Testing all possible scissor effects (every combination of x, x', x'', with x' between x and x'') is not guaranteed to reveal the true λ, or even to unambiguously identify any single implicit preference. Suppose we have just two attributes. Testing two orthogonal pairs is guaranteed to reveal at least one implicit preference, but you cannot know in advance which attribute it will work on. With 3 or more dimensions it is possible that you will not learn any unambiguous fact about any dimension's implicit preferences.

8.4 Sequential Evaluations

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

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

Given this assumption, and our results on evaluation, there are some simple tests for implicit preferences derived from Proposition 5. We illustrate with an example. Given data on sequential sentencing, an implicit bias against black defendants predicts the following three effects:

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

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

to when they are less similar), and higher if x^t is white and x^t, x^{t-1} are more similar (compared to when they are less similar).

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

Two assumptions underlie (SEQ). First, that only the prior case is relevant as comparator. Second, that the evaluation of an outcome x will be affected in the same way by a comparator x', whether x' is evaluated at the same time as x, or prior to x. The second assumption will hold in the implicit-knowledge model: in that model, System 2 conditions on signals received from System 1 which are independent of the choice set, and there is no reason to treat them differently based on the timing of those signals (unless System 2 is more likely to forget earlier signals). Turning to the signaling model, our earlier derivation found that $\frac{d\hat{\omega}_i}{dy}$ is independent of y'. This implies that when choosing an evaluation y the marginal signaling incentives do not depend on the evaluation given to the comparator, y' (though they do depend on the identity of the comparator, x'). Thus the assumption above, (SEQ), will hold in our signaling model. However the assumption will not hold in the *ceteris-paribus* model: suppose that, when evaluated separately, a man is given a salary of \$60K, and a woman is given a salary of \$50K, but they are both given \$55K salaries when evaluated jointly. Then, if the pair is evaluated sequentially, both will be given the same salary, determined by whoever is evaluated first.[74]

9 Appendix: 3-valued attributes

In this Appendix we extend the model to allow 3-valued attributes, allowing us to infer implicit preferences from an additional class of intransitivities. As before we make assumptions about the relationship between the geometry of the choice set and the revealingness of choices. However, for this extension, we do not further justify the assumptions with deeper models. We do however conjecture that there exists a relatively simple extension of the signaling model which could justify our assumptions.

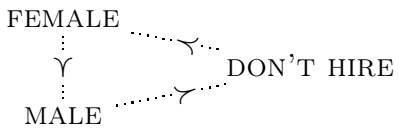
Suppose we observe a hiring decision with the 3 elements {FEMALE, MALE, DON'T HIRE}.

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

It seems natural to think of this choice set as having a similarity structure, which we might graphically represent as follows:

FEMALE	DON'T HIRE
MALE	

The horizontal dimension has a natural binary representation (whether or not to hire), but the vertical dimension has 3 distinct values, and we could describe the "don't hire" outcome as being "neutral" on the gender dimension. Our assumption on revealingness will be that, given two cases x and x' which differ only on attribute i, then if a third case z is neutral on that attribute (i.e., has the new 3rd value), then the choice set $\{x, x'\}$ is relatively more revealing about dimension i, than either $\{x, z\}$ or $\{x', z\}$, and vice versa for all other attributes. This allows us to conclude, from the following cycle, that the decision-maker has an implicit preference for male candidates:



Let $X = \{0, 1, \varnothing\}$. We make the following additional assumption about revealingness:

Definition 4. For any $x, x', z \in X$, we say that z **is a vertex of** $\{x, x'\}$ if and only if there exists some j such that

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

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

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

We finally extend the definition of implicit preferences to the 3-valued space, by assuming an ordering among features $(0, \varnothing, 1)$, and assuming that changes in implicit

preferences will respect this ordering (i.e., the relatively higher value will come to be relatively preferred).

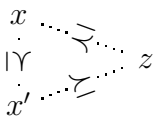
Definition 5. We say that $u(x, A)$ has **implicit preferences** $\lambda \in \{-1, 0, 1\}^n$, if, for any $x, x' \in X$, and $A, B \in \mathcal{A}$, normalizing such that $\forall j, x_j \in \{\varnothing, 1\}, x'_j \in \{-1, \varnothing\}$, and for every i with $x'_i \neq x_i$,

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

then

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

Proposition 6. *If z is a vertex of x, x', and we observe choices such that,*



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

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

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

While we also know that,

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

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

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

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

attribute j. $\square$

10 Appendix: Foundations for Implicit Preference

In this Appendix we give three formal foundations for implicit preference (*ceteris paribus*, signaling, and implicit knowledge). We begin by defining linear implicit preferences (a class of menu-dependent utility functions, $u(x, A)$), and then show that the three parametric foundations all exhibit linear implicit preferences.

10.1 Linear Implicit Preferences

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

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

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

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

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

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

and

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

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

in the utility functions we get:

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

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

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

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

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

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

We wish to show that this implies,

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

We can express the difference:

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

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

Finally we wish to show monotonicity: if, for every $x, x' \in X$, $A, B \in \mathcal{X}$, with

normalization $x_i = 1$, $\forall i$, and $x'_i = 0 \Leftrightarrow m \leq i \leq n$, and for all $i \in \{m, \ldots, n\}$:

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

then

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

The antecedent implies:

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

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

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

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

Aggregation of the parallel construction test. Before moving on, linearity allows us to prove a useful aggregation result for implicit preferences. Suppose we collect data on evaluations on a set of outcomes, and we wish to check for data which passes the "parallel construction" test (Proposition 5) - i.e. two outcomes which differ in one attribute, and which move in opposite directions when their comparators change in equivalent ways. If we have between-subjects data, and there are many attributes, then it is likely that we will not have enough observations to test a single parallel construction (a single pair $\bar{x}$ and $\underline{x}$), but we could check if the condition holds on *average*. The following proposition shows that, with linear preferences, if the parallel

construction test is true on average, then it must hold in at least one individual case, and therefore it will establish an unambiguous implicit preference .

Proposition 7 (average evaluations)**.** *If $y(\cdot,\cdot)$ satisfies LIN, then for some set of outcomes $\{\bar{x}^k, \bar{x}'^k, \bar{x}''^k, \underline{x}^k, \underline{x}'^k, \underline{x}''^k\}_{k=1}^l$, with, for all $k \in \{1,\ldots,l\}$, $\bar{x}_i^k = 1 = 1 - \underline{x}^k$, and $\bar{x}_j^k = \underline{x}_j^k$ for $j \neq i$, and $\bar{x}'^k$ between $\bar{x}^k$ and $\bar{x}''^k$, and*

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

and if,

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

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

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

Proof. [COMING] □

10.2 Foundation for Implicit Preferences: *Ceteris Paribus* Rules

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

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

Proposition 8. *The linear order induced by the function*

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

satisfies strong betweenness and equivalence.

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

Equivalence follows directly. □

This proposition shows that a *ceteris-paribus* decision-maker will have implicit preferences, in our sense, and so the techniques we describe can be used to identify their implicit preferences.

There are two extensions of this model worth considering. First, to take into account the interaction of rules. Suppose you have two rules - one against choosing men over women (all else equal), and one against choosing white over black candidates (all else equal). As written, the model does not disallow the choice of a white man over a black woman. A fuller version of this model would disallow any choices which imply, by transitive extension, that at least one of the ceteris paribus rules is broken.[75]

Second, using the model on *evaluation* data requires some additional assumptions. Assume that the decision-maker incurs the penalty of κ_i if they evaluate an outcome x more highly than an outcome x' when x and x' differ only in attribute i, and there is a *ceteris paribus* rule against choosing in favor of attribute i. This would predict that, for example, an implicitly sexist decision-maker would pay an otherwise-identical man and a woman exactly the same salary when evaluated side by side, but would pay them different salaries when evaluated separately, or when there was some difference in their qualifications.[76]

[75] It is not clear whether this extended version can be written as having linear implicit preferences (LIN).

[76] This property would be achieved by assuming a separable intrinsic utility function, $v(x) = \sum \omega_i x_i$, and letting $\kappa_i = \omega_i$ when there was a ceteris paribus rule.

10.3 Foundation: Gaussian Signaling and Choice

In this section we present a linear-Gaussian signaling model of choice, and show that it can be described as having linear implicit preferences (LIN), with a comparison function $\Phi_i(A)$ that satisfies strong betweenness and equivalence.

We start with a linear utility function over x, $v(x) = \sum_{i=1}^{n} \omega_i(x_i - \frac{1}{2})$, with $\omega_1, \ldots, \omega_n$ being the weights on each attribute. We additionally assume the existence of an observer who, upon observing the choice of x from a choice-set A, updates beliefs regarding the weights $\omega_1, \ldots, \omega_n$, with the observer's mean posteriors represented by $\hat{\omega}_i(x, A)$. And we assume that the decision-maker cares about the observer's beliefs, so the decision-maker's complete preferences are represented by the function:

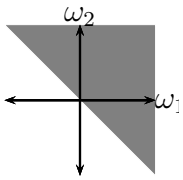
$$u(x, A) = \sum_{i=1}^{n} \omega_i(x_i - \frac{1}{2}) + \sum_{i=1}^{n} \kappa_i \hat{\omega}_i(x, A), \tag{3}$$

where $\kappa_i > 0$ implies that the decision-maker prefers for the observer to have a more positive belief about the DM's preferences over attribute i, and vice versa. We assume that the observer is *naïve*, in the sense that they believe the decision-maker has no signaling motive (i.e., $\kappa_i = 0, \forall i$). It would be more satisfying to use a model with a sophisticated observer, but the derivation would be more complicated and we expect that the basic results would remain the same.

The observed choice function will therefore be:

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

We assume that the observer has Gaussian priors over the weights, $\omega_i \sim N(\mu_i, \sigma_i^2)$. Once a choice is observed, the posteriors will therefore be a truncated multivariate normal. For example, if we observe that $x \in c(\{x, x'\})$, with $x = \binom{1}{1}$ and $x' = \binom{0}{0}$, this implies that $\omega_1 + \omega_2 \geq 0$. Graphically, the posteriors will be contained in the shaded region:



Unfortunately the mean marginals of this distribution, $\hat{\omega}_i(x, A)$, do not have a simple

expression. However we need only a qualitative result: that the *difference* in mean posteriors is smaller when the two alternatives differ among a strictly larger set of attributes. Define $\Delta_i^{\{x,x'\}}$ as the difference in posteriors regarding ω_i , between that generated by the choice of x, and that generated by the choice of x', i.e.:

$$\Delta_i^{\{x,x'\}} \equiv \left| E\left[\omega_i | \sum_{j=1}^{n} \omega_j (x_j - \frac{1}{2}) \geq \sum_{j=1}^{n} \omega_j (x'_j - \frac{1}{2})\right] - E\left[\omega_i | \sum_{j=1}^{n} \omega_j (x_j - \frac{1}{2}) \leq \sum_{j=1}^{n} \omega_j (x'_j - \frac{1}{2})\right] \right|.$$

Proposition 9. *For binary choice sets, choice maximizes the following utility function,*

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

which satisfies LIN.

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

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

The left-hand side implies that,

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

which completes the proof. □

We now turn to determining whether $\Delta_i^{\{x,x'\}}$ satisfies strong betweenness: we wish to show that the size of the effect on posteriors (Δ_i^A) will decrease when the attribute-wise distance between x and x' increases. This requires an additional assumption: that the observer does not have too-strong priors about the value of any attribute.

To explain the intuition, consider choice among breakfasts: the choice set $\{\binom{\text{APPLE}}{\text{COFFEE}}, \binom{\text{BANANA}}{\text{COFFEE}}\}$, seems more informative about the relative preference for apple over banana than is the choice set $\{\binom{\text{APPLE}}{\text{COFFEE}}, \binom{\text{BANANA}}{\text{TEA}}\}$. However, consider another pair of choice sets: $\left\{\binom{\text{APPLE}}{\$0}, \binom{\text{BANANA}}{\$0}\right\}$, and $\left\{\binom{\text{APPLE}}{\$20}, \binom{\text{BANANA}}{\$0}\right\}$. If you choose a banana from the first choice set, I will infer a mild preference for bananaas; but if you choose a banana from the second choice set I will infer a very strong preference for bananas over apples. This happens because we already have strong priors about the difference in value between \$0 and \$20, so seeing you choose a banana over \$20 causes an extreme revision of my beliefs about your fruit preferences. This implies a violation of strong betweenness: the second choice set could be *more* revealing about fruit preferences, despite the alternatives differing in a greater number of attributes.[77]

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

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

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

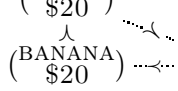
We first give a lemma to show the sufficiency of symmetric probability density functions (for a Gaussian distribution, a zero mean implies that the distribution is symmetric around zero).

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

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

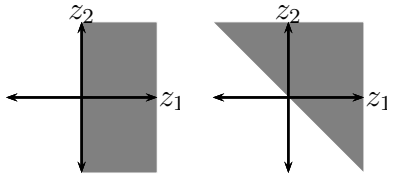
Proof. When we observe that $z_1 > 0$, our posteriors will be truncated by a vertical line,

[77] This could generate a right-triangle cycle such as the following:



by someone who prefers bananas to apples, but would get disutility from people thinking that they have a strong preference for apples over bananas.

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



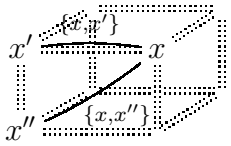
We can write the difference between the two expectations as follows, using $f(z_1)$ and $g(z_2)$ to represent probability density functions:

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

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

Proposition 10. *Ordering binary choice sets by Δ_i^A satisfies strong betweenness and equivalence.*

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



Let $x_i = 1$ for all i, and for convenience reorder the attributes such that $x'_i = 0$ if and only if $m < i \leq n$, and $x''_i = 0$ if and only if $l < i \leq n$ (by betweenness $m < l$). Then

we can write the following:

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

From these equations, and Lemma 2, we can see that strong betweenness will hold for Δ_i^A, because:[78]

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

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

10.4 Foundation: Gaussian Signaling and Evaluation

The signaling model can also be applied to evaluation. The implications are somewhat different because the observer will make different inferences when observing an evaluation, than when observing a choice.

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

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

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

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

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

The solution to this problem is:

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

From the logic of Gaussian signal extraction, we get

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

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

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

Proposition 11. *Evaluation according to 5 can be written as LIN.*

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

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

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

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

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

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

We regard this assumption as merely technical, mainly necessary because of the stylized way in which we have treated attributes, where the values of each realization are opposites. In an alternative version of the model, where the two realizations of

the attributes are treated as independent dummy variables, with independent weights, this assumption is not necessary. Results for this version of the model are available on request.

Proposition 12. *Ranking binary choice sets by* $\left|\frac{d\hat{\omega}_i}{dy}\right|$ *satisfies strong betweenness and equivalence.*

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

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

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

10.5 Foundation: Implicit Knowledge

We now present a model of implicit knowledge and show that it will generate implicit preferences - a more extensive version of this model is given in Cunningham (2013). A decision-maker with implicit knowledge will satisfy betweenness, but not strong betweenness, meaning that our techniques for identifying implicit preferences will work in choice but not evaluation. This is because the interactions among different aspects of implicit knowledge are complex, and become difficult to disentangle.

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

Implicit preferences in choice can reflect one of two different effects. First, suppose there is some attribute that we consciously wish to ignore, then it will tend to have a bigger effect in less revealing comparisons, because in those comparisons it will become harder to infer the contribution of that attribute. For example suppose that I regard

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

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

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

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

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

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

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

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

Proposition 13. *The decisions and evaluations of an implicit-knowledge decision-*

maker can both be represented with a linear implicit preference utility function,

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

where

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

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

Proof. First consider what happens when evaluating a single outcome, x: the logic of Gaussian signal extraction implies that inferring $\omega_1, \ldots, \omega_n$ from v simply involves splitting up the sum v into its parts, weighted by their respective variances (σ_i^2). When we have two outcomes (x, x') and two evaluations (v, v'), we can use the same principle but now we are inferring the common attributes ($i \in S$) from the average ($\frac{v+v'}{2}$), and the distinctive attributes ($i \in D$) from the difference ($\frac{v-v'}{2}$):

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

Substituting this into the expression for y we get:

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

We can write the last expression as,

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

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

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

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

Ideally System 2 wishes to weight each association ω_i according to its corresponding factor π_i. Because it does not know the associations it cannot do that, and it instead uses the *average* correction factors $\bar{\pi}_S$ and $\bar{\pi}_D$. It will be able to perfectly identify π_i when the two outcomes differ only in attribute i (because then $\bar{\pi}_D = \pi_i$). But when the outcomes differ in more than one attribute we will tend to over-react to attributes we wish to ignore ($\kappa_i = 0$), and under-react to associations we wish to follow ($\kappa_i = 1$).

Proposition 14. *The function $\Gamma_i(A)$ satisfies betweenness and equivalence among choice sets which differ in two or fewer attributes (i.e., for any $A = \{x, x'\}$ such that $\sum_{i=1}^{n} |x_i - x'_i| \leq 2$).*

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

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

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

Betweenness may be violated when outcomes differ on more than two attributes. This is because, when bundling some attribute i with additional attributes, the direction of the effect on $\bar{\pi}_D$ is ambiguous: it will increase if the new attribute is relevant ($\pi_i = 1$), but it will decrease if the new attribute is irrelevant ($\pi_i = 0$). For example suppose that someone is trying to ignore race: then when outcomes differ on both race and education

this will increase the effect of race, but increasing the number of differences by adding variation in age has an ambiguous effect. For this reason the model does not generate implicit preferences, in our sense, when outcomes differ in more than 2 attributes.

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

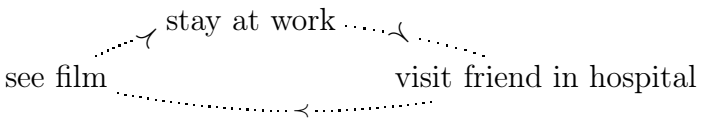
These qualifications imply that, under the implicit knowledge model, implicit preferences can be identified endogenously only when choice involves variation in 2 attributes. However the model still remains useful in other sense .

The model can also be applied to data on evaluations, but it requires ancillary assumptions. Suppose that we have data on sentencing. Then if we assume that people are trying to ignore race ($\pi_i = 0$), but follow their instincts for other attributes ($\pi_i = 1$), then this will generate an array of testable implications: (1) evaluations of both black and white defendants should increase when the comparator is white; and (2) evaluations of black defendants should decrease, and evaluations of white defendants should increase, when the comparator becomes same-race.

11 Additional Notes on Interpretation

11.1 Other types of cycle

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



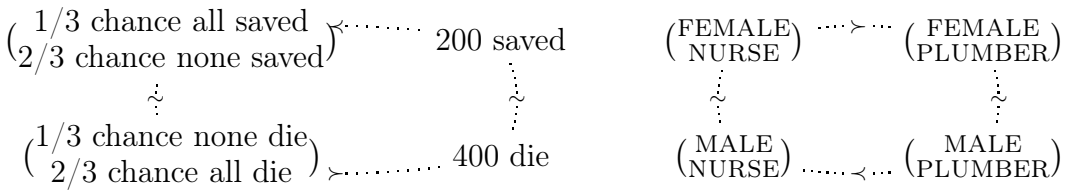
Here there seems to be some sort of implicit preference, but each alternative is idiosyncratic - i.e. it is not obvious that the cycle can be described as due to any one pair of outcomes sharing an attribute that the remaining outcome lacks (as occurs in a right-triangle cycle). We could call this type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

An intuitive way to rationalize this cycle in the signaling model would be if the

value of staying at work tended to be highly uncertain to an observer, due to you having private information about how urgent your work is. In our signaling model the revealingness of a choice declines as the variance of prior beliefs over the value of each alternative increases: i.e., if I have very broad priors over the value of attribute i, then when I observe you choosing i over j, I will not update much about the value of j. Thus one interpretation of why this hypothetical cycle seems compelling is that we feel the decision-maker to be motivated to visit their friend by, in part, signaling considerations, and so when making a choice against staying at work the signaling incentive is relatively weaker, i.e. people will not infer much about your dedication to your friend from this choice.

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

One other type of cycle is worth noting: cases where the *separability* of implicit preferences is violated. The following two examples show intransitive cycles which appear to be driven by some kind of implicit preference, but which violate separability. On the left we show choices from the "Asian disease" problem described in Tversky and Kahneman (1981), represented in a space with two attributes (safe/risky and gain/loss), and supplemented with the assumption that people would be indifferent between outcomes which differ only in whether described as gains or losses. On the right we show choices from a decision-maker who implicitly prefers nurses to be female, and implicitly prefers plumbers to be male.



In both of the cases discussed we could design an experiment to isolate a single implicit preference by holding one of the attributes fixed: e.g., test for implicit prefer-

ences for women just among nurses (either by adding a third attribute, or by using an isosceles cycle).

11.2 Larger choice sets.

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

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

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

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

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

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

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

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

4. An interesting implication of the implicit-knowledge model is that discrimination should be a U-shaped function of the composition of a group. Consider a decision-maker who must assign salaries to a set of employees, each of whom differs in some idiosyncratic attribute, as well as varying in gender. The model predicts that gender bias in salaries will be decreasing and then increasing in the fraction of male employees, because the decision-maker learns the most about her bias when the fraction is 50-50.[79] It is not clear whether this will hold in the signaling model: on the one hand, a 50-50 composition will reveal, to the observer, the most information about the decision-maker's gender bias; but on the other hand, it is more costly to implement equal salaries when there is a 50-50 composition.

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