Choosing for the Right Reasons

Sarah Ridout*

October 8, 2021

Abstract

I propose a model of decision making when some preferences are acceptable to act on and others are not. Through an axiomatic characterization, I demonstrate that the model has significantly stronger identification properties than existing models with a similar scope of application, and admits tractable extensions to between-subject data and information choice. The between-subject extension is motivated by the concern, not addressed in existing work, that subjects with unacceptable preferences may exhibit consistency motives in within-subject data. I propose a simple method for determining, in between-subject data, whether there is a gap between what subjects prefer and what they choose. I then extend the model to information choice and show that it can accommodate and build on the findings from the large "moral wiggle room" literature.

1 Introduction

People care about the reasons behind choices as well as the choices themselves. To determine whether a decision maker is rational, has good moral character, is doing his duty, or is following the law, it is often necessary to investigate his reasons as well as his actions. For instance:

- It is blameworthy to withhold a charitable donation out of stinginess rather than concerns about the effectiveness of the charity.
- It unreasonable to evaluate a policy on the basis of partisan loyalty rather than the effects of the policy.
- It is illegal to fire an employee because of personal animosity rather than dissatisfaction with her work.

*sarah.e.ridout@vanderbilt.edu. For helpful comments and suggestions, I am indebted to Ned Augenblick, Tom Cunningham, Jon de Quidt, Christine Exley, Drew Fudenberg, Ed Glaeser, Ben Golub, Jerry Green, Yoram Halevy, Peter Klibanoff, David Laibson, Shengwu Li, Yusufcan Masatlioglu, Efe Ok, Pietro Ortoleva, Fernando Payró Chew, Dmitri Taubinsky, and participants at numerous seminars and conferences. For years of guidance, I am especially indebted to Matthew Rabin and Tomasz Strzalecki.

When a decision maker wants to take an action for a reason that is unacceptable to himself or others, he may search for a better reason to do the same thing. He may ask himself whether there is any reasonable combination of principles, values, tastes and beliefs that would provide an acceptable motivation for his desired action. If so, he may take the action even if he does not actually share those principles, values, tastes and beliefs, and would not act on them in other contexts.

The following simple example, loosely based on an experiment in Exley (2016), shows how the search for a good reason can affect choice behavior.

Example 1. *The domain of choice consists of risky payments to the decision maker and risky payments to a charity. All payments are made by the experimenter. Let*

$$a = \$10 \text{ to self with probability } 1/2$$
$$b = \$6 \text{ to charity with probability } 1$$
$$d = \$24 \text{ to charity with probability } 1/4$$

Consider a decision maker who does not care much about the charity, but wants to avoid appearing selfish. Since the choice of a over b can result from risk loving rather than selfishness, the decision maker may choose a from $\{a, b\}$. Similarly, since the choice of a over d can result from risk aversion rather than selfishness, the decision maker may choose a from $\{a, d\}$. Given both these choices, preference maximization would imply that a is chosen from $\{a, b, d\}$. Neither risk loving nor risk aversion justifies this choice, though; risk aversion implies that b should be chosen, while risk loving implies that d should be chosen. Since choosing a from $\{a, b, d\}$ would reveal a lack of generosity, the decision maker may forego his preferred option in favor of one of the donations.

As Example 1 demonstrates, decision makers who care about justifying their actions sometimes fail to maximize their preferences. However, their preferences still shape their behavior through their choice between different justifications. For instance, the decision maker above exploits flexibility in the range of acceptable risk attitudes to get as close to his preferred outcomes as possible. This suggests that it might be possible to disentangle the decision maker's underlying preferences from his notion of what is acceptable. The first two-thirds of this paper are primarily devoted to that task, first in deterministic and then in stochastic settings.

Readers familiar with the rationalization model of Cherepanov et al. (2013) (CFS) or the limited attention model of Masatlioglu et al. (2012) (MNO) will wonder whether this task has not already been completed, at least for the deterministic case. Indeed, those models are consistent with examples like the above, and can be used to partially identify the decision maker's preferences. However, that identification is often too limited to be of much practical use, even under ideal circumstances with no constraints on data collection. Choice behavior is often consistent with multiple representations, which offer completely different views of the decision maker's preferences and justifications. The leading empirical example in CFS provides an illustration.

Example 2. *The decision maker must choose a room in which to watch a movie. Some rooms are empty, and others are occupied by a physically disabled stranger. Let*

$$a = \textit{watch movie } A \textit{ alone}$$
$$b = \textit{watch movie } B \textit{ alone}$$
$$d = \textit{watch movie } A \textit{ with disabled stranger.}$$

Suppose the decision maker chooses

$$a = c(a, b) \quad b = c(b, d) = c(a, b, d) \quad d = c(a, d).$$

Notice that the decision maker chooses to sit alone if and only if there is an empty room playing a movie that differs from the movie playing in the occupied room. This suggests that the decision maker prefers not to sit with the disabled stranger, but only indulges his prejudice if he can use a desire to see a particular movie as an excuse. Although CFS agree with this interpretation, their model (and that of MNO) is consistent with a preference for d over a. This preference is the opposite of prejudiced, as the decision maker prefers to sit with the disabled person when all else is equal. In fact, the only conclusion the analyst can draw from the CFS model is that the decision maker prefers movie A to movie B. I show in Section 3 that Example 2 is prototypical rather than unusual, as exactly the same choice pattern must occur whenever the decision maker cannot justify choosing one alternative over another. Thus, identification problems in this simple example will propagate to more complex data sets.

In the first third of the paper, I show that a natural special case of the CFS and MNO models greatly improves on the identification properties of both. I refer to that special case as "the justification model." A decision maker in the justification model is endowed with two objects: a standard preference $\succsim$ on the domain of choice, and a non-empty set of preferences $\mathcal{M}$ on the same domain. Given any menu of alternatives, the decision maker maximizes $\succsim$ over the subset of alternatives that maximize at least one of the preferences in $\mathcal{M}$. Since the decision maker does not select any alternative that would not be selected by any of the preferences in $\mathcal{M}$, the members of $\mathcal{M}$ may be interpreted as the preferences that the decision maker considers it acceptable to act on. I refer to the members of $\mathcal{M}$ as "justifiable preferences," or "justifications" for short. The model differs from preference maximization because the decision maker need not use the same justification on different choice sets. Since the choice between justifications is governed by $\succsim$, that preference may be interpreted as capturing the decision maker's desires or inclinations. The model is interesting when $\succsim$ does not belong to $\mathcal{M}$. Since the decision maker considers some of his desires or inclinations to be unacceptable, he cannot simply maximize his preference. Instead, he strategically chooses between justifications to get as close to his ideal outcome as possible.

The first main advantage of the justification model is as follows: for any choice data consistent with the model, there exists a unique "focal preference" consistent with the data that attributes choice to preference as far as possible. I defer the formal definition to Section 3 and give only an intuition here. The justification model is intended to explain choice data that cannot be explained by preference maximization, not to reinterpret data that is consistent with preference maximization. Thus, the sensible policy is to assume that preferences are consistent with choices in the absence of evidence to the contrary. Applying this rule to the CFS and MNO models may eliminate some representations, but need not deliver a unique preference nor eliminate all counter-intuitive representations. For instance, neither prejudiced nor unprejudiced preferences can be eliminated in Example 2, so it is still impossible to draw any conclusions about the decision maker's attitude to disabled people. The justification model solves this problem because there is a unique focal preference for every data set consistent with the model. Consistent with intuition, the focal preference in Example 2 is indeed prejudiced.

The focal preference is recovered by way of an axiomatic representation theorem for the justification model. The representation theorem is based on two simple choice patterns, which are generalizations of Examples 1 and 2. As it turns out, all the information that the analyst could hope to recover about the decision maker's justifications is contained in those two choice patterns. Thus, they can be used to identify the alternatives that are unjustifiable in a given menu. The main axiom that characterizes the model requires that choice be unchanged when alternatives revealed to be unjustifiable are made unavailable. Intuitively, this condition is necessary because an unjustifiable alternative can neither be chosen nor affect whether another alternative is justifiable. When choice is single-valued, this condition is also sufficient. Otherwise, it is sufficient in conjunction with the requirement that the set of selected alternatives does not shrink when unchosen alternatives are made unavailable. The construction for the focal preference emerges from the sufficiency proof. The only inputs to the construction are choice on two-element menus and any three-element menus on which pairwise choices form a cycle. Thus, the construction is simple and can be executed with very limited data. While additional data is useful for testing the model and learning more about the justifications, it is not needed to obtain the focal preference.

The second third of the paper extends the justification model to address a significant difficulty in collecting evidence of justification. For an experiment on justification to yield interesting results, subjects need to make choices that are inconsistent with preference maximization. Those inconsistencies allow the analyst to disentangle subjects' preferences from the constraints they perceive. If subjects are wary of having their preferences revealed (which is quite likely in the types of situations the model is intended to study), they may avoid or limit inconsistencies that they would commit if they did not expect their choices to be reviewed by a single observer. This prevents the analyst from understanding the full extent of justifying behavior.

The simplest way to eliminate this problem is to collect data from a population of subjects, each

of whom makes only one choice.[1] However, existing models do not provide any tools for analyzing this data. It is not appropriate to apply a single-person model to modal choices, as this can lead to misleading results. For instance, Example 2 is actually based on the modal choices in an experiment by Snyder et al. (1979). This experiment is commonly used as an example of justifying behavior, but the data from the experiment is consistent with a standard random utility model.[2] It is an open question how data from a population of justifiers would depart from random utility and what those departures might imply about the distribution of preferences and justifications in that population.

To answer this question, I extend the justification model to random choice. I take the standard step of extending the domain of choice to lotteries and assuming that preferences have an expected-utility form. I then study a particular class of departures from random expected utility, which I call anomalies. Anomalies are easy to test for because they involve only simple perturbations to three-element menus. Roughly, a three-element menu is anomalous if the probability of choosing one alternative falls when another alternative is mixed with it. Anomalies can be used to draw inferences about the distribution over justifications in the subject population. More precisely, each anomaly implies a statement of the form "some people lack a justification for choosing alternative a over alternative b." In situations where there are natural ex ante restrictions on the justifications people are likely to perceive, this result can also be used to test the model. Anomalies should only be observed where they have sensible implications for the distribution over justifications.

Anomalies are even more useful in situations where different people are unlikely to perceive opposing obligations. Under a formal version of this assumption, which is called Limited Disagreement, anomalies can be used to identify the full range of obligations that people can perceive. That is, anomalies can be used to determine which choices are always justifiable and which are sometimes unjustifiable. Even when it is not possible to collect enough data to carry out this identification, this result is still useful. Whenever it is sometimes unjustifiable to choose one alternative over another, there must exist an anomaly involving those two alternatives plus a third alternative. The discussion following the result provides some guidance about how to choose the third alternative to observe the anomaly.

The first two results about the random justification model are simple and practical, but at the expense of limiting the analyst to qualitative inferences. The final result requires (much) more data and more complicated constructions, but proves that quantitative conclusions can be drawn from between-subject data. This result focuses on the behavior of people whose preferences are the furthest from justifiable, i.e. the people who always want to violate the constraints imposed by the justifications. Although it is not straightforward, it is possible to construct menus that allow the behavior of this group of people to be observed separately from the behavior of others. For any

[1] An alternative would be to impose a significant time lag between choices. Even if this is feasible, it is still important to understand how randomness in preferences and justifications affects choice, as these objects are likely to fluctuate over time.

[2] I am grateful to Tom Cunningham for pointing this out.

given pair of alternatives, these constructions can be used to identify the probability that the people in this group are unable to justify choosing one alternative over the other. Under the benchmark assumption that preferences and sets of justifications are drawn independently, this result delivers full identification on binary menus. Under the more plausible assumption that the people with the least justifiable preferences are the least constrained, it delivers a lower bound on the proportion of constrained people. That is, it delivers a *conservative* estimate of the extent to which behavior departs from preference maximization.

The final third of the paper extends the random justification model to a richer choice environment, in which decision makers choose whether to acquire information about their options before making a selection. This exercise is important because some of the strongest evidence to date of a conflict between preferences and justifications comes from experiments on information choice in ethical settings. There is a large empirical literature demonstrating that people often avoid information about the effects of their actions on others.[3] It is widely understood that this behavior results from a tension between what people would like to do and what they feel they ought to do—precisely the mechanism behind the justification model. However, there is no existing model that links the literature on information choice to other work on justification.

I use the extension to information choice to address some questions raised by, but beyond the scope of, existing experimental work. First, when people feel free to avoid information they do not want, how much scope remains for using information to encourage socially desirable behavior? Consider a benevolent policymaker[4] who wishes to influence a group of people faced with a binary choice. It turns out that information can be a useful tool for the policymaker, even if no one feels obligated to acquire that information and no one shares the policymaker's preferences on the original choice set. Information avoidance limits, but does not collapse, the range of people the policymaker can hope to reach. If people already know their justifications when choosing whether to observe the signal offered by the policymaker, then the policymaker can only reach people whose preferences match his own on at least one signal realization. This constraint is relaxed when people are sufficiently uncertain about their justifications at the information choice stage. Then, the policymaker can reach some people who disagree with him on every realization of that signal, although it still cannot reach people whose preferences are especially misaligned with his own.

Given that free information can improve behavior, is it optimal for a benevolent policymaker to provide as much information as possible? In an interesting class of situations that I call "ambiguous," the answer is no. A situation is ambiguous if some people have access to sets of justifications that allow them to justify more than they could if they were limited to a single justification. In ambiguous situations, there are signals that the policymaker would strictly prefer to withhold from people whose preferences are sufficiently misaligned with his own. Intuitively, those signals en-

[3]The popularity of information choice experiments might be explained, in part, by the absence of tools for working with other types of data—an absence this paper aims to help remedy.

[4]That is, a policymaker whose preferences are always justifiable.

courage socially undesirable behavior because they allow people to exploit disagreement between justifications.

Finally, I turn to the question of why justifiers feel able to avoid information in the first place. I assume that people need to justify their information choices, but allow for the possibility that the set of available justifications depends on the choice stage (information choice vs. choice over final outcomes). I show that information avoidance *does not* occur if the same set of justifications governs both stages, but *does* occur if the set of justifications governing information choice is strictly more permissive than the set of justifications governing final choice. When the set of justifications governing information choice is so permissive that it does not rule anything out, people avoid any information that makes them worse off. However, this extreme assumption is not necessary for avoidance. Some avoidance occurs whenever people are willing to accept weaker justifications for choices that do not directly determine final outcomes.

The paper is organized as follows. Section 2 reviews related experimental and theoretical work not treated elsewhere in the paper. Section 3 introduces the deterministic justification model, characterizes it, and details its identification properties. Section 4 extends the model to stochastic choice and explains how it can be used to analyze between-subject data. Section 5 studies information choice in the stochastic model and connects it to the literature on moral wiggle room. Section 6 concludes.

2 Related work

2.1 Related experiments

2.2 Related models

3 Identification from within-subject data

This section studies the behavior of an individual decision maker on an arbitrary domain of choice, $\mathcal{A}$. Choices are made from menus, which are nonempty finite subsets of $\mathcal{A}$. The collection of all menus is denoted $\mathcal{F}(\mathcal{A})$. For all but one result in this section, the observable is a choice correspondence $c : \mathcal{F}(\mathcal{A}) \rightrightarrows \mathcal{F}(\mathcal{A})$. As usual, c is non-empty-valued and satisfies $c(A) \subseteq A$ for all $A \in \mathcal{F}(\mathcal{A})$.

A preference is a reflexive, complete and transitive binary relation on $\mathcal{A}$. A generic preference is denoted $\succsim$. A justification model consists of a preference $\succsim$ and a non-empty set of preferences $\mathcal{M}$. A generic member of $\mathcal{M}$ is called a justification and denoted $\succsim_m$.

Two additional pieces of notation are used throughout the paper. For an alternative a and menu b, the notation $a \succsim B$ means that $a \succsim b$ for all $b \in B$. Given a set of justifications $\mathcal{M}$, the notation

$A \succ_{\mathcal{M}} b$ means that, for each $\succsim_m \in \mathcal{M}$, there exists $a \in A$ such that $a \succ_{\mathcal{M}} b$. Similarly, $a \succ_{\mathcal{M}} b$ means that $a \succ_m b$ for all $\succsim_m \in \mathcal{M}$.

Definition 1 (Justification representation)**.** *Justification model* $(\succsim, \mathcal{M})$ *is a justification representation for choice correspondence* c *if, for each menu* A*,*

$$c(A) = \max(\succsim, \mathcal{M}(A)) \text{ where } \mathcal{M}(A) := \bigcup_{\succsim_m \in \mathcal{M}} \max(\succsim_m, A).$$

The members of $\mathcal{M}(A)$ are said to be justifiable in A, and the members of A that do not belong to $\mathcal{M}(A)$ are said to be unjustifiable in A. Whether an alternative is justifiable always depends on the context; for instance, any alternative is justifiable in the singleton menu containing itself. Nevertheless, I sometimes shorten "justifiable in A" to "justifiable" for brevity.

The justification model is a special case of both the CFS and MNO models discussed in the introduction, which are not themselves nested. The equation $c(A) = \max(\succsim, \mathcal{M}(A))$ appears in the definitions of both CFS and MNO representations. In CFS, $\mathcal{M}(A)$ is the set of alternatives that maximize at least one member of a set of binary relations (as opposed to a set of preferences). In MNO, $\mathcal{M}(A)$ is simply taken as primitive. It is assumed to obey the following restriction: if $\mathcal{M}(A) \subseteq B \subseteq A$, then $\mathcal{M}(B) = \mathcal{M}(A)$. This restriction is satisfied in the justification model. Since it is easy to see that the justification model is closely related to existing work, the model definition should not be seen as a contribution of the paper. Instead, the paper contributes by establishing previously unknown properties of the model and using it to attack previously intractable problems.

Given any model where a decision maker maximizes a preference subject to a constraint, it is natural to wonder whether, or to what extent, the underlying preference can be recovered from choice data. Full identification on every data set is clearly too much to hope for, though. Consider any dataset consistent with preference maximization. While this dataset has a representation in which the decision maker is unconstrained and his choices reflect his preferences, it has other representations in which the decision maker is highly constrained and his choices fail to reflect his preferences. The justification model is no different.

There are two ways to respond to this difficulty. First, try to obtain a unique preference on a narrower range of datasets that are "far" from consistent with preference maximization. Second, try to find a criterion for selecting between representations that will deliver a unique preference on any dataset consistent with the model. It turns out that both can be done within the justification model, although neither can be done in the more general CFS and MNO models. Thus, the justification model has desirable identification properties.

The formal identification results are derived from an axiomatic representation theorem, which is the focus of the following section. The representation theorem may also be of independent interest because it clarifies the choice behavior that justifiers can exhibit.

3.1 Representation theorem

Before presenting the main representation theorem, I consider an easier exercise that builds intuition (and delivers a key lemma used in the proof of the main theorem). Suppose that the analyst already has a plausible candidate $\succsim$ for the decision maker's preference. Under what conditions on the choice correspondence c does there exist a set of justifications $\mathcal{M}$ such that $(\succsim, \mathcal{M})$ together represent c? The answer can be expressed as a pair of axioms. The main axiom is Irrelevance of Unjustifiable Alternatives (IUA). It says that an alternative a that the decision maker would like to choose from menu A, but does not actually choose, can be removed from any superset of A without changing behavior. Since the decision maker would like to choose a but does not, a must not maximize any of the justifications on A. Thus, it cannot be chosen on any superset of A. Moreover, it cannot prevent any other alternative from being chosen: if alternative b fails to maximize any justification on $B \supseteq A$, then the same is true when a is removed. The second axiom, Optimization, is vacuous when choice is single-valued. It says that the decision maker must be indifferent between any two alternatives he chooses from the same menu. Together, Optimization and IUA are necessary and sufficient for the existence of a justification representation conditional on the candidate preference.

Axiom 1 (Optimization)**.** *If $\{a, b\} \subset c(A)$, then $a \sim b$.*

Axiom 2 (Irrelevance of Unjustifiable Alternatives (IUA))**.** *If $a \in A$, $a \succsim c(A)$ and $a \notin c(A)$, then for all $B \supseteq A$, $c(B) = c(B \setminus \{a\})$.*

Theorem 1. *c has a justification representation $(\succsim, \mathcal{M})$ if and only if it satisfies IUA and Optimization conditional on $\succsim$.*

The main step in the sufficiency proof is to show that, for any choice observed in the data, there exists some preference on $\mathcal{A}$ that justifies that choice, but does not justify any choice known to be unjustifiable. If there are no almost-consistent sets in the data, the problem of finding a justification reduces to completing an incomplete binary relation. The well-known Szpilrajn Extension Theorem provides conditions under which a solution exists, and these conditions can easily be verified given the axioms. Almost-consistent sets make the problem harder because they introduce conditions of the following form: "the justification must prefer at least one of $a_1, a_2, \ldots, a_n$ to b." I state and prove a novel version of the Szpilrajn Extension Theorem that can handle these types of conditions.

Now I turn to characterizing the model in the absence of any candidate preference. It turns out that the required conditions are closely related to IUA and Optimization. The main difference is that choice data, rather than a candidate preference, must be used to determine whether a particular alternative is unjustifiable. It turns out that two simple choice patterns can be used to identify all the unjustifiable choices. While the definitions of these choice patterns may initially appear strange and unfamiliar, both patterns have straightforward interpretations within the model. The first choice pattern, a chain, appears when the decision maker cannot justify choosing a over b.

Example 2 is the simplest instance of a chain. The second choice pattern, an almost-consistent set, appears when the decision maker cannot justify choosing a over everything in non-singleton menu B simultaneously, but can justify choosing a over any proper subset of B. Intuitively, this happens because the decision maker's justification for choosing a over one element of B is incompatible with his justification for choosing a over the rest of B. Example 1 is the simplest instance of an almost-consistent set.

Formally, chains are built up from (three-element) cycles. A cycle occurs whenever there are three items a, b, d such that $a \in c(a, b)$, $b \in c(b, d)$, and $d \in c(a, d)$, and at least one of those choices is single-valued. Cycles are very informative in the justification model, but it is important to keep track of choice from the triple $\{a, b, d\}$ as well as the pairwise choices. Definition 2 uses that information to impose an order on the alternatives in the cycle. That order turns out to match the decision maker's preference. That is, if (a, b, d) is a cycle, then $a \succsim b \succsim d$ in any justification representation for c. The preference is strict wherever choice is single-valued.

Definition 2 (Cycle)**.** *(a, b, d) is a cycle if*

$$a \in c(a, b) \quad b \in c(b, d) \quad d \in c(a, d),$$

at least one of the above choices is single-valued, and

$$a \notin c(a, b, d) \quad b \in c(a, b, d).$$

A chain is a sequence of alternatives such that every adjacent pair in the sequence is also an adjacent pair in some cycle.

Definition 3 (Chain)**.** *$(a_1, \ldots, a_n)$ is a chain if, for each $1 \leq i < n$, there exists x such that (x, a_i, a_{i+1}) or (a_i, a_{i+1}, x) is a cycle. A chain is strict if, for some $i < n$, $\{a_i\} = c(a_i, a_{i+1})$.*

Since the decision maker must weakly prefer a to b whenever (a, b) is an adjacent pair in some cycle, he must also prefer a to b whenever a precedes b in a chain. Since single-valued choice corresponds to strict preference, he must strictly prefer a to b whenever a precedes b in a strict chain. Thus, it must be unjustifiable to choose a over b if a precedes b in a chain but $a \notin c(a, b)$, or if a precedes b in a strict chain but $b \in c(a, b)$.

Now we turn to the second pattern used to identify unjustifiable choices: almost-consistent sets. We need the notion of (pairwise) consistency. A set is consistent if an available alternative that is chosen over every other available alternative in pairwise comparison is actually chosen from the set. A set is almost-consistent if it is not consistent, but all of its proper subsets are.

Definition 4 (Consistency)**.** *A menu A is consistent if*

$$a \in A \text{ and } a \in c(a, b) \text{ for all } b \in A \implies a \in c(A).$$

Definition 5 (Almost-consistent)**.** *A menu A is almost-consistent if all nonempty proper subsets of A are consistent, but A is not.*

It turns out that the decision maker's preference on any almost-consistent set matches pairwise choice on that set. Thus, if A is almost-consistent, then the decision maker would like to choose any alternative in A that pairwise-beats everything else in A. The definition of almost-consistency implies that the decision maker foregoes some such alternative. Thus, it is unjustifiable to choose a from A if a pairwise-beats everything else in A, but a is not chosen from A.

Definition 6 summarizes the procedure for recovering unjustifiable choices from chains and almost-consistent sets.

Definition 6 (Revealed unjustifiable)**.** *It is revealed unjustifiable to choose a from $A \ni a$ if:*

1. *For $A = \{a, b\}$: a precedes b in a strict chain and $b \in c(a, b)$, or a precedes b in a chain and $a \notin c(a, b)$.*
2. *For $|A| > 2$: A is almost-consistent, $a \notin c(A)$, and $a \in c(a, b)$ for all $b \in A$.*

This definition makes possible a version of IUA that does not rely on any candidate preference. The new version is called Irrelevance of Revealed Unjustifiable Alternatives (IRUA) because the unjustifiable alternatives are revealed by choice data alone. Aside from the definition of "unjustifiable," IRUA and IUA are the same.

Axiom 3 (Irrelevance of revealed unjustifiable alternatives (IRUA))**.** *If it is revealed unjustifiable to choose a from A, then for any $B \supseteq A$, $c(B) = c(B \setminus \{a\})$.*

Optimization can also be modified to remove any reference to a candidate preference. The new version, Revealed Optimization, says that a and b are tied in pairwise comparison if they are tied on some larger menu. As before, this axiom is vacuous if choice is single-valued.

Axiom 4 (Revealed optimization)**.** *If $\{a, b\} \subset c(A)$, then $c(a, b) = \{a, b\}$.*

Together, IRUA and Revealed Optimization characterize the justification model.

Theorem 2. *c has a justification representation if and only if it satisfies IRUA and Revealed Optimization.*

The sufficiency proof proceeds in two steps. The first step is to construct a candidate preference for the decision maker. Since the decision maker's preference is fully identified on cycles, the candidate preference is constructed so that it delivers the correct ordering on cycles (and, by extension, chains). For any two alternatives that are not linked by a chain, the candidate preference agrees with pairwise choice. This ensures that the candidate preference delivers the correct ordering on almost-consistent sets, where the decision maker's preference is pinned down by the pairwise

choices. The bulk of the proof consists of verifying that the construction suggested above is actually a preference.

The second step is to check that c satisfies IUA conditional on the candidate preference. I sketch the argument here for the case of single-valued choice. Suppose that $a \in A$ and $a \succ c(A)$, so it is unjustifiable to choose a from A according to the candidate preference. There are two ways to get $a \succ c(A)$. One is that $c(A)$ is chosen over a in pairwise comparison, but a precedes $c(A)$ in a chain. In that case, it is revealed unjustifiable to choose a over $c(A)$. The other way is that a is chosen over $c(A)$ in pairwise comparison. In that case, the menu A violates WARP. It is easy to check that a WARP-violating menu must contain a cycle or an almost-consistent set. Thus, it is revealed unjustifiable to choose some item in A from some subset of A. Suppose that it is not revealed unjustifiable to choose a from any subset of A. By IRUA, the revealed unjustifiable item can be removed without changing choice. Since a is not removed, the WARP violation persists. We can iterate the argument until everything is removed but a and $c(A)$. Since choice is unchanged at every step, $c(A)$ must be chosen over a in pairwise comparison, which contradicts the starting assumption. Thus, it is revealed unjustifiable to choose a from some subset of A. By IRUA, a is irrelevant in any superset of A, so IUA holds. Theorem 1 delivers a representation.

3.2 Identification properties

With Theorem 2 in hand, we now return to the identification properties of the justification model. We have already seen that the justification model delivers a unique preference on cycles and almost-consistent sets. Since any dataset that is consistent with the justification model but not with preference maximization must contain at least one cycle or almost-consistent set, cycles and almost-consistent sets are the *minimal* instances of justifying behavior. Since experiments are typically designed to minimize data collection requirements, full identification on these minimal instances implies full identification in standard experimental settings. This property is unique to the justification model, as the CFS and MNO models do not deliver a unique preference on either cycles or almost-consistent sets. (Non-uniqueness on cycles was discussed on Example 2.)

When dealing with more complicated datasets, there may be more than one preference consistent with choice. However, there is a natural criterion for selecting among these possible preferences. To motivate it, consider a decision maker whose choices are entirely consistent with maximizing a preference $\succsim$. Although those choices are also consistent with a justification model in which the preference is the opposite of $\succsim$, and $\succsim$ is the sole justification, this alternative model is not very appealing. It is a gratuitous departure from the classical model, and it is not parsimonious: it imposes constraints on the decision maker that are not needed to explain his behavior. More generally, a preference is undesirable if using that preference to explain choice would require imposing unnecessary constraints on the decision maker. The most desirable preference is the one that requires the

fewest possible constraints on what the decision maker is allowed to choose. Equivalently, it is the preference that makes choice as close to preference maximization as possible. It is not obvious that such a preference must exist; in fact, it does not exist in the CFS or MNO models. However, it always exists (and is unique) in the justification model. I refer to this maximally desirable preference as the "focal preference" and denote it $\succsim_{\text{foc}}$.

The formal definition of the focal preference is as follows. Say that a preference $\succsim$ is "possible" for c if there is some $\mathcal{M}$ such that $(\succsim, \mathcal{M})$ is a justification representation for c. Formally, $\succsim_{\text{foc}}$ is the focal preference for c if, for any possible preference $\succsim$, any menu A, and any $a \in A$,

$$c(A) \succsim a \implies c(A) \succsim_{\text{foc}} a \quad \text{and} \quad c(A) \succ a \implies c(A) \succ_{\text{foc}} a. \tag{1}$$

No more than one preference can satisfy (1), hence the use of "the."[5]

(1) captures the idea that the focal preference attributes choice to preference maximization as far as possible. If it is possible (in some representation) to say that the decision maker chose $c(A)$ over a because he preferred everything in $c(A)$ to a, then the focal preference must also rank everything in $c(A)$ above a. To see why this definition delivers a preference that requires as few constraints on the decision maker as possible, suppose that $c(A) \succsim_{\text{foc}} a$ but $a \succ c(A)$ for some preference $\succsim$ in some representation. In that representation, the decision maker would like to choose a from A, but does not. Thus, he must lack a justification for choosing a from A. If the preference is $\succsim_{\text{foc}}$, no such constraint is required.

Existence of the focal preference is a corollary to Theorem 2. As discussed in the previous section, the sufficiency proof of Theorem 2 constructs a justification representation for an arbitrary choice correspondence satisfying IRUA and Revealed Optimization. The preference in that representation turns out to be the focal preference, although additional work is needed to verify that it satisfies (1). Corollary 1 confirms that the focal preference exists and provides the algorithm for constructing it.

Corollary 1. *If c has a justification representation, then there is a unique possible preference $\succsim_{foc}$ that satisfies (1). Moreover, $a \succsim_{foc} b$ if and only if (1) a precedes b in a chain or (2) neither item precedes the other in a strict chain and $a \in c(a, b)$.*

To illustrate why Corollary 1 is useful, let us return to Example 2. Recall that CFS and MNO are consistent with both $a \succ d$ and $d \succ a$—that is, with both prejudice and the opposite of prejudice. Neither model admits a focal preference on this example, so it is not possible to use (1) to select among representations. In the justification model, $a \succ_{\text{foc}} d$; the focal preference is prejudiced.[6] This is consistent with the intuitive interpretation of Example 2, which is shared by CFS.

Moreover, Example 2 is not an isolated case. Choice in Example 2 takes the form of a cycle.

[5] Suppose $a \succsim_1 b$ and $b \succ_2 a$ for possible preferences $\succsim_1$ and $\succsim_2$. Since $\succsim_2$ is possible, a and b cannot be tied in pairwise comparison. If a is chosen over b, then $\succsim_2$ violates (1), and if b is chosen over a, then $\succsim_1$ violates (1).

[6] In fact, the focal preference is the unique possible preference here.

Recall from Theorem 2 that cycles are used to form chains, and that chains are used to infer that it is unjustifiable to choose one alternative over another. In fact, chains are the *only* way to draw this inference. In any representation that minimizes constraints on the decision maker, it is unjustifiable to choose a over b only if a precedes b in a chain.[7] Thus, a dataset that provides evidence of justification is very likely to contain one or more instances of Example 2. The advantages of the justification model on Example 2 will carry over to any such dataset.

Inspection of Corollary 1 reveals that only two inputs are required for constructing the focal preference: pairwise choice and cycles. Once pairwise choice has been observed, the analyst can observe all cycles by eliciting choice on triples for which pairwise choice is cyclic. The remainder of the choice correspondence c is not required for constructing the focal preference. In fact, the remainder of c is not required for verifying the existence of a justification representation. Suppose that the analyst only observes the data required for constructing the focal preference. Then, a justification representation exists if and only if IRUA and Revealed Optimization are satisfied on the observed data. Corollary 2 summarizes.

Corollary 2. *Suppose that the domain of c_{inc} consists of all pairs and all triples on which pairwise choice is cyclic.*[8] *If c_{inc} satisfies IRUA and Revealed Optimization on its domain, then*

1. *There is an extension of c_{inc} that has a justification representation.*
2. *Every extension of c_{inc} that has a justification representation has the same focal preference, which can be recovered from c_{inc} as described in Corollary 1.*

Corollary 2 does not imply that additional data is useless for identification. Notice that the data described in Corollary 2 will not include any almost-consistent sets. Since almost-consistent sets are informative about the decision maker's justifications, some information is lost by ignoring them. Moreover, additional data can strengthen identification by ruling out some non-focal preferences. Consider Example 1. If the analyst only observes pairwise choices, then she learns that the focal preference is $a \succ_{\text{foc}} b \succ_{\text{foc}} d$, but she cannot rule out other preferences. Once she observes choice on the triple, she can rule out everything but the focal preference. Thus, data not needed to construct the focal preference can still be informative.

After so much discussion of the focal preference, the reader may wonder whether there is any analogous concept for the justifications. The answer is a qualified yes. Whenever c has a justification representation, it has a unique representation in which the set of justifications is maximal in the sense of set inclusion. (Incidentally, the preference in that representation is the focal preference.) It turns out that a preference belongs to the maximal set of justifications if and only if it does not justify any choice that is revealed unjustifiable. Thus, the maximal set of justifications is useful because it leads

[7] This follows from Corollary 3 below.

[8] Pairwise choice is cyclic on $\{a, b, d\}$ if $a \in c(a, b)$, $b \in c(b, d), d \in c(a, d)$, and at least one of those choices is single-valued.

to a precise characterization of the choices that the decision maker must consider unjustifiable. Since it may contain many superfluous justifications, the maximal set of justifications is *not* very useful for understanding precisely which preferences the decision maker considers acceptable. For this reason, Corollary 3 focuses on the set of unjustifiable choices rather than the justifications themselves. It says that there is a representation in which the set of unjustifiable choices is minimal. In that representation, it is unjustifiable to choose a from A if and only if it is revealed unjustifiable to choose a from a subset of A. Thus, the analyst can use chains and almost-consistent sets to recover the minimal set of unjustifiable choices as well as the focal preference.

When reading Corollary 3, recall that the notation "$A \succ_{\mathcal{M}} b$" means "there is no justification in $\mathcal{M}$ for choosing b over everything in A."

Corollary 3. *If c has a justification representation, then it has a representation $(\succsim_{foc}, \mathcal{M}_{foc})$ such that*

$$A \succ_{\mathcal{M}_{foc}} a \quad \iff \quad \textit{it is revealed unjustifiable to choose } a \textit{ from some } B \subseteq A. \tag{2}$$

Moreover, the backward direction of (2) holds for any $(\succsim, \mathcal{M})$ that represents c.

4 Identification from between-subject data

The results of the previous section assume that the decision maker always chooses the most advantageous justification for the situation at hand. Thus, his decision in any given situation is not affected by any decisions he may have made previously. This is a standard approach, shared by the CFS and MNO models. It is fairly reasonable if the stakes are not too high and subjects can be distracted between related choices, as in Exley (2016). It is less reasonable when subjects have a strong incentive to conceal their preferences, i.e. in experiments on prejudice. In that case, subjects may avoid or limit the WARP violations that constitute evidence of justification.

The results of this section will be entirely robust to consistency motives on the part of subjects. They apply to situations where data is collected from a large population of subjects, each of whom makes only one choice. I model this type of data by using a stochastic choice function ρ. In random utility models, it is well known that extending the choice domain to lotteries and requiring preferences to be expected utility improves tractability and identification. The same turns out to be true for the justification model. Thus, I assume that all preferences and all justifications have an expected-utility form.

The domain of choice, denoted $\Delta(Z)$, is the set of lotteries on a finite set Z. The observable is $\rho : \Delta(Z) \times \mathcal{F}(\Delta(Z)) \to [0,1]$ such that $\sum_{a \in A} \rho(a|A) = 1$ for all $A \in \mathcal{F}(\Delta(Z))$.[9] The set of expected-utility preferences on $\Delta(Z)$ is denoted $\mathcal{U}$. A subset of $\mathcal{U}$ is said to be closed and convex

[9] In fact, it is enough for ρ to be defined on the collection menus with two or three elements. This is in contrast to the deterministic case, where menus with more than three elements can convey additional information.

if it can be represented by a closed and convex set of utilities. The collection of closed, convex, nonempty subsets of $\mathcal{U}$ is denoted $\mathcal{C}(\mathcal{U})$. A random justification model is a probability measure P on $\mathcal{U} \times \mathcal{C}(\mathcal{U})$.

Definition 7 (Random justification representation)**.** *Random justification model P is a random justification representation for stochastic choice function ρ if, for each menu A and each item $a \in A$,*

$$\rho(a|A) = P\left(\left\{(\succsim, \mathcal{M}) : a = c_{(\succsim,\mathcal{M})}(A)\right\}\right)$$

where $c_{\succsim,\mathcal{M}}$ is the choice function represented by $(\succsim, \mathcal{M})$.

4.1 Qualitative inferences

This section explains how to draw qualitative inferences about justification from between-subject data. Like the rest of the paper, the material in this section applies to any situation where desires or inclinations conflict with principles, norms, or duties, I focus on moral decision-making to simplify the exposition. Suppose the analyst wants to determine whether a society considers one alternative morally worse than another, while allowing for the possibility that people differ in their moral principles and their adherence to those principles. One reasonable (and tractable) approach is to say that b is morally worse than b if the event "it is unjustifiable to choose b over a" has positive probability.[10] Definition 8 formalizes this idea, extending it to allow an alternative to be morally worse than a menu.

Definition 8. *Given a random justification model P: lottery b is morally worse than menu A if $P(A \succ_{\mathcal{M}} b) > 0$. b is weakly morally worse than A if*

$$b \in cl(\{x \in \Delta(Z) : x \text{ is morally worse than } A\}).$$

When choice is deterministic, unjustifiable choices are pinned down by a particular class of departures from utility maximization. When choice is stochastic, morally worse choices are pinned down by a particular class of departures from random expected utility. Example 3 will be used to illustrate that class of departures.

[10] This definition is especially appealing when it generates an asymmetric relation on the domain of choice. Then, the statement "a is morally worse than b" can be interpreted as "everyone agrees that choosing b is morally better, but not everyone feels obligated to do it." I impose an assumption to that effect later on.

Example 3. *The decision maker must choose whether to volunteer for a task. Let*

$$a = \textit{no task}$$
$$b = \textit{difficult and very important task}$$
$$d = \textit{easy and somewhat important task.}$$

In the first treatment, the choice set is $\{a, b, d\}$*. In the second treatment, the decision maker is told that there may be too many volunteers for the difficult task, so anyone who volunteers for that task will be reassigned to the easy task with probability* $\epsilon \in (0, 1)$*. Thus, the choice set is* $\{a, (1-\epsilon)b + \epsilon d, d\}$*.*

Random expected utility predicts that the probability of d is identical across the two treatments. However, this is not the case in the justification model. There are exactly two groups of people whose probability of choosing d is affected when b is mixed with d. One group has $a \succ d \succ b$ and can justify choosing d but not a. The other group has $b \succ d \succ a$ and can justify choosing d but not b. It turns out that the first group must be weakly *less* likely to choose d in the second treatment, while the second group must be weakly *more* likely to choose d. Thus, if the probability of d falls from the first treatment to the second, the analyst learns that there are people who cannot justify choosing a over b and d. (In fact, the size of the fall provides a lower bound on the proportion of such people.) If the inequality across treatments continues to hold as ϵ is made arbitrarily small, the effect is entirely driven by people who cannot justify choosing a over alternatives arbitrarily close to b. Thus, a is weakly morally worse than b. Proposition 1 formalizes this discussion.

Definition 9 (Anomalous)**.** (a, b, d) *is anomalous if*

$$\rho(d|a, b, d) > \rho(d|a, (1-\epsilon)b + \epsilon d, d) \tag{3}$$

for all sufficiently small $\epsilon > 0$*.*

Proposition 1. *Suppose that* ρ *has a random justification representation* P*. If* (a, b, d) *is anomalous, then* a *is weakly morally worse than* b*.*

Proposition 1 is useful in two ways. First, it explains how one might design a between-subject experiment to obtain evidence of justification. Despite the advantages of between-subject data discussed at the beginning of this section, this question has not been addressed in existing work. In the absence of tools for analyzing between-subject data, single-person models have sometimes been inappropriately applied to modal choices. As mentioned in the introduction, this practice has led data consistent with preference maximization to be misinterpreted as evidence of justification. Proposition 1 solves this problem because it relies on departures from preference maximization. Second, Proposition 1 can be used to test the justification model in conjunction with natural

restrictions on the "morally worse than" relation. Example 3 illustrates. Since it is implausible that volunteering for an important task is morally worse than doing nothing, the triple (b, a, d) should not be anomalous.

The weakness of Proposition 1 is that it does not specify when anomalies are likely to occur. In particular, it does not guarantee that (a, b, d) is anomalous for some d whenever a is morally worse than b. To see why, return to Example 3, and recall that there are two opposing forces on the probability of choosing d when b is mixed with d. One force, which pushes in the right direction for an anomaly, is nonzero if a is morally worse than b. The other force, which pushes in the wrong direction, is nonzero if b is morally worse than a. If the latter force is large enough, (a, b, d) will fail to be anomalous even if a is morally worse than b.

Of course, this is quite unlikely in the particular context of Example 3. It makes sense for a to be morally worse than b, but not vice versa. More generally, there are many situations in which the "morally worse" relation is likely to be asymmetric. When we restrict attention to these situations and impose some additional technical conditions, we obtain the desired converse of Proposition 1. The main assumption is called Limited Disagreement, and the additional technical assumptions are together called Regularity. These assumptions will also be required for the quantitative results in the next section.

Definition 10 (Limited disagreement)**.** *For any menu A and lottery b: if each $a \in A$ is weakly morally worse than b, then b is not weakly morally worse than A.*

Limited Disagreement is easy to interpret when A is a singleton. Extending it to non-singleton A ensures that b is not morally worse than any mixture of alternatives that are morally worse than b itself.

Regularity consists of three assumptions. The first assumption, which says that the marginal distribution on preferences admits a density, is only required for the next section. Since it seems unobjectionable, I have included it here. The second assumption says that any expected-utility preference is possible, and that any possible preference can occur together with any possible set of justifications. The first half of this assumption can be relaxed at the expense of additional notation.[11] The third assumption is a richness condition on the sets of justifications. It says that, for any possible set of justifications that does not already permit everything, there is another possible set of justifications that is larger than but arbitrarily close to the original set.

Definition 11 (Regularity)**.** *Suppose that P is a random justification model. P is regular if*

1. *P_{pref} admits a density.*
2. *$supp(P) = \mathcal{U} \times supp\,(P_{just})$.*

[11] For instance, it would be enough to assume that the support of P_{just} includes the set of preferences that are unjustifiable on $\Delta(Z)$ in the sense of Definition 12.

3. *For any* $\mathcal{M} \in supp\,(P_{just})$ *and any closed, convex* $U \subseteq \mathcal{U}$ *such that* $\mathcal{M} \subset int(U)$*, there exists* $\mathcal{N} \in supp\,(P_{just})$ *such that* $\mathcal{M} \subset int(\mathcal{N}) \subset \mathcal{N} \subset int(U)$*.*

Given Limited Disagreement and Regularity, there is an anomaly corresponding to every pair of alternatives that are strongly morally ranked. This implies that the "morally worse than" relation is fully pinned down by anomalies.

Theorem 3. *If* ρ *has a regular random justification representation* P *that satisfies Limited Disagreement, then*

$$\{a : a \text{ is morally worse than } b\} = int\,(\{a : (a, b, d) \text{ is anomalous for some } d\})\,.$$

The most practically useful implication of Theorem 3 is the following: if a is morally worse than b, then there must be some d for which (a, b, d) is anomalous. The discussion of Example 3 provides some guidance for finding an appropriate d. There should people who like a best and d second best, and can justify choosing d over b but not a over b. Thus, d should be moderately morally compelling, and moderately appealing to someone whose preferences are not particularly moral.

4.2 Quantitative inferences

This section turns from qualitative to quantitative inferences about justification. The results of the previous section were intended to be simple and practical. By contrast, the results of this section are intended to probe the limits of the model by showing what can be identified in principle.

As in the deterministic case, full identification is not possible in general. To see this, consider a preference that always prefers b to a if a is morally worse than b. Since a person with this preference never wants to violate the constraints imposed by the justifications, it is impossible to determine what distribution over justifications he faces. More generally, if a is morally worse than b, it it is impossible to learn how often someone who prefers b is able to justify a.

To avoid this problem, I focus the identification exercise on people whose preferences are the furthest from justifiable, i.e. people who always want to violate the constraints imposed by the justifications. These preferences are called "unjustifiable."

Definition 12 (Unjustifiable)**.** *For any subspace* X *of* $\Delta(Z)$ *such that the restriction of the morally worse than relation to* X *is nonempty: a preference* $\succsim \in \mathcal{U}$ *is unjustifiable on* X *if*

$$y \text{ is morally worse than } x \quad \Longrightarrow \quad y \succ x$$

for all $x, y \in X$*.*

In between-subject data, it is not possible to determine precisely who the people with unjustifiable preferences are. Although it is not straightforward, it *is* possible to disentangle their aggregate

behavior from that of others by examining the right collection of choice problems. Theorem 4 says that it is possible to determine, for any given binary choice, the proportion of such people who cannot justify choosing the morally worse alternative.

Theorem 4. *If ρ has regular random justification representations P and Q that satisfy limited disagreement, then for any two-dimensional subspace X of $\Delta(Z)$ and lotteries $a, b \in X$ such that a is morally worse than b,*

$$P(b \succ_{\mathcal{M}} a | \succsim \text{ is unjustifiable on } X) = Q(b \succ_{\mathcal{M}} a | \succsim \text{ is unjustifiable on } X).$$

If preferences and sets of justifications are drawn independently, Theorem 4 delivers full identification on binary menus. That is, for any given binary choice, the analyst can determine what proportion of people are choosing in accordance with their preferences and what proportion are not.

While the independence assumption is a reasonable starting point for a model of conflict between preferences and justifications, it is not particularly plausible. It is more plausible that people with unjustifiable preferences are less likely than the average person to see a given alternative as unjustifiable. Then, for any given binary choice, the fraction of such people who cannot justify the morally worse alternative provides a lower bound on the overall probability that the morally worse alternative is unjustifiable. This lower bound can be used to obtain an estimate of the proportion of people who prefer the morally better alternative. That upper bound falls somewhere in between the proportion of people who choose the morally better alternative and the proportion who actually prefer it. Thus, it provides a *conservative* estimate of the extent to which choice departs from preference maximization. The quality of the estimate depends on the degree to which people with unjustifiable preferences depart from the rest of the population.

5 Information choice

This section studies information choice in the random justification model. It is motivated by the large experimental literature on information choice in ethical settings, which is often called the "moral wiggle room" literature. As discussed in Section 2, many experiments study the effects of information about a possible negative externality to an appealing action. Subjects are less likely to choose the action when they learn that it will cause the externality, but a substantial fraction of subjects choose not to learn about the externality when given the option. The standard interpretation of this behavior is that avoiding information somehow allows subjects to feel better about indulging their desires. This interpretation suggests a conflict between desires and standards for acceptable behavior, which is the same mechanism that drives the justification model. However, the moral wiggle room literature is not closely linked to other work on justification. This may be

due to the absence of a model that can accommodate both.[12]

The setup of this section is modeled after the moral wiggle room literature. Choice takes place in two stages. In the second stage, subjects choose from a menu of lotteries as in Section 4. In the first stage, subjects are offered a choice between two lotteries over second-stage menus. One lottery, denoted δ_A, is degenerate. The other, denoted S, is a finite-support distribution over $\mathcal{F}(\mathcal{A})$ such that, for each $a \in A$ and each $\tilde{A} \in \text{supp}(S)$, there exists $\tilde{a} \in \tilde{A}$ such that

$$a = \sum_{\tilde{A} \in \text{supp}(S)} S\left(\tilde{A}\right) \tilde{a}.$$

I refer to S as a "signal" about A.

All results in this section take as given a random justification model P. In its current form, the justification model does not say what a given person will *do* at the information choice stage, but it does say what that person *prefers*. For the first two results in this section, information preferences will be enough. I return to the question of information choice in the final result.

The moral wiggle room literature provides strong evidence that many people avoid information that might induce them to behave more morally. This finding raises concerns about the effectiveness of public information campaigns intended to promote pro-social behavior. The first result in this section offers some limited reassurance. Consider a situation in which people are broadly aligned in their moral values. Suppose people must choose between two alternatives, one of which is morally worse than the other. Suppose further that there is a policymaker who wishes to encourage morally good behavior. It turns out that the policymaker can design a signal that will improve average behavior among the people who prefer the morally worse alternative, even if they are entirely free to avoid unwanted information. While the policymaker is not as well off as he would be if he could compel people to observe the signal, he is still better off than he would be if he had not offered the signal.

Before stating the result, we need to formalize the notion that people are broadly aligned in their moral values. Consider a set of binary choice problems, where the alternatives in each choice problem are morally ranked. Say that the set of binary choice problems is orderly if people can be ranked by the strictness of the obligations they perceive. That is, if one person can justify the morally worse alternative in choice problem 1 but not choice problem 2, then there cannot be another person who can justify the morally worse alternative in choice problem 2 but not choice problem 1. Definition 13 captures this requirement. Orderliness is useful for the policymaker because it ensures that a signal realization that induces one person to behave more morally will not induce another person to behave less morally. Thus, it makes it easy to construct a signal with unambiguously good effects.

[12] For instance, Spiekermann and Weiss (2016) and Grossman and Van Der Weele (2017) propose models of information choice in ethical settings, but do not consider any other type of choice problem.

Definition 13 (Orderly)**.** *Let A and B be open convex nonempty subsets of $\Delta(Z)$ such that each $a \in A$ is morally worse than each $b \in B$. (A, B) is orderly if the family of sets*

$$\{\{\mathcal{M} \in supp(P_{just}) : b \succ_{\mathcal{M}} a\} : a \in A, b \in B\} \tag{4}$$

is totally ordered by set inclusion.

Although orderliness is a strong assumption, it is satisfied in an important class of justification models. If people disagree on how strict to be but do not have more substantive disagreements, then different sets of justifications will be ordered by set inclusion. In that case, (4) will be satisfied vacuously.

Example 4. *Let*

$$\begin{aligned} a_1 &= \textit{order beef from sustainable farm} \\ a_2 &= \textit{order beef from unusustainable farm} \\ a &= \frac{1}{2}a_1 + \frac{1}{2}a_2 \\ b &= \textit{do not order beef.} \end{aligned}$$

If $Z = \{a_1, a_2, b\}$, then $supp(P_{just})$ is likely to be totally ordered by set inclusion.

Proposition 2 says that a policymaker faced with an orderly situation can use free information to improve behavior.

Proposition 2. *Suppose that (A, B) is orderly and that $\succ_{pol} \in \mathcal{U}$ satisfies $b \succ_{pol} a$ for all $a \in A$ and $b \in B$. For almost any $(a, b) \in A \times B$, there exists a signal S about (a, b) such that*

$$\sum_{supp(S)} S(a_i, b_i) c_{(\succ, \mathcal{M})}(a_i, b_i) \succsim_{pol} c_{(\succ, \mathcal{M})}(a, b)$$

for every $(\succ, \mathcal{M})$ such that $a \succ b$. The above holds with strict preference for a positive-probability set of $(\succ, \mathcal{M})$ such that $a \succ b$ and

$$\sum_{supp(S)} S(a_i, b_i) c_{(\succ, \mathcal{M})}(a_i, b_i) \succ c_{(\succ, \mathcal{M})}(a, b).$$

Proposition 2 says that the policymaker can reach people who disagree with him on the original choice set, but does not say that he can reach people who disagree with him on every realization of the signal. The stronger statement would not be correct. As the proof of Proposition 2 shows, the benefits of the signal are confined to people who agree with the policymaker on one of the

signal realizations. This is easy to see in Example 4. Consider a policymaker who objects only to unsustainable beef. Diners who already feel obligated not to eat beef may seek information about sustainability, but this will not affect the policymaker. The benefit of the information comes from diners who prefer eating beef in the absence of information, but not when they know the beef is unsustainable.

While this point may appear obvious, it relies on the assumption that people know their justifications when choosing their information. When people are uncertain about their justifications, free information is even more useful, as the policymaker can reach some people who disagree with him on every realization of the signal. In Example 4, consider a diner who does not yet know whether he will be able to justify ordering beef. Suppose that the diner prefers eating beef regardless of its source, but enjoys it much less when he knows that the beef is unsustainable. He may be willing to demand information about sustainability so that he will be free to indulge when the information is good. The formal result, which is very similar to Proposition 2, is in Appendix B.

Proposition 2 is closely aligned with the moral wiggle room literature: it studies information that induces more moral behavior, leading some people with less-than-moral preferences to choose ignorance. This may give the impression that information is always beneficial from a moral policymaker's point of view, and that avoidance is the only obstacle to achieving the full benefits of information. However, that impression turns out to be mistaken. The mistake comes from neglecting a key feature of the justification model: decision makers shift between justifications to get closer to their desired outcome. In ambiguous situations, different justifications may exhibit substantive disagreements. Some signals allow decision makers to exploit that disagreement by appealing to different justifications on different signal realizations.

Before stating the result, we need to formalize the notion an ambiguous situation. As before, consider a set of binary choice problems, where the alternatives in each choice problem are morally ranked. Say that a situation is unambiguous if every possible set of justifications contains a single justification that is maximally permissive, meaning it prefers the morally better alternative only if every other justification does as well. Otherwise, say that the situation is ambiguous. In ambiguous situations, there are people who benefit by using different justifications on different choice problems.

Definition 14 (Ambiguous)**.** *Let A and B be open convex nonempty subsets of $\Delta(Z)$ such that each $a \in A$ is morally worse than each $b \in B$. (A, B) is ambiguous if, for some $\mathcal{M} \in supp(P_{just})$, there does not exist $\succ_m^* \in \mathcal{M}$ such that*

$$b \succ_{\mathcal{M}} a \quad \iff \quad b \succ_m^* a. \tag{5}$$

Example 5 illustrates.

Example 5. *This example is based on Norton et al. (2004). The DM must hire someone for a*

managerial role at a construction company. Let

$a =$ *male candidate*

$b_1 =$ *female candidate with more education, less experience than the male candidate*

$b_2 =$ *female candidate with more experience, less education than the male candidate*

$b = \frac{1}{2}b_1 + \frac{1}{2}b_2$.

Let $A := B_\epsilon(a)$*, and* $B := B_\epsilon(b)$ *for some* $\epsilon > 0$*. (5) will hold if some decision makers have access to different justifications that place very different weights on education vs. experience.*

Although Definitions 13 and 14 may sound like opposites, they are not. Orderliness constrains how different sets of justifications relate to one another, while ambiguity constrains how different justifications in the same set relate to one another. For instance, Example 5 is consistent with the possibility that different people can be ranked by the strength of the obligation they perceive to prioritize underrepresented candidates. That would make (A, B) orderly as well as ambiguous.

Proposition 3 says that some people facing ambiguous situations can find signals that are unambiguously good from their point of view, and unambiguously bad from the point of view of a moral policymaker. Despite the literature's focus on information avoidance, it is not the only interesting behavior that justifiers exhibit in information choice problems.

Proposition 3. *Suppose that* (A, B) *is ambiguous and* $\succ_{pol} \in \mathcal{U}$ *satisfies* $b \succ_{pol} a$ *for all* $a \in A$ *and* $b \in B$*. There exists a signal* S *about some* $(a, b) \in A \times B$ *and a positive-probability set of* $(\succ, \mathcal{M})$ *such that*

$$c_{(\succ,\mathcal{M})}(a_i, b_i) \prec_{pol} c_{(\succ,\mathcal{M})}(a, b)$$
$$c_{(\succ,\mathcal{M})}(a_i, b_i) \succ c_{(\succ,\mathcal{M})}(a, b)$$

for every $(a_i, b_i) \in supp(S)$ *and every* $(\succ, \mathcal{M})$ *in the set.*

To illustrate, consider a decision maker in Example 5 who prefers to hire a male candidate but cannot justify doing so in the absence of further information. He may then seek out detailed information about both candidates' qualifications. By choosing between justifications that place different weights on different qualifications, he may be able to hire the male candidate in most states of the world.

In fact, Bleemer (2019) finds that universities engage in a similar practice, although the aim is to advantage rather than disadvantage applicants from underrepresented groups. In "holistic review," universities collect multidimensional information about applicants and vary the weight on each dimension from application to application. Proposition 3 predicts that holistic review would

increase the proportion of students from underrepresented groups, and Bleemer (2019) finds that this is the case. This example is a good reminder that the desires-vs.-morals interpretation of the justification model is not applicable in every situation. While I have focused on that interpretation to simplify the terminology and exposition, the model can also apply to situations where legal or social constraints stop decision makers from acting on preferences that are morally unobjectionable.

For the final result of the paper, I return to the question of precisely how justifiers choose between information and ignorance. One possibility is simply that justifiers choose information in accordance with their preferences. However, this does not fit well with the premise behind the justification model. I propose a flexible extension of the model in which information choices do have to be justified, but may be subject to a different set of justifications than choices over final outcomes.

Formally, a two-stage justification model consists of a preference $\succsim$ and two (possibly equal) sets of justifications, $\mathcal{M}_{\text{info}}$ and $\mathcal{M}_{\text{final}}$. I assume that the set of justifications governing information is weakly larger than the set of justifications governing final choice. The motivating idea is that information choices do not directly determine outcomes, so they may be subject to less thought or scrutiny, or seen as less informative about the decision maker's character. I also impose a maximality condition on the sets of justifications: if adding a justification to the set would not change what is justifiable, then that justification must already belong to the set. This assumption ensures that a gap between $\mathcal{M}_{\text{info}}$ and $\mathcal{M}_{\text{final}}$ is meaningful; it does not reflect the presence or absence of superfluous justifications.

Definition 15 (Two-stage justification model)**.** *A two-stage justification model is a triple* $(\succsim, \mathcal{M}_{info}, \mathcal{M}_{final}) \in \mathcal{U} \times \mathcal{C}(\mathcal{U})^2$ *such that*

1. $\mathcal{M}_{info} \supseteq \mathcal{M}_{final}$.
2. *(Maximality) For* $\mathcal{M} \in \{\mathcal{M}_{info}, \mathcal{M}_{final}\}$*: if* $\succsim^* \in \mathcal{U}$ *has* $a \succ^* b$ *for all items* a *and* b *such that* $a \succ_{\mathcal{M}} b$*, then* $\succsim^* \in \mathcal{M}$.

The choice procedure is as follows. Consider a decision maker facing a two-stage choice problem, where the first stage is information choice. A strategy consists of a signal and a selection from each signal realization. For a given strategy to be justifiable, it must be justifiable from both the first-stage and second-stage points of view. A strategy is justifiable from the second-stage point of view if, for each signal realization, there is a second-stage justification for selecting the alternative specified by the strategy. (This is the usual notion of "justifiable.") A strategy is justifiable from the first-stage point of view if there is a first-stage justification for carrying out the entire strategy.

When the decision maker's first-stage choice is between ignorance and information, any strategy that picks information in the first stage and picks justifiable alternatives in the second stage is justifiable. Thus, the definition of "justifiable" is only distinctive for a strategy that picks ignorance in the first stage.

Definition 16. *Fix a two-stage justification model. Given a menu A, an item $a \in A$, and a signal S about A, the pair (δ_A, a) is justifiable if*

$$a = \mathbb{E}_S \left[\max \left(\succsim_{info}, \tilde{A} \right) \right] \text{ for some } \succsim_{info} \in \mathcal{M}_{info}$$

and

$$a = \max \left(\succsim_{final}, A \right) \text{ for some } \succsim_{final} \in \mathcal{M}_{final}.$$

Proposition 4 says that the justification model predicts some information avoidance if and only if the set of justifications governing information choice is strictly larger than the set of justifications governing choice over final outcomes. Thus, the justification model can account for the main finding of the moral wiggle room literature if people are willing to use somewhat weaker justifications for their information choices than choices that directly affect others. Admittedly, this is only a partial account, as the justification model takes the justifications as given rather than explaining where they come from.

Proposition 4. *For any two-stage justification model, $\mathcal{M}_{info} \supset \mathcal{M}_{final}$ if and only if there exist a menu A, item $a \in A$, and signal S about A such that (δ_A, a) is justifiable and*

$$a = \max(\succsim, \mathcal{M}_{final}(A)) \succ \mathbb{E}_S \left[\max \left(\succsim, \mathcal{M}_{final} \left(\tilde{A} \right) \right) \right] \text{ for some } \succsim \in \mathcal{U}. \tag{6}$$

6 Conclusion

We seem to be living in a time when people are especially concerned with having, or being seen to have, the right values, principles, and beliefs. The goal of this paper has been to provide a broad range of researchers (both theoretically and empirically oriented) with a common framework for thinking about choice in these sorts of contexts. While the model owes a great deal to prior work on rationalization and limited attention, I believe that it is unique in its strong identification properties, its ability to handle subjects' attempts to conceal their preferences, and its connections to a broad range of experimental work.

One question this paper leaves open is that of welfare. While decision makers in the justification model do have preferences that guide their actions, it would likely be a mistake to take those preferences as a measure of welfare, at least where they are known to conflict with the justifications. It is very plausible that people are sometimes better off when a moral principle, desire to be virtuous, or sense of obligation stops them from indulging one of their desires.

References

Bleemer, Zachary, "Diversity in University Admissions: Affirmative Action, Percent Plans, and

Holistic Review. Research & Occasional Papers Series: CSHE. 6.2019.," *Center for Studies in Higher Education*, 2019.

Cherepanov, Vadim, Timothy Feddersen, and Alvaro Sandroni, "Rationalization," *Theoretical Economics*, 2013, *8* (3), 775–800.

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

Grossman, Zachary and Joel J Van Der Weele, "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y Ozbay, "Revealed attention," *American Economic Review*, 2012, *102* (5), 2183–2205.

Norton, Michael I, Joseph A Vandello, and John M Darley, "Casuistry and social category bias.," *Journal of personality and social psychology*, 2004, *87* (6), 817.

Snyder, Melvin L, Robert E Kleck, Angelo Strenta, and Steven J Mentzer, "Avoidance of the handicapped: an attributional ambiguity analysis.," *Journal of personality and social psychology*, 1979, *37* (12), 2297.

Spiekermann, Kai and Arne Weiss, "Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'," *Games and Economic Behavior*, 2016, *96*, 170–183.

A Proofs of results in text

A.1 Proof of Theorem 1

Throughout this proof, justifications are assumed to be strict preferences. This is without loss of generality.

First, we show necessity. Necessity of Optimization follows because the items that maximize a preference over a set must all be indifferent. For IUA, fix A, a such that $a \in A$. Suppose $a \succsim c(A)$ and $a \notin c(A)$, and fix $B \supseteq A$. For all $\succ_m \in \mathcal{M}$, we have $a \not\succsim_m A$, so $a \not\succsim_m B$. To confirm that $c(B) = c(B \setminus \{a\})$, it suffices to show that $\mathcal{M}(B) = \mathcal{M}(B \setminus \{a\})$. Take any $b \in \mathcal{M}(B)$. Since $b \succsim_m B$ for some $\succ_m \in \mathcal{M}$, $b \succsim_m A$ for some $\succsim_m \in \mathcal{M}$, so $b \neq a$. Since $b \succsim_m B$ implies $b \succsim_m B \setminus \{a\}$, we have $b \in \mathcal{M}(B \setminus \{a\})$. Now take any $b \in \mathcal{M}(B \setminus \{a\})$. There exists $\succ_m \in \mathcal{M}$ such that $b \succsim_m B \setminus \{a\}$. Since it cannot be that $a \succsim_m b \succsim_m B \setminus \{a\}$, we have $b \succsim_m B$, so $b \in \mathcal{M}(B)$.

Notation: for a menu X and item $y \notin X$, write $X \triangleright y$ if $y \succsim X$ and $y \notin c(X \cup \{y\})$.

Let

$$\mathcal{M} := \{\succ_m : y \not\succ_m X \text{ if } X \triangleright y\}.$$

We need to show that

$$c(A) = \max\left(\succsim, \mathcal{M}(A)\right)$$

where $\mathcal{M}(A) = \bigcup_{\succ_m \in \mathcal{M}} \max(\succ_m, A)$.

Fix some menu A and item $b \notin A$ such that $b \in c(A \cup \{b\})$. We need to show that $b \in \mathcal{M}(A)$. That is, we need to find a preference $\succ_m$ such that $b \succ_m A$ and $y \not\succ_m X$ if $X \triangleright y$. We will start by extending $\triangleright$. First, we define several useful properties of $\triangleright$.

Definition 17 (Menu-item relation)**.** *A menu-item relation is a subset of* $(\mathcal{F}(\mathcal{A}) \cup \{\emptyset\}) \times \mathcal{A}$.

Definition 18 (Properness)**.** *A menu-item relation R is proper if $X\ R\ x \implies X \neq \emptyset$.*

Definition 19 (Irreflexivity)**.** *A menu-item relation R is irreflexive if $X\ R\ x \implies x \notin X$.*

Definition 20 (Transitivity)**.** *A menu-item relation R is transitive if*

$$X\ R\ y \in Y\ R\ z \quad \implies \quad X \cup Y \setminus \{y, z\}\ R\ z.$$

As usual, the transitive closure of a menu-item relation R is the smallest transitive menu-item relation that includes R. It is denoted $\text{tr}(R)$.

Lemma 1. *$\triangleright$ is transitive.*

Proof. If $X \triangleright y \in Y \triangleright z$, then $z \succsim Y$ and $y \succsim X$. Since $y \in Y$, we have $z \succsim X \cup Y \setminus \{y\}$. By IUA,

$$c(\{z\} \cup X \cup Y \setminus \{y\}) = c(\{z\} \cup X \cup Y) = c(X \cup Y \setminus \{z\}).$$

Thus, $z \notin c(\{z\} \cup X \cup Y \setminus \{y\})$. Conclude that $X \cup Y \setminus \{y, z\} \triangleright z$. □

Definition 21 (Consistency with $b \succ A$)**.** *A menu-item relation R is consistent with $b \succ A$ if it is not the case that $A'\ R\ b$ for any nonempty $A' \subseteq A$.*

Lemma 2. *If $b \in c(\{b\} \cup A)$, then $\triangleright$ is consistent with $b \succ A$.*

Proof. Suppose that $A' \triangleright b$ for some $A' \subseteq A$. Then, $b \succsim A'$, so $b \succsim c(\{b\} \cup A')$, and $b \notin c(\{b\} \cup A')$. By IUA, $c(\{b\} \cup A) = c(A)$, so $b \notin c(\{b\} \cup A)$—contradiction. □

The following two lemmas will be useful for extending $\triangleright$.

Lemma 3. *For any irreflexive, transitive and proper menu-item relation R and any distinct items x, and y such that $\neg(\{y\}\ R\ x)$, the menu-item relation $tr(R \cup (\{x\}, y))$ is irreflexive and proper.*

Proof. Let $R^0 := R$. For $i > 0$, let R^i be the extension of R^{i-1} obtained by imposing

$$\left(\bigcup_{j=1}^{k} X_j \cup \{y_{k+1}, \ldots, y_n\} \right) \setminus \{y\} \; R^i \; y$$

whenever

$$\{y_1, \ldots, y_n\} \; R^0 \; y \text{ and, for all } j \leq k, \; X_j \; R^{i-1} \; y_j.$$

Then, $\mathrm{tr}(R) = \bigcup_{i=0}^{\infty} R^i$. This is a standard result about the transitive closure. The usual proof goes through with the version of transitivity used here.

Repeated applications of transitivity will not lead to a violation of irreflexivity, so we only need to check whether $\mathrm{tr}(R \cup (\{x\}, y))$ is proper. To keep track of repeated applications of transitivity, we introduce the notion of a tree.

Definition 22 (Q-tree)**.** *For a menu-item relation Q, a Q-tree from $W \in \mathcal{F}(\mathcal{A}) \cup \{\emptyset\}$ to $w \in \mathcal{A}$ is inductively defined as follows:*

- *$z_0 := w$ is mapped to $Z_1(z_0)$ such that $Z_1(z_0) \; Q \; z_0$.*
- *For $k > 0$: each $z_k(z_{k-1}(z_{k-2}(\cdots)))$ that does not belong to*

  $$W \cup \{z_{k-1}(z_{k-2}(\cdots)), z_{k-2}(\cdots), \ldots, z_1(z_0), z_0\}$$

  is mapped to $Z_{k+1}(z_k(z_{k-1}(\cdots)))$ such that $Z_{k+1}(z_k(z_{k-1}(\cdots))) \; Q \; z_k(z_{k-1}(\cdots))$.
- *For some finite $K > 0$: each $z_K(z_{K-1}(\cdots))$ belongs to*

  $$W \cup \{z_{K-1}(z_{K-2}(\cdots)), z_{K-1}(\cdots), \ldots, z_1(z_0), z_0\}.$$

$Z_k(z_{k-1}(\cdots))$ is called the set of parents of $z_{k-1}(\cdots)$. If $z_k(\cdots)$ does not have parents, it is called a top node. A branch of a tree is a sequence $(z_0, z_1(z_0), \ldots, z_k(z_{k-1}(\cdots)))$ such that $z_k(\cdots)$ is a top node. For any $i < k$, we refer to $z_0, \ldots, z_{i-1}(\cdots)$ as descendants of $z_i(\cdots)$, and to $z_{i+1}(z_i(\cdots)), \ldots, z_k(z_{k-1}(\cdots))$ as ancestors of $z_i(\cdots)$.

It is not difficult to see that (W, w) belongs to $\mathrm{tr}(Q)$ if and only if there is a Q-tree from W to w. If $\mathrm{tr}(R \cup (\{x\}, y))$ is improper, there is a $R \cup (\{x\}, y)$-tree from $\emptyset$ to w for some $w \in \mathcal{A}$. Notice that there must be at least one instance of y in the tree that has x as its sole parent. (Otherwise, there would be any R-tree from $\emptyset$ to w, contradicting properness of R.) Suppose there are n such instances. Let x_i and y_i denote the relevant instances of x and y. Let T_1 denote the tree obtained from the original tree by removing everything except x_1 and its ancestors. We claim that T_1 is an R-tree. If not, there is a branch or the original tree that contains two instances of (y, x). But the

second instance of y would have to be a top node by definition of a tree, so it could not have x as its parent. Let V_1 be the set of top nodes of T_1 that are not repetitions of any of their descendants in T_1. T_1 is an R-tree from V_1 to x. Since R is transitive, $V_1\ R\ x$. Since each $v \in V_1$ is a top node of the original tree from $\emptyset$ to w, and since all the top nodes of the original tree are repetitions of their descendants, each $v \in V_1$ must be a descendant of y_1.

Now construct a new tree from the original tree by removing everything except the ancestors of w. Also remove all the ancestors of y_j for each j. The result is an R-tree from $\{y\}$ to w. (All the other top nodes must be top nodes of the original tree, which are repetitions of their descendants.) Since R is transitive, $\{y\}\ R\ w$. Now fix any parent z_1 of w. Construct a new tree by removing everything except the ancestors of w. Also remove all the ancestors of y_j for each j. The result is an R-tree from $\{y, w\}$ to z_1. (All the top nodes besides y must be repetitions of their descendants in the original tree. All the descendants except w are present in the new tree.) Since $\{y\}\ R\ w$, there is an R-tree from $\{y\}$ to z_1, so $\{y\}\ R\ z_1$. We can iterate this process to get $y\ R\ z$ for any descendant z of any y_i. Since each $v \in V_1$ is a descendant of y_1, we have $\{y\}\ R\ v$ for each $v \in V_1$. Since $V_1\ R\ x$, we have an R-tree from $\{y\}$ to x, so $\{y\}\ R\ x$—contradiction. □

Lemma 4. *Fix an irreflexive, transitive, proper menu-item relation R that is consistent with $b \succ A$. For any $a \in A$, $tr(R \cup (\{b\}, a))$ is consistent with $b \succ A$.*

Proof. Suppose that $\mathrm{tr}(R \cup (\{b\}, a))$ is inconsistent with $b \succ A$, so $(A', b) \in \mathrm{tr}(R \cup (\{b\}, a))$ for some nonempty $A' \subseteq A$. Then, there must be an $R \cup (\{b\}, a)$-tree from A' to b. Construct a new tree by removing all the ancestors of a wherever b is the sole parent of a. The result is an R-tree from $A'' \subseteq A' \cup \{a\}$ to b. Since R is transitive and proper, we have $A''\ R\ b$ for some nonempty $A'' \subseteq\subseteq A$. This contradicts consistency of R with $b \succ A$. □

Let $A = \{a_1, \ldots, a_n\}$. Let $\triangleright^0 := \triangleright$. For $i \in \{1, \ldots, n\}$, let $\triangleright^i = \mathrm{tr}(\triangleright_{i-1} \cup (\{b\}, a_i))$. Since $\triangleright$ is irreflexive, proper, transitive, and consistent with $b \succ A$, we can use Lemmas 3 and 4 to show that the same is true of each $\triangleright^i$. Notice that $\{b\} \triangleright^n a$ for all $a \in A$.

Now we use Lemma 3 to show that $\triangleright^n$ can be extended to an irreflexive, proper and transitive relation $\triangleright^+$ such that, for all distinct $x, y \in \mathcal{A}$, $\{x\} \triangleright^+ y$ or $\{y\} \triangleright^+ x$. The proof is similar to that of the Szpilrajn Extension Theorem. Consider the set of irreflexive, proper and transitive relations that extend $\triangleright^n$, ordered by set inclusion. Take any chain in the partially ordered set. The union of its elements is clearly irreflexive, proper and transitive, so it is an upper bound for the chain. By Zorn's Lemma, the partially ordered set must have a maximal element $\triangleright^+$. Suppose that, for some distinct x, y, neither $\{x\} \triangleright^+ y$ nor $\{x\} \triangleright^+ y$. By Lemma 3, $\triangleright^+$ can be extended to another irreflexive, proper and transitive relation containing $(\{x\}, y)$. Then $\triangleright^+$ cannot be maximal, a contradiction. Moreover, for each $X \in \mathcal{F}(\mathcal{A})$ and $y \in \mathcal{A}$, $\triangleright^+$ must satisfy

$$X \triangleright y \Longrightarrow \left(\exists x \in X \text{ s.t. } \{x\} \triangleright^+ y\right). \tag{7}$$

Suppose not. Then $\{y\} \rhd^+ x$ for all $x \in X$, as well as $X \rhd^+ y$. Since $\rhd^+$ is transitive, $\emptyset \rhd^+ y$. Since $\rhd^+$ is proper, this is a contradiction. Similarly, suppose that $\{x\} \rhd^+ y$ and $\{y\} \rhd^+ x$. By transitivity, $\emptyset \rhd^+ x$, a contradiction.

We can use $\rhd^+$ to define a strict preference $\succ_m$ on $\mathcal{A}$:

$$\{x\} \rhd^+ y \quad \iff \quad x \succ_m y.$$

It is easy to see that $\succ_m$ is ansymmetric and transitive, and ranks every distinct pair of items. Because of (7), it belongs to $\mathcal{M}$. It also satisfies $b \succ_m A$ because $\rhd^+$ extends $\rhd^n$, and $\{b\} \rhd^n a$ for all $a \in A$.

We have now shown that

$$c(A) \subset \max(\succsim, \mathcal{M}(A)).$$

Now suppose that $a \in \max(\succsim, \mathcal{M}(A))$. Since $a \succsim \mathcal{M}(A)$ and $c(A) \subseteq \mathcal{M}(A)$, we have $a \succsim c(A)$. Suppose $a \notin c(A)$.

Lemma 5. *Fix a menu A and item $a \notin A$. If $a \succsim c(A \cup \{a\})$ and $a \notin c(A \cup \{a\})$, then*

$$\{b \in A : a \succsim b\} \rhd a.$$

Proof. Enumerate the items in A from $\succsim$-best to $\succsim$-worst, breaking ties arbitrarily. Suppose $a \succsim a_1$. Then, $A = \{b \in A : a \succsim b\} \rhd a$, and we are done. Now suppose $a_i \succ a$ for $i \leq i^*$, but $a \succsim i$ for $i > i^*$. Since $a_1 \succ a \succsim c(A \cup \{a\})$, we have $a_1 \succ c(A \cup \{a\})$. By Optimization, $a_1 \notin c(A \cup \{a\})$. By IUA, $c(\{a\} \cup A \setminus \{a_1\}) = c(\{a\} \cup A)$. If $i^* = 1$, we are done. Otherwise, we have $a_2 \succ c(\{a\} \cup A) = c(\{a\} \cup A \setminus \{a_1\})$, which implies $a_2 \notin c(\{a\} \cup A \setminus \{a_1\})$. By IUA, $c(\{a\} \cup A \setminus \{a_1, a_2\}) = c(\{a\} \cup A)$. We can iterate this argument until we arrive at

$$c(A) = c(\{a\} \cup \{a_{i^*+1}, \ldots, a_n\}).$$

Since $a \notin c(A)$, we have $a \notin c(\{a\} \cup \{a_{i^*+1}, \ldots, a_n\})$ as well as $a \succsim \{a_{i^*+1}, \ldots, a_n\}$. Since $\{a_{i^*+1}, \ldots, a_n\} = \{b \in A : a \succsim b\}$, we have $\{b \in A : a \succsim b\} \rhd a$ as desired. □

By Lemma 5,

$$\{b \in A \setminus \{a\} : a \succsim b\} \rhd a.$$

By definition of $\mathcal{M}$, we have

$$a \not\succ_m \{b \in A \setminus \{a\} : a \succsim b\}$$

for all $\succ_m \in \mathcal{M}$. This implies $a \not\succ_m A \setminus \{a\}$ for all $\succ_m \in \mathcal{M}$, so $a \notin \mathcal{M}(A)$—contradiction. Conclude that

$$c(A) \subset \max(\succsim, \mathcal{M}(A)),$$

so $\mathcal{M}$ represents c.

A.2 Proof of Theorem 2

Notation:

- $a\ C\ b$ means there exists d such that (a,b,d) or (d,a,b) is a cycle.
- $a\ \bar{C}\ b$ means that a precedes b in a chain.
- $a\ S\ b$ means that a precedes b in a strict chain.
- $a \to b$ means that neither $a\ S\ b$ nor $b\ S\ a$, and $\{a\} = c(a,b)$.
- $a \leftrightarrow b$ means that $\{a,b\} = c(a,b)$.

Sufficiency: For each a, let

$$E(a) := \{b \in \mathcal{A} : a \leftrightarrow x_1 \leftrightarrow \cdots \leftrightarrow x_n \leftrightarrow b \text{ for some } (x_1,\ldots,x_n) \text{ and } n \in \mathbb{N}\}.$$

$\{E(a)\}_{a\in\mathcal{A}}$ partitions the domain into equivalence classes. Notice that if $E(a) \neq E(b)$, then $\{a,b\} \neq c(a,b)$.

Now define a binary relation $\succ_E$ on $\{E(a)\}_{a\in\mathcal{A}}$ as follows: for distinct E_a and E_b, impose $E_a \succ_E E_b$ if

1. $a\ S\ b$ for some $a \in E_a$ and $b \in E_b$, or
2. For all $a \in E_a$ and $b \in E_b$, it is not the case that $a\ S\ b$ or $b\ S\ a$, and for some $a \in E_a$ and $b \in E_b$, $\{a\} = c(a,b)$.

Clearly, $\succ_E$ ranks every distinct E_a, E_b. To show that $\succ_E$ is asymmetric and transitive, we require the following lemmas.

Lemma 6. *If $a \in c(a,b)$, $b \in c(b,d)$ and $d \in c(a,d)$, and if at least one of $c(a,b)$, $c(b,d)$ and $c(a,d)$ is a singleton, then one of the following is a cycle: (a,b,d), (b,d,a), (d,a,b).*

Proof. If $\{a,b,d\} = c(a,b,d)$, then revealed optimization implies that none of $c(a,b)$, $c(b,d)$, or $c(a,d)$ is a singleton. Thus, $\{a,b,d\} \neq c(a,b,d)$. If $a \notin c(a,b,d)$ then (a,b,d) is a cycle if $b \in c(a,b,d)$. Otherwise, $d \in c(a,b,d)$, so (b,d,a) is a cycle. If $b \notin c(a,b,d)$, then (b,d,a) is a cycle if $d \in c(a,b,d)$. Otherwise, $a \in c(a,b,d)$, so (d,a,b) is a cycle. If $d \notin c(a,b,d)$, then (d,a,b) is a cycle if $a \in c(a,b,d)$. Otherwise, $b \in c(a,b,d)$, so (a,b,d) is a cycle. □

Lemma 7. *It is not the case that $x\ S\ x$.*

Proof. If $x\ S\ x$, then we have a sequence such that $x_n\ C\ x_1\ C\ x_2\ \cdots\ x_n$ and $\{x_n\} = c(x_n, x_1)$. Since $x_n\ C\ x_1$, we have $x_n \neq x_1$ by definition of a cycle. For each $i > 1$, we have $x_i\ S\ x_{i-1}$ and $x_{i-1} \in c(x_{i-1}, x_i)$. Thus, it is revealed unjustifiable to choose x_i from $\{x_{i-1}, x_i\}$. By IRUA, $x_i \notin c(x_1, \ldots, x_n)$. Thus, $\{x_1\} = c(x_1, \ldots, x_n)$. But $x_1\ \bar{C}\ x_n$ and $x_1 \notin c(x_1, x_n)$, so it is revealed unjustifiable to choose x_1 from $\{x_1, x_n\}$. By IRUA, $x_1 \notin c(x_1, \ldots, x_n)$—contradiction. □

Lemma 8. *If $x_1\ S\ x_2$, then $E(x_1) \neq E(x_2)$.*

Proof. We begin by proving a series of claims.

Claim 1: If $x_1\ S\ x_2 \leftrightarrow x_3\ \cdots\ x_{n-1} \leftrightarrow x_n \leftrightarrow x_1$ or $x_1\ S\ x_2 \leftrightarrow x_3\ \cdots\ x_{n-1} \leftrightarrow x_n \rightarrow x_1$, then $x_i\ S\ x_j \leftrightarrow x_k \leftrightarrow x_i$ or $x_i\ S\ x_j \leftrightarrow x_k \rightarrow x_i$ for some i, j, k. The proof is by induction on n. If $n = 3$, the result is automatic. Suppose the result holds for all $m < n$ and that $x_1\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-1} \leftrightarrow x_n \rightarrow x_1$. If $x_{n-1} \leftrightarrow x_1$, then the result follows from the inductive hypothesis. Suppose $\{x_1\} = c(x_1, x_{n-1})$. Then, one of the following is a cycle: (x_1, x_{n-1}, x_n), (x_{n-1}, x_n, x_1), (x_n, x_1, x_{n-1}). The first case implies $x_1\ S\ x_n$, and the second and third imply $x_n\ S\ x_1$. Both contradict $x_1 \rightarrow x_n$. Finally, suppose $\{x_{n-1}\} = c(x_1, x_{n-1})$. If $x_{n-1}\ S\ x_1$, then $x_{n-1}\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-1}$, so the result follows from the inductive hypothesis. If $x_1\ S\ x_{n-1}$, then $x_1\ S\ x_{n-1} \leftrightarrow x_n \rightarrow x_1$, which is the desired result. If $x_{n-1} \rightarrow x_1$, then $x_1\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-2} \leftrightarrow x_{n-1} \rightarrow x_1$, so the result follows from the inductive hypothesis. This covers all the cases. Now suppose that the result holds for all $m < n$ and that $x_1\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-1} \leftrightarrow x_n \leftrightarrow x_1$. If $x_{n-1} \leftrightarrow x_1$, the result follows from the inductive hypothesis. Suppose $\{x_1\} = c(x_1, x_{n-1})$. One of the following is a cycle: (x_1, x_{n-1}, x_n), (x_{n-1}, x_n, x_1), (x_n, x_1, x_{n-1}). The first and third cases imply $x_1\ S\ x_n$. Since x_1 comes before x_n in a strict chain and $x_n \in c(x_1, x_n)$, IRUA implies $x_1 \notin c(x_1, x_n)$—contradiction. Thus, $x_{n-1}\ C\ x_n\ C\ x_1$. We have $x_{n-1}\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-1}$, so the result follows from the inductive hypothesis. Now suppose $\{x_{n-1}\} = c(x_1, x_{n-1})$. If $x_1\ S\ x_{n-1}$, then $x_1\ S\ x_{n-1} \leftrightarrow x_n \leftrightarrow x_1$, which is the desired result. If $x_{n-1}\ S\ x_1$, then $x_{n-1}\ S\ x_2 \leftrightarrow x_3 \cdots x_{n-1}$, so the result follows from the inductive hypothesis. If $x_{n-1} \rightarrow x_1$, the result again follows from the inductive hypothesis. This covers all the cases.

Claim 2: If $x_1\ C\ x_2\ \cdots\ x_n \leftrightarrow x_{n+1} \leftrightarrow x_1$ or $x_1\ C\ x_2\ \cdots\ x_n \leftrightarrow x_{n+1} \rightarrow x_1$, and if $\{x_i\} = c(x_i, x_{i+1})$, then $x_i\ C\ x_{i+1}\ \cdots\ x_n \leftrightarrow x_{n+1} \leftrightarrow x_i$ or $x_i\ C\ x_{i+1}\ \cdots\ x_n \leftrightarrow x_{n+1} \rightarrow x_i$. The proof is by induction on i. If $i = 1$, the result is automatic. Suppose the result holds for all $j < i$. If $x_{n+1} \leftrightarrow x_2$, the result follows from the inductive hypothesis. Suppose $\{x_2\} = c(x_2, x_{n+1})$. Then, one of the following is a cycle: (x_2, x_{n+1}, x_1), (x_{n+1}, x_1, x_2), (x_1, x_2, x_{n+1}). In the first and third cases, $x_2\ S\ x_{n+1}$, so $x_1\ S\ x_{n+1}$, which contradicts both $x_{n+1} \rightarrow x_1$ and $x_{n+1} \leftrightarrow x_1$. In the second case, we have $x_{n+1}\ S\ x_n$, which contradicts $x_n \leftrightarrow x_{n+1}$. Suppose $\{x_{n+1}\} = c(x_2, x_{n+1})$. If $x_{n+1}\ S\ x_2$, we have $x_{n+1}\ S\ x_n$, which contradicts $x_n \leftrightarrow x_{n+1}$. If $x_2\ S\ x_{n+1}$, we have $x_1\ S\ x_{n+1}$, which contradicts both $x_{n+1} \rightarrow x_1$ and $x_{n+1} \leftrightarrow x_1$. Thus, $x_{n+1} \rightarrow x_2$, and the result follows from the inductive hypothesis.

Claim 3: If $x_1\ C\ x_2\ \ldots x_n \leftrightarrow x_{n+1} \rightarrow x_1$ or $x_1\ C\ x_2\ \ldots x_n \rightarrow x_{n+1} \rightarrow x_1$, and if $\{x_1\} = c(x_1, x_2)$, then $x_1\ C\ x_2 \leftrightarrow x_{n+1} \rightarrow x_1$ or $x_1\ C\ x_2 \rightarrow x_{n+1} \rightarrow x_1$. Similarly, if $x_1\ C\ x_2\ \ldots x_n \leftrightarrow x_{n+1} \leftrightarrow x_1$ or $x_1\ C\ x_2\ \ldots x_n \rightarrow x_{n+1} \leftrightarrow x_1$, and if $\{x_1\} = c(x_1, x_2)$, then $x_1\ C\ x_2 \leftrightarrow x_{n+1} \leftrightarrow x_1$ or $x_1\ C\ x_2 \rightarrow x_{n+1} \leftrightarrow x_1$. The proof is by induction on n. If $n = 2$, the result is automatic. If $x_{n-1} \leftrightarrow x_{n+1}$, the result follows from the inductive hypothesis. Suppose $\{x_{n+1}\} = c(x_{n-1}, x_{n+1})$. Then, one of the following is a cycle: (x_{n+1}, x_{n-1}, x_n), (x_{n-1}, x_n, x_{n+1}), (x_n, x_{n+1}, x_{n-1}). In the first case, $x_{n+1}\ S\ x_n$, which contradicts both $x_n \rightarrow x_{n+1}$ and $x_n \leftrightarrow x_{n+1}$. In the second case, $x_1\ S\ x_n\ C\ x_{n+1}$, so $x_1\ S\ x_{n+1}$, which contradicts both $x_1 \leftrightarrow x_{n+1}$ and $x_{n+1} \rightarrow x_1$. In the third case, $x_n\ S\ x_{n-1}$, so $x_n\ S\ x_n$. This contradicts Lemma 7. Suppose $\{x_{n-1}\} = c(x_{n-1}, x_{n+1})$. If $x_{n+1}\ S\ x_{n-1}$, then $x_{n+1}\ S\ x_n$, which contradicts both $x_n \rightarrow x_{n+1}$ and $x_n \leftrightarrow x_{n+1}$. If $x_{n-1}\ S\ x_{n+1}$, then $x_1\ S\ x_{n+1}$, which contradicts both $x_1 \leftrightarrow x_{n+1}$ and $x_{n+1} \rightarrow x_1$. We have $x_{n-1} \rightarrow x_{n+1}$, so the result follows from the inductive hypothesis.

Claim 4: If $\{x_1\} = c(x_1, x_2)$, then none of the following can happen: $x_1\ C\ x_2 \leftrightarrow x_3 \leftrightarrow x_1$, $x_1\ C\ x_2 \rightarrow x_3 \leftrightarrow x_1$, $x_1\ C\ x_2 \leftrightarrow x_3 \rightarrow x_1$, $x_1\ C\ x_2 \rightarrow x_3 \rightarrow x_1$. In all of these cases, one of the following is a cycle: (x_1, x_2, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2). The first case implies $x_1\ S\ x_3$, which contradicts both $x_1 \leftrightarrow x_3$ and $x_3 \rightarrow x_1$. The second case implies $x_2\ C\ x_3\ C\ x_1\ S\ x_2$, so $x_2\ S\ x_2$. This contradicts Lemma 7. The third case implies $x_3\ S\ x_2$, which contradicts both $x_2 \leftrightarrow x_3$ and $x_2 \rightarrow x_3$.

To complete the argument, suppose $x_1\ S\ x_2$ and $E(x_1) = E(x_2)$. We have $x_1\ S\ x_2 \leftrightarrow x_3\ \cdots\ x_n \leftrightarrow x_1$. By Claim 1, there exists $(y_1, \ldots, y_{n+1})$ such that $y_1\ C\ y_2\ \cdots\ y_n \leftrightarrow y_{n+1} \leftrightarrow y_1$ or $y_1\ C\ y_2\ \cdots\ y_n \leftrightarrow y_{n+1} \rightarrow y_1$, where $\{y_i\} = c(y_i, y_{i+1})$ for some i. By Claim 2, we have $y_i\ C\ y_{i+1}\ \cdots\ y_n \leftrightarrow y_{n+1} \leftrightarrow y_i$ or $y_i\ C\ y_{i+1}\ \cdots\ y_n \leftrightarrow y_{n+1} \rightarrow y_i$. By Claim 3, we have $y_i\ C\ y_{i+1} \leftrightarrow y_{n+1} \leftrightarrow y_i$ or $y_i\ C\ y_{i+1} \leftrightarrow y_{n+1} \rightarrow y_i$ or $y_i\ C\ y_{i+1} \rightarrow y_{n+1} \leftrightarrow y_i$ or $y_i\ C\ y_{i+1} \rightarrow y_{n+1} \rightarrow y_i$. By Claim 4, all lead to contradictions. □

Lemma 9. *If $E(x_1) = E(x_2)$ and $\{x_2\} = c(x_1, x_2)$, then $x_1\ \bar{C}\ x_2$.*

Proof. Suppose that $y_1 \leftrightarrow y_2 \leftrightarrow \cdots \leftrightarrow y_n$ and that $\{y_n\} = c(y_1, y_n)$. Suppose also that, for all i and all $j > i + 1$, it is not the case that $y_i \leftrightarrow y_j$. (This assumption is without loss, since we can always get to it from $y_1 \leftrightarrow y_2 \leftrightarrow \cdots \leftrightarrow y_n$ by deleting alternatives strictly between y_1 and y_n.) We will show by induction on n that $y_1\ C\ y_2\ C\ \cdots\ C\ y_n$.

Consider $n = 3$. Since $\{y_3\} = c(y_1, y_3)$, one of the following is a cycle: (y_1, y_2, y_3), (y_2, y_3, y_1), (y_3, y_1, y_2). In the first case, $y_1\ C\ y_2\ C\ y_3$ as desired. The second and third cases imply $y_3\ S\ y_1$, which violates Lemma 8.

Now suppose the result holds for $n - 1$. By assumption, it is not the case that $y_1 \leftrightarrow y_{n-1}$. Suppose $\{y_1\} = c(y_1, y_{n-1})$. Then, one of the following is a cycle: (y_1, y_{n-1}, y_n), (y_{n-1}, y_n, y_1), (y_n, y_1, y_{n-1}). In the first case, $y_1\ S\ y_{n-1}$, which violates Lemma 8. In the second and third cases, $y_n\ S\ y_1$, which also violates Lemma 8. Thus, $\{y_{n-1}\} = c(y_1, y_{n-1})$. By the inductive

hypothesis, $y_1\ C\ y_2\ C\ \cdots\ C\ y_{n-1}$. A parallel argument establishes $y_2\ C\ \cdots\ C\ y_n$. Conclude that $y_1\ C\ y_2\ C\ \cdots\ C\ y_n$. □

Lemma 10. *If $x_1 \to x_2\ \bar{C}\ x_n$ or $x_1 \leftrightarrow x_2\ \bar{C}\ x_n$, and if $E(x_1) \neq E(x_i)$ for some i, then $E(x_1) \neq E(x_j)$ for all $j > i$.*

Proof. Suppose $x_1 \leftrightarrow x_2\ C\ \cdots\ C\ x_n$. Let i be the smallest index such that $E(x_i) \neq E(x_1)$. Since $x_{i-1}\ C\ x_i$ and $E(x_{i-1}) = E(x_1) \neq E(x_i)$, we have $\{x_{i-1}\} = c(x_{i-1}, x_i)$. Thus, $x_{i-1}\ S\ x_j$ for all $j > i$. By Lemma 8, we have $E(x_{i-1}) \neq E(x_j)$, which implies $E(x_1) \neq E(x_j)$.

Now suppose $x_1 \to x_2\ C\ \cdots\ C\ x_n$. If $E(x_1) = E(x_2)$, then the argument above goes through unchanged. Now suppose $E(x_2) \neq E(x_1)$. Suppose also that $E(x_1) = E(x_3)$. If $x_3 \in c(x_1, x_3)$, then one of the following is a cycle: (x_3, x_1, x_2), (x_1, x_2, x_3), (x_2, x_3, x_1). The first and second cases imply $x_1\ S\ x_2$, which contradicts $x_1 \to x_2$. Since $E(x_2) \neq E(x_3)$, we have $\{x_2\} = c(x_2, x_3)$, so the third case implies $x_2\ S\ x_1$. This also contradicts $x_1 \to x_2$, so $\{x_1\} = c(x_1, x_3)$. By Lemma 9, $x_3\ \bar{C}\ x_1$. Since $x_2\ S\ x_3$, we have $x_2\ S\ x_1$, which contradicts $x_1 \to x_2$.

Let j be the lowest index above 1 such that $E(x_j) = E(x_1)$, and suppose $j > 3$. Since $E(x_1) \neq E(x_3)$, we cannot have $x_1 \leftrightarrow x_3$. Suppose $\{x_3\} = c(x_1, x_3)$. Then, one of the following is a cycle: (x_1, x_2, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2). In the first and third cases, $x_1\ S\ x_2$, which contradicts $x_1 \to x_2$. In the second case, $x_3\ S\ x_1$, so $x_2\ S\ x_1$, which contradicts $x_1 \to x_2$. Thus, $\{x_1\} = c(x_1, x_3)$. We already ruled out $x_3\ S\ x_1$. If $x_1\ S\ x_3$, then $x_1\ S\ x_j$, which contradicts Lemma 8. Thus, $x_1 \to x_3\ C\ \cdots\ C\ x_n$. We can iterate the argument until we arrive at $x_1 \to x_{j-1}\ C\ x_j$. Since $E(x_1) = E(x_j) \neq E(x_{j-1})$, the argument in the previous paragraph delivers a contradiction. □

Lemma 11. *If $x_1 \to x_2\ C\ \cdots\ C\ x_n$ or $x_1 \leftrightarrow x_2\ C\ \cdots\ C\ x_n$, and if $E(x_1) \neq E(x_i)$ for at least one i, then $x_1 \to x_n$ or $x_1\ S\ x_n$. Similarly, if $x_1\ C\ x_2\ \cdots\ x_{n-1} \to x_n$ or $x_1\ C\ x_2\ \cdots\ x_{n-1} \leftrightarrow x_n$, and if $E(x_1) \neq E(x_i)$ for at least one i, then $x_1 \to x_n$ or $x_1\ S\ x_n$.*

Proof. We prove the first sentence, as the proof of the second sentence is very similar. The proof is by induction on n. The base case is $n = 3$. Suppose $x_1 \to x_2\ C\ x_3$ or $x_1 \leftrightarrow x_2\ C\ x_3$ and that $E(x_1) \neq E(x_i)$ for $i = 2$ or $i = 3$. By Lemma 10, it is not the case that $x_1 \leftrightarrow x_3$. Suppose $\{x_3\} = c(x_1, x_3)$. Then, one of the following is a cycle: (x_1, x_2, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2). In the first and third cases, $x_1\ S\ x_2$, which contradicts both $x_1 \to x_2$ and $x_1 \leftrightarrow x_2$. In the second case, $x_2\ S\ x_1$, which also contradicts $x_1 \to x_2$ and $x_1 \leftrightarrow x_2$. Thus, $\{x_1\} = c(x_1, x_3)$. If $x_3\ S\ x_1$, then $x_2\ S\ x_1$, which contradicts $x_1 \to x_2$ and $x_1 \leftrightarrow x_2$. Thus, we have either $x_1\ S\ x_3$ or $x_1 \to x_3$.

Suppose that the result holds for all $m < n$. Suppose $x_1 \to x_2\ C\ \cdots\ C\ x_n$ or $x_1 \leftrightarrow x_2\ C\ \cdots\ C\ x_n$. If $E(x_1) \neq E(x_i)$ for some $i < n$, then the inductive hypothesis gives $x_1\ S\ x_{n-1}$ or $x_1 \to x_{n-1}$. In the first case, $x_1\ S\ x_n$, which is the desired result. In the second case, we have $x_1 \to x_{n-1}\ C\ x_n$ and $E(x_1) \neq E(x_{n-1})$. The result follows from the inductive hypothesis. Now suppose that $E(x_1) = E(x_2) = \cdots = E(x_{n-1}) \neq E(x_n)$. If $x_1 \leftrightarrow x_{n-1}$, then we have $x_1 \leftrightarrow x_{n-1}\ C\ x_n$,

where $E(x_n) \neq E(x_1)$. The result follows from the inductive hypothesis. If $\{x_1\} = c(x_1, x_{n-1})$, Lemma 8 implies $x_1 \to x_{n-1}$. We have $x_1 \to x_{n-1}\ C\ x_n$, where $E(x_n) \neq E(x_1)$. The result again follows from the inductive hypothesis. If $\{x_{n-1}\} = c(x_1, x_{n-1})$, then $x_1\ \bar{C}\ x_{n-1}$ by Lemma 9. Since $x_{n-1}\ S\ x_n$, we have $x_1\ S\ x_n$, which is the desired result. □

Lemma 12. *If $a \to b$ and $b \notin E(a)$, or if $a\ S\ b$, then for any $\hat{b} \in E(b)$, either $a\ S\ \hat{b}$ or $a \to \hat{b}$. Similarly, for any $\hat{a} \in E(a)$, either $\hat{a}\ S\ b$ or $\hat{a} \to b$.*

Proof. We prove the first sentence, as the proof of the second sentence is very similar. It suffices to show the following: if $a \to b_1 \leftrightarrow b_2\ \cdots\ b_n$ or $a\ S\ b_1 \leftrightarrow b_2\ \cdots\ b_n$, and if $E(a) \neq E(b_1)$, then $a \to b_n$ or $a\ S\ b_n$. The proof is by induction on n. If $n = 1$, the result is automatic. Suppose the result holds for $n-1$ and that $a\ S\ b_1 \leftrightarrow\ \cdots\ b_n$. By the inductive hypothesis, either $a\ S\ b_{n-1} \leftrightarrow b_n$ or $a \to b_{n-1} \leftrightarrow b_n$. Suppose the former, so $z_1\ C\ z_2\ \cdots\ C\ z_m \leftrightarrow b_n$ for some $(z_1, \ldots, z_m)$ such that $z_1 = a$ and $z_m = b_{n-1}$. Suppose that $b_n \leftrightarrow z_j$ for some $j < m$. Let j denote the smallest index i such that $b_n \leftrightarrow z_i$. Since $E(a) \neq E(b_n)$, $1 < j \leq n-1$. We have $z_1\ C\ z_2\ \cdots\ z_j \leftrightarrow b_n$. By definition of j, it is not the case that $z_{j-1} \leftrightarrow b_n$. Suppose that $\{b_n\} = c(z_{j-1}, b_n)$. Then, one of the following is a cycle: (b_n, z_{j-1}, z_j), (z_{j-1}, z_j, b_n), (z_j, b_n, z_{j-1}). In the first and third cases, $b_n\ S\ z_{j-1}$, so $b_n\ S\ z_j$, which contradicts $z_j \leftrightarrow b_n$. In the second case, $z_j\ C\ b_n$. Since $a\ \bar{C}\ z_j$ and $E(z_j) = E(b_n) \neq E(a)$, we have $a\ S\ z_j$, so $a\ S\ b_n$. This is the desired result. Now suppose $\{z_{j-1}\} = c(z_{j-1}, b_n)$. We already ruled out $b_n\ S\ z_{j-1}$. If $z_{j-1}\ S\ b_n$, then $a\ S\ b_n$, which is the desired result. If $z_{j-1} \to b_n$, we have $a\ C\ \cdots\ z_{j-1} \to b_n$. Since $E(a) \neq E(b_n)$, Lemma 11 delivers $a\ S\ b_n$ or $a \to b_n$, which is the desired result.

Now suppose $a \to b_{n-1} \leftrightarrow b_n$. Since $E(a) \neq E(b_n)$, we cannot have $a \leftrightarrow b_n$. Suppose $\{b_n\} = c(a, b_n)$. Then, one of the following is a cycle: (b_n, a, b_{n-1}), (a, b_{n-1}, b_n), (b_{n-1}, b_n, a). In the first case, $b_n\ S\ b_{n-1}$, which contradicts $b_n \leftrightarrow b_{n-1}$. In the second case, $a\ S\ b_{n-1}$, which contradicts $a \to b_{n-1}$. In the third case, $b_{n-1}\ S\ a$, which contradicts $a \to b_{n-1}$. Thus, $\{a\} = c(a, b_n)$. If $b_n\ S\ a$, then $b_{n-1} \to a$ or $b_{n-1}\ S\ a$ by Lemma 11, both of which contradict $a \to b_{n-1}$. The only remaining possibilities are $a\ S\ b_n$ and $a \to b_n$. This is the desired result. □

Lemma 13. *If $E_a \neq E_b$ and neither $a\ S\ b$ nor $b\ S\ a$ for all $a \in E_a$ and $b \in E_b$, then*

$$a \to b \text{ for some } a \in E_a \text{ and } b \in E_b \implies a \to b \text{ for all } a \in E_a \text{ and } b \in E_b.$$

Proof. Suppose that $a \to b$ and $\hat{a} \to \hat{b}$ for some $a, \hat{a} \in E_a$ and $b, \hat{b} \in E_b$. Since $\neg(a\ S\ \hat{b})$ and $\neg(\hat{b}\ S\ a)$, we have $a \to \hat{b}$ and $\hat{b} \to a$ by Lemma 12—contradiction. □

Now we show that $\succ_E$ is asymmetric. Suppose that $E_a \succ_E E_b$ and $E_b \succ_E E_a$, where $E_a \neq E_b$. There are two ways to have $E_a \succ_E E_b$. One is $a\ S\ b$ for some $a \in E_a$ and $b \in E_b$. Then, $E_b \succ_E E_a$ implies that $\hat{b}\ S\ \hat{a}$ for some $\hat{a} \in E_a$ and $\hat{b} \in E_b$. By Lemma 12, either $a\ S\ \hat{b}$ or $a \to \hat{b}$. By Lemma

12 again, either $\hat{a}\ S\ \hat{b}$ or $\hat{a} \to \hat{b}$. The former implies $\hat{a}\ S\ \hat{a}$, which contradicts Lemma 8. The latter contradicts $\hat{b}\ S\ \hat{a}$. Conclude that $\neg(a\ S\ b)$ for all $a \in E_a$ and $b \in E_b$. The other way to have $E_a \succ_E E_b$ is $\neg(a\ S\ b)$ and $\neg(b\ S\ a)$ for all $a \in E_a$ and $b \in E_b$, and $a \to b$ for some $a \in E_a$ and $b \in E_b$. In this case, the only way to have $E_a \succ_E E_b$ is $\hat{b} \to \hat{a}$ for some $\hat{a} \in E_a$ and $\hat{b} \in E_b$. But $\hat{a} \to \hat{b}$ by Lemma 13—contradiction.

Now we show that $\succ_E$ is transitive. Suppose that $E_a \succ_E E_b \succ_E E_d$. If it is not the case that $E_a \succ_E E_d$, then we have $E_a \succ_E E_b \succ_E E_d \succ_E E_a$. Suppose that we have $a\ S\ b$, $\hat{b}\ S\ \hat{d}$, and $\tilde{d}\ S\ \tilde{a}$ for some $a, \tilde{a} \in E_a$, $b, \hat{b} \in E_b$, and $\hat{d}, \tilde{d} \in E_d$. By Lemma 12, $a\ S\ \hat{b}$ or $a \to \hat{b}$. Also, $\hat{b}\ S\ \tilde{d}$ or $\hat{b} \to \tilde{d}$. If $a\ S\ \hat{b}\ S\ \tilde{d}$, then $a\ S\ \tilde{d}$. If $a\ S\ \hat{b} \to \tilde{d}$ or $a \to \hat{b}\ S\ \tilde{d}$, then $a \to \tilde{d}$ or $a\ S\ \tilde{d}$ by Lemma 11. Thus, the three possibilities are $a\ S\ \tilde{d}$, $a \to \tilde{d}$, and $a \to \hat{b} \to \tilde{d}$. By Lemma 12, $\tilde{d}\ S\ a$ or $\tilde{d} \to a$. In both cases, it cannot be that $a\ S\ \tilde{d}$ or $a \to \tilde{d}$, so the only remaining possibility is $a \to \hat{b} \to \tilde{d}$. If $\tilde{d}\ S\ a$, then Lemma 11 implies $\hat{b}\ S\ a$ or $\hat{b} \to a$, both of which contradict $a \to \hat{b}$. If $a \to \tilde{d}$, then one of the following is a cycle: $(a, \hat{b}, \tilde{d})$, $(\hat{b}, \tilde{d}, a)$, $(\tilde{d}, a, \hat{b})$. The first and third cases imply $a\ S\ \hat{b}$, which contradicts $a \to \hat{b}$. The second case implies $\hat{b}\ S\ a$, which also contradicts $a \to \hat{b}$.

Now suppose that, for all $a \in E_a$, $b \in E_b$, and $d \in E_d$, we have $\neg(a\ S\ b)$, $\neg(b\ S\ a)$, $\neg(b\ S\ d)$, $\neg(d\ S\ b)$, $\neg(a\ S\ d)$, and $\neg(d\ S\ a)$. Then, we have $a \to b$, $\hat{b} \to \hat{d}$, and $\tilde{d} \to \tilde{a}$ for some $a, \tilde{a} \in E_a$, $b, \hat{b} \in E_b$, and $\hat{d}, \tilde{d} \in E_d$. By Lemma 13, we have $a \to \hat{b}$, $\hat{b} \to \tilde{d}$, and $\tilde{d} \to a$. We already derived a contradiction from these conditions.

It is now without loss to assume that $a\ S\ b$ for some $a \in E_a$ and $b \in E_b$, that $\neg(b\ S\ d)$ and $\neg(d\ S\ b)$ for all $b \in E_b$ and $d \in E_d$, and that $\hat{b} \to \hat{d}$ for some $\hat{b} \in E_b$ and $\hat{d} \in E_d$. By Lemma 13, we have $b \to \hat{d}$. By Lemma 11, we have $a\ S\ \hat{d}$ or $a \to \hat{d}$. In the former case, $E_a \succ_E E_d$, which contradicts $E_d \succ_E E_a$. Thus, $a \to \hat{d}$. There are two ways to have $E_d \succ_E E_a$. One is $\neg(\tilde{a}\ S\ \tilde{d})$ and $\neg(\tilde{d}\ S\ \tilde{a})$ for all $\tilde{a} \in E_a$ and $\tilde{d} \in E_d$, and $\tilde{d} \to \tilde{a}$ for some $\tilde{a} \in E_a$ and $\tilde{d} \in E_d$. Since $a \to \hat{d}$, this contradicts Lemma 13. The other way to have $E_d \succ_E E_a$ is $\tilde{d}\ S\ \tilde{a}$ for some $\tilde{a} \in E_a$ and some $\tilde{d} \in E_d$. Applying Lemma 12 twice, we get $\hat{d}\ S\ a$ or $\hat{d} \to a$. Both contradict $a \to \hat{d}$. Conclude that $\succ_E$ is transitive.

We can now use $\succ_E$ to define a preference $\succsim$ on $\mathcal{A}$: $a \succsim b$ if $E(a) = E(b)$ or $E(a) \succ E(b)$. It remains to show that c satisfies IUA conditional on $\succsim$ if c satisfies IRUA.

Lemma 14. *If c satisfies IRUA on A but violates WARP on A, then A contains a cycle or almost-consistent set.*

Proof. Suppose that WARP is violated on A. Since WARP is always satisfied on a pair, there exists $B \subseteq A$ such that WARP is violated on B, but not on any proper subset of B. Let P be the relation on B defined by pairwise choice: $a\ P\ b$ if $a \in c(a, b)$. Clearly, P is reflexive and complete. If P is not transitive, then there exist a, b, d in B such that $a \in c(a, b)$, $b \in c(b, d)$, and $\{d\} = c(a, d)$. By Lemma 6, there is a cycle containing a, b and d.

Now suppose that P is transitive, so it is a preference. Consider any nonempty proper subset D

of B. WARP is satisfied on D, so $c(D) = \max(P, D)$, which implies that D is consistent. Suppose that B is not almost-consistent. Then, B is consistent, so $c(B) \supseteq \max(P, B)$. Since WARP is violated on B, the inclusion is strict: there exists $b \in c(B)$ such that $b \notin \max(P, B)$. That is, $\neg(b \; P \; a)$ for some $a \in B$. Since P is a preference, $\max(P, B)$ is nonempty. Fix any $d \in \max(P, B)$, and suppose that $b \; P \; d$. Since $d \; P \; a$, we have $b \; P \; a$—contradiction. Thus, $\{d\} = c(b, d)$. Since $\{b, d\} \in C(B)$, revealed optimization implies $\{b, d\} = c(b, d)$—contradiction. Conclude that B is almost-consistent. □

For IUA, suppose that $a \in A$, $a \gtrsim c(A)$, and $a \notin c(A)$. It suffices to show that it is revealed unjustifiable to choose a from some $B \subset A$. First, suppose $a \notin c(a^*, a)$ for some $a^* \in c(A)$. If $a \sim a^*$, then $E(a) = E(a^*)$, so $a \; \bar{C} \; a^*$ by Lemma 9. Since a precedes a^* in a chain and $a \notin c(a, a^*)$, it is revealed unjustifiable to choose a from $\{a, a^*\}$. If $a \succ a^*$, there are two possibilities. Case 1: for all $b \in E(a)$ and $b^* \in E(a^*)$, it is not the case that $b \; S \; b^*$ or $b^* \; S \; b$, and for some $b \in E(a)$ and $b^* \in E(a^*)$, $\{b\} = c(b, b^*)$. We have $b \to b^*$. By Lemma 12, $a \to b^*$. Applying Lemma 12 again, we have $a \to a^*$, which contradicts $a \notin c(a^*, a)$. Case 2: for some $b \in E(a)$ and $b^* \in E(a^*)$, we have $b \; S \; b^*$. By Lemma 12, $a \; S \; b^*$ or $a \to b^*$. By Lemma 12 again, $a \; S \; a^*$ or $a \to a^*$. The latter is ruled out by $a \notin c(a^*, a)$, so $a \; S \; a^*$. Since a precedes a^* in a strict chain and $a^* \in c(a^*, a)$, it is revealed unjustifiable to choose a from $\{a, a^*\}$.

Now suppose that $a \in c(a^*, a)$ for some $a^* \in c(A)$. This is a WARP violation, so A contains at least one cycle or almost-consistent set by Lemma 14. We show that it is revealed unjustifiable to choose at least one item in A from some subset of A. Suppose that A contains the cycle (x, y, z). Since x precedes z in a chain, it is revealed unjustifiable to choose x from $\{x, z\}$ if $x \notin c(x, z)$. Suppose to the contrary that $\{x, z\} = c(x, z)$. Then, either $\{x\} = c(x, y)$ or $\{y\} = c(y, z)$, so x precedes z in a strict chain. Since $z \in c(x, z)$, it is revealed unjustifiable to choose x from $\{x, z\}$. Now suppose that A contains the almost-consistent set B. Since B is not consistent, there exists $b^* \in B$ such that $b^* \in c(b, b^*)$ for all $b \in B$, but $b^* \notin c(B)$. Thus, it is revealed unjustifiable to choose b^* from B.

Toward a contradiction, suppose it is not revealed unjustifiable to choose a from any subset of A. Fix any $a_1 \in A$ such that it is revealed unjustifiable to choose a_1 from some subset of A, and let $A_1 = A \setminus \{a_1\}$. By IRUA, $c(A_1) = c(A)$, so $a^* \in c(A)$ and $a \notin c(A)$. Thus, A_1 still violates WARP. We can iterate the argument, removing one unjustifiable alternative at each stage, until we arrive at $A_n = c(A) \cup \{a\}$. It is not revealed unjustifiable to choose anything in A_n from any subset of A_n, but A_n violates WARP—contradiction.

Necessity: Necessity of revealed optimization is obvious. For necessity of IRUA, we use the following lemmas.

Lemma 15. *If (a, b, d) is a cycle, then for any justification representation $(\gtrsim, \mathcal{M})$, $a \gtrsim b \gtrsim d$ and $d \succ_m a$ for all $\succ_m \in \mathcal{M}$. In addition, if $\{a\} = c(a, b)$, then $a \succ b$. If $\{b\} = c(b, d)$, then*

$b \succ d$. *Additionally, if* (a, b, d) *is a cycle, then for every justification representation* $(\succsim, \mathcal{M})$*, for every* $\succ_m \in \mathcal{M}$*,* $d \succ_m a$.

Proof. Suppose $b \succ a$ in some representation $(\succsim, \mathcal{M})$. Since $a \in c(a, b)$, we have $a \succ_m b$ for all $\succ_m \in \mathcal{M}$, so $b \notin c(a, b, d)$, which contradicts the definition of a cycle. Now suppose $\{a\} = c(a, b)$ and $b \succsim a$. Again, we have $a \succ_m b$ for all $\succ_m \in \mathcal{M}$, which leads to a contradiction. Now suppose $d \succ b$. Since $b \in c(b, d)$, we have $b \succ_m d$ for all $\succ_m \in \mathcal{M}$. By necessity of IUA, $c(a, b, d) = c(a, b)$. But $a \in c(a, b)$ and $a \notin c(a, b, d)$, so we have a contradiction. Now suppose $\{b\} = c(b, d)$ and $b \succsim d$. Again, we have $b \succ_m d$ for all $\succ_m \in \mathcal{M}$, which leads to a contradiction.

If $\{a\} = c(a, b)$ or $\{b\} = c(b, d)$, we have $a \succ d$. Thus, $d \in c(a, d)$ implies $d \succ_m a$ for all $\succ_m \in \mathcal{M}$. If $\{a, b\} = c(a, b)$ and $\{b, d\} = c(b, d)$, then we have $\{d\} = c(a, d)$ by definition of a cycle. Thus, $d \succ_m a$ for all $\succ_m \in \mathcal{M}$. □

Lemma 16. *If* A *is almost-consistent, then for any representation* $(\succsim, \mathcal{M})$ *and any* $a, b \in A$*,*

$$a \succsim b \iff a \in c(a, b).$$

For any $a^* \in A$ *such that* $a^* \in c(a, a^*)$ *for all* $a \in A$*,* $\neg(a^* \succsim_m A)$ *for all* $\succsim_m \in \mathcal{M}$.

Proof. Suppose $\{a, b\} = c(a, b)$. Then, any representation must have $a \sim b$. Now suppose $\{a\} = c(a, b)$ but $b \succsim a$ in some representation $(\succsim, \mathcal{M})$. Then, $a \succ_m b$ for all $\succ_m \in \mathcal{M}$. By necessity of IUA, $c(A) = c(A \setminus \{b\})$. Since A is almost-consistent, there exists $a^* \in A$ such that $a^* \in c(a^*, \tilde{a})$ for all $\tilde{a} \in A$, but $a^* \notin c(A)$. Additionally, $a^* \in c(B)$ for any $B \subset A$ such that $a^* \in B$. Since $b \notin c(a, b)$, it cannot be that $b = a^*$. But then $a^* \in A \setminus \{b\}$. Since $a^* \notin c(A)$, we have a contradiction to $c(A) = c(A \setminus \{b\}$.

We just showed: if $a^* \in c(a^*, a)$ for all $a \in A$, then $a^* \succsim A$. If $a^* \notin c(A)$, then $\neg(a^* \succsim_m A)$ for all $\succsim_m \in \mathcal{M}$. □

Suppose that a precedes b in a strict chain and $b \in c(a, b)$. By definition of a strict chain and Lemma 15, we have $a \succ b$ in any representation $(\succsim, \mathcal{M})$. Since $b \in c(a, b)$, we have $b \succ_m a$ for all $\succ_m \in \mathcal{M}$. Now suppose that a precedes b in a weak chain and $a \notin c(a, b)$. By Lemma 15, we have $a \succsim b$ in any representation $(\succsim, \mathcal{M})$. Since $a \notin c(a, b)$, we have $b \succ_m a$ for all $\succ_m \in \mathcal{M}$. Since IUA is necessary, $c(D) = c(D \setminus \{a\})$ for any $D \supset \{a, b\}$.

Now suppose that $B \ni a$ is almost-consistent, $a \notin c(B)$, and $a \in c(a, b)$ for all $b \in c(B)$. By Lemma 16, $a \succsim B$ in any representation $(\succsim, \mathcal{M})$. Since $a \notin c(B)$, we have $\neg(a \succsim_m B)$ for all $\succ_m \in \mathcal{M}$. Since IUA is necessary, $c(D) = c(D \setminus \{a\})$ for any $D \supseteq B$.

A.3 Proof of Corollary 1

We show that the procedure for constructing $\succsim$ in the proof of Theorem 2 coincides with the procedure given in Corollary 1. Suppose that $a \sim b$. We have $E(a) = E(b)$. By Lemma 8, neither item can precede the other in a strict chain. If $\{a, b\} = c(a, b)$, we have $a \sim_{\text{foc}} b$. Suppose that $\{a\} = c(a, b)$. We have $a \succsim_{\text{foc}} b$ by (2). By Lemma 9, b precedes a in a chain. Thus, $b \succsim_{\text{foc}} a$ by (1). Conclude that $a \sim_{\text{foc}} b$.

Now suppose that $a \succ b$. We have $E(a) \neq E(b)$. There are two ways to get $a \succ b$. One is $\tilde{a}\ S\ \tilde{b}$ for some $\tilde{a} \in E(a)$ and $\tilde{b} \in E(b)$. By Lemma 12, we have either $a\ S\ b$ or $a \to b$. If $a\ S\ b$, then $a \succsim_{\text{foc}} b$. By Lemma 8, it cannot be that $b\ S\ a$. Thus, $a \succ_{\text{foc}} b$. If $a \to b$, then $a \succsim_{\text{foc}} b$. Since $E(a) \neq E(b)$, if b precedes a in a chain, then $b\ S\ a$ by Lemma 8. This contradicts $a \to b$, so b does not precede a in a chain. Since $b \notin c(a, b)$, we have $a \succ_{\text{foc}} b$. The other way to get $a \succ b$ is $\neg(\tilde{a}\ S\ \tilde{b})$ and $\neg(\tilde{b}\ S\ \tilde{a})$ for all $\tilde{a} \in E_a$ and $\tilde{b} \in E_b$, and $\tilde{a} \to \tilde{b}$ for some $\tilde{a} \in E_a$ and $\tilde{b} \in E_b$. By Lemma 13, we have $a \to b$. Thus, $a \succsim_{\text{foc}} b$. We already saw that b cannot precede a in a chain. Since $b \notin c(a, b)$, we have $a \succ_{\text{foc}} b$.

We have shown that $a \sim b$ implies $a \sim_{\text{foc}} b$, and that $a \succ b$ implies $a \succ_{\text{foc}} b$. Suppose $a \sim_{\text{foc}} b$. Since it cannot be that $a \succ b$ or $b \succ a$, we have $a \sim b$. Now suppose $a \succ_{\text{foc}} b$. It cannot be that $a \sim b$ or that $b \succ a$, so $a \succ b$. Conclude that $\succsim = \succsim_{\text{foc}}$.

Now we show that $\succsim$ satisfies (1). Suppose that $c(A)\ \hat{\succsim}\ a$ for some possible $\hat{\succsim}$, but $a \succ c(A)$. Let E_c denote the equivalence class containing the items in $c(A)$. Since $a \succ c(A)$, it cannot be that $E(a) = E_c$. There are two ways to get $a \succ c(A)$. One is $\tilde{a}\ S\ \tilde{c}$ for some $\tilde{a} \in E(a)$ and $\tilde{c} \in E_c$. By Lemma 15, $\tilde{a}\ \hat{\succ}\ \tilde{c}$. Since items in the same equivalence class must be indifferent in any representation, we have $a\ \hat{\succ}\ c(A)$—contradiction. The other way to get $a \succ c(A)$ is to have $\neg(\tilde{a}\ S\ \tilde{c})$ and $\neg(\tilde{c}\ S\ \tilde{a})$ for all $\tilde{a} \in E(a)$ and $\tilde{c} \in E_c$, and $\tilde{a} \to \tilde{c}$ for some $\tilde{a} \in E(a)$ and $\tilde{c} \in E_c$. By Lemma 13, we have $a \to a^*$ for all $a^* \in c(A)$. Let $\hat{\mathcal{M}}$ be any set of justifications such that $(\hat{\succsim}, \hat{\mathcal{M}})$ represents c. Since $a^*\ \hat{\succ}\ a$ but $a = c(a, a^*)$ for all $a^* \in c(A)$, we have $a \succ_m a^*$ for all $a^* \in c(A^*)$ an all $\succ_m \in \hat{\mathcal{M}}$. Since $a \notin c(A^*)$ but $a = c(a, a^*)$ for all $a^* \in c(A)$, an argument from the proof of Theorem 2 implies that it is revealed unjustifiable to choose a from some $B \subseteq A$. Thus, for each $\succ_m \in \hat{\mathcal{M}}$, there exists $b \in B$ such that $b \succ_m a$. Conclude that, for each $a^* \in c(A)$ and $\succ_m \in \hat{\mathcal{M}}$, there exists $b \in B$ such that $b \succ_m a^*$. Since $B \subseteq A$, this contradicts $a^* \in c(A)$. Conclude that $c(A)\ \hat{\succsim}\ a$ implies $c(A) \succsim a$.

Now suppose that $c(A)\ \hat{\succ}\ a$ for some possible $\hat{\succsim}$, but $a \succsim c(A)$. If $a \sim c(A)$, then a and $c(A)$ are in the same equivalence class, so $c(A)\ \hat{\sim}\ a$—contradiction. Thus, $a \succ c(A)$. The rest of the argument goes through as above.

A.4 Proof of Corollary 3

For any representation, the backward direction of (2) follows directly from the necessity part of the proof of Theorem 2. We show that, for some representation $(\succsim_{\text{foc}}, \mathcal{M}_{\text{foc}})$,

$$a \not\succsim_m A \text{ for all } \succ_m \in \mathcal{M}_{\text{foc}} \quad \Longrightarrow \quad \text{it is revealed unjustifiable to choose } a \text{ from some } B \subseteq A.$$

In the proof of Theorem 2, we first constructed $\succsim_{\text{foc}}$ such that c satisfies IUA conditional on $\succsim_{\text{foc}}$. We then appealed to the proof of Theorem 1, which implies that c has a representation $(\succsim_{\text{foc}}, \mathcal{M}_{\text{foc}})$ with the following property. For any item a and $A \ni a$,

$$\text{there is no } B \subseteq A \text{ s.t. } a \in B, a \succsim_{\text{foc}} B \text{ and } a \notin c(B) \quad \Longrightarrow \quad a \succsim_m A \text{ for some } \succ_m \in \mathcal{M}_{\text{foc}}. \tag{8}$$

Suppose that $a \not\succsim_m A$ for all $\succ_m \in \mathcal{M}_{\text{foc}}$. By (8), there exists $B \subseteq A$ such that $a \in B$, $a \succsim_{\text{foc}} B$, and $a \notin c(B)$. We showed in the proof of Theorem 2 that $a \in B$, $a \succsim c(B)$, and $a \notin c(B)$ together imply that there exists $D \subseteq B$ such that it is revealed unjustifiable to choose a from D. Thus, it is revealed unjustifiable to choose a from some subset of A.

A.5 Proof of Corollary 2

We first prove a stronger result, which is useful for more general incomplete data.

Start with a choice correspondence c_{inc} defined on $\mathcal{D} \subseteq \mathcal{F}(\mathcal{A})$, where

$$A \in \mathcal{D} \quad \Longrightarrow B \in \mathcal{D} \text{ for any nonempty } B \subseteq A.$$

For a (possibly empty) menu A and item $b \notin A$, write $A \; R \; b$ if it is revealed unjustifiable, given c_{inc}, to choose b from a subset of $A \cup \{b\}$. Let $\text{tr}(R)$ denote the transitive closure of R, where we use the notion of transitivity from Definition 20.

Axiom 5 (Incomplete IRUA)**.** *If A $tr(R)$ a and if c_{inc} is defined on $A \cup \{a\}$, then $c_{inc}(A) = c_{inc}(A \cup \{a\})$.*

Proposition 5. *c_{inc} satisfies revealed optimization and incomplete IRUA if and only if there is an extension of c_{inc} that has a justification representation.*

Proof. **Necessity:** Suppose c_{inc} has an extension c that has a justification representation. By Corollary 1, c has a representation with preference $\succsim_{\text{foc}}$.

For a (possibly empty) menu A and item $b \notin A$, we write $A \; R_c \; b$ if it is revealed unjustifiable, given c, to choose b from a subset of $A \cup \{b\}$. Since c extends R, we have $R_c \supseteq R$. We show that R_c is transitive. Suppose that $X \; R_c \; y \in Y \; R_c \; z$. We want to show that $X \cup Y \setminus \{y\} \; R_c \; z$, i.e that it is revealed unjustifiable, given c, to choose z from a subset of $\{z\} \cup X \cup Y \setminus \{y\}$. It suffices to show

that $z \succsim_{\text{foc}} Z$ and $z \notin c(\{z\} \cup Z)$ for some $Z \subseteq X \cup Y \setminus \{y\}$. (We showed in the proof of Theorem 2 that $z \succsim_{\text{foc}} Z$ and $z \notin c(\{z\} \cup Z)$ implies that it is revealed unjustifiable to choose z from a subset of $\{z\} \cup Z$.)

Since $X \; R_c \; y$, it is revealed unjustifiable, given c, to choose y from $\underline{X} \cup \{y\}$, where $\underline{X} \subseteq X$. We showed in the necessity part of the proof of Theorem 2 that $y \succsim_{\text{foc}} \underline{X}$. Similarly, since $Y \; R_c \; z$, it is revealed unjustifiable, given c, to choose z from $\underline{Y} \cup \{z\}$, where $\underline{Y} \subseteq Y$ and $z \succsim_{\text{foc}} \underline{Y}$. If $y \notin \underline{Y}$, then $\underline{Y} \subseteq \underline{X} \cup \underline{Y} \setminus \{y\}$, so we are done. Suppose $y \in \underline{Y}$. Since $y \succsim_{\text{foc}} \underline{X}$, we have $z \succsim_{\text{foc}} \underline{X} \cup \underline{Y} \setminus \{y\}$. Since c must satisfy IRUA, we also have

$$c(\{z\} \cup \underline{X} \cup \underline{Y} \setminus \{y\}) = c(\{z\} \cup \underline{X} \cup \underline{Y}) = c(\underline{X} \cup \underline{Y} \setminus \{z\}).$$

Since z cannot belong to the right-hand side, we have $z \notin c(\{z\} \cup \underline{X} \cup \underline{Y} \setminus \{y\})$. Conclude that R_c is transitive.

Fix any menu A and item a such that $A \; \text{tr}(R) \; a$. Since $R \supset R_c$ and R is transitive, $A \; R \; a$. Equivalently, it is revealed unjustifiable, given c, to choose a from a subset of $\{a\} \cup A$. Since c satisfies IRUA, we have $c(A) = c(A \cup \{a\})$. Since c extends c_{inc}, we have $c_{\text{inc}}(A) = c_{\text{inc}}(A \cup \{a\})$ if c_{inc} is defined on $A \cup \{a\}$.

Sufficiency: For any menu A, let

$$c(A) := \max(\succsim_{\text{foc}}, \{a \in A : \neg(A \setminus \{a\} \; \text{tr}(R) \; a)\}).$$

We show that $c(A) = c_{\text{inc}}(A)$ if $c_{\text{inc}}(A)$ is defined. Fix any $a \in c_{\text{inc}}(A)$. Since c_{inc} satisfies incomplete IRUA, we have $\neg(A \setminus \{a\} \; \text{tr}(R) \; a)$. It remains to show that $a \succsim_{\text{foc}} b$ for any $b \in A$ such that $\neg(A \setminus \{b\} \; \text{tr}(R) \; b)$. Fix any $b \in A$ such that $b \succ_{\text{foc}} a$. We will show that $A \setminus \{b\} \; R \; b$. First, suppose $b \in c_{\text{inc}}(A)$. Then, revealed ptimization implies $\{a, b\} = c_{\text{inc}}(a, b)$, which contradicts $b \succ_{\text{foc}} a$. Conclude that $b \notin c_{\text{inc}}(A)$. Take any $\tilde{a} \in c_{\text{inc}}(A)$. By revealed optimization, $c_{\text{inc}}(a, \tilde{a}) = \{a, \tilde{a}\}$, so $a \sim_{\text{foc}} \tilde{a}$. We have $b \succ_{\text{foc}} c_{\text{inc}}(A)$ as well as $b \notin c(A)$. Since c_{inc} is defined on all subsets of A, we can use the argument in the proof of Theorem 2 to show that it is revealed unjustifiable to choose b from a subset of A. We have $A \setminus \{b\} \; R \; b$ as desired.

Now suppose that $\neg(A \setminus \{a\} \; \text{tr}(R) \; a)\})$ and $a \succsim_{\text{foc}} \{b \in A : \neg(A \setminus \{b\} \; \text{tr}(R) \; a)\})$. We show that $a \in c_{\text{inc}}(A)$. Fix any $a_1 \in A$ such that $A \setminus \{a_1\} \; \text{tr}(R) \; a_1$, and let $A_1 = A \setminus \{a_1\}$. Since c_{inc} satisfies incomplete IRUA, we have $c_{\text{inc}}(A) = c_{\text{inc}}(A_1)$. Now fix any $a_2 \in A$ distinct from a_1 such that $A \setminus \{a_2\} \; \text{tr}(R) \; a_2$. Since $A \setminus \{a_1\} \; \text{tr}(R) \; a_1$, we have $A_1 \setminus \{a_2\} \; \text{tr}(R) \; a_2$. Let $A_2 = A_1 \setminus \{a_2\}$. Since c_{inc} satisfies incomplete IRUA, we have $c_{\text{inc}}(A_1) = c_{\text{inc}}(A_2)$. We can iterate this process until we arrive at

$$c_{\text{inc}}(A) = c_{\text{inc}}\left(\{b \in A : \neg(A \setminus \{b\} \; \text{tr}(R) \; a\}\right).$$

Suppose $a \notin c_{\text{inc}}(A)$. Since $a \succsim_{\text{foc}} \{b \in A : \neg(A \setminus \{b\} \; \text{tr}(R) \; a\}$ and $a \notin c_{\text{inc}}\left(\{b \in A : \neg(A \setminus \{b\} \; \text{tr}(R) \; a\}\right)$,

we have $A \setminus \{a\}\ R\ a$ as above—contradiction. Conclude that $a \in c_{\text{inc}}(A)$.

The equality $c(A) = c_{\text{inc}}(A)$ implies that c is well defined wherever c_{inc} is defined. Take any menu A on which c_{inc} is not defined, and suppose that $\{a \in A : \neg(A \setminus \{a\}\ \text{tr}(R)\ a)\})$ is empty. Let $A := \{a_1, \ldots, a_n\}$. As above, we have $A \setminus \{a_1, \ldots, a_i\}\ \text{tr}(R)\ \{a_i\}$. In particular, $\emptyset\ \text{tr}(R)\ a_n$. Since c_{inc} satisfies incomplete IRUA, we have $c_{\text{inc}}(a_n) = c_{\text{inc}}(\emptyset)$. Since c_{inc} is not even defined on $\emptyset$, this is a contradiction.

Finally, we show that c satisfies IUA conditional on $\succsim_{\text{foc}}$. Suppose that $a \in A$, $a \succsim_{\text{foc}} c(A)$ and $a \notin c(A)$. By definition of c, we have $A \setminus \{a\}\ \text{tr}(R)\ a$. Now take any $B \supseteq A$. We want to show that $c(B) = c(B \setminus \{a\})$. It suffices to show that $B \setminus \{b\}\ \text{tr}(R)\ b$ implies $B \setminus \{a, b\}\ \text{tr}(R)\ b$. Suppose $B \setminus \{b\}\ \text{tr}(R)\ b$. Since $A \setminus \{a\}\ \text{tr}(R)\ a$, and $A \subseteq B$, we have $B \setminus \{a, b\}\ \text{tr}(R)\ b$ by definition of transitivity. □

To prove the result in the text, suppose that c_{inc} is defined only on pairs and triples for which pairwise choice is cyclic, and satisfies IRUA on its domain. We show that c_{inc} satisfies incomplete IRUA. It suffices to show that $R = \text{tr}(R)$ in this case. Notice that c_{inc} is not defined on any almost-consistent sets. Thus, $B\ R\ a$ implies $B = \{b\}$ where a precedes b in a chain. Suppose that $B\ \text{tr}(R)\ a$. Then, there exists a finite sequence $(x_1, \ldots, x_n)$ such that $x_1 = b$, $x_n = a$, and $\{x_i\}\ R\ x_{i+1}$ for all $i < n$. Since x_i precedes x_{i+1} in a chain for each i, a precedes b in a chain. Since c_{inc} satisfies IRUA on its domain, and since $\{b\}\ \text{tr}(R)\ a$, we have $\{b\} = c(b, a)$. Thus, it is revealed unjustifiable to choose a from $\{a, b\}$. Equivalently, $B\ R\ a$. Conclude that $R = \text{tr}(R)$.

A.6 Notation and lemmas for random model

For any menu X, any lottery x, and any $\mathcal{M} \in \text{supp}(P_{\text{just}})$,

$$\begin{aligned} W_{\mathcal{M}}(X) &:= \{y \in \Delta(Z) : X \succ_{\mathcal{M}} y\} \\ B_{\mathcal{M}}(x) &:= \{y \in \Delta(Z) : y \succ_{\mathcal{M}} x\} \\ W_P(X) &:= \{y \in \Delta(Z) : P(y \in W_{\mathcal{M}}(X)) > 0\} \\ B_P(x) &:= \{y \in \Delta(Z) : P(y \in B_{\mathcal{M}}(x)) > 0\}. \end{aligned}$$

We often write $W(X)$ instead of $W_{\mathcal{M}}(X)$ when the identity of $\mathcal{M}$ is not important.

The following lemmas are frequently needed.

Lemma 17. *For any menu B, lottery a, and $\mathcal{M} \in supp(P_{just})$: if $B \succ_{\mathcal{M}} a$, then there is some $b \in co(B)$ such that $b \succ_{\mathcal{M}} a$.*

Proof. Fix a compact, convex set of utility functions, $\mathcal{V}$, that represents precisely the preferences in $\mathcal{M}$. Normalize each $v \in \mathcal{V}$ so that $v(a) = 0$. (This will not affect the properties of $\mathcal{V}$.) Write $B = \{b_1, \ldots, b_n\}$. For each $v \in \mathcal{V}$, let v_B be the vector in $\mathbb{R}^n$ that has $v(b_i)$ as its ith element. Let

$\mathcal{V}_B := \{v_B : v \in \mathcal{V}\}$. Like $\mathcal{V}$, $\mathcal{V}_B$ is compact and convex. Let $N := \mathbb{R}^n_-$, which is closed and convex. Since $v(a) = 0 < \max_{b \in B} v(b)$ for all $v \in \mathcal{V}$, we have $N \cap \mathcal{V}_B = \emptyset$. By the separating hyperplane theorem, there is a nonzero $\alpha \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that $\alpha' v_B > c > \alpha' n$ for all $v_B \in \mathcal{V}_B$ and all $n \in N$. Since the zero vector is in N, we have $c > 0$. Suppose the ith element of α is strictly negative. By choosing n with a sufficiently negative number in ith position and zeros elsewhere, we obtain $\alpha' n > c$, a contradiction. Thus, each element of α is weakly positive. If we rescale α to a unit sum, we still have $\alpha' v_B > 0$ for all $v_B \in \mathcal{V}_B$. This can be rewritten $v(b) > v(a)$ for all $v \in \mathcal{V}$, where $b := \sum_{i=1}^n \alpha_i v(b_i)$. □

Lemma 18. *If $y \in W_P(X)$, then there exists $x \in co(X)$ such that $y \in W_P(x)$.*

Proof. If $y \in W_P(X)$, then there exists $\mathcal{M} \in \text{supp}(P_{\text{just}})$ such that $y \in W_{\mathcal{M}}(X)$. By Lemma 17, there exists $x \in \text{co}(X)$ such that $y \in W_{\mathcal{M}}(x)$. This implies $y \in W_{\tilde{\mathcal{M}}}(x)$ for all $\tilde{\mathcal{M}}$ in an open neighborhood of $\mathcal{M}$, so $y \in W_P(x)$. □

Lemma 19. *For any menus X_1, X_2, lottery $y \in X_1 \cap X_2$, and $\mathcal{M} \in supp(P_{just})$:*

$$B_\epsilon(y) \cap co(X_1) = B_\epsilon(y) \cap co(X_2)$$

for some $\epsilon > 0$, then

$$y \in W_{\mathcal{M}}(X_1) \iff y \in W_{\mathcal{M}}(X_2).$$

Proof. Suppose $y \in W_{\mathcal{M}}(X_1)$. By Lemma 17, there exists $x_1 \in \text{co}(X_1)$ such that $y \in W_{\mathcal{M}}(x_1)$. For each $\alpha \in (0,1)$, let $x_\alpha := \alpha x_1 + (1-\alpha)y$. Since $y \in X_1$, $x_\alpha \in \text{co}(X_1)$. For $\alpha > 0$ sufficiently small, $x_\alpha \in B_\epsilon(y) \cap \text{co}(X_1)$. Thus, $\alpha x_1 + (1-\alpha)y \in \text{co}(X_2)$. Since there exists $x_2 \in \text{co}(X_2)$ such that $y \in W_{\mathcal{M}}(x_2)$, we have $y \in W_{\mathcal{M}}(X_2)$. The other direction is symmetric. □

Lemma 20. *If P satisfies Limited Disagreement, then there exists $\succsim \in \mathcal{U}$ such that $cl(W(y)) \setminus \{y\} \succ y$ for all interior y.*

Proof. It suffices to show that $y \notin \text{co}(\text{cl}(W_P(y))) \setminus \{y\}$ for any interior y. Suppose that there exists a finite menu X such that $y \in \text{co}(X)$, $y \notin X$, and $x \in \text{cl}(W_P(y)) \setminus \{y\}$ for each $x \in X$.

By limited disagreement, $y \notin \text{cl}(W_P(X))$. Since $y \in \text{cl}(W_P(y))$, there exists a sequence $y_n \to y$ such that $y_n \in W_P(y)$ for each n. That is, $P(y \succ_{\mathcal{M}} y_n) > 0$ for all n. Since $y \in \text{co}(X)$, we have $P(X \succ_{\mathcal{M}} y_n) > 0$ for all n. That is, $y_n \in W_P(X)$ for all n, so $y \in \text{cl}(W_P(X))$—contradiction. □

A.7 Proof of Proposition 1

If ρ has a random justification representation P, then

$$\begin{aligned}\rho(d|a,b,d) =& P(d \succ a,b \text{ and } d \notin W(a,b))\\ & P(a \succ d \succ b \text{ and } a \in W(b,d) \text{ and } d \notin W(b))\\ & P(b \succ d \succ a \text{ and } b \in W(a,d) \text{ and } d \notin W(a))\\ & P(a,b \succ d \text{ and } a,b \in W(d)).\end{aligned}$$

For each $\epsilon \in (0,1)$, let $b_\epsilon := (1-\epsilon)b + \epsilon d$. Since

$$\begin{aligned} d \succ a,b \text{ and } d \notin W(a,b) \quad &\iff \quad d \succ a,b_\epsilon \text{ and } d \notin W(a,b_\epsilon)\\ a,b \succ d \text{ and } a,b \in W(d) \quad &\iff \quad a,b_\epsilon \succ d \text{ and } a,b_\epsilon \in W(d),\end{aligned}$$

the first two terms do not change when b is replaced with b_ϵ. Also,

$$\begin{aligned} a \succ d \succ b_\epsilon \text{ and } a \in W(b_\epsilon,d) \text{ and } d \notin W(b_\epsilon) \quad &\iff \quad a \succ d \succ b \text{ and } a \in W(b_\epsilon,d) \text{ and } d \notin W(b)\\ &\implies \quad a \succ d \succ b \text{ and } a \in W(b,d) \text{ and } d \notin W(b)\\ b \succ d \succ a \text{ and } b \in W(a,d) \text{ and } d \notin W(a) \quad &\iff \quad b_\epsilon \succ d \succ a \text{ and } b \in W(a,d) \text{ and } d \notin W(a)\\ &\implies \quad b_\epsilon \succ d \succ a \text{ and } b_\epsilon \in W(a,d) \text{ and } d \notin W(a).\end{aligned}$$

This implies

$$\begin{aligned}&\rho(d|a,b,d) - \rho(d|a,b_\epsilon,d)\\ &\qquad = P(a \succ d \succ b \text{ and } a \in W(b,d) \text{ and } a \notin W(b_\epsilon,d) \text{ and } d \notin W(b))\\ &\qquad\qquad - P(b \succ d \succ a \text{ and } b \notin W(a,d) \text{ and } b_\epsilon \in W(a,d) \text{ and } d \notin W(a)) \quad (9)\end{aligned}$$

If $\rho(d|a,b,d) > \rho(d|a,b_\epsilon,d)$ for all sufficiently small $\epsilon > 0$, then the first term above must be strictly positive, which implies

$$P(a \in W(b,d) \text{ and } a \notin W(b_\epsilon,d)) > 0.$$

By Lemma 18, this implies that for all $\epsilon > 0$, there exists $\delta \in (0,\epsilon)$ such that $P(a \in W(b_\delta)) > 0$. Since

$$a \in W(b_\delta) \quad \iff \quad a + (b - b_\delta) \in W(b),$$

we can find $\tilde{a}$ arbitrarily close to a such that $P(\tilde{a} \in W(b)) > 0$. Conclude that $a \in \text{cl}(W_P(b))$.

A.8 Proof of Theorem 3

Suppose that $a \in W_P(b)$. We show that (a, b, d) is anomalous for some d. Fix λ such that $b+\lambda(b-a)$ is interior. There exists $\epsilon_1 > 0$ such that $\tilde{b}+\lambda(\tilde{b}-a) \in \Delta(Z)$ for all $\tilde{b} \in B_{\epsilon_1}(b)$. Since $b+\lambda(b-a) \in B_P(b)$ and $B_P(b)$ is open, there exists $\epsilon_2 \in (0, \epsilon_1]$ such that $\tilde{b}+\lambda(\tilde{b}-a) \in B_P(b)$ for all $\tilde{b} \in B_{\epsilon_2}(b)$.

We will show that there exists $\tilde{b} \in B_{\epsilon_2}(b)$ such that

$$P(b \in B(a) \text{ and } \alpha\tilde{b} + (1-\alpha)b \notin B(a)) > 0 \tag{10}$$

for all $\alpha \in (0,1)$. Suppose that, for each $\tilde{b} \in B_{\epsilon_2}(b)$, there exists $\alpha \in (0,1)$ such that

$$b \in B_{\mathcal{M}}(a) \quad \Longrightarrow \quad \alpha\tilde{b} + (1-\alpha)b \in B_{\mathcal{M}}(a)$$

for all $\mathcal{M} \in \text{supp}(P_{\text{just}})$. Since each $B_{\mathcal{M}}(a)$ is convex, conclude that there exists $\epsilon_3 \in [0, \epsilon_2)$ such that

$$b \in B_{\mathcal{M}}(a) \quad \Longrightarrow \quad B_{\epsilon_3}(b) \subset B_{\mathcal{M}}(a). \tag{11}$$

There exists $\mathcal{M}^* \in \text{supp}(P_{\text{just}})$ such that $B_{\epsilon_3}(b) \subset B_{\mathcal{M}^*}(a)$ but $B_{\epsilon_3}(b) \nsubseteq B_{\mathcal{N}}(a)$ for all $\mathcal{N} \in \text{supp}(P_{\text{just}})$ such that $\mathcal{N} \supset \mathcal{M}^*$. (Existence of $\mathcal{M}^*$ is by Zorn's Lemma.) By the third part of Regularity and $b \in B_{\mathcal{M}^*}(a)$, there exists $\mathcal{N} \in \text{supp}(P_{\text{just}})$ such that $b \in B_{\mathcal{N}}(a)$ and $\mathcal{N} \supset \mathcal{M}^*$. By definition of $\mathcal{M}^*$, it cannot be that $B_{\epsilon_3}(b) \subset B_{\mathcal{N}}(a)$. This contradicts (11). Conclude that, for some $b^* \in B_{\epsilon_2}(b)$, for all $\alpha \in (0,1)$, there exists $\mathcal{M}^\alpha \in \text{supp}(P_{\text{just}})$ such that

$$b \in B_{\mathcal{M}^\alpha}(a) \text{ and } \alpha b^* + (1-\alpha)b \notin B_{\mathcal{M}^\alpha}(a).$$

We can always choose $\mathcal{M}^\alpha$ so that $\alpha b^* + (1-\alpha)b$ is not on the boundary of $B_{\mathcal{M}^\alpha}(a)$. Then,

$$b \in B_{\mathcal{M}}(a) \text{ and } \alpha b^* + (1-\alpha)b \notin B_{\mathcal{M}}(a)$$

for all $\mathcal{M}$ in an open neighborhood of $\mathcal{M}^\alpha$. This implies (10).

Let $d := b^* + \lambda(b^* - a)$. Since

$$\alpha d + (1-\alpha)b \notin B(a) \quad \Longleftrightarrow \quad \frac{\alpha(1+\lambda)}{1+\alpha\lambda}b^* + \frac{1-\alpha}{1+\alpha\lambda}b \notin B(a),$$

we have

$$P(b \in B(a) \text{ and } \alpha d + (1-\alpha)b \notin B(a)) > 0$$

for all $\alpha \in (0,1)$. Since $d \in B_{\epsilon_2}(b)$, we have $d \in B_P(b)$. By Limited Disagreement, $d \notin W_P(b)$. Thus,

$$P(a \in W(b) \text{ and } a \notin W(\alpha d + (1-\alpha)b) \text{ and } d \notin W(b)) > 0.$$

By the second part of Regularity,

$$P(a \succ d \succ b \text{ and } a \in W(b) \text{ and } a \notin W(\alpha d + (1-\alpha)b) \text{ and } d \notin W(b)) > 0.$$

This is the first term in (9). To complete the argument that (a, b, d) is anomalous, we need to show that the second term is zero. It suffices to show that, for all $\alpha > 0$ sufficiently small,

$$P(\alpha d + (1-\alpha)b \in W(a, d) \text{ and } b \notin W(a, d)) = 0.$$

Suppose not. Then, there exist arbitrarily small $\alpha > 0$ such that

$$P\left(b \in W\left(\frac{a - \alpha d}{1-\alpha}, d\right) \text{ and } b \notin W(a, d)\right) > 0.$$

By Lemma 18, this implies $b \in W_P(\tilde{a})$ for $\tilde{a}$ arbitrarily close to a. This is equivalent to $b + \tilde{a} - a \in W_P(a)$, so $b \in \text{cl}(W_P(a))$. Since $a \in W_P(b)$, this is a violation of limited disagreement. Conclude that (a, b, d) is anomalous.

We have shown that (a, b, d) is anomalous for some d whenever $a \in W_P(b)$. Since $W_P(b)$ is open, $a \in W_P(b)$ implies $\tilde{a} \in W_P(b)$ for all $\tilde{a}$ sufficiently close to a. For each such $\tilde{a}$, there exists $\tilde{d}$ such that $(\tilde{a}, b, \tilde{d})$ is anomalous. Conclude that

$$W_P(b) \subseteq \text{int}\left(\{a \in \Delta(Z) : (a, b, d) \text{ is anomalous for some } d\}\right).$$

Now we show the reverse inclusion. Suppose that, for all $\tilde{a}$ sufficiently close to a, there exists $\tilde{d}$ such that $(\tilde{a}, b, \tilde{d})$ is anomalous. By Proposition 1, we have $a \in \text{cl}(W_P(b))$. Suppose that $a \notin W_P(b)$. Then, there exists $\tilde{a}$ arbitrarily close to a such that $\tilde{a} \notin \text{cl}(W_P(b))$. By Proposition 1, there cannot be any $\tilde{d}$ such that $(\tilde{a}, b, \tilde{d})$ is anomalous. This is a contradiction, so $a \in W_P(b)$.

A.9 Proof of Theorem 4

Since any two-dimensional subspace of $\Delta(Z)$ is simply $\Delta(Z')$ for some $|Z'| = 3$, it suffices to show the result for $|Z| = 3$.

Fix any interior lotteries x, y such that $x \in W_P(y)$. By Lemma 20, there exists $\succsim$ such that $\text{cl}(W_P(y)) \setminus \{y\} \succ y$. We first show that $P(x \in W(y) | \succsim)$ is identified from ρ.

Fix a lottery $a \neq x$ such that $a \sim x$. Let $b := 2x - a$. We can always choose a close enough to

x that $\hat{a} := x + a - b$, $\hat{b} := x + b - a$, $y + a - b$, and $y + b - a$ are all interior. For $n \in \mathbb{N}$, let

$$\begin{aligned} x_n &:= \frac{1}{n}x + \left(1 - \frac{1}{n}\right) d \\ a_n &:= \frac{1}{n}a + \left(1 - \frac{1}{n}\right) d \\ b_n &:= \frac{1}{n}b + \left(1 - \frac{1}{n}\right) d \\ \hat{a}_n &:= x_n + a - b \\ \hat{b}_n &:= x_n + b - a. \end{aligned}$$

For each $n \in \mathbb{N}$ and $m \in \mathbb{M}$, let

$$\begin{aligned} a_n^m &:= \frac{1}{m}a_n + \left(1 - \frac{1}{m}\right) x_n \\ b_n^m &:= \frac{1}{m}b_n + \left(1 - \frac{1}{m}\right) x_n. \end{aligned}$$

To simplify notation, we write a^m and b^m instead of a_1^m and b_1^m.

We claim that, for n large enough, $\mathrm{cl}(W_P(y)) \cap \mathrm{co}(\hat{a}_n, y + a - b) = \emptyset$. Suppose otherwise: for n arbitrarily large, there exists $q_n \in \mathrm{co}(\hat{a}_n, y + a - b)$ such that $q_n \in \mathrm{cl}(W_P(y))$. Since (q_n) must converge to $y+a-b$, we have $y+a-b \in \mathrm{cl}(W_P(y))$. Since $y+a-b \sim y$, this contradicts the definition of $\succsim$. A parallel argument establishes that $\mathrm{cl}(W_P(y)) \cap \mathrm{co}(\hat{b}_n, y + b - a) = \emptyset$ for n sufficiently large. Thus, there exists $N \in \mathbb{N}$ such that $W_{\mathcal{M}}(y) \cap \mathrm{cl}(\hat{a}_n, y + a - b) = \emptyset$ and $W_{\mathcal{M}}(y) \cap \mathrm{cl}(\hat{b}_n, y + b - a) = \emptyset$ for all $n \geq N$ and all $\mathcal{M} \in \mathrm{supp}(P_{\mathrm{just}})$. Since we must have $q \succ y$ for any $q \in W_{\mathcal{M}}(y)$, we know $W_{\mathcal{M}}(y)$ must intersect at least one of $\mathrm{co}(\hat{a}_n, y + a - b)$, $\mathrm{co}(\hat{b}_n, y + b - a)$, or $\mathrm{co}(\hat{a}_n, \hat{b}_n)$ for each $n \in \mathbb{N}$. Since we have already ruled out the first two possibilities for $n \geq N$, $W(y) \neq \emptyset$ implies $W(y) \cap \mathrm{co}(\hat{a}_n, \hat{b}_n) \neq \emptyset$ for all $n \geq N$.

The following manipulations are needed to identify $P(x \in W(y)|\succsim)$:

$$\begin{aligned}
P(x \in W(y)|\succsim) &= P(x \in W(y)|a \sim b \succ y) \\
&= \lim_{n\to\infty} P\left(x \in W(y)|\hat{a}_n, \hat{b}_n \succ y\right) \\
&= \lim_{n\to\infty} P\left(\lim_{m\to\infty} \{a^m, b^m \in W(y)\}\,|\hat{a}_n, \hat{b}_n \succ y\right) \\
&= \lim_{n\to\infty} \lim_{m\to\infty} P\left(a^m, b^m \in W(y)|\hat{a}_n, \hat{b}_n \succ y\right) \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \Big\{P\left(a^m \in W(b^m, y)|\hat{a}_n, \hat{b}_n \succ y\right) + P\left(b^m \in W(a^m, y)|\hat{a}_n, \hat{b}_n \succ y\right) \\
&\quad -P\left(a^m_N \in W(b^m_N, y) \text{ or } b^m_N \in W(a^m_N, y)|\hat{a}_n, \hat{b}_n \succ y\right)\Big\} \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \Big\{P\left(a^m \in W(b, y)|\hat{a}_n, \hat{b}_n \succ y\right) + P\left(b^m \in W(a, y)|\hat{a}_n, \hat{b}_n \succ y\right) \\
&\quad +P\left(a^m_N \notin W(b_N, y) \text{ and } b^m_N \notin W(a_N, y)|\hat{a}_n, \hat{b}_n \succ y\right) - 1\Big\} \\
&= \lim_{n\to\infty} \Big\{P\left(\lim_{m\to\infty} \{a^m \in W(b, y)\}\,|\hat{a}_n, \hat{b}_n \succ y\right) + P\left(\lim_{m\to\infty} \{b^m \in W(a, y)\}\,|\hat{a}_n, \hat{b}_n \succ y\right) \\
&\quad +P\left(\lim_{m\to\infty} \{a^m_N \notin W(b_N, y) \text{ and } b^m_N \notin W(a_N, y)\}\,|\hat{a}_n, \hat{b}_n \succ y\right) - 1\Big\} \\
&= \lim_{n\to\infty} \Big\{P\left(x \in W(b, y)|\hat{a}_n, \hat{b}_n \succ y\right) + P\left(x \in W(a, y)|\hat{a}_n, \hat{b}_n \succ y\right) \\
&\quad +P\left(W(y) = \emptyset\big|\hat{a}_n, \hat{b}_n \succ y\right) - 1\Big\} \\
&= P(x \in W(b, y)|a \sim b \succ y) + P(x \in W(a, y)|a \sim b \succ y) + P(W(y) = \emptyset|a \sim b \succ y) - 1.
\end{aligned}$$

The first equality uses the fact that $|Z| = 3$, so $\succsim$ is pinned down by $a \sim b \succ y$. The fourth uses the fact that $\{a^m, b^m \in W(y) \text{ and } \hat{a}_n, \hat{b}_n \succ y\}_{m\in\mathbb{N}}$ is an increasing sequence of events for any n. The sixth uses Lemma 19. The seventh uses the fact that $\{a^m \in W(b, y) \text{ and } \hat{a}_n, \hat{b}_n \succ y\}_{m\in\mathbb{N}}$ $\{b^m \in W(a, y) \text{ and } \hat{a}_n, \hat{b}_n \succ y\}_{m\in\mathbb{N}}$ are increasing sequences of events for any n, and that $\{a^m_N \notin W(b, y) \text{ and } b^m_N \notin W(a, y) \text{ and } \hat{a}_n, \hat{b}_n \succ y\}_{m\in\mathbb{N}}$ is a decreasing sequence of events for any n. To show the latter, notice that $a^m_N \notin W(b_N, y)$ implies $W(y) \cap \text{co}(\hat{a}_N, a^m_N) = \emptyset$ and that $b^m_N \notin W(a_N, y)$ implies $W(y) \cap \text{co}(b^m_N, \hat{b}_N) = \emptyset$. Since $W(y) \neq \emptyset$ implies $W(y) \cap \text{co}(\hat{a}_N, \hat{b}_N) \neq \emptyset$, we have

$$a^m_N \notin W(b_N, y) \text{ and } b^m_N \notin W(a_N, y) \quad \Longleftrightarrow \quad W(y) = \emptyset \text{ or } W(y) \cap \text{co}(\hat{a}_N, \hat{b}_N) \subset \text{co}(a^m_N, b^m_N).$$

The sequence of sets $\{W(y) = \emptyset \text{ or } W(y) \cap \text{co}(\hat{a}_N, \hat{b}_N) \subset \text{co}(a^m_N, b^m_N) \text{ and } \hat{a}_n, \hat{b}_n \succ y\}_{m\in\mathbb{N}}$ is decreasing. The eighth equality uses the fact that the limit of this sequence is $\{W(y) = \emptyset\}$.

First, we show how to identify $P(x \in W(b, y)|b \sim x \succ y)$. We claim that

$$\rho(\hat{b}_n|\hat{a}_n, \hat{b}_n, y) = P(\hat{b}_n \succsim x_n, y) \tag{12}$$

for all $n \geq N$. Notice that $\hat{b}_n \succsim x_n$ if and only if $\hat{b}_n \succsim \hat{a}_n$, so the term on the right-hand side is simply $P(\hat{b}_n \succsim \hat{a}_n, y)$. It suffices to show that $\hat{b}_n \notin W_P(\hat{a}_n, y)$, that $\hat{a}_n \notin W_P(\hat{b}_n, y)$, and that $y \notin W_P(\hat{a}_n, \hat{b}_n)$. The third part follows from the definition of $\succsim$ and the fact that $y \succ \hat{a}_n, \hat{b}_n$. Toward a contradiction, suppose $\hat{b}_n \in W_P(\hat{a}_n, y)$. By Lemma 18, there exists $\alpha \in [0,1]$ such that $\hat{b}_n \in W_P(\alpha\hat{a}_n + (1-\alpha)y)$. Rearranging, we get

$$\frac{1-\alpha}{1+\alpha}\hat{b}_n + \frac{2\alpha}{1+\alpha}(y+b-a) \in W_P(y),$$

which contradicts the fact that $W_P(y) \cap \text{co}(\hat{b}_N, y+b-a) = \emptyset$. A very similar argument establishes $\hat{a}_n \notin W_P(\hat{b}_n, y)$ for all $n \geq N$.

Next, we claim that

$$\rho(\hat{b}_n | x_n, \hat{b}_n, y) = P(\hat{b}_n \succsim x_n, y) + (\hat{a}_n \succsim \hat{b}_n \succsim y \text{ and } x \in W(b,y)) \tag{13}$$

for all $n > N$. Notice that $x \in W(b,y)$ if and only if $x_n \in W(\hat{b}_n, y)$ by Lemma 19, and that $\hat{a}_n \succsim \hat{b}_n$ if and only if $x_n \succsim \hat{b}_n$. Thus, the second term on the right-hand side is simply $P(x_n \succsim \hat{b}_n \succsim y \text{ and } x_n \in W(\hat{b}_n, y))$. It remains to show that $y \notin W_P(\hat{b}_n, x_n)$ and $\hat{b}_n \notin W(x_n, y)$. The former follows from $y \notin W_P(\hat{b}_n, \hat{a}_n)$ and Lemma 18, and the latter from $\hat{b}_n \notin W(\hat{a}_n, y)$ and Lemma 19. Combining (12) and (13), we get

$$\rho(\hat{b}_n | x_n, \hat{b}_n, y) - \rho(\hat{b}_n | \hat{a}_n, \hat{b}_n, y) = P(x \in W(b,y) \text{ and } \hat{a}_n \succsim \hat{b}_n \succsim y) \tag{14}$$

for all $n \geq N$.

Fix any $n \geq N$. We show how to identify $P(\hat{a}_n \succsim \hat{b}_n \succsim y)$. Let $b_\epsilon := \hat{b}_n + \epsilon(\hat{b}_n - \hat{a}_n)$ for each $\epsilon > 0$ small enough to make b_ϵ a lottery. Notice that $\hat{a}_n \succsim \hat{b}_n \succsim y$ if and only if $\hat{b}_n \succsim y, b_\epsilon$ for some $\epsilon > 0$. Thus, it suffices to identify $P(\hat{b}_n \succsim y, b_\epsilon)$ for some $\epsilon > 0$. We will do this by identifying $P(y \succsim \hat{b}_n, b_\epsilon)$ and $P(b_\epsilon \succsim y, \hat{b}_n)$. We show that

$$\rho(b_\epsilon | b_\epsilon, \hat{a}_n, y) = P(b_\epsilon \succsim \hat{b}_n, y)$$

for all $\epsilon > 0$ for which b_ϵ is defined. Notice that $b_\epsilon \succsim \hat{b}_n, y$ if and only if $b_\epsilon \succsim \hat{a}_n, y$. Thus, the term on the right-hand side is just $P(b_\epsilon \succsim \hat{a}_n, y)$. It suffices to show that $y \notin W_P(b_\epsilon, \hat{a}_n)$, that $\hat{a}_n \notin W_P(b_\epsilon, y)$, and that $b_\epsilon \notin W_P(\hat{a}_n, y)$. We have $y \notin W_P(b_\epsilon, \hat{a}_n)$ because $b_\epsilon, \hat{a}_n \succ y$. By Lemma 19 and $\hat{a}_n \notin W_P(\hat{b}_n, y)$, we have $\hat{a}_n \notin W_P(b_\epsilon, y)$. Also by Lemma 19, $b_\epsilon \in W_P(\hat{a}_n, y)$ if and only if $b_\epsilon \in W_P(\hat{b}_n, y)$. Toward a contradiction, suppose $b_\epsilon \in W_P(\hat{b}_n, y)$. By Lemma 18, there exists $\alpha \in [0,1]$ such that $b_\epsilon \in W_P(\alpha\hat{b}_n + (1-\alpha)y)$. Rearranging, we get

$$\frac{1-\alpha}{1-\alpha+2\epsilon}\hat{b}_n + \frac{2\epsilon}{1-\alpha+2\epsilon}(y+b-a) \in W_P(y),$$

which contradicts the fact that $W_P(y) \cap \text{co}(\hat{b}_N, y + b - a) = \emptyset$.

Since $\hat{b}_n$ is interior, we can define $\tilde{b} := \hat{b}_n + \lambda(\hat{b}_n - y)$ for some $\lambda > 0$. We show that

$$\rho(y|y, b_\epsilon, \tilde{b}) = P(y \succsim \hat{b}_n, b_\epsilon)$$

for $\epsilon > 0$ sufficiently small. Since $y \succsim \hat{b}_n$ if and only if $y \succsim \tilde{b}$, the term on the right-hand side is just $P(y \succsim b_\epsilon, \tilde{b})$. It suffices to show that $y \notin W_P(b_\epsilon, \tilde{b})$, that $b_\epsilon \notin W_P(\tilde{b}, y)$, and that $\tilde{b} \notin W(b_\epsilon, y)$. We have $y \notin W_P(b_\epsilon, \tilde{b})$ because $b_\epsilon, \tilde{b} \succ y$. By Lemma 18, $b_\epsilon \succ \text{int}(\text{co}(\hat{b}_n, \tilde{b}))$, and $b_\epsilon \notin W_P(\hat{b}_n, y)$, we have $b_\epsilon \notin W_P(\tilde{b}, y)$. Toward a contradiction, suppose $\tilde{b} \in W_P(b_\epsilon, y)$ for ϵ arbitrarily small. By Lemma 18, for ϵ arbitrarily small, there exists $\alpha_\epsilon \in [0, 1]$ such that $\tilde{b} \in W_P(\alpha_\epsilon b_\epsilon + (1 - \alpha_\epsilon)y)$. Rearranging, we get

$$\hat{b}_n - \frac{2\epsilon\alpha_\epsilon}{1 + \lambda - \alpha_\epsilon}(b - a) \in W_P(y).$$

Taking a limit as $\epsilon \to 0$, we obtain $\hat{b}_n \in \text{cl}(W_P(y))$, which contradicts $\text{cl}(W_P(y)) \cap \text{co}(\hat{b}_N, y + b - a) = \emptyset$. This completes the argument that $P(\hat{a}_n \succsim \hat{b}_n \succsim y)$ is identified for all $n \geq N$. Combining this with (14), we can now identify $P(x \in W(b, y)|\hat{a}_n \succsim \hat{b}_n \succsim y)$ for all $n \geq N$. Taking a limit as $n \to \infty$ gives $P(x \in W(b, y)|a \sim b \succ y)$. A very similar argument suffices to show that $P(x \in W(a, y)|a \sim b \succ y)$ is identified.

Next, we identify $P(W(y) \neq \emptyset|a \sim b \succ y)$. Let

$$x_m := x + \frac{1}{m}(x - y)$$

for each m small enough that x_m is a lottery. We show that there exists $M \in \mathbb{N}$ such that

$$W_{\mathcal{M}}(y) \neq \emptyset \quad \iff \quad x_m \in W(a, b)$$

for all $m \geq M$. The arguments used to show $\text{cl}(W_P(y)) \cap \text{co}(\hat{a}_n, y + a - b) = \emptyset$ and $\text{cl}(W_P(y)) \cap \text{co}(\hat{b}_n, y + b - a) = \emptyset$ for n sufficiently large can be used to show $\text{cl}(W_P(a)) \cap \text{co}(\hat{a}, x_m + a - b) = \emptyset$ and $\text{cl}(W_P(a)) \cap \text{co}(x_m, x) = \emptyset$ for m sufficiently large. The arguments used to show that $W(y) \neq \emptyset$ implies $W(y) \cap \text{co}(\hat{a}_n, \hat{b}_n)$ for n sufficiently large can then be used to show that $W(a) \neq \emptyset$ implies $W(a) \cap \text{co}(x_m + a - b), x_m) \neq \emptyset$ for m sufficiently large. Similarly, $W_P(b) \cap \text{co}(\hat{b}, x_m + b - a) = \emptyset$ for m sufficiently large, and $W(b) \neq \emptyset$ implies $W(b) \cap \text{co}(x_m + b - a, x_m)$ for m sufficiently large. Thus, there exists $M \in \mathbb{N}$ such that $x_m \in W(a, b)$ for all $m \geq M$ whenever $W(a) \neq \emptyset$ and $W(b) \neq \emptyset$—equivalently, whenever $W(y) \neq \emptyset$. This implies

$$\rho(x_m|a, b, x_m) = P(W(y) = \emptyset \text{ and } x_m \succsim a, b) \tag{15}$$

for all $m \geq M$.

We now show how to identify $P(x_m \succsim a, b)$ for m sufficiently large. We do this by identifying $P(a \succsim x_m, b)$ and $P(b \succsim x_m, a)$. Let

$$\hat{b}_m := x_m + x_m - a$$
$$\hat{a}_m := x_m + x_m - b$$

for m sufficiently large that $\hat{b}_m, \hat{a}_m$ are lotteries. We show that

$$\rho(a|a, x, \hat{b}_m) = P(a \succsim x_m, b)$$

for m sufficiently large. Since $a \succsim x_m, b$ if and only if $a \succsim \hat{b}_m, x$, the term on the right-hand side is simply $P(a \succsim x, \hat{b}_m$. It suffices to show that $a \notin W_P(x, \hat{b}_m)$, $x \notin W_P(\hat{b}_m, a)$, and $\hat{b}_m \notin W_P(a, x)$ for m sufficiently large. The first two claims follow because $a \sim x \succ \hat{b}_m$. Suppose $\hat{b}_m \in W_P(a, x)$ for arbitrarily large m. By Lemma 18, for arbitrarily large m, there exists $\alpha_m \in [0, 1]$ such that $\hat{b}_m \in W_P(\alpha a + (1 - \alpha)x)$. Rearranging, we have

$$x + \frac{2}{m(1 + \alpha)}(x - y) \in W_P(a),$$

which implies $x \in \text{cl}(W_P(a))$. This contradicts the fact that $\text{cl}(W_P(a)) \cap \text{co}(x, x_m) = \emptyset$. A parallel argument establishes that

$$\rho(b|b, x, \hat{a}_m) = P(b \succsim x_m, a)$$

for m sufficiently large. Combining this with (15) allows us to identify $P(W(y) = \emptyset | x_m \succsim a, b)$ for all m sufficiently large. Taking a limit as $m \to \infty$ delivers $P(W(y) = \emptyset | a \sim b \succ y)$.

By the first part of regularity, P_{pref} admits a density. We can use the construction above to compute this density at $\succsim$:

$$\lim_{m \to \infty} \frac{P(x_m \succsim a, b)}{P_{\text{unif}}(x_m \succsim a, b)},$$

where $P_{\text{unif}}(x_m \succsim a, b)$ is the probability of the event $\{\succsim \in \mathcal{U} : x_m \succsim a, b\}$ given a uniform distribution on $\mathcal{U}$.

We have shown how to obtain $P(x \in W(y) | \succsim)$ and the density of P_{just} at $\succsim$ for any member of $\{\succsim \in \mathcal{U} : \text{cl}(W_P(y)) \setminus \{y\} \succ y\}$. Notice that this set is equal to $\text{int}\,(\{\succsim \in \mathcal{U} : \succsim \text{ is unjustifiable}\})$. Thus, we can now compute

$$P_{\text{pref}}(\succsim \text{ is unjustifiable}) = \int_{\{\succsim \text{ is unjustifiable}\}} dP_{\text{pref}}(\succsim).$$

This allows us to obtain the density of P_{pref} conditional on the event that $\succsim$ is unjustifiable. Using

this density, we can compute

$$P(x \in W(y)|\succsim \text{ is unjustifiable}) = \int_{\{\succsim \text{ is unjustifiable}\}} P(x \in W(y)|\succsim) dP_{\text{pref}}(\succsim | \succsim \text{ is unjustifiable}).$$

A.10 Proof of Proposition 3

Since (A, B) is ambiguous, there exists $\mathcal{M}^* \in \text{supp}(P_{\text{just}})$ for which there is no $\succ_m^*$ satisfying (5). If $b \succ_{\mathcal{M}^*} a$ for all $a \in A$ and $b \in B$, then every $\succ_m \in \mathcal{M}^*$ satisfies (5)—contradiction. Thus, there exist $\bar{a} \in A$ and $\bar{b} \in B$ such that $\bar{a} \precsim_m \bar{b}$ for some $\precsim_m \in \mathcal{M}^*$.

Suppose that, for all $a \in A$ and all $b \in B$, there exists $\succ_m \in \mathcal{M}^*$ such that $a \precsim_m b$. Fix some $a^* \in A \cap \text{int}(\Delta(Z))$. For sufficiently small $\epsilon > 0$, we have $a^* + \epsilon(b - a) \in \Delta(Z)$ for all $a \in A$ and $b \in B$. Let $B^* := \{a^* + \epsilon(b - a) : a \in A, b \in B\}$ for some such ϵ. For each $b^* \in B^*$, there exists $\succ_m \in \mathcal{M}^*$ such that $a^* \precsim_m b^*$. Notice that B^* inherits convexity from A and B.

Fix a countable dense subset of B^*: $\{b_1^*, b_2^*, \ldots\}$. Suppose that, for some $n \in \mathbb{N}$, $\{b_1^*, \ldots, b_n^*\} \succ_{\mathcal{M}^*} a^*$. By Lemma 17, there exists $b^* \in B^*$ such that $b^* \succ_{\mathcal{M}^*} a^*$—contradiction. Conclude that, for each n, there exists $\succ^n \in \mathcal{M}^*$ such that $a^* \precsim^n \{b_1^*, \ldots, b_n^*\}$. We can pass to a convergent subsequence of $(\succ^n)$; let $\succ^*$ be the limit. Since $\mathcal{M}^*$ is closed, it contains $\succ^*$. Since $a^* \precsim^j b_i^*$ for all $j \geq i$, we have $a^* \precsim^* b_i^*$ for all i. This implies $a^* \precsim^* B^*$, which implies $a \precsim^* b$ for all $a \in A$ and $b \in B$. Conclude that $\succ^*$ satisfies (5)—contradiction. Thus, there exist $\underline{a} \in A$ and $\underline{b} \in B$ such that $\underline{b} \succ_{\mathcal{M}^*} \underline{a}$.

By mixing $(\bar{a}, \bar{b})$ and $(\underline{a}, \underline{b})$, we obtain $a \in A$ and $b \in B$ such that $b \precsim_m a$ for all $\succ_m \in \mathcal{M}^*$ and $a \sim_1 b$ for some $\succ_1 \in \mathcal{M}^*$. It is without loss to assume that a and b are interior lotteries. Since $\succ_1$ does not satisfy (5), there exist $a_2 \in A$ and $b_2 \in B$ such that $b_2 \succ_1 a_2$ but $a_2 \precsim_2 b_2$ for some $\succ_2 \in \mathcal{M}^*$. It is without loss to assume $a_2 \succ_2 b_2$. Let

$$(a_1, b_1) := \frac{1}{\lambda}(a, b) - \frac{1 - \lambda}{\lambda}(a_2, b_2).$$

for some $\lambda \in (0, 1)$ such that $(a_1, b_1) \in A \times B$. Since $a \sim_1 b$ and $b_2 \succ_1 a_2$, we have $a_1 \succ_1 b_1$.

For each $\epsilon \in (0, 1)$, we have

$$\epsilon\underline{b} + (1 - \epsilon)b \succ_{\mathcal{M}^*} \epsilon\underline{a} + (1 - \epsilon)a.$$

Thus, we can find $(\tilde{a}, \tilde{b})$ arbitrarily close to (a, b) such that $\tilde{b} \succ_{\mathcal{M}^*} \tilde{a}$. If we choose $(\tilde{a}, \tilde{b})$ sufficiently close to (a, b), we will have $(a_i, b_i) + (\tilde{a}, \tilde{b}) - (a, b) \in A \times B$ and $a_i + \tilde{a} - a \succ_i b_i + \tilde{b} - b$ for $i = 1, 2$. Fix some $(\tilde{a}, \tilde{b})$ that satisfies these conditions, and let $(\tilde{a}_i, \tilde{b}_i) := (a_i, b_i) + (\tilde{a}, \tilde{b}) - (a, b)$. Let S be the lottery over pairs that puts weight λ on $(\tilde{a}_1, \tilde{b}_1)$ and weight $1 - \lambda$ on $(\tilde{a}_2, \tilde{b}_2)$. By construction, S is a signal for $(\tilde{a}, \tilde{b})$.

By construction, $\tilde{b} \succ_{\mathcal{M}^*} \tilde{a}$. For each $i = 1, 2$, for some $\succ_m \in \mathcal{M}^*$, we have $\tilde{a}_i \succ_m \tilde{b}_i$. The same conditions hold for all $\mathcal{M}$ in an open neighborhood N_{just} of $\mathcal{M}^*$. Let $\succ_{\text{anti}}$ be the opposite of the

policy preference: that is, $x \succsim_{\text{anti}} y$ if and only if $y \succsim_{\text{pol}} x$. Since $b \succ_{\text{pol}} a$ for all $a \in A$ and $b \in B$, and since $(\tilde{a}_i, \tilde{b}_i) \in A \times B$ for $i = 1, 2$, we have $\tilde{a}_i \succ_{\text{anti}} \tilde{b}_i$ for $i = 1, 2$. The same conditions hold for all $\succ$ in an open neighborhood N_{pref} of $\succ_{\text{anti}}$. Let $N := N_{\text{pref}} \times N_{\text{just}}$. We have $P(N) > 0$ by the second part of regularity.

For any $(\succ, \mathcal{M}) \in N$,

$$c_{(\succ,\mathcal{M})}(\tilde{a}_i, \tilde{b}_i) = \tilde{a}_i \succ \tilde{b} \succ c_{(\succ,\mathcal{M})}(\tilde{a}, \tilde{b})$$

for $i = 1, 2$. Since $\tilde{b} \succ_{\text{pol}} \tilde{a}_i$ for $i = 1, 2$, we are done.

A.11 Proof of Proposition 2

Let X^* be the (possibly empty) set of all $(a, b) \in A \times B$ such that, for all $(\tilde{a}, \tilde{b}) \in A \times B$,

$$\tilde{b} \succ_{\mathcal{M}} \tilde{a} \quad \Longrightarrow \quad b \succ_{\mathcal{M}} a.$$

Suppose that X^* has non-empty interior. Then, there exists $(a^*, b^*) \in X^*$ and $\epsilon > 0$ such that $\{a^*\} \times B_\epsilon(b^*) \subset X^*$. By definition of X^*,

$$b^* \in B_{\mathcal{M}}(a^*) \quad \Longrightarrow \quad B_\epsilon(b^*) \subset B_{\mathcal{M}}(a^*).$$

This is the same as (11) in the proof of Theorem (3). We showed that (11) contradicts regularity. Conclude that X^* has empty interior. We show that X^* is convex. Fix any $(a, b) \in X^*$, and let

$$N := \{\mathcal{M} \in \text{supp}(P_{\text{just}}) : b \succ_{\mathcal{M}} a\}.$$

For any $(\tilde{a}, \tilde{b}) \in X^*$,

$$\tilde{b} \succ_{\mathcal{M}} \tilde{a} \quad \Longleftrightarrow \quad \mathcal{M} \in N.$$

Fix any $(a_1, b_1), (a_2, b_2) \in X^*$, and any $\alpha \in (0, 1)$. Since A and B are convex, $\alpha(a_1, b_1) + (1 - \alpha)(a_2, b_2) \in A \times B$. Fix any $\mathcal{M} \in N$. Since $b_1 \succ_{\mathcal{M}} a_1$ and $b_2 \succ_{\mathcal{M}} a_2$, we have $\alpha b_1 + (1 - \alpha) b_2 \succ_{\mathcal{M}} \alpha a_1 + (1 - \alpha) a_2$. This implies $\alpha(a_1, b_1) + (1 - \alpha)(a_2, b_2) \in X^*$, so X^* is convex. Since X^* has empty interior, it is contained in a hyperplane of $\Delta(Z)^2$. Thus, X^* has measure zero.

For the rest of the proof, we will restrict attention to $(a, b) \in A \times B$ such that $(a, b) \notin X^*$ and $a, b \in \text{int}(\Delta(Z))$. The set of all such (a, b) has the same measure as $A \times B$. Fix any such (a, b). Since $(a, b) \notin X^*$, we can find $(\tilde{a}, \tilde{b}) \in A \times B$ such that

$$a \in W_{\mathcal{M}}(b) \quad \Longrightarrow \quad \tilde{a} \in W_{\mathcal{M}}(\tilde{b}),$$

but not vice versa. It is without loss to assume $\tilde{b} = b$. Since there exists $\mathcal{M}^*$ such that $\tilde{a} \in W_{\mathcal{M}^*}(b)$

but $a \notin W_{\mathcal{M}^*}(b)$, and since $W_{\mathcal{M}^*}(b)$ is open, there exists $\epsilon > 0$ such that $B_\epsilon(\tilde{a}) \subset W_{\mathcal{M}^*}(b)$. Take any $\mathcal{M}$ such that $a \in W_{\mathcal{M}}(b)$. Since (A, B) is orderly, we must have $B_\epsilon(\tilde{a}) \subset W_{\mathcal{M}}(b)$. (If not, there is $x \in B_\epsilon(\tilde{a})$ such that $x \notin W_{\mathcal{M}}(b)$ but $x \in W_{\mathcal{M}^*}(b)$. Since $a \in W_{\mathcal{M}}(b)$ but $a \notin W_{\mathcal{M}^*}(b)$, this is a contradiction.) Thus, $B_\epsilon(\tilde{a}) \subset W_{\mathcal{M}}(b)$ whenever $a \in W_{\mathcal{M}}(b)$. Suppose that there exists $n \in \mathbb{N}$ such that

$$a \in W_{\mathcal{M}}(b) \quad \Longrightarrow \quad a + (1/n)(a - \tilde{a}) \in W_{\mathcal{M}}(b).$$

Since $a \in W_{\mathcal{M}}(b)$ also implies $B_\epsilon(\tilde{a}) \subset W_{\mathcal{M}}(b)$, we have

$$a \in W_{\mathcal{M}}(b) \quad \Longrightarrow \quad \mathrm{co}\left(\{a + (1/n)(a - \tilde{a})\} \cup B_\epsilon(\tilde{a})\right) \subset W_{\mathcal{M}}(b).$$

This implies

$$a \in W_{\mathcal{M}}(b) \quad \Longrightarrow \quad B_\delta(a) \subset W_{\mathcal{M}}(b)$$

for some $\delta > 0$. This is the same as (11) in the proof of Theorem (3), which led to a contradiction. Conclude that, for all $n \in \mathbb{N}$, there exists $\mathcal{M}_n \in \mathrm{supp}(P_{\mathrm{just}})$ such that

$$a \in W_{\mathcal{M}_n}(b) \text{ and } a + (1/n)(a - \tilde{a}) \notin W_{\mathcal{M}_n}(b).$$

We can pass to a convergent subsequence of the $\mathcal{M}_n$ with limit $\bar{\mathcal{M}}$. Since $a \in W_{\mathcal{M}_n}(b)$ for all n, we have $b \not\succsim_m a$ for all $\succ_m \in \bar{\mathcal{M}}$. For each n, there exists $\succ_m^n \in \mathcal{M}_n$ such that $a + (1/n)(a - \tilde{a}) \not\succsim_m^n b$. Passing to a convergent subsequence of $\succ_m^n$, we have $a \not\succsim_m b$ for some $\not\succsim_m \in \bar{\mathcal{M}}$. Thus, $a \notin W_{\bar{\mathcal{M}}}(b)$. Since $a \in W_{\mathcal{M}_n}(b)$ for all n, we have $B_\epsilon(\tilde{a}) \subset W_{\mathcal{M}_n}(b)$ for all n, so $\tilde{a} \in W_{\bar{\mathcal{M}}}(b)$. This implies $\alpha a + (1-\alpha)\tilde{a} \in W_{\bar{\mathcal{M}}}(b)$ for all $\alpha \in (0, 1)$. By the third part of Regularity, we can find $\mathcal{N}$ arbitrarily close to $\bar{\mathcal{M}}$ such that $\bar{\mathcal{M}} \subset \mathrm{int}(\mathcal{N})$. Since $a \notin W_{\bar{\mathcal{M}}}(b)$, we have $a \notin W_{\mathcal{N}}(b)$ for any $\mathcal{N} \supset \bar{\mathcal{M}}$. Fix any $\alpha \in (0, 1)$. Since $\alpha a + (1 - \alpha)\tilde{a} \in W_{\bar{\mathcal{M}}}(b)$, the same will be true of $\mathcal{N}$ if it is close enough to $\mathcal{M}$. Conclude that, for each $\epsilon > 0$, there exists $a_\epsilon \in B_\epsilon(a)$ such that

$$a \in W_{\mathcal{M}}(b) \quad \Longrightarrow \quad a_\epsilon \in W_{\mathcal{M}}(b).$$

but not vice versa.

Since $a \in A \cap \mathrm{int}(\Delta(Z))$, there exists $\lambda > 0$ such that $a + \lambda(a - b) \in A \cap \mathrm{int}(\Delta(Z))$. Let u be a representation of $\succ_{\mathrm{pol}}$. Let

$$b_\epsilon := a + \left(\frac{u(b) - u(a)}{u(a) - u(a_\epsilon + \lambda(a_\epsilon - b))}\right)(a - a_\epsilon - \lambda(a_\epsilon - b))$$

for all $\epsilon > 0$ small enough that b_ϵ is a lottery. We have $\lim_{\epsilon \to 0} b_\epsilon = b$. Fix some ϵ for which b_ϵ is defined, and let $a_1 := a_\epsilon + \lambda(a_\epsilon - b)$ and $a_2 := b_\epsilon$. Since $a \in \mathrm{co}(a_1, a_2)$, there exists a signal S about (a, b) with support $\{(a_1, b), (a_2, b)\}$.

Suppose that there is some $(\succ, \mathcal{M}) \in \text{supp}(P)$ such that $a \succ b$ and

$$S(a_1, b)c_{(\succ,\mathcal{M})}(a_1, b) + S(a_2, b)c_{(\succ,\mathcal{M})}(a_2, b) \prec_{\text{pol}} c_{(\succ,\mathcal{M})}(a, b). \tag{16}$$

In that case, $c_{(\succ,\mathcal{M})}(a, b) = b$. Since $a \succ b$, we have $a \in W_{\mathcal{M}}(b)$. By construction of a_1, we have $a_1 \in W_{\mathcal{M}}(b)$, so $c(a_1, b) = b$. But since $a_2 \sim_{\text{pol}} b$, (16) cannot hold.

Let

$$N_{\text{just}} := \{\mathcal{M} \in \text{supp}(P_{\text{just}}) : a_1 \in W_{\mathcal{M}}(b) \text{ and } a \notin W_{\mathcal{M}}(b)\}.$$

By construction of a_1, there exists $\mathcal{M} \in \text{supp}(P_{\text{just}})$ such that $a_1 \in W_{\mathcal{M}}(b)$ and $a \notin W_{\mathcal{M}}(b)$. By the third part of Regularity, it is without loss to assume that a is not on the boundary of $W_{\mathcal{M}}(b)$. Then, for any $\tilde{M}$ in an open neighborhood of $\mathcal{M}$, we have $a_1 \in W_{\tilde{\mathcal{M}}}(b)$ and $a \notin W_{\tilde{\mathcal{M}}}(b)$. Conclude that $P_{\text{just}}(N_{\text{just}}) > 0$. Similarly, let $N_{\text{pref}} = \{\succ \in \mathcal{U} : a \succ b \succ a_1\}$. Since a_1, a and b cannot be collinear, N_{pref} is an open nonempty subset of $\mathcal{U}$. Let $N := N_{\text{pref}} \times N_{\text{pref}}$. By the second part of Regularity, $P(N) > 0$.

We show that

$$S(a_1, b)c_{(\succ,\mathcal{M})}(a_1, b) + S(a_2, b)c_{(\succ,\mathcal{M})}(a_2, b) \succ c_{(\succ,\mathcal{M})}(a, b) \tag{17}$$

for any $(\succ, \mathcal{M}) \in N$. Since $a \notin W_{\mathcal{M}}(b)$ and $a \succ b$, we have $c_{(\succ,\mathcal{M})}(a, b) = a$. Since $a_1 \in W_{\mathcal{M}}(b)$, we have $c_{(\succ,\mathcal{M})}(a_1, b) = b$. Since $W_{\mathcal{M}}(b)$ is convex and contains a_1 but not a, we have $a_2 \notin W_{\mathcal{M}}(b)$. Since $a \succ b \succ a_1$, we must have $a_2 \succ b$. Thus, $c_{(\succ,\mathcal{M})}(a_2, b) = a_2$. Since

$$S(a_1, b)b + S(a_2, b)a_2 \succ S(a_1, b)a_1 + S(a_2, b)a_2 = a,$$

(17) holds. Finally, (17) holds with $\succ_{\text{pol}}$ in place of $\succ$ because $b \sim a_2 \succ_{\text{pol}} a$, so $b \succ_{\text{pol}} a_1$.

A.12 Proof of Proposition 4

Suppose that $\mathcal{M}_{\text{info}} \supset \mathcal{M}_{\text{final}}$. Fix any interior a. We must have $W_{\text{info}}(a) \subset W_{\text{final}}(a)$. (Otherwise, maximality would imply $\mathcal{M}_{\text{info}} = \mathcal{M}_{\text{final}}$.) Fix any interior b on the boundary of $W_{\text{final}}(a)$ that is not in $\text{cl}(W_{\text{info}}(a))$. Suppose $a \succsim_{\text{info}} b$ for all $\succsim_{\text{info}} \in \mathcal{M}_{\text{info}}$. If $\mathcal{M}_{\text{info}}$ contains a constant preference, then $\mathcal{M}_{\text{info}} = \mathcal{U}$ by maximality. Thus, $\mathcal{M}_{\text{info}}$ does not contain a constant preference. By Lemma 17, $W_{\text{info}}(a) \neq \emptyset$. Fix some $x \in W_{\text{info}}(a)$. For all $\epsilon > 0$ sufficiently small, we have $a \succ_{\text{info}} b + \epsilon(x - a)$ for all $\succsim_{\text{info}} \in \mathcal{M}_{\text{info}}$. Thus, $b \in \text{cl}(W_{\text{info}}(a))$—contradiction. Conclude that $b \succ_{\text{info}} a$ for some $\succsim_{\text{info}} \in \mathcal{M}_{\text{info}}$.

Since $b \in \text{cl}(W_{\text{final}}(a))$, we can choose $\underline{b} \in W_{\text{final}}(a)$ arbitrarily close to b. By choosing $\underline{b}$

sufficiently close to b, we get $\underline{b}, 2b - \underline{b} \succ_{\text{info}} a$. Let $\bar{b} := 2b - \underline{b}$, and let

$$b^* := \frac{2}{3}\bar{b} + \frac{1}{3}\underline{b}.$$

Since $b^* \in \text{co}(\bar{b}, \underline{b})$, there exists a signal S about $\{a, b^*\}$ with support $\{\{a, \bar{b}\}, \{a, \underline{b}\}\}$.

Since $b \notin W_{\text{final}}(a)$, there exists $\succsim_{\text{final}} \in \mathcal{M}_{\text{final}}$ such that $b \succsim_{\text{final}} a$. Since $a \succ_{\text{final}} \underline{b}$, we have $\bar{b} \succ_{\text{final}} b \succsim_{\text{final}} a \succ_{\text{final}} \underline{b}$. Since $b^* \in \text{co}(\bar{b}, b)$, we have $b^* \succ_{\text{final}} a$. Thus, $b^* = \max(\succsim_{\text{final}}, \{a, b*\})$. Since $\bar{b}, \underline{b} \succ_{\text{info}} a$, we have

$$b^* = S\left(a, \bar{b}\right) \max\left(\succsim_{\text{info}}, \{a, \bar{b}\}\right) + S\left(a, \underline{b}\right) \max\left(\succsim_{\text{info}}, \{a, \underline{b}\}\right).$$

Conclude that $\left(\delta_{\{b^*, a\}}, b^*\right)$ is feasible.

Consider any preference $\succsim \in \mathcal{U}$ such that $\underline{b}, \bar{b} \succ a$. We have

$$\max\left(\succsim, \mathcal{M}_{\text{final}}\left(a, b^*\right)\right) = \max\left(\succsim, \{a, b^*\}\right) = b^*.$$

Since $\underline{b} \in W_{\text{final}}(a)$ and $\bar{b} \notin W_{\text{final}}(a)$,

$$S\left(a, \underline{b}\right) \max\left(\succsim, \mathcal{M}_{\text{final}}\left(a, \underline{b}\right)\right) + S\left(a, \bar{b}\right) \max\left(\succsim, \mathcal{M}_{\text{final}}\left(a, \bar{b}\right)\right) = S\left(a, \underline{b}\right) a + S\left(a, \bar{b}\right) \bar{b} \prec b^*.$$

Thus, (6) holds for $\succsim$.

Now suppose that $\mathcal{M}_{\text{info}} = \mathcal{M}_{\text{final}}$. Toward a contradiction, suppose that (6) holds for some menu A, item $a \in A$, and signal S about A. By (6), it cannot be possible to obtain the lottery a by choosing an appropriate element from $\mathcal{M}_{\text{final}}(\tilde{A})$ for each $\tilde{A} \in \text{supp}(S)$. This contradicts feasibility of (δ_A, a), which says that it is possible to obtain the lottery a by maximizing some member of $\mathcal{M}_{\text{final}}$ over each $\tilde{A} \in \text{supp}(S)$.

B Uncertainty about justifications

Definition 23 (Extended random justification model)**.** *An extended random justification model P^{ext} is a probability measure P^{ext} on $\mathcal{U} \times \Delta(\mathcal{C}(\mathcal{U}))$.*

Let P^{ext}_{μ} denote the marginal of P^{ext} on $\Delta(\mathcal{C}(\mathcal{U}))$, and let μ denote an arbitrary member of $\text{supp}(P^{\text{ext}}_{\mu})$. Let P denote the measure on $\mathcal{U} \times \mathcal{C}(\mathcal{U})$ induced by P^{ext}, and let P_{just} denote the marginal of P on $\mathcal{C}(\mathcal{U})$.

Definition 24 (Regularity)**.** *P^{ext} is regular if*

1. *P is a regular random justification model.*

2. $supp(P^{ext}) = \mathcal{U} \times supp(P^{ext}_{\mu})$.

3. $supp(\mu) = supp(P_{just})$ *for any* $\mu \in supp(P^{ext}_{\mu})$.

Proposition 6. *Fix an extended random justification model* P *and* $\succ_{pol} \in \mathcal{U}$ *such that* $b \succ_{pol} a$ *whenever* $a \in W_P(b)$. *Suppose that* (A, B) *is locally improvable. For almost any* $(a, b) \in A \times B$, *there exists a signal* S *about* (a, b) *such that*

$$\sum_{supp(S)} S(a_i, b_i) \int_{\mathcal{M}} c_{(\succ,\mathcal{M})}(a_i, b_i)\, d\mu \succsim_{pol} \int_{\mathcal{M}} c_{(\succ,\mathcal{M})}(a, b)\, d\mu \tag{18}$$

for every $(\succ, \mu)$ *such that* $a \succ b$. *The above holds with strict preference for a positive-probability set of* $(\succ, \mathcal{M})$ *such that* $a_i \succ b_i$ *for each* $(a_i, b_i) \in supp(S)$ *and*

$$\sum_{supp(S)} S(a_i, b_i) \int_{\mathcal{M}} c_{(\succ,\mathcal{M})}(a_i, b_i)\, d\mu \succ \int_{\mathcal{M}} c_{(\succ,\mathcal{M})}(a, b)\, d\mu. \tag{19}$$

Proof. Since P is a regular random justification model, we can borrow the construction of S from Proposition 2.

Recall that a_2 is chosen so that $a_2 \sim_{\text{pol}} b$. Since $x \succ_{\text{pol}} y$ whenever $y \in W_P(x)$, we have $a_2 \notin W_P(b)$, so $P_{\text{just}}(a_2 \in W(b)) = 0$. If $a_2 \in W_{\mathcal{M}}(b)$ for some $\mathcal{M} \in \text{supp}(P_{\text{just}})$, then the same is true for all $\tilde{\mathcal{M}}$ in an open neighborhood of $\mathcal{M}$, so $P_{\text{just}}(a_2 \in W(b)) > 0$—contradiction. Since $\text{supp}(\mu) = \text{supp}(P_{\text{just}})$ for all $\mu \in \text{supp}(P^{\text{ext}}_{\mu})$, we have $a_2 \notin W_{\mathcal{M}}(b)$ for all $\mathcal{M} \in \text{supp}(\mu)$. This implies $\mu(a_2 \notin W_{\mathcal{M}}(b))$ for all $\mu \in \text{supp}(P^{\text{ext}}_{\mu})$.

Fix any $\mu^* \in \text{supp}(P^{\text{ext}}_{\mu})$. Since $P_{\text{just}}(a \in W(b)) > 0$, there exists an open $N \subset \text{supp}(P_{\text{just}})$ such that $a \in W_{\mathcal{M}}(b)$. By the third part of Regularity for P^{ext}, $N \subset \text{supp}(\mu^*)$, so $\mu^*(a \in W(b)) > 0$. Similarly, since $P_{\text{just}}(a \notin W(b) \text{ and } a_1 \in W(b)) > 0$, we have $\mu^*(a \notin W(b) \text{ and } a_1 \in W(b)) > 0$. Since a_1 was chosen so that $a_1 \notin W(b)$ implies $a \notin W(b)$, we have $\mu^*(a \in W(b)) < \mu^*(a_1 \in W(b))$.

We claim that there exists $\succ^* \in \mathcal{U}$ such that, for some representation u of $\succ^*$,

$$u(a_2) > u(a_1) > u(b) = 0 \tag{20}$$

$$S(a_1, b)u(a_1)\mu^*(a_1 \notin W(b)) + S(a_2, b)u(a_2) > u(a)\mu^*(a \notin W(b)). \tag{21}$$

To see why, take any u_1 such that $u_1(a_1) > u_1(b) = 0$ and any u_2 such that $u_2(a_2) > u_2(a_1) = u_2(b) = 0$. The utility $\epsilon u_1 + (1-\epsilon)u_2$ will satisfy (19) and exhibit the required ordering for sufficiently small $\epsilon > 0$.

By the third part of Regularity for P^{ext}, we have $(\succ^*, \mu^*) \in \text{supp}(P^{\text{ext}})$. For any $(\succ, \mu)$ sufficiently close to $(\succ^*, \mu^*)$, $\succ$ will have a representation u that satisfies (20) and (21) with μ in place of μ^*. Thus, there is an open neighborhood N of $(\succ^*, \mu^*)$ in $\text{supp}(P^{\text{ext}})$ such that each $(\succ, \mu) \in N$ satisfies $a_2, a_1 \succ b$ and (19).

Consider any representation u_{pol} of $\succ_{\text{pol}}$ such that $u_{\text{pol}}(b) = 0$. Using $0 = u_{\text{pol}}(b) = u_{\text{pol}}(a_2) > u_{\text{pol}}(a_1)$ and $\mu(a_1 \notin W(b)) < \mu(a \notin W(b))$ for all μ such that $(\succ, \mu) \in N$, we have

$$\begin{aligned} S(a_1, b)u_{\text{pol}}(a_1)\mu(a_1 \notin W(b)) + S(a_2, b)u(a_2)\mu(a_2 \notin W(b)) &= S(a_1, b)u_{\text{pol}}(a_1)\mu(a_1 \notin W(b)) \\ &> S(a_1, b)u_{\text{pol}}(a_1)\mu(a \notin W(b)) \\ &= u_{\text{pol}}(a)\mu(a \notin W(b)) \end{aligned}$$

for all $(\succ, \mu) \in N$. This implies (18) with strict preference for any $(\succ, \mu) \in N$. □