Revealed Deliberate Preference Changes

Boissonnet, Niels and Ghersengorin, Alexis and Gleyze, Simon

Paris School of Economics, Bielefeld University

19 May 2020

Revealed Deliberate Preference Change[*]

Niels BOISSONNET[†] Alexis GHERSENGORIN[‡] Simon GLEYZE[§]

Abstract

We propose a model of chosen preferences together with conditions on choice data that falsify and identify our model. Preferences on alternatives are defined on attributes—e.g. candidates for a job may be experienced or inexperienced. Choice behavior is driven by a subset of attributes. Whenever an attribute becomes salient, the decision maker chooses to make it relevant or irrelevant for her future choices—e.g. employers may deliberately ignore race in the future to prevent discrimination. We identify when this decision is based on the maximization of a meta-preference, implying that preference changes are deliberate. This shows that theories of endogenous preferences, motivated reasoning, evolving attention, changing awareness, etc. can be empirically founded. Moreover, the model can rationalize heterogeneity in choice behavior even under the testable hypothesis that agents' preferences and meta-preferences are identical.

Keywords: Revealed Preference Theory, Reason-Based Choice, Endogenous Preferences, Awareness, Inattention, Changing Tastes

JEL classification codes: D01, D60, D90

[*] First draft: May 19th 2020. This draft: May 19th 2020. We are grateful to Franz Dietrich, Douglas Bernheim, Jean-Marc Tallon and seminar participants at PSE, Stanford and TUS-VI for helpful conversations and comments.

[†] Bielefeld University. Email: niels.boissonnet@gmail.com

[‡] Paris School of Economics, Paris 1 Panthéon-Sorbonne. Email: a.ghersen@gmail.com

[§] Paris School of Economics, Paris 1 Panthéon-Sorbonne. Email: gleyze.simon@gmail.com

1 Introduction

The objective of this paper is to make progress toward a *testable* model of preference change. Economic models incorporating preference changes have typically been criticized for their lack of empirical power[1]—e.g. theories of endogenous preferences, human capital, intergenerational transmission of traits, motivated reasoning, etc. To address this problem, we propose a model of deliberate preference changes that is sufficiently structured to be falsified yet quite general. A preference change is *deliberate* if: (1) the decision maker (DM) is *aware* of what has triggered her change of behavior; (2) DM considers that her new behavior is "*better*" than her previous one.

Deliberate preference changes are essential to understand the emergence of new political preferences and consumption behaviors as well as lobbying or activism. For instance, legal grounds for abortion have expanded in a growing number of countries since the 1990's, a phenomenon that is hardly explained by belief updating or psychological biases and heuristics. More plausibly, the emergence of new values such as women rights together with the increased labor force participation of women lead to important changes in our distributional preferences regarding labor income, labor supply and bargaining power in the household, etc. Similarly, deradicalization programs rely on the assumption that individuals can deliberately modify their preferences. By involving the subjects in questioning the drivers of their behaviors, these programs help them willfully adopt nonviolent views.[2] More generally, we want to understand whether some choice reversals are the result of a deliberate behavior change, as opposed to a "mistake" or non-rational behavior.

We propose a model of deliberate preference changes together with conditions on choice data that falsify it. At a given period, DM makes her choices by maximizing her preferences on the alternatives. Each alternative is defined by a set of attributes. DM's preferences are induced by the subset of attributes she deems *relevant* to compare alternatives. Between two periods, DM can *deliberately* modify her preferences by changing her set of *relevant* attributes. Formally, such a change is triggered by (1) the *awareness* of new attributes, and (2) the *maximization* of an ordering on preferences themselves—we call this ordering the *meta-preference*. We now describe in more details each component of the model.

[1] We also use the expression "empirical content" of a model to refer to the collection of datasets that falsify a given model.

[2] See Grune-Yanoff and Hansson (2009) for a review of evidence of preference changes.

Preferences are represented by (i) an ordering on the alternatives' attributes—we call it the *attribute ordering*—, and (ii) DM's set of *relevant* attributes.[3] Let us consider a simple labor market example: when asked "Why did you hire candidate x over candidate y?" an employer would explain that x is more experienced, or that y is too shy for this job. Hence, the employer justifies her choice invoking each candidate's set of attributes.[4] Typically, DM is only comparing alternatives through the lens of a subset of attributes that are *relevant* for her decision. Hence, an attribute is irrelevant if DM does not take it into account to rank alternatives. The employer might for instance overlook the candidates' gender for her hiring decision. Importantly, we do not impose that DM is fully conscious that some attributes are relevant: in the above example, race may be relevant to explain the employer's choice even if she does not *consciously* discriminate against Black candidates.

Preference changes take the following form: whenever DM becomes *aware* of an attribute — through education, social interactions, medias or introspection — she can decide to make it relevant or irrelevant for the next period, inducing a preference change. For instance, if the employer becomes aware of the discrimination against minorities (making her aware of the attribute race) and if she does not want to be racist in her hiring decision, she can deliberately ignore race in her future choices by making it irrelevant. Note that the attribute ordering remains stable, only the set of relevant attributes changes; which implies that if DM deems relevant the same set of attributes from one period to another, she must make exactly the same choices. We investigate when a succession of such preference changes is consistent with the maximization of a *meta-preference* relation, capturing DM's moral values, social objectives, norms, etc. See Figure 1 for a high-level representation of the model.

Note that the *constraint* of changing awareness on the meta-choice is a key feature of the model. Would DM be unconstrained in the maximization of the meta-preference, she would directly reach her most preferred set of relevant attributes at birth. Moreover, the constraint gives rise to interesting path-dependent dynamics of preference changes.

We assume that the analyst observes a choice correspondence on all menus in each period, and the set of all attributes the objects possess. The analyst

[3] Perhaps disappointingly, the model is agnostic as to *how* DM evaluates these attributes. For instance, DM could evaluate actions based on their consequences (i.e. consequentialism). Alternatively, actions could be compared based on whether their characteristics obey some rules (i.e. deontological ethics).

[4] There is a connection between what we call attributes and the philosophical works on reasons; see Dietrich and List (2013a).

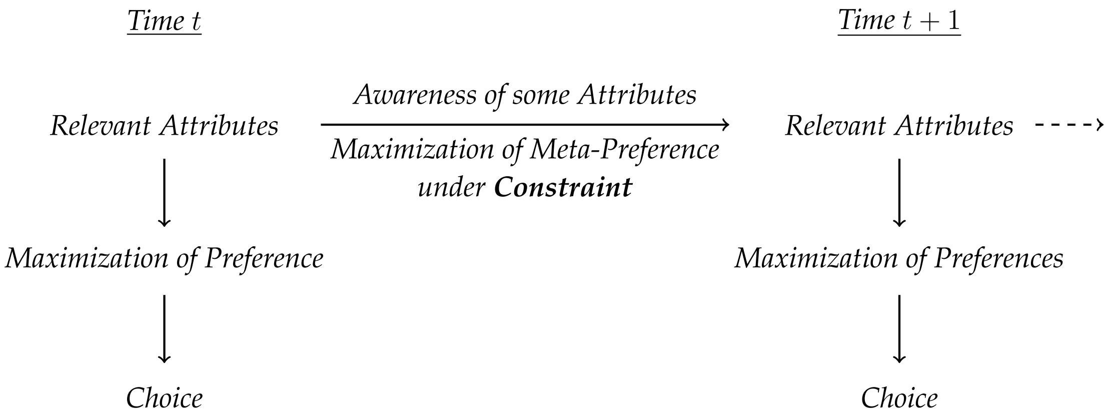
Time t
Time $t+1$
Awareness of some Attributes
Relevant Attributes
Maximization of Meta-Preference under **Constraint**
Relevant Attributes
Maximization of Preference
Maximization of Preferences
Choice
Choice

Figure 1: *The Dynamics of Deliberate Preference Change.*

does not observe (and wants to identify) DM's preferences on the alternatives, the sequence of relevant attributes, the sequence of awareness and the meta-preference. We also consider the case where the awareness is observed[5]—which is reasonably feasible in experiments—as it substantially increases the empirical content of the model.

The falsification exercise is difficult because of an *indeterminacy problem*: we can always rationalize an indifference between two alternatives either by an indifference of the preference relation itself (i.e. the attribute ordering), or by the fact that the attributes that differ between the two alternatives are irrelevant (but had DM thought they were relevant, her preference would be strict). For instance, take two employers whose choices reveal an indifference between Black and non-Black candidates (ceteris paribus). As an observer, we cannot distinguish whether it is caused by an indifference of the attribute ordering or by the irrelevance the attribute race. We say that race is *not revealed relevant*. This is important because the two scenarios may lead to different preference change. Suppose that the first employer considers race *relevant* but is indifferent toward this attribute (e.g. because she thinks race has no impact on productivity), whereas race is not relevant for the hiring decision of the second employer (e.g. as a matter of principle, she does not use race in her hiring decision). If they both become aware of an affirmative action policy in favor of minorities, they may react very differently. For example, the first one may integrate this new attribute to promote Black candidates in the future, whereas the

[5] In this case, we argue that the awareness can be observed when some attributes are explicitly made *salient* to DM.

second employer may not change her behavior because she thinks race is irrelevant. Therefore the indeterminacy problem relates to the under-identification of DM's preferences that we overcome in our model by observing the dynamics of preference change.

Nevertheless, we show that despite this indeterminacy the model has a lot of structure due to the *time-independence* of the attribute ordering. This requirement of *stability* implies that preference changes occur only if DM modifies which attributes are relevant. Therefore, by observing DM's choices at each period, the analyst accumulates some partial knowledge about the attribute ordering, and the subsequent choices must be consistent with this knowledge. This constraint is a key feature of our model. Without this restriction, an observer would have too many degrees of freedom to rationalize a preference change, hence making the model non-falsifiable. This constraint is imposed at the expense of some explanatory power. Note however that our model does not prevent choice reversals—even multiple times—, but it requires such reversals to be triggered by a complementarity with new attributes DM becomes aware of. For instance, if the employer prefers the non-Black candidate and then the Black one, the awareness of a new attribute such as an affirmative action must explain this change.

Our contribution is twofold: we show that models incorporating changing preferences can have empirical content.[6] Moreover, our model rationalizes heterogeneity in choice behavior even under the testable assumption that agents' preferences on attributes and meta-preferences on relevant attributes are identical. Indeed, due to the constraint of awareness, the set of relevant attributes at period t is *path-dependent*. Hence if two identical decision makers change awareness on the exact same attributes but in a different order, their choice behavior and how they justify it will typically differ in the end.

It should be noted that our approach is *complementary* to a model of belief updating on some underlying state. Throughout we rationalize DM's behavior as if she only changes awareness and not beliefs. This distinction is clearly described in Dekel et al. (1998): "an 'uninformative' statement—such as 'event x might or might not happen'—can change the agent's decision." Strictly speaking, we refer to this kind of statement when we mention *changing awareness* (or sometimes *salience*) about the attributes. In Section 4 we elaborate on how one could integrate beliefs in our model.

[6] This may include models of chosen preferences, endogenous preferences, motivated reasoning, evolving attention, changing awareness, etc., depending on the interpretation of the objects of our model.

The idea of representing objects by their attributes goes back to Lancaster (1966). Moreover we draw on an important literature in philosophy and decision theory introducing reason-based theories of choice, most notably Simonson (1989), Shafir et al. (1993), and Tversky and Simonson (1993). We also draw upon Dietrich and List (2013a, 2016)'s model of reason-based choice. In particular, we use their characterization to link DM's set of relevant attributes with her preferences on alternatives. Boissonnet (2019) provides a decision theoretic characterization of our model, and Dietrich and List (2013b) propose a related theory of non-informational preference change. Our paper should be seen as the first counterpart of these models within the revealed preference theory.

A recent literature on behavioral revealed preference theory relaxes the assumption of stability of preferences across menus by introducing context dependence. Kalai et al. (2002), Manzini and Mariotti (2007) and Cherepanov et al. (2013) consider rationalization by multiple linear ordering (possibly applied sequentially). Salant and Rubinstein (2008) and Bernheim and Rangel (2009) explicitly introduce a context variable attached to menus. Masatlioglu et al. (2012) consider a decision maker who may not be fully attentive of all alternatives in a menu. De Clippel and Eliaz (2012) introduce a model of intrapersonal bargaining between different selves. Bernheim and Rangel (2009), Chambers and Hayashi (2012), Apesteguia and Ballester (2015) and Nishimura (2018) investigate robust welfare analysis when the decision maker may not exhibit standard rational preferences. Ok et al. (2015) provide a revealed preference theory for reference dependent choices. Instead of investigating choice inconsistencies *across menus*, our model focuses on explaining inconsistencies *across time*. Hence these models typically explore relaxations of WARP, whereas our representation imposes within-period WARP with respect to the attribute ordering, but weaker consistency requirements between periods.

We also emphasize that there is an important literature on "changing tastes" understood as time inconsistency. Strotz (1955) is the first to uncover the problem of consistent planning and to investigate how should individuals with non-exponential discounting make dynamically consistent choices. Peleg and Yaari (1973) propose a solution based on Markov Perfect equilibrium play against one's future selves. Gul and Pesendorfer (2001, 2005) and Dekel et al. (2009) provide behavioral foundations of preferences for commitment, namely choosing a smaller choice set for one's future self to avoid temptation. Sarver (2008) provides an alternative representation of preference for commitment based on regret aversion. The main differences with our paper is that they consider departures between expected behavior and actual behavior which is typically *not* *deliberate* (inconsistent) from the point of view of past selves. Instead, we look at preference changes that are deliberate but completely myopic, meaning that DM is unaware of what may change in the distant future.

In the applied theory literature, the closest paper is Bernheim et al. (2019). Their model and ours share two important ideas. First, they argue that DM can choose "worldviews" which determine her valuation of future consumption streams. This is related to our concept of relevant attributes. Second, in their model DM is constrained by her "mindset flexibility" when changing worldviews. This echoes our constraint on awareness. Despite the differences in modelling assumptions, their paper is largely complementary with ours as we focus on the identification and falsification of deliberate preference changes. Other theories of chosen preferences include Becker and Mulligan (1997), Akerlof and Kranton (2000), and Palacios-Huerta and Santos (2004).

2 The Model

There is a finite number of time periods $t = 1, \dots, T$ and a finite set of **binary attributes**[7] $\mathcal{M}$. Denote $\mathscr{M} = 2^{\mathcal{M}}$ the power set of attributes. An **alternative** is defined as the subset of attributes it satisfies. Let $X \subseteq \mathscr{M}$ be the finite set of alternatives. Denote $K \subseteq X$ a **menu**, and let $\mathscr{K} = 2^X \setminus \emptyset$ be the set of all non-empty menus. A dataset is represented by a **choice correspondence** for each period $C_t : \mathscr{K} \longrightarrow \mathscr{K}$ such that $C_t(K) \subseteq K$ for all $K \in \mathscr{K}$.[8] In the baseline model, the analyst only observes the set of alternatives X and the dataset $(C_t)_t$.

Assumption 1. *(Complete Dataset) We observe $C_t(K)$ for each menu $K \in \mathscr{K}$ at each t.*

Moreover, we assume that we can always find an object which satisfies any given subset of attributes.[9]

[7] A binary attribute is a property that the object either has or does not have. For instance, "color" is not a binary attribute but "red" is a binary attribute. Continuous attributes such as age must be divided in finitely many intervals. Whether the objects belongs to each interval is then a binary attribute.

[8] It is important that we observe a choice correspondence instead of a choice function because most of the axioms rely on the evolution of indifference classes. See Bouacida (2019) for an experimental method to elicit correspondences. An alternative approach would be to observe a choice function for each pair of alternatives, with a third option in case DM is indifferent.

[9] This assumption can be relaxed as long as the set of alternatives has a nested structure—e.g. X can be a lattice.

ASSUMPTION 2. *(Perfect Instantiation) We have $X = \mathscr{M} \setminus \emptyset$.*

Denote by $M_t \in \mathscr{M}$ the set of **relevant attributes** at period t: this is the set of attributes that determine DM's choice behavior at period t. Importantly, DM may not be conscious of the relevant attributes. For instance, values such as paternalism, racism, or sexism, may impact her choice behavior without her realizing it.

Although this is not the only interpretation, our model naturally embeds inattention to products' characteristics. In this case the set of relevant attributes would be DM's *consideration set* regarding the alternatives' attributes.

Application 1: Labor Market Discrimination. *An employer wants to hire a worker. Workers are represented by the following attributes: (m_1) "college educated" or not, (m_2) "experienced" or not, and (m_3) "Black" or not. The employer's decision is based on all attributes, hence the set of relevant attributes is $M_t = \{m_1, m_2, m_3\}$.*

Denote $\succsim_{M_t} \subseteq X^2$ the **preference relation** on alternatives induced by the set of relevant attributes M_t at period t. We need structure on how preferences are linked to relevant attributes, otherwise the model cannot be falsified. Put differently, DM's choices must be at least *partially informative* on the relevant attributes. We adopt the following representation: for any alternatives $x, y \in X$ and relevant attributes M_t,

$$x \succsim_{M_t} y \iff x \cap M_t \geqslant y \cap M_t.$$

where $\geqslant \subseteq \mathscr{M}^2$ is an **attribute ordering**, which is an unrestricted weak order (whose symmetric part is denoted $\simeq$). Dietrich and List (2016) provide a behavioral characterization for this attribute ordering. Importantly, this relation *need not* agree with the inclusion ordering or the cardinality ordering, i.e. rank higher an alternative that has more attributes. For instance, an employer may prefer a candidate who *does not have* the attribute "Black", because she implicitly associates it to lower productivity due to statistical discrimination. Note also that the attribute ordering can capture various forms of complementarities between the attributes. If the employer thinks that human capital depreciates with age, she may prefer a young educated candidate to an older educated candidate. At the same time, she may think that without any diploma experience is crucial, in which case she would prefer an old uneducated to a younger uneducated.

Throughout, we assume that the attribute ordering is **stable**, i.e. it is *time independent*. Therefore choice reversals must be explained by DM changing which attributes are relevant. In particular, DM cannot change her "attitude" toward an attribute—for instance, because she learns new consequences of the attribute—while keeping the attribute relevant or irrelevant. Preference changes must be justified by making relevant or irrelevant an attribute, and this is an all-or-nothing decision. This is arguably a strong assumption, though the added structure is essential for the representation.

Application 1 (continued): *Ceteris paribus, the employer prefers non-Black candidates to Black ones (attribute m_3). Hence conditional on any subset of attributes, Black candidates are ranked lower according to the attribute ordering: $\{m_1, m_2\} > \{m_1, m_2, m_3\}$, $\{m_1\} > \{m_1, m_3\}$ and $\{m_2\} > \{m_2, m_3\}$.*

Between each period, DM becomes *aware* of a set of attributes $A_t \in \mathscr{M}$—through education, social interactions, medias, etc. This event enables DM to modify the set of relevant attributes inducing a behavior change at the next period. The constraint that DM is able to modify the relevant attributes only after changing awareness is an essential element of the representation as otherwise she would directly reach her most preferred set of relevant attributes and we could not observe variations in her choices. Hence, we impose the following constraint on meta-choice: DM can only adopt or reject attributes in A_t.

Formally, given the *relevant attributes* M_t and the *awareness* A_t, the set of **reachable attributes** is:

$$R(M_t, A_t) \equiv \{M \in \mathscr{M} : M_t \setminus A_t \subseteq M \subseteq M_t \cup A_t\}.$$

This can be interpreted as a "menu" for the meta-choice which consists of all subsets of attributes that can be obtained from M_t by adding or removing attributes DM becomes aware of. DM deliberately changes her preferences when she maximizes a **meta-preference** relation $\rhd \subseteq \mathscr{M}^2$ on the reachable attributes. In Section 4, we discuss the multiple interpretations of the meta-preference relation. At this point, it is useful to simply interpret $\rhd$ as DM's preference for consistency in the justifications of her behavior with respect to a set of values or norms. In our labor market example, suppose the employer is aware of the attribute "Black" and realizes that her hiring decision is discriminatory. If she considers this behavior inappropriate, she may want to remove this attribute from the relevant ones for her future decisions.

DEFINITION: WEAK DELIBERATE PREFERENCE CHANGE. *$(C_t)_t$ satisfies Weak Deliberate Preference Change if there exists a stable (i.e. time-independent) weak order $\geqslant$, a sequence of relevant attributes $(M_t)_t$, a sequence of awareness $(A_t)_t$ and a strict order $\rhd$, such that, for any menu $K \in \mathscr{K}$, and for any t,*

$$\begin{aligned} &C_t(K) = \max(\{x \cap M_t : x \in K\}, \geqslant). \\ &M_{t+1} = \max(R(M_t, A_t), \rhd). \end{aligned}$$

Rationalization by *Strict* Deliberate Preference Change is defined similarly, except that the attribute ordering $\geqslant$ must be strict, in which case we denote it $>$. As one might expect, the identification with a strict attribute ordering is substantially easier than with a weak attribute ordering. The latter is quite challenging because the model is *under-determined:* we can explain an indifference between two alternatives either because DM is indifferent between their attributes (i.e. indifference with respect to $\geqslant$), or because the attributes that differ between the two alternatives are irrelevant (i.e. attributes that differ are not in M_t). Clearly, if we impose that the attribute ordering is strict this problem does not arise.

Application 1 (continued): *DM is aware of the attribute "Black" between t and $t+1$. She does not want to discriminate against Black candidates, hence she makes race irrelevant for her future self. Formally, $A_t = \{m_2\}$ and $M_{t+1} = M_t \setminus A_t \rhd M_t$.*

In the next section we consider the point of view of an analyst who observes a sequence of choice correspondences $(C_t)_t$. We address the following questions: (1) Which choice correspondences are compatible with the model of Deliberate Preference Change? (2) How can we identify DM's preference, meta-preference, changing awareness, and relevant attributes through her choice? The objects that are observed by the analyst and those that need to be identified are summarized in the following table.

What the Analyst Observes	*What the Analyst Does not Observe*
Set of alternatives $X = \mathscr{M}$	Relevant Attributes $(M_t)_t$
Choice Correspondence $(C_t)_t$	Salient Attributes $(A_t)_t$
	Attribute Ordering $\geqslant$
	Meta-Preference $\trianglerighteq$

3 Representation Theorems

In this section we investigate the indeterminacy problem and we provide two representation theorems. This problem relates to the fact that, in our model, indifferences can be rationalized in two ways. It raises technical difficulties because verifying that preference changes are consistent with the maximization of a meta-preference usually requires to identify *a unique* sequence of relevant attributes. The first representation theorem by-passes this problem altogether by assuming a strict attribute ordering—in this case the sequence of relevant attributes is unique. This representation is simple and provides a new explanation of indifference in standard decision theory: DM is indifferent between two alternatives if and only if the attributes by which they differ are irrelevant. The second representation accommodates a *weak* attribute ordering. Such a representation is much richer and allows for "background attributes" that only impact choice *indirectly* through preference change. This model, however, faces the problem of indeterminacy, hence we provide conditions on partially identified objects which are typically not unique.

3.1 The Indeterminacy Problem

The indeterminacy problem relates to the fact that an indifference can be rationalized either by an indifference of the attribute ordering, or by making irrelevant the attributes that differ between the alternatives. For instance, an employer may be indifferent between a Black and a non-Black candidate (ceteris paribus) because she thinks race has no impact on productivity, or because she does not use race for her hiring decision (even though race could have an impact on productivity).

To define this problem more formally, we first need conditions that guarantee the existence of a transitive attribute ordering $\geqslant$ at each period. This is obtained by imposing the Weak Axiom of Revealed Preferences (WARP) at each period.

Weak Axiom of Revealed Preferences. *The choice correspondence C_t satisfies WARP at period t if for every $K, K' \in \mathscr{K}$ such that $K' \subset K$,*

$$C_t(K) \cap K' \neq \emptyset \implies C_t(K) \cap K' = C_t(K').$$

This axiom guarantees the existence of *at least* one set of relevant attributes

M_t such that DM's behavior can be represented by the maximization of an attribute ordering $\succcurlyeq_t$ together with this set.[10] Note that $\succcurlyeq_t$ is possibly time-dependent here, and latter we will impose extra conditions to guarantee its stability.

Let us illustrate formally the indeterminacy problem with three attributes $\mathcal{M} = \{m_1, m_2, m_3\}$ (see Figure 2). If C_t satisfies WARP, then DM's choices can be represented using indifference classes. Say that there are two indifference classes: $\{x_{123}, x_{12}, x_{13}, x_1\}$ and $\{x_{23}, x_2, x_3\}$ where the subscript on each alternative describes its attributes. This implies that, irrespective of the attribute ordering, the attribute m_1 *must* be relevant as otherwise DM could not possibly exhibit strict preference between x_{123} and x_{23}. Call such an attribute **revealed relevant**. In the example, we observe that only m_1 is revealed relevant because it is the attribute that always differs between any pair of alternative from each indifference class. Therefore, the set of relevant attributes $M_t = \{m_1\}$ together with the attribute ordering $\{m_1\} \not\preccurlyeq \emptyset$ are candidates to rationalize DM's choice. Notice, however, that the grand set $\mathcal{M}$ *could also* be used to rationalize DM's behavior as we can fine tune the attribute ordering to reproduce the indifference classes. That is, $M_t = \{m_1, m_2, m_3\}$ and the attribute ordering $\{m_1, m_2, m_3\} \simeq \{m_1, m_2\} \simeq \{m_1, m_3\} \simeq \{m_1\} \not\preccurlyeq \{m_2, m_3\} \simeq \{m_2\} \simeq \{m_3\}$ can *also* rationalize DM's behavior. We then face an indeterminacy problem: DM's choice behavior does not permit complete identification of the relevant attributes and the attribute ordering. This is problematic to find axioms that guarantee various form of consistencies between periods, as the *reachable attributes* (i.e. the *meta-menu*) depends on the set of *relevant attributes*.

The identified set, fortunately, has a lot of structure. We show that any superset of the *revealed relevant attributes*—here any superset of $\{m_1\}$—is a possible candidate for the set of relevant attributes. To prove this formally, the following notation will prove useful: for any attribute $m \in \mathcal{M}$ and for any alternative $x \in X$ such that $m \notin x$, denote $x + m \in X$ the alternative whose attributes are $x \cup \{m\} \in \mathscr{M}$.[11] Given a dataset C_t, the attribute m **is revealed relevant** at t if for some x, $C_t(\{x, x+m\}) \neq \{x, x+m\}$.[12] Namely, the attribute m makes a difference in DM's choices. Denote $\underline{M}_t$ the set of **revealed relevant attributes** at t: these are the attributes that are necessarily relevant to explain DM's behavior.

[10] Slightly abusing notation, we will use M_t to denote a *candidate* set of relevant attributes at a given period t, not necessarily DM's "true" set of relevant attributes.

[11] The alternative $x + m$ always exists by Assumption 2 (Perfect Instantiation).

[12] The menu $K = \{x, x+m\}$ always exists by Assumption 1 (Complete Dataset).

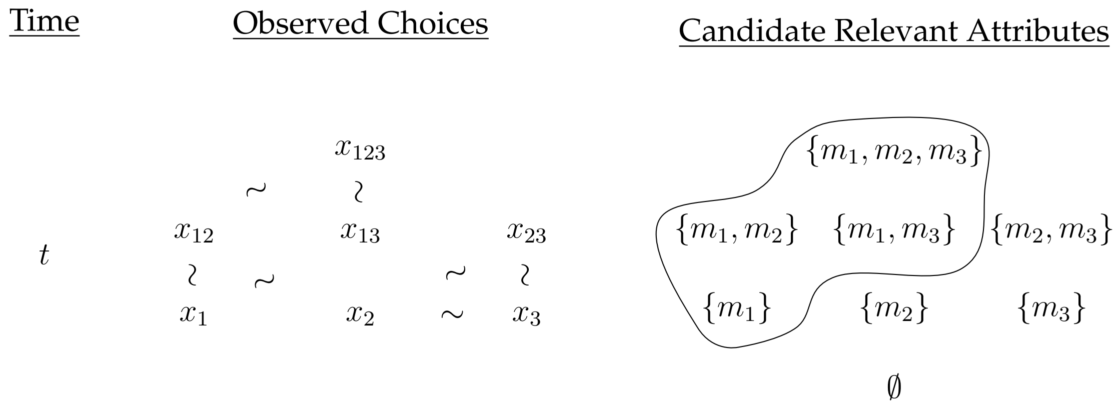

Figure 2: *An example of indeterminacy. There are two indifference classes (column 2) and four sets of relevant attributes (together with a specific attribute ordering) that can rationalize DM's behavior (column 3).*

Proposition 1 shows that: (1) WARP is necessary and sufficient for a choice correspondence to be rationalized by a set of relevant attributes together with an attribute ordering; (2) the collection of candidates of relevant attributes is a lattice.

Proposition 1 (Characterization: Identified Set)**.** *The choice correspondence C_t satisfies WARP at period t if and only if the collection of sets of relevant attributes M_t which rationalize C_t (together with an attribute ordering $\geqslant_t$) is a non-empty lattice denoted $M(C_t)$ (ordered by inclusion). Its infimum is $\underline{M}_t$ and its supremum is $\mathcal{M}$.*

It is sometimes important to rationalize DM's behavior at period t by a set of relevant attributes M_t which is strictly bigger than the *revealed relevant* attributes, that is to have $M_t \supset \underline{M}_t$. The attributes $M_t \setminus \underline{M}_t$ do not have have a direct impact on choice at t, but they impact how DM changes preferences. Hence attributes in $M_t \setminus \underline{M}_t$ are referred to as **background attributes**. For instance, consider an employer who is indifferent toward race—e.g. because she thinks it has no impact on productivity. Still, the employer thinks that race is relevant because Black are known to be discriminated against on the labor market. Suppose that the employer becomes aware of affirmative action policies that favor minorities at several colleges, thus reducing the informativeness of the attribute "college educated" for Black candidates. The employer does not want to discriminate against minorities, hence she decides to make education irrelevant altogether. In this example, race acts as a *background attribute* because hiring decisions are insensitive to race throughout, but it impacts how the employer modifies her behavior towards all candidates.

Despite its importance for a positive description of behavior, this indeterminacy challenges the falsification exercise of *Weak Deliberate Preference Change*. Hence, before turning to this issue, we study a more restrictive model in which we impose the attribute ordering to be strict. In this case, the unique set of relevant attributes that rationalize C_t is $\underline{M}_t$ (Proposition 2). In the next section, we provide our first representation theorem which assumes a strict attribute ordering. The second representation theorem relaxes this assumption, hence it accommodates background attributes.

3.2 Strict Deliberate Preference Change

The representation consists of four axioms which guarantee various forms of consistency. The set of revealed relevant attributes $\underline{M}_t$ is uniquely identified by choice data. Hence it will prove convenient to write some axioms directly using this set. Obviously, we could have equivalently written the axioms in terms of choice by substituting the definition of $\underline{M}_t$.

We first characterize, with an additional axiom, when we can rationalize a choice correspondence by a strict attribute ordering, in which case the set of relevant attributes must be $\underline{M}_t$. WARP is indeed not sufficient to guarantee the strictness of the attribute orderings $(\geqslant_t)_t$. The latter is obtained if DM is not indifferent between any pair of alternatives that differ by at least one *revealed relevant* attribute. Conversely, DM will be indifferent between two alternatives only if these alternatives differ by attributes that are *not revealed relevant*.

JUSTIFIED INDIFFERENCE. *The choice correspondence $(C_t)_t$ satisfies Justified Indifference at period t if for any alternatives $x, y \in X$,*[13]

$$C_t(\{x,y\}) = \{x,y\} \implies (x \triangle y) \cap \underline{M}_t = \emptyset.$$

Together with WARP, Justified Indifference characterizes the rationalization of a choice correspondence by a strict attribute ordering and identifies the set of *revealed relevant attributes* $\underline{M}_t$ as the unique possible set of relevant attributes.

Proposition 2. *The dataset C_t satisfies WARP and Justified Indifference at period t if and only if the set of revealed relevant attributes $\underline{M}_t$ is the unique set which rationalizes C_t together with a strict attribute ordering $>_t$.*

[13] Recall that $x \triangle y = (x \cup y) \setminus (x \cap y)$ are the attributes that belong to x or y but not both. Hence $(x \triangle y) \cap \underline{M}_t$ are the *revealed relevant* attributes that *differ* between x and y.

The stability (across periods) of the attribute ordering is obtained by observing that choices must be consistent between alternatives that are *equally relevant* between t and t'. A useful analogy is that, when alternatives are "perceived"[14] equivalently through the lens of the set of *revealed relevant attributes*, the choice must be the same. Conversely, if we observe choice reversals between t and t' it must be that at least one attribute of the alternative was relevant at t and irrelevant at t' (or vice versa). Therefore, choice reversals are *only* due to changes in the attributes DM deems relevant.

BETWEEN CONSISTENCY. *The dataset $(C_t)_t$ satisfies Between Consistency if for any t, t', and for any $x, x', y, y' \in X$ such that $x \cap \underline{M}_t = x' \cap \underline{M}_{t'}$ and $y \cap \underline{M}_t = y' \cap \underline{M}_{t'}$,*

$$x \in C_t(\{x, y\}) \iff x' \in C_{t'}(\{x', y'\}).$$

Finally, DM's choices must be compatible with the maximization of a meta-preference constrained by changing awareness. Characterizing this aspect is challenging because (1) we do not observe the meta-choices, and (2) only DM observes and controls the sequence of meta-menus.

At first sight, it might be tempting to simply impose that the sequence of choice correspondences is acyclic. Namely, that there are no $t < t' < t''$ such that $C_t \neq C_{t'}$ and $C_{t''} = C_t$. This condition is necessary though not sufficient. For instance, consider the following sequence of revealed relevant attributes: $\underline{M}_1 = \{m_1\}$, $\underline{M}_2 = \{m_1, m_2, m_3\}$, and $\underline{M}_3 = \{m_1, m_2\}$. We know (from Proposition 2) that the sequence of choice correspondences is acyclic. Such a sequence, however, requires DM to become aware of attributes m_2 and m_3 between the first and the second period so that $\underline{M}_2 \subseteq \underline{M}_1 \cup A_1$. But then, $\{m_1, m_2\} = \underline{M}_3$ was also accessible in the first period, which contradicts the fact that DM is maximizing a meta-preference.

A necessary and sufficient condition is that, whenever we observe "cycles" of relevant attributes, other (complementary) attributes must have become relevant as well. For instance, if we observe $\underline{M}_1 = \{m_1\}$, $\underline{M}_2 = \{m_2\}$, and $\underline{M}_3 = \{m_1, m_2, m_3\}$ then m_1 is forming a "cycle". Though, this is compatible with the maximization of the meta-preference because m_1 appears to be complementary with m_3, and DM's awareness can be represented *as if* m_3 was not in the awareness of the first period.

[14] Here "perception" should not be interpreted literally but as a simple way to to illustrate how the set of relevant attributes acts on an alternative: $M_t \cap x$.

Discovery Constrained Cycles. *The dataset $(C_t)_t$ satisfies Discovery Constrained Cycles if for any t, t', $t' > t+1$: $\underline{M}_{t'} \neq \underline{M}_{t+1}$ implies that for any $\tau \geq t'$,*

$$\underline{M}_t \triangle \underline{M}_\tau \not\subseteq \underline{M}_t \triangle \underline{M}_{t+1}.$$

If this condition is satisfied, we can always find a sequence of changing awareness $(A_t)_t$ such that DM's behavior can be rationalized by the maximization of a meta-preference. Importantly, note that $(A_t)_t$ need not be "growing," namely we need not have $A_t \subseteq A_{t+1}$ for all t. This is consistent with a decision maker who does not keep track of everything she has ever been aware of, hence optimizing her behavior only with what she is *currently* aware of. Even though growing awareness seems intuitive, it may not best describe our thought processes—e.g. it is well-known that the number of issues that drive a political campaign is quite limited. Ultimately, the plausibility (or lack thereof) of growing awareness is an empirical question.

The four axioms are necessary and sufficient for the representation and Proposition 2 identifies $(\underline{M}_t)_t$ as the unique possible sequence of relevant attributes.

Theorem 1 (Characterization: Strict DPC)**.** *The dataset $(C_t)_t$ is rationalizable by Strict Deliberate Preference Change if and only if it satisfies WARP and Justified Indifference at each period t, Between Consistency and Discovery Constrained Cycles. Moreover, $(\underline{M}_t)_t$ is the unique sequence of relevant attributes that rationalizes the dataset.*

3.3 Weak Deliberate Preference Change

We now allow the attribute ordering to be a weak order. This is important to capture the effect of "background attributes" on DM's preference changes. Think of the previous example of two employers whose choices reveal an indifference toward race. Say that they become aware of preferential treatment for minorities in college admissions, thus reducing the informativeness of the attribute "college educated" for Black candidates. Suppose that the observed indifference is due to an indifference in the attribute ordering, but that "race" is relevant to the first employer—i.e race is a background attribute. She may then decide to make the attribute "education" irrelevant to avoid discriminating against minorities. This impacts all candidates, and the hiring decision remains independent of race. On the contrary, if the observed indifference is due to the irrelevance of "race", the second employer may now decide to make race relevant and discriminate against educated Black candidates. This impacts only black candidates, and the hiring decision that was independent of race now depends on this attribute. Therefore, seemingly identical employers may react very differently due to the presence of background attributes. The main objective of this section is to understand how background attributes impact revealed preferences, and how we can adapt the previous axioms accordingly.

The previous axioms are neither necessary nor sufficient due to the indeterminacy problem. Let us illustrate the previous axioms with an example and see why they fail here (see Figure 3). Assume that C_t and C_{t+1} satisfy WARP. At time t, DM's choices can be represented by two indifference classes, whereas, at $t+1$, they are represented by four indifference classes. These indifference classes imply that the revealed relevant attributes are $\underline{M}_t = \{m_1\}$ and $\underline{M}_{t+1} = \{m_1, m_2\}$. Due to the indeterminacy problem, any superset of $\underline{M}_t$ and $\underline{M}_{t+1}$ can be used to represent DM's choices (see Proposition 1).

It is easy to see that Justified Indifference is not necessary here: the alternatives x_{12} and x_{23} differ by a revealed relevant attribute at time $t+1$, $(x_{12} \triangle x_{23}) \cap \underline{M}_{t+1} = \{m_1\} \neq \emptyset$, but we can rationalize $C_{t+1}(\{x_{12}, x_{23}\}) = \{x_{12}, x_{23}\}$ (which violates Justified Indifference) with $M_{t+1} = \{m_1, m_2, m_3\}$ and $\{m_1, m_2\} \simeq \{m_2, m_3\}$ because we allow for indifference. Conversely, Between Consistency is not sufficient: at period t, the axiom applied to the revealed relevant attribute $\underline{M}_t = \{m_1\}$ only puts restriction on the stability of $\{m_1\} \not\simeq \emptyset$. Instead, if the "truly" relevant attributes at t were $M_t = \{m_1, m_3\}$ (i.e m_3 is a background attribute), we should put restrictions on the stability of $\{m_1\} \simeq \{m_1, m_3\} \not\simeq \{m_3\} \simeq \emptyset$.

The indeterminacy introduces a "combinatorial" aspect to the identification that significantly complicates our objective of finding axioms on choice that characterize the model. To solve this problem, we provide conditions on *partially* identified objects, i.e. on "candidate" changes of relevant attributes that are typically not unique.

As an observer, we want to find out whether *changes* in choice can be consistent with a *meta-maximization*. The indeterminacy problem, however, prevents us from keeping track of a unique sequence $(M_t)_t$ of relevant attributes. Therefore, we shall adopt a more general approach and ask: what are the attributes that must have changed between any two periods? Instead of working on sets of relevant attributes, we build our axioms on "candidate" *sets of changing attributes* to explain DM's behavior between any two periods t and t'. Such a candidate typically does not identify a unique sequence of relevant attributes $(M_t)_t$. This is desirable, as if we were to work directly on candidate sets of relevant attributes, then the conditions would trivially coincide with the definition

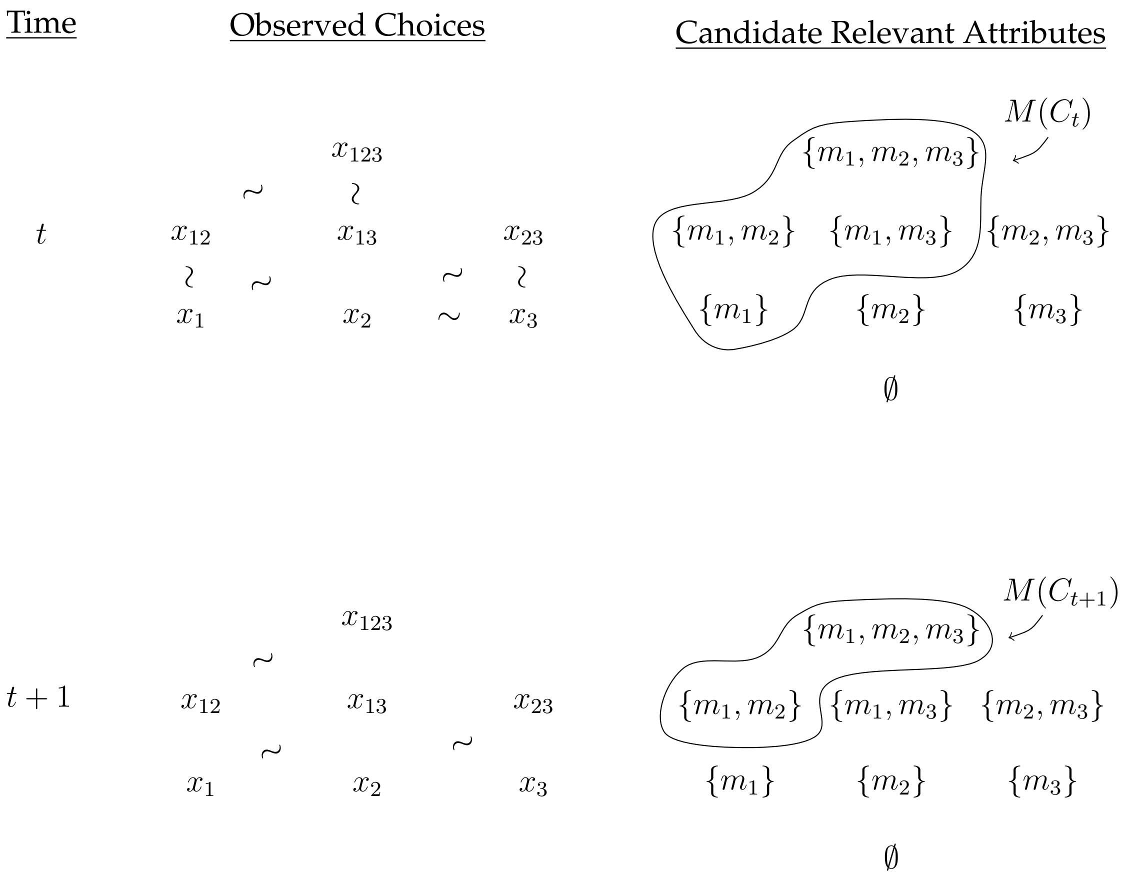

Figure 3: *An example of the non-necessity of Justified Indifference and of the non-sufficiency of Between Consistency.*

of the model.

Formally, we define an **explanation** $E = (E_{t,t'})_{t<t'}$ of the dataset $(C_t)_t$ as an upper triangular matrix whose element $E_{t,t'}$ represents a change in relevant attributes between period t and t'. That is, $E_{t,t'} = M \triangle M'$ for some $M \in M(C_t)$ and for some $M' \in M(C_{t'})$. Elements of the row t of this matrix correspond to relevant attribute changes between t and any $t' > t$. Importantly, an explanation is compatible with *multiple* sequences of relevant attributes $(M_t)_t$, hence it is a non-trivial exercice to find conditions on explanations that characterize the model. Conversely, an explanation is typically not unique given a dataset hence these conditions cannot be interpreted as axioms on choice directly.

A dataset together with an explanation are compatible with a stable attribute ordering if choices between alternatives that are *equally relevant* between

t and t' are consistent. Determining which attributes are relevant, however, is not straightforward due to the indeterminacy problem. In particular, if an attribute m is a background attribute for *all* periods then the analyst can never discover the preference ranking of this attribute. It follows that we should not impose any form of consistency in the ranking of alternatives that possess such an attribute. At the other extreme, the analyst directly observes the ranking of the attributes that are *revealed relevant* for all periods. Therefore we must impose consistency across periods of DM's choices with respect to these attributes; this implies that Between consistency is necessary. In-between, some background attributes are *sometimes* revealed relevant. Whenever these attributes are revealed relevant, the analyst observes DM's ranking on these attributes and therefore we must impose consistency of these preferences. For a given explanation E, let $V_{t,t'}$ be the background attributes at t that are revealed relevant at t':

$$V_{t,t'} = \{m \notin E_{t,t'} : m \text{ revealed relevant at } t' \text{ but not at } t\}$$

Formally, an attribute m is in $V_{t,t'}$ implies that m is relevant at t'. Because it is not in $E_{t,t'}$, it did not change between t and t'. This means that either m is also revealed relevant at t, or is a background attribute at t. In any case, it is relevant for the choice at both periods. Therefore, choices that are made involving this attribute must be consistent between t and t'.

The following condition extends Between Consistency to our new framework and guarantees stability of the attribute ordering with respect to (1) revealed relevant attributes, and (2) background attributes that are sometimes revealed relevant.

EXTENDED BETWEEN CONSISTENCY. *An explanation E of the dataset $(C_t)_t$ satisfies Extended Between Consistency if, for every t, t', and every $x, x', y, y' \in X$ such that:*

$$x \cap (V_{t,t'} \cup \underline{M}_t) = x' \cap (V_{t',t} \cup \underline{M}_{t'})$$
$$y \cap (V_{t,t'} \cup \underline{M}_t) = y' \cap (V_{t',t} \cup \underline{M}_{t'})$$

we have,

$$x \in C_t(\{x,y\}) \iff x' \in C_{t'}(\{x',y'\}).$$

We now impose some form of coherency on the analyst's explanation. The explanation should not exhibit "gaps" in the sense that any sequence of *local* changes of attributes between t and $t+1$, $t+1$ and $t+2$, ... until τ and $\tau+1$ should be consistent with the *global* explanation from t to $\tau+1$.[15] For instance, if m becomes relevant between t and $t+1$ in the analyst's explanation, and then becomes irrelevant between $t+1$ and $t+2$, then m cannot be used to rationalize DM's behavior from t to $t+2$.

No Explanatory Gap. *An explanation E of the dataset $(C_t)_t$ satisfies No Explanatory Gap if, for every $t < t' < t''$,*

$$E_{t,t'} \triangle E_{t',t''} = E_{t,t''}.$$

Finally, verifying that choices are consistent with the maximization of a meta-preference is captured by an acyclicity condition. If the analyst formulates a *global* explanation between t and $\tau > t+1$ that is included in a local explanation from t to $t+1$, then no attributes should change between $t+1$ and τ. If this were the case, the explanation would violate the maximization of a meta-preference relation.

Acyclic Explanation. *An explanation E of the dataset $(C_t)_t$ is Acyclic if, for all $\tau > t+1$,*

$$E_{t,\tau} \subseteq E_{t,t+1} \implies E_{t+1,\tau} = \emptyset.$$

If the dataset satisfies WARP and if we can find an explanation of the dataset that satisfy the above three conditions, then the dataset is rationalizable. The status of this representation theorem, however, is slightly different than Theorem 1 as conditions on explanations cannot be interpreted as axioms on choice directly.

Theorem 2 (Characterization: Weak DPC). *A dataset $(C_t)_t$ is rationalizable by Weak Deliberate Preference Change if and only if it satisfies WARP at every period t, and there exists an explanation E of $(C_t)_t$ which satisfies Extended Between Consistency, No Explanatory Gap and Acyclic Explanation.*

[15] There is a formal connection between our concept of explanation and the theory of *dynamic systems*. In a wide sense, a dynamic system is an arbitrary action of a group (say, the real numbers) on a set called the phase space. That the evolution follows from a *group action* means that the state of the system at any moment of time is fully determined by its initial state: for a given initial value, the evolution of the system during t periods followed by the evolution during t' periods is identical to the evolution of the system during $t+t'$ periods.

3.4 Representation with Observable Awareness

Throughout we assumed that the analyst only observes choices, hence the sequence of *salient* attributes[16] $(A_t)_t$ was part of the representation. Yet, in some circumstances, observing this sequence is a valid assumption, which strengthens the empirical content of the model. This can be achieved in a controlled environment where various attributes are made salient throughout the experiment. For instance, Chetty et al. (2009) show that consumers under-react to taxes that are not salient by posting tax-inclusive price tags in a grocery store. In such a context, we arguably observe a dataset consisting of choices *and* salient attributes $(C_t, A_t)_t$. This behavior could then be interpreted as a change of preference: the salience of taxes makes people question their choice and their willingness-to-pay.

Similarly in development economics, Dutta et al. (2014) implemented an "awareness intervention". The poor results of an employment scheme in India were possibly driven by the individuals' lack of knowledge about the existence and the functioning of the program. Thus, they tested the impact of raising their awareness about the program—simply by showing them a video clip explaining the details of the scheme. Again, we arguably observe a dataset consisting of choices *and* salient attributes $(C_t, A_t)_t$.

The representation of Weak Deliberate Preference Change is obtained by modifying the condition of Acyclic Explanation.

ACYCLIC EXPLANATION*. *An explanation E of the dataset $(C_t, A_t)_t$ is Acyclic* if for any $t < t' < t''$ such that $E_{t',t''} \neq \emptyset$,*

$$E_{t,t''} \not\subseteq A_t.$$

Moreover, we need an extra condition which guarantees that the explanation does not violate the constraint on meta-choice. Indeed, the representation requires DM to only add or remove *salient* attributes to modify her relevant attributes. Therefore, this excludes "associative memory" or "complementarity" between salient attributes and other attributes—for instance, we exclude that the salience of discrimination on race allows DM to change her behavior toward other (non-salient) attributes such as gender discrimination.

[16] In this section, we use the concept of salience instead of awareness as it seems more plausible to observe the former than the latter. The two notions are valid interpretations of $(A_t)_t$.

CONSTRAINED REACHABILITY. *An explanation E of the dataset $(C_t, A_t)_t$ satisfies Constrained Reachability if, for any t, $E_{t,t+1} \subseteq A_t$.*

Theorem 3 (Characterization: Weak DPC with Non-Choice Data)**.** *A dataset $(C_t, A_t)_t$ is rationalizable by Weak Deliberate Preference Change if and only if it satisfies WARP at every period t, and there exists an explanation E of $(C_t, A_t)_t$ which satisfies Dynamically Consistent Explanation, No Explanatory Gap, Acyclic Explanation*, and Constrained Reachability.*

This result is a straightforward adaptation of Theorem 2, hence its proof is omitted.

4 DISCUSSION

4.1 Interpretation of the Meta-Preference

Strictly speaking, the meta-preference is a purely behavioral object: DM's behavior can be represented *as if* she were maximizing $\rhd$ on sets of relevant attributes. Further interpretation of this object is necessarily speculative, and this exercise is less obvious than interpreting a preference relation on alternatives.

A possible interpretation of the meta-preference is that DM prefers *coherent justifications* of her actions. For instance, whenever DM becomes aware that her actions lead to contradictory consequences, she might decide to change her behavior to make it internally consistent. If DM wants her consumption behavior to be sustainable and she discovers the environmental impact of meat production, she could decide to make the attribute "meat" relevant and reduce her consumption accordingly. The idea of coherent justification can explain the experimental evidence in Nielsen and Rehbeck (2020). They show that whenever participants violate stated normative criteria on choice—such as first-order stochastic dominance or transitivity—subjects are willing to modify their choices on lotteries to resolve the contradiction. Alternatively, the meta-preference could represent a form of *external* consistency with respect to a norm or a culture.

At the same time, the meta-preference relation may capture *motivated reasoning*—which can be seen as the opposite of a preference for coherent justification. Whenever DM becomes aware of new consequences, she decides to make it relevant or not depending on her preferences over the alternatives. For instance, if DM wants her consumption behavior to be sustainable but she discovers the environmental impact of meat production, she could decide to *ignore* this aspect because she likes meat. This suggests that the attribute ordering and the meta-preference are possibly related. This interpretation can explain the experimental evidence in Exley (2016). She shows that lack of charitable donations involve excuse-driven responses to charity risk when participants use their *own* money, whereas subjects are insensitive to charity risk when using *other* people's money.

Therefore, our model provides a unified and testable framework for thinking the trade-off between coherency and motivated reasoning in preference change.

4.2 Universality of Preferences and Meta-Preferences

Our model rationalizes heterogeneity in choice behavior even under the assumption that all individuals share the same preferences and meta-preferences. The reason is that preference change is *path-dependent*, meaning that if two agents become aware of the same attributes in a different order, they would end up with different sets of relevant attributes. Let us take a simple example. Two voters share the same attribute ordering:

$$\{\text{corrupt}\} < \emptyset < \{\text{corrupt, right-wing}\} < \{\text{right-wing}\}$$

Namely, they prefer a non corrupt right-wing candidate, to a corrupt right-wing candidate, to a left-wing candidate, to a corrupt left-wing candidate. Moreover, they share the same meta-preference:

$$\emptyset \lhd \{\text{right-wing}\} \lhd \{\text{corrupt, right-wing}\} \lhd \{\text{corrupt}\}$$

Namely, knowing that a right-wing candidate is corrupt is relevant, but if the candidate is known to be corrupt then her ideology is irrelevant. Say that voters become aware of the *same* attributes but in a different order: for the first voter $A_1^i = \{\text{right-wing}\}$ and $A_2^i = \{\text{corrupt}\}$; whereas for the second voter $A_1^j = \{\text{corrupt}\}$ and $A_2^j = \{\text{right-wing}\}$. It is easy to see that after the second period, the relevant attributes for the first voter are $M_2^i = \{\text{corrupt, right-wing}\}$ whereas the relevant attributes for the second voter are $M_2^j = \{\text{corrupt}\}$. Therefore the former has a strictly better view of the candidate than the latter.

The hypothesis that all individuals can be represented as sharing the same preferences and meta-preferences is easily testable: identify the model for all individuals in a population N and check whether the intersection of the identified sets $\bigcap_{i \in N} \geqslant^i$ and $\bigcap_{i \in N} \rhd^i$ are non-empty. This hypothesis is potentially relevant for welfare analysis as we show in the next section.

4.3 Normative Analysis

Welfare analysis with changing preferences is a notoriously hard problem. In this section, our objective is simply to clarify why normative analysis is difficult in this context, and to propose avenues for future research.

The first obstacle is the question of commensurability: is there a common standard (or unit) to compare improvements along the preference relation and along the meta-preference relation? Put in decision theoretic terms, is there a utility representation of Deliberate Preference Change that aggregates the preference ordering and the meta-preference ordering? It is not obvious how the two should interact in a functional representation as improvements along the meta-preference relation could reduce DM's satisfaction. For instance, moving toward sustainable consumption is certainly not enjoyable for most people, but they still think it is important to protect the environment.

The second obstacle is the question of time-separability: should welfare include current consumption, future consumption as well as the timing of preference change? This last point is key when we think of education: when should we make students aware of a new concept or problem? Common sense suggests that the timing of preference change is not welfare neutral, and this is partly captured by the path-dependence of preference change in our model. This suggests that two different paths $(M_t)_t$ and $(M'_t)_t$ that are equivalent up to a reordering and such that $M_1 = M'_1$ and $M_T = M'_T$ need not be welfare equivalent.

The most established normative theory which accounts for preference changes might be the "equal opportunity for welfare" proposed by Arneson (1989). This is a theory of distributive justice that tolerates inequalities of the positions individuals reach only if individuals are *responsible* for these inequalities (through their choices). If DM decides to stay unemployed, and if she is responsible for this choice, then DM being poorer than the average worker is acceptable according to this theory. That being said, how can society guarantee that people's responsibility is *effective*? Arneson argues that this is the case if (1) individuals face equivalent menus of options according to their potential preferences, and (2) individuals are equally aware of these options. Our model allows—at least theoretically—an implementation of the equal opportunity for welfare.

4.4 Belief Updating

We assume throughout that DM only changes awareness and not beliefs. Of course this is quite unrealistic. In order to accommodate belief updating, one might formally incorporate consequences and attribute-consequence links—namely, DM's perception of how attributes map to consequences. We can then extend the concept of relevant attributes to *relevant attribute-consequence links*. DM entertains different beliefs as to what is the "correct" set of attribute-consequence links—in this sense DM is Bayesian—but when she becomes aware of new consequences she decides whether to make it relevant or not according to her meta-preference—in this sense DM is not Bayesian. In the latter case, DM needs to reweight her beliefs as the state space has expanded, and similarly if the state space shrinks.

These ideas are reminiscent of Karni and Vierø (2013) and Dietrich (2018) who extend the Savagian model to accommodate changing awareness. They propose axioms which guarantee that the *relative* probabilities are kept constant when DM changes awareness, however they lack a theory of how DM chooses the *absolute* probabilities on the new state space. The meta-preference relation extended to the domain of attribute-consequence links could possibly fill this gap.

5 Conclusion

Empirical evidence on choice variations across time and choice heterogeneity across individuals seem incompatible with the stability of preferences or a simple model of belief updating. Various theories have been proposed to address this issue, but they typically lack empirical content. In this paper, we make progress toward a *testable* theory of preference change that embeds the act of discovery which is key in the formation of preferences. Each alternative is represented by a set of binary attributes, and DM's preferences over alternatives are directly defined on sets of attributes. When comparing two alternatives, DM's behavior is determined by a subset of relevant attributes. Whenever DM is aware of a set of attributes, she can decide to make it relevant or irrelevant for the her next period choices. This preference change is rational if it is consistent with the maximization of a meta-preference relation under the constraint that preference changes are driven by awareness.

We provide a behavioral characterization of this model consisting of four axioms. First, choices must satisfy WARP at each stage so as to be consistent with the maximization of a preference relation. The attribute ordering is strict if and only if DM does not exhibit indifference between alternatives that differ by revealed relevant attributes. The stability of the attribute ordering requires consistency of choice across time for alternatives that are equally relevant. Finally, the sequence of revealed relevant attributes must be acyclic with respect to previously accessible sets of relevant attributes so as to be consistent with the maximization of a meta-preference.

We then argued that extending the representation to weak attribute orderings is important because it allows "background attributes." These are attributes that drive preference change, even though they do not impact choice directly. This general model creates an indeterminacy problem as DM's indifference can be explained either by an indifference of the attribute ordering, or by the irrelevance of some attributes. The indeterminacy problem makes it impossible to obtain axioms on choice directly, hence we provide conditions on partially identified objects.

Our paper opens new avenues to empirically test models of chosen preferences, endogenous preferences, motivated reasoning, evolving attention, changing awareness, etc. Moreover, our model sheds new light on the debate about the universality of preferences. Indeed, due to the constraint on awareness the set of relevant attributes at period t is path-dependent. Hence, if two identical decision makers change awareness on the exact same attributes but in a different order, their choice behavior and how they justify their changes will typically differ.

APPENDIX A PROOFS OF PROPOSITIONS 1 AND 2

For any set S, we denote $\Delta_{S^2} = \{(x, x) \in S^2\}$ its diagonal.

Proof of Proposition 1. (Sufficiency) Suppose that C_t satisfies WARP at period t. First, we show that $\mathcal{M}$ can rationalize C_t. In this case, our representation at t coincides with standard preference maximization because for any $x \in X$, $x \cap \mathcal{M} = x$. By WARP, there exists a preference relation (weak order) $\succsim_t \subset X^2$ such that C_t can be represented by the maximization of $\succsim_t$. Identifying $\geqslant_t$ with $\succsim_t$ yields the desired result.

Second, we show that $\underline{M}_t$ can rationalize C_t. By WARP, C_t partitions the set of outcomes into indifference classes: for all x, let $I_t(x) \equiv \{y \in X : C_t(\{x, y\}) =$

$\{x, y\}\}$. Set $\geqslant'_t = \geqslant_t \cap \underline{M}_t^2$, i.e:

$$\begin{aligned}
&>'_t = \{(x,y) \in X^2 : x, y \subseteq \underline{M}_t \text{ and } C_t(\{x,y\}) = \{x\}\} \\
&\simeq'_t = \{(x,y) \in X^2 : x, y \subseteq \underline{M}_t \text{ and } C_t(\{x,y\}) = \{x,y\}\} \cup \Delta_{(2^{\underline{M}_t})^2}.
\end{aligned}$$

Let $x, y \in X$ two alternatives and denote $x' = x \cap \underline{M}_t$ and $y' = y \cap \underline{M}_t$. We denote $x \backslash x' = \{x_1, \dots, x_p\}$ and $y \backslash y' = \{y_1, \dots, y_n\}$. We have that $x \backslash x' \cap \underline{M}_t = \emptyset$ and $y \backslash y' \cap \underline{M}_t = \emptyset$, hence by definition of $\underline{M}_t$:

$$\begin{aligned}
&C_t(\{x', x' + x_1\}) = \{x', x' + x_1\} \\
&C_t(\{x' + x_1, x' + x_1 + x_2\}) = \{x' + x_1, x' + x_1 + x_2\} \\
&\quad \vdots \\
&C_t(\{x' + x_1 + \cdots + x_{p-1}, x\}) = \{x' + x_1 + \cdots + x_{p-1}, x\}
\end{aligned}$$

$$\begin{aligned}
&C_t(\{y', y' + y_1\}) = \{y', y' + y_1\} \\
&C_t(\{y' + y_1, y' + y_1 + y_2\}) = \{y' + y_1, y' + y_1 + y_2\} \\
&\quad \vdots \\
&C_t(\{y' + y_1 + \cdots + y_{n-1}, y\}) = \{y' + y_1 + \cdots + y_{n-1}, y\}
\end{aligned}$$

By the transitivity of the revealed preference relation induced by WARP, this implies that $C_t(\{x', x\}) = \{x', x\}$ and $C_t(\{y', y\}) = \{y', y\}$. Therefore, by WARP, we get that $x \in C_t(\{x, y\}) \iff x' \in C_t(\{x', y'\})$. By definition of $\geqslant'_t$, $x' \in C_t(\{x', y'\}) \iff x' \geqslant'_t y'$, which means that the pair $(\underline{M}_t, \geqslant'_t)$ rationalizes C_t. But because $\geqslant'_t$ is the restriction of $\geqslant_t$ to subsets of $\underline{M}_t$, it implies that $(\underline{M}_t, \geqslant_t)$ also rationalizes C_t.

Moreover, $\underline{M}_t$ is the smallest set of relevant attributes that can rationalize C_t. By contradiction, suppose that $M \subset \underline{M}_t$ rationalizes C_t. Then, for any x, DM is indifferent between x and $x + m$ for any $m \notin M$ as $x \cap M = x + m \cap M$. This directly contradicts the definition of $\underline{M}_t$, i.e., the existence of a z such that $C_t(\{z + m, z\}) \neq \{z + m, z\}$ for any $m \in (\underline{M}_t \setminus M)$.

Finally, we prove the lattice property. For any $m \notin \underline{M}_t$, let $M = \underline{M}_t + m$. Set $\geqslant''_t = \geqslant_t \cap M^2$. But because m is not revealed this implies that:

$$\begin{aligned}
&>''_t = >'_t \\
&\simeq''_t = \simeq'_t \cup \{(x,y) \in X^2 : x, y \subseteq \underline{M}_t + m \text{ and } x \triangle y = \{m\}\} \cup \{(x + m, x + m) : x \subseteq \underline{M}_t\}
\end{aligned}$$

Then, by a similar argument as above, one shows that $(\underline{M}_t + m, \geqslant''_t)$, hence $(\underline{M}_t + m, \geqslant_t)$, rationalizes C_t. Then, by induction, we get that for any $M \supset \underline{M}_t$, $(M, \geqslant_t)$ rationalizes C_t.

(Necessity) Let $K, K' \in \mathcal{K}$ with $K' \subseteq K$, and such that,

$$C^{(M_t,\geqslant_t)}(K) \cap K' = \{x \in K' : x \cap M_t \geqslant_t y \cap M_t \text{ for all } y \in K\} \neq \emptyset.$$

We need to show that $C^{(M_t,\geqslant_t)}(K) \cap K' = C^{(M_t,\geqslant_t)}(K')$.

Let $x \in C^{(M_t,\geqslant_t)}(K) \cap K'$, i.e $x \in K'$ such that $x \cap M_t \geqslant_t y \cap M_t$ for all $y \in K$, but because $K' \subseteq K$, we have that $x \cap M_t \geqslant_t y \cap M_t$ for all $y \in K'$, so $x \in C^{(M_t,\geqslant_t)}(K')$.

Conversely, let $x \in C^{(M_t,\geqslant_t)}(K')$, then suppose that $x \notin C^{(M_t,\geqslant_t)}(K) \cap K'$. Given that $C^{(M_t,\geqslant_t)}(K) \cap K' \neq \emptyset$, there exist $y \in K'$ such that $y \cap M_t >_t x \cap M_t$, but then x cannot be in $C^{(M_t,\geqslant_t)}(K')$, a contradiction.

□

Proof of Proposition 2. (Necessity) The necessity of WARP is the same as for the proof of proposition 1. Let C a choice correspondence induced by the pair $(M, >)$. For any x and any m such that $m \in M$ and $m \notin x$, $>$ ranks strictly $x \cap M$ and $(x + m) \cap M$, so $C(\{x + m, x\}) \neq \{x + m, x\}$, meaning that every $m \in M$ is *revealed relevant*, i.e $M \subseteq \underline{M}$. By proposition 1, the set of revealed relevant is the minimal set that rationalizes C, so $\underline{M} \subseteq M$, hence $M = \underline{M}$. From that, it is easy to check that Justified Indifference is verified.

(Sufficiency) By the proof of proposition 1, we know that $(\underline{M}_t, \geqslant'_t)$ rationalizes C_t. We show that

$$\simeq'_t = \{(x, x') : x \cap \underline{M}_t = x' \cap \underline{M}_t\}$$

That $\{(x, x') : x \cap \underline{M}_t = x' \cap \underline{M}_t\} \subseteq \simeq'_t$ is a trivial consequence from the fact that $\underline{M}_t$ rationalizes C_t.

Let $(x, y) \notin \{(x, x') : x \cap \underline{M}_t = x' \cap \underline{M}_t\}$. Then, $\emptyset \neq x \triangle y \subset \underline{M}_t$. By the contraposition of Justified Indifference, this implies that that $x \not\simeq'_t y$. Hence, since $\geqslant'_t$ is complete, it ranks *strictly* x and y and $(x, y) \notin \simeq'_t$. Thus,

$$\simeq'_t \subseteq \{(x, x') : x \cap \underline{M}_t = x' \cap \underline{M}_t\}$$

Furthermore, for any $M \supset \underline{M}_t$, as we show in the proof of proposition 1, by the definition of revealed relevant, some indifferences must be added to $\geqslant'_t$ to rationalize the choices between alternatives that differ by some attributes not in $\underline{M}_t$. This shows that $\underline{M}_t$ is the unique set of relevant attributes that rationalizes C_t with a strict attribute ordering. □

Appendix B Proof of Theorem 1

Proof of Theorem 1. (Sufficiency). Suppose that $(C_t)_t$ satisfies all the axioms. By the proof of proposition 2, we know that $(\underline{M}_t)_t$ is the unique sequence of relevant attributes such that, for each period t, C_t can be rationalized by $\underline{M}_t$ together with a strict attribute ordering $>_t$ defined as follows:

$$>_t = \{(x,y) \in X^2 : x, y \subseteq \underline{M}_t \text{ and } \{x\} = C_t(\{x,y\})\}$$

Therefore $>_t$ corresponds to the most incomplete attribute ordering that ranks only subsets of $\underline{M}_t$. This attribute ordering, however, is possibly time dependent.

We verify that the attribute ordering is stable. Let t and t' be two distinct periods, and let x, x' and y, y' be such that $x \cap \underline{M}_t = x' \cap \underline{M}_{t'}$ and $y \cap \underline{M}_t = y' \cap \underline{M}_{t'}$. Then by Between Consistency we obtain:

$$x \in C_t(\{x,y\}) \iff x \in C_{t'}(\{x,y\})$$

which is equivalent to:

$$\begin{aligned} x \cap \underline{M}_t >_t y \cap \underline{M}_t &\iff x' \cap \underline{M}_{t'} >_{t'} y' \cap \underline{M}_{t'} \\ &\iff x \cap \underline{M}_t >_{t'} y \cap \underline{M}_t \\ &\iff x' \cap \underline{M}_{t'} >_t y' \cap \underline{M}_{t'} \end{aligned}$$

where the second and third equivalence follow from the assumption on x, x', y, y'. Therefore $>_t$ and $>_{t'}$ agree on all the subset of attributes that they both rank. Therefore if we define $>_{t,t'} = >_t \cup >_{t'}$, we get that $>_{t,t'} \cap 2^{\underline{M}_t} \times 2^{\underline{M}_t} = >_t$ (because either it ranks sets not ranked by $>_{t'}$ or it ranks sets on which $>_t$ and $>_{t'}$ agree), and similarly for $>_{t'}$. Therefore C_t can be rationalized by $(\underline{M}_t, >_{t,t'})$ and $C_{t'}$ can be rationalized by $(\underline{M}_{t'}, >_{t,t'})$. Let us now define $>^\star = \bigcup_t >_t$. From the previous argument, we get that for any t, $>^\star \cap 2^{\underline{M}_t} \times 2^{\underline{M}_t} = >_t$, therefore for any t we indeed have that C_t can be rationalized by $(\underline{M}_t, >^\star)$.

Finally, we construct one particular sequence of changing awareness which, together with Discovery Constrained Cycles, makes the sequence $(\underline{M}_t)_t$ consistent with the maximization of a meta-preference. Define $A_t = \underline{M}_t \triangle \underline{M}_{t+1}$, hence the set of reachable relevant attributes reduces that:

$$R(\underline{M}_t, \underline{M}_t \triangle \underline{M}_{t+1}) = \{M : \underline{M}_t \cap \underline{M}_{t+1} \subseteq M \subseteq \underline{M}_t \cup \underline{M}_{t+1}\}.$$

Define the revealed metapreference relation $\triangleright$ by: $M \triangleright M'$ if and only if $M \neq M'$ and there exists t, such that $M = \underline{M}_t$ and,

$$M' \in \bigcup_{t':t'<t} R(\underline{M}_{t'}, A_{t'}).$$

We verify that $\triangleright$ is asymmetric. Suppose that $M \triangleright M'$. Let $t > t'$, such that $M = \underline{M}_t$ and $M' \in R(\underline{M}_{t'}, A_{t'})$.

First, Discovery Constrained Cycles (DCC) implies that there cannot be any $t'' > t$ such that $\underline{M}_{t''} = M'$, otherwise this would mean that $\underline{M}_{t'} \triangle \underline{M}_{t''} \subseteq \underline{M}_{t'} \triangle \underline{M}_{t'+1}$, a violation of DCC.

Second, let suppose that exists $t'' < t$ such that $M' = \underline{M}_{t''}$. Then if there exists $t''' < t''$ such that $M \in R(\underline{M}_{t'''}, A_{t'''})$, this would imply that $\underline{M}_{t'''} \triangle \underline{M}_t \subseteq \underline{M}_{t'''} \triangle \underline{M}_{t'''+1}$, a clear violation of DCC.

Therefore, we conclude that $\neg(M' \triangleright M)$, which proves the asymmetry of $\triangleright$.

We now verify that $\triangleright$ is transitive. Assume $M \triangleright M'$ and $M' \triangleright M''$. Then there exist t, t', $t > t'$, such that, $M = \underline{M}_t$ and $M' = \underline{M}_{t'}$, and,

$$M'' \in \bigcup_{t'':t''<t} R(\underline{M}_{t''}, A_{t''}) \text{ and } M'' \in \bigcup_{t''':t'''<t'} R(\underline{M}_{t'''}, A_{t'''})$$

Noting that since $t < t'$,

$$M'' \in \bigcup_{t'':t''<t} R(\underline{M}_{t''}, A_{t''}) \subseteq \bigcup_{t''':t'''<t'} R(\underline{M}_{t'''}, A_{t'''})$$

We conclude that $M \triangleright M''$, implying the transitivity of $\triangleright$. We can complete it in any way on subsets that are not ranked yet.

Finally, by definition we have that for every t, $\underline{M}_{t+1} = \max(R(\underline{M}_t, A_t), \triangleright)$.

(Necessity). Suppose that the dataset $(C_t)_t$ is rationalisable by Strict Deliberate Preference Change, and let $(M_t)_t$ be the sequence of relevant attributes and $>$ be the attribute ordering. By Proposition 2, the dataset satisfies WARP and Justified Indifference. By the proof of the same proposition, we also know that the for every t, $M_t = \underline{M}_t$.

Let t, t', and $x, x', y, y' \in X$ be such that $x \cap \underline{M}_t = x' \cap \underline{M}'_t$ and $y \cap \underline{M}_t = y' \cap \underline{M}'_t$.

We have:

$$\begin{aligned} x \in C_t(\{x,y\}) &\iff x \cap \underline{M}_t > y \cap \underline{M}_t \\ &\iff x' \cap \underline{M}_{t'} > y' \cap \underline{M}_{t'} \\ &\iff x' \in C'_t(\{x',y'\}) \end{aligned}$$

where the first equivalence is by definition of rationalizability, the second equivalence follows from the assumption on x, x', y, y', and the last equivalence is again by definition of rationalizability. Therefore, the dataset satifies Between Consistency.

Let t be a given period. We denote $t' = \min\{\tau > t+1 : M_\tau \neq M_{t+1}$. Hence t' is the first period period after $t+1$ such that $M_{t'} \neq M_{t+1}$. Let $\tau \geq t'$, by contradiction suppose that $M_t \triangle M_\tau \subseteq M_t \triangle M_{t+1}$, which violates DCC. The constraint of awareness implies that $M_t \triangle M_{t+1} \subseteq A_t$, hence $M_t \triangle M_\tau \subseteq A_t$, i.e. $M_\tau \in R(M_t, A_t)$. But by the transitivity of the meta-preference $\rhd$, we know that for any two periods p, p': $p' > p \implies M_{p'} \rhd M_p$. Hence $M_\tau \rhd M_{t+1}$, which contradicts that $M_{t+1} = \max(R(\underline{M}_t, A_t), \rhd)$. Hence DCC is satisfied.

□

APPENDIX C PROOF OF THEOREM 2

We first need to establish a couple of lemmas that show how the stability of the ordering provides some structure to the set of candidates we can have.

Let $\mathcal{C}$ be the set of choice correspondences on X satisfying WARP. As above, for each $C \in \mathcal{C}$, $M(C)$ is the lattice of attributes sets that rationalize C (cf Proposition 1). We denote $\underline{M}$ the infimum of $M(C)$.

We also denote by $F : \mathcal{C}^2 \mapsto \mathscr{M}^2$ the function that maps pairs of choice correspondences into the collection of pairs of sets of relevant attributes that do not yield choice reversals that can only result from a change of the ordering, i.e. that satisfy the condition of Between Consistency. $(M, M') \in F(C, C')$ if $M \in M(C)$, $M' \in M(C')$, and for all x, x', y, y', such that

$$x \cap M = x' \cap M' \text{ and } y \cap M = y' \cap M'$$

we have that

$$x \in C(\{x,y\}) \iff y \in C'(\{x',y'\})$$

The following lemma is a classical implication of the transitivity of the revealed preference under WARP. Hence its proof is omitted.

Lemma 1. *For all $C \in \mathcal{C}$ and all x, x', y, if $C(\{x, x'\}) = \{x, x'\}$ and $x \in C(\{x, y\})$, then $x' \in C(\{x', y\})$.*

Lemma 2. *Let $C, C' \in \mathcal{C}$. For all $M \in M(C)$ and $M' \in M(C')$, if $M_* \in M(C), M_* \subseteq M$ and $M'_* \in M(C'), M'_* \subseteq M'$, then*

$$(M, M') \in F(C, C') \implies (M_*, M'_*) \in F(C, C')$$

Proof. Let $(M, M') \in F(C, C')$. We need to show that $(M_*, M'_*) \in F(C, C')$. Let x, x', y, y' such that

$$x \cap M_* = x' \cap M'_* \text{ and } y \cap M_* = y' \cap M'_*$$

By Perfect Instantiation, there exists $\alpha, \alpha', \beta, \beta' \in X$ such that, $\alpha = x \cap M_*$, $\alpha' = x' \cap M'_*$, $\beta = y \cap M_*$ and $\beta' = y' \cap M'_*$. Then, since $M_* \subseteq M$ and $M'_* \subseteq M'$, we have that:

$$\alpha \cap M = x \cap M_* = x' \cap M'_* = \alpha' \cap M' \tag{1}$$

$$\beta \cap M = y \cap M_* = y' \cap M'_* = \beta' \cap M' \tag{2}$$

Hence, by (1), (2) and the fact that $(M, M') \in F(C, C')$ we get that

$$\alpha \in C(\{\alpha, \beta\}) \iff \alpha' \in C'(\{\alpha', \beta'\}) \tag{3}$$

Moreover, since M_* rationalize C, $\alpha \cap M_* = x \cap M_*$ and $\beta \cap M_* = y \cap M_*$ we have that,

$$C(\{x, \alpha\}) = \{x, \alpha\} \text{ and } C(\{y, \beta\}) = \{y, \beta\} \tag{4}$$

Similarly, since M'_* rationalize C', $\alpha' \cap M'_* = x' \cap M'_*$ and $\beta' \cap M'_* = y' \cap M'_*$ we have that,

$$C'(\{x', \alpha'\}) = \{x', \alpha'\} \text{ and } C'(\{y', \beta'\}) = \{y', \beta'\} \tag{5}$$

Applying WARP to both equality of (4) we get that $x \in C(\{x,y\}) \iff x \in C(\{x,y,\beta\}) \iff \alpha \in C(\{\alpha,y,\beta\}) \iff \alpha \in C(\{\alpha,\beta\})$. Similarly with (5) we get that $x' \in C'(\{x,y\}) \iff \alpha' \in C'(\{\alpha',\beta'\})$.

So by (3) we have that $x \in C(\{x,y\}) \iff x' \in C'(\{x,y\})$. □

Lemma 3. *For all $C, C' \in \mathcal{C}$, if $[(M,M') \in F(C,C')]$ and $[m \notin M$, or $m \in M \cap M']$, we have that $(M, M'+m) \in F(C,C')$.*

Proof. If $m \in M'$, then $M'+m = M'$, hence, since $(M,M') \in F(C,C')$, $(M, M'+m) \in F(C,C')$.

Assume now that $m \notin M'$. Thus, $m \notin M \cap M'$ and $m \notin M$. Let x, x', y, y' such that,

$$x \cap M = x' \cap (M'+m) \text{ and } y \cap M = y' \cap (M'+m) \tag{6}$$

The, we must have that $m \notin x' \cup y'$. Otherwise, contradicting the fact that $m \notin M$,

$$m \in x' \implies x \cap M = x' \cap (M'+m) \ni m \implies m \in M \tag{7}$$

$$m \in y' \implies y \cap M = y' \cap (M'+m) \ni m \implies m \in M \tag{8}$$

Thus, $x \cap M = x' \cap (M'+m) = x' \cap M'$ and $y \cap M = y' \cap (M'+m) = y' \cap M'$, and the fact that $(M,M') \in F(C,C')$ implies that $x \in C(\{x,y\})$ if and only if $x' \in C'(\{x',y'\})$. This shows as desired that $(M, M'+m) \in F(C,C')$ □

Lemma 4. *For all $C, C' \in \mathcal{C}$, $M, M' \in F(C,C')$, and all $m \notin M'$, if $(M, M' \cup \neg M) \in F(C,C')$, then $(M+m, M' \cup \neg M) \in F(C,C')$.*

Proof. First of all, note that, given that $M' \cup \neg(M+m) \subseteq M' \cup \neg M$, by Lemma 2, $(M, M' \cup \neg M) \in F(C,C')$ implies that:

$$(M, M' \cup \neg(M+m)) \in F(C,C') \tag{9}$$

If $m \in M$, then $M+m = M$, hence, since $(M, M' \cup \neg M) \in F(C,C')$, $(M+m, M' \cup \neg M) \in F(C,C')$.

Assume now that $m \notin M$. Let x, x', y, y' such that,

$$(M+m) \cap x = (M' \cup \neg M) \cap x' \text{ and } (M+m) \cap y = (M' \cup \neg M) \cap y' \tag{10}$$

We have two cases: either $m \in x \cup y$, or $m \notin x \cup y$.

Case 1: $m \in x \cup y$. W.l.o.g we can assume that $m \in x$ and since $m \notin M'$ we have that

$$(11) \qquad (M+m) \cap x = (M' \cup \neg M) \cap x' \implies M \cap x = (M' \cup \neg(M+m)) \cap x'$$

Now, we have that either if $m \in y$ or not.
Subcase 1.a: $m \in y$. Since $m \notin M'$ we have,

$$(12) \qquad (M+m) \cap y = (M' \cup \neg M) \cap y' \implies M \cap y = (M' \cup \neg(M+m)) \cap y'$$

Combining (11) and (12) with (9), we obtain, as desired,

$$x \in C(\{x,y\}) \iff x' \in C'(\{x',y'\})$$

Subcase 1.b: $m \notin y$. Since $m \in \neg M$, then by (10), we must have that $m \notin y'$. Thus,

$$(13) \qquad (M+m) \cap y = (M' \cup \neg M) \cap y' \implies M \cap y = (M' \cup \neg(M+m)) \cap y'$$

Combining (11) and (13) with (9), we obtain, as desired,

$$x \in C(\{x,y\}) \iff x' \in C'(\{x',y'\})$$

Case 2: If $m \notin x \cup y$. Since $m \notin M$, this implies that $m \notin x' \cup y'$, thus,

$$M \cap x = (M' \cup \neg M) \cap x' \text{ and } M \cap y = (M' \cup \neg M) \cap y'$$

Since $(M, M' \cup \neg M) \in F(C, C')$, we obtain, as desired,

$$x \in C(\{x,y\}) \iff x' \in C'(\{x',y'\})$$

□

Lemma 5. *For all $C, C' \in \mathcal{C}$, $(M, M') \in F(C, C')$, and all B such that $B \cap M' = \emptyset$, if $(M, M' \cup \neg M) \in F(C, C')$, then $(M \cup B, M' \cup \neg M) \in F(C, C')$.*

Proof. We prove it by induction. For $|B| = 1$, then there exists $m \notin M'$ such that $B = \{m\}$. Applying Lemma 4, we obtain the desired result.

To prove the induction step, assume that for all B such that $B \cap M' = \emptyset$ and $|B| = n$, we have that if $(M, M' \cup \neg M) \in F(C, C')$, then $(M \cup B, M' \cup \neg M) \in F(C, C')$.

Let B' such that $B' \cap M' = \emptyset$, and $|B'| = n+1$ and,

$$(M, M' \cup \neg M) \in F(C, C') \tag{14}$$

Note that, given $M' \subseteq M' \cup \neg(B' \cup M) \subseteq M' \cup \neg M$ and $(M, M') \in F(C, C')$, by Lemma 2, (14) implies:

$$(M, M' \cup \neg(B' \cup M)) \in F(C, C') \tag{15}$$

Similarly, let $m \in B'$. Given that $M' \cup \neg((B' - m) \cup M) \subseteq M' \cup \neg M$, by Lemma 2, (14) implies:

$$\Big(M, M' \cup \neg((B' - m) \cup M)\Big) \in F(C, C') \tag{16}$$

Thus, since $|B' - m| = n$ by the induction hypothesis we have,

$$\Big(M \cup (B' - m), M' \cup \neg M\Big) \in F(C, C')$$

Since $m \in B'$ with $B' \cap M' = \emptyset$, $m \notin M'$. Hence, we can apply Lemma 4 and we obtain, as desired,

$$\Big(M \cup B', M' \cup \neg M\Big) \in F(C, C')$$

□

Lemma 6. *$(M_*, M'_*) \in F(C, C')$ if, and only if, for all (M, M') such that*

$$M_* \subseteq M \subseteq M_* \cup (\mathcal{M} \backslash M'_*) \text{ and } M'_* \subseteq M' \subseteq M'_* \cup (\mathcal{M} \backslash M_*) \tag{17}$$

we have $(M, M') \in F(C, C')$.

Proof. Assume $(M_*, M'_*) \in F(C, C')$. Let (M, M') such that (17) is satisfied. Then, $M' = M'_* \cup \hat{M}'$ for some $\hat{M}' \subseteq \mathcal{M} \backslash M_*$. We first have by induction on the size of $\hat{M}'$ that

$$(M_*, M') \in F(C, C') \tag{18}$$

Indeed, for $|\hat{M}'| = 0$, we have have by assumption $(M_*, M'_*) \in F(C, C')$; and the inductive step is given by Lemma 3.

Similarly, there exists $\hat{M} \subseteq \neg M'_*$ such that $M = M_* \cup \hat{M}$. Each condition of lemma 5 is satisfied, hence:

$$(M, M') \in F(C, C')$$

$\square$

Proof of Theorem 2. (Necessity) Suppose that the dataset $(C_t)_t$ is rationalizable by Weak Deliberate Preference Change. Let, for all t, t' such that $t < t'$,

$$E_{t,t'} := M_t \triangle M_{t'}$$

First, by definition of the sequence $(M_t)_t$, $(E_{t,t'})_{t<t'}$ is an explanation of $(C_t)_t$. Furthermore, given its definition, No Explanatory Gap is satisfied.

To show that $(E_{t,t'})_{t<t'}$ satisfies Extended Between Consistency we first show that for all t, t', $(M_t, M_{t'}) \in F(C_t, C_{t'})$, then we show that $\underline{M}_t \cup V_{t,t'} \subseteq M_t$, $\underline{M}_{t'} \cup V_{t',t'} \subseteq M'_t$, and apply lemma 2.

$(M_t, M_{t'}) \in F(C_t, C_{t'})$ is implied by the fact that $C_t = C^{(M_t, \geqslant)}$ and $C_{t'} = C^{(M_{t'}, \geqslant)}$ for the same $\geqslant$. Indeed, if x, x', y, y' are such that $x \cap M_t = x' \cap M_{t'}$ and $y \cap M_t = y' \cap M_{t'}$. We have:

$$x \in C_t(\{x, y\}) \iff x \cap M_t \geqslant y \cap M_t \iff x' \cap M_{t'} \geqslant y' \cap M_{t'} \iff x' \in C'_t(\{x', y'\})$$

For any t, by definition, $\underline{M}_t \subseteq M_t$. Furthermore, for any t', let $m \in V_{t,t'}$; then $m \in \underline{M}_{t'} \subseteq M_{t'}$ and $m \notin E_{t,t'}$, thus, $m \in M_t$. Which implies that $V_{t,t'} \subseteq M_t$. Hence $\underline{M}_t \cup V_{t,t'} \subseteq M_t$. A similar reasoning implies that $\underline{M}_{t'} \cup V_{t',t} \subseteq M_{t'}$.

Then, by Lemma 2, given that $(M_t, M_{t'}) \in F(C_t, C_{t'})$, we obtain that $(\underline{M}_t \cup V_{t,t'}, \underline{M}_{t'} \cup V_{t',t}) \in F(C_t, C_{t'})$. This completes the proof that Extended Between Consistency is satisfied.

To show that $(E_{t,t'})_{t<t'}$ satisfies Acyclic Explanation, consider t and $\tau > t+1$. Suppose that $E_{t,\tau} \subset E_{t,t+1}$. Given that $E_{t,t+1} \subseteq A_t$ (because $M_{t+1} \in R(M_t, A_t)$), this means that $M_\tau \in R(M_t, A_t)$. Because, $M_{t+1} = \max(R(M_t, A_t), \rhd)$, this implies that $M_\tau = M_{t+1}$, hence $E_{t+1,\tau} = \emptyset$.

Sufficiency: Suppose that the dataset $(C_t)_t$ satisfies WARP at every period t and there is an explanation E that satisfies Extended Between Consistency, No Explanatory Gap and Acyclic Explanation.

Step 1: We first build the candidate sequences (M^*_t, A_t) that rationalizes $(C_t)_t$.

We define for all t

$$E(M(C_t), E_{t,t+1}) = \bigcup_{M \in M(C_t)} \{M' \in M(C_{t+1}) : M \triangle M' = E_{t,t+1}\}$$

From these objects, we define recursively a sequence $\mathcal{L}_t$ as follows:

$$\begin{aligned} & \mathcal{L}_1 = M(C_1) \\ \forall t \geq 2 \qquad & \mathcal{L}_t = E(\mathcal{L}_{t-1}, E_{t-1,t}) \end{aligned}$$

Note that these sets are never empty since $(E_{t,t'})_{t<t'}$ explains $(C_t)_t$.

We can define the following sequence $(M_t^*)_t$:

$$\begin{aligned} & M_T^* \in \arg\min_{M \in \mathcal{L}_T} (\# M) \\ \forall\, t \leq T-1 \qquad & M_t^* \in \arg\min_{M \in \mathcal{L}_t \text{ s.t. } M \triangle M_{t+1}^* = E_{t,t+1}} (\# M) \end{aligned}$$

Let $A_t = M_t^* \triangle M_{t+1}^* = E_{t,t+1}$ for all t. Furthermore, given that $(E_{t,t'})_{t<t'}$ satisfies No Explanatory Gap, for all t, t', $t' > t$, $E_{t,t'} = M_t^* \triangle M_{t'}^*$.

Step 2: We show that for every t, t', $(M_t^*, M_{t'}^*) \in F(C_t, C_{t'})$

Note that, similarly to what was proved for the necessity part:

$$\underline{M}_t \cup V_{t,t'} \subset M_t^* \text{ and } \underline{M}_{t'} \cup V_{t',t} \subset M_{t'}^* \tag{19}$$

Indeed, consider $m \in V_{t,t'}$. Thus $m \notin E_{t,t'}$ and $m \in \underline{M}_{t'} \subseteq M_{t'}^*$, therefore, $m \in M_t^*$. Hence we get that $\underline{M}_t \cup V_{t,t'} \subset M_t^*$. The same reasoning applies to show $\underline{M}_{t'} \cup V_{t',t} \subset M_{t'}^*$. Which implies that $M_t^* = (\underline{M}_t \cup V_{t',t}) \cup \big(M_t^* \backslash (\underline{M}_t \cup V_{t,t'})\big)$ and $M_{t'}^* = (\underline{M}_t \cup V_{t',t}) \cup \big(M_t^* \backslash (\underline{M}_{t'} \cup V_{t',t})\big)$

Furthermore, let $m \in M_t^* \backslash (\underline{M}_t \cup V_{t,t'})$. If $m \in E_{t,t'}$, then $m \notin M_{t'}^*$ hence $m \notin \underline{M}_{t'}$. If $m \notin E_{t,t'}$, since $m \notin V_{t,t'}$ we have $m \notin \underline{M}_{t'}$. Thus,

$$m \in M_t^* \backslash (\underline{M}_t \cup V_{t,t'}) \implies m \notin \underline{M}_{t'} \tag{20}$$

Similarly, one can show that,

$$m \in M_{t'}^* \backslash (\underline{M}_{t'} \cup V_{t',t}) \implies m \notin \underline{M}_t \tag{21}$$

Extended Between Consistency implies that $(\underline{M}_t \cup V_{t,t'}, \underline{M}_{t'} \cup V_{t',t}) \in F(C_t, C_{t'})$.

Therefore, by (20) and (21) and Lemma 6 we infer that

$$(\underline{M}_t \cup V_{t,t'}, \underline{M}_{t'} \cup V_{t',t}) \in F(C_t, C_{t'}) \implies (M_t^*, M_{t'}^*) \in F(C_t, C_{t'})$$

Step 3: We build a stable attributes ordering $\geqslant^\star$ such that for any t, $C_t = C^{(M_t^*, \geqslant^\star)}$.

We proceed similarly as in the proof of Theorem 1. We denote by $\geqslant_t$ the most incomplete attribute ordering such that $C_t = C^{(M_t^*, \geqslant_t)}$, that is $\geqslant_t$ only ranks subsets of M_t^* (only non-empty ones if M_t^* is the grand set).

Let t and t' be two distinct periods, and let x, x', y, y' be such that $x \cap M_t^* = x' \cap M_{t'}^*$ and $y \cap M_t^* = y' \cap M_{t'}^*$. Then by *step* 2 we know that $(M_t^*, M_{t'}^*) \in F(C_t, C_{t'})$, so

$$x \in C_t(\{x, y\}) \iff x' \in C_{t'}(\{x', y'\})$$

which is equivalent to:

$$x \cap M_t^* \geqslant_t y \cap M_t^* \iff x' \cap M_{t'}^* \geqslant_{t'} y' \cap M_{t'}^*.$$

This means that $\geqslant_t$ and $\geqslant_{t'}$ agree on all the sets of attributes that they both rank. Hence, if we define $\geqslant_{t,t'} = \geqslant_t \cup \geqslant_{t'}$, we get that $\geqslant_{t,t'} \cap 2^{M_t^*} \times 2^{M_t^*} = \geqslant_t$ (because either it ranks sets not ranked by $\geqslant_{t'}$ or it ranks sets on which $\geqslant_t$ and $\geqslant_{t'}$ agree), and similarly for $\geqslant_{t'}$. This means that $C_t = C^{(M_t^*, \geqslant_{t,t'})}$ and $C_{t'} = C^{(M_{t'}^*, \geqslant_{t,t'})}$.

Let us now define $\geqslant^\star = \bigcup_t \geqslant_t$. From the previous argument, we get that for any t, $\geqslant^\star \cap 2^{M_t^*} \times 2^{M_t^*} = \geqslant_t$, therefore for any t we indeed have that $C_t = C^{(M_t^*, \geqslant^\star)}$.

Step 4: Show that for all t, t', with $t < t'$, if $M_{t'}^* \in R(M_t^*, A_t)$, then for all t'', with $t < t'' < t'$, $M_{t''}^* = M_{t'}^*$.

We know that $E_{t,t'} = M_t^* \triangle M_{t'}^*$, hence, $M_{t'}^* \in R(M_t^*, A_t)$ implies that $E_{t,t'} \subseteq A_t = E_{t,t+1}$. By Acyclic Explanation, this implies that $E_{t+1,t'} = \emptyset$. Hence $E_{t+1,t'} \subset E_{t+1,t+2}$, and by applying Acyclic Explanation, we obtain that $E_{t+2,t'} = \emptyset$. By iterating Acyclic Explanation, we obtain that $E_{t'',t'} = \emptyset$ which means that $M_{t''}^* = M_{t'}^*$.

Step 5: We build the meta-preference and show that it is asymmetric and transitive.

Define the metapreference $\rhd$ as follows: $M' \lhd M$ if $M \neq M'$ and $\exists t$, such that $M = M_t^*$ and,

$$M' \in \bigcup_{t':t'<t} R(M_{t'}^*, A_{t'})$$

Assume that $M' \lhd M$, and let t such that $M = M_t^*$. We know from *Step 4* that if there exists $t' < t$ such that $M \in R(M_{t'}^*, A_{t'})$ then $M_t^* = M_{t'}^*$. Furthermore, for any $t'' > t$, $M_{t''}^* \neq M'$, otherwise, given that $M' \in \bigcup_{t':t'<t} R(M_{t'}^*, A_{t'})$, by iterating *Step 4*, this would imply that $M' = M_t^*$, a contradiction.

So we can conclude that $\neg(M \rhd M')$, which establishes the asymmetry of $\rhd$.

Now assume $M'' \lhd M'$ and $M' \lhd M$. Then there exist t, t' such that, $M = M_t^*$ and $M' = M_{t'}^*$, and,

$$M'' \in \bigcup_{t'':t''<t'} R(M_{t''}^*, A_{t''}) \text{ and } M' \in \bigcup_{t''':t'''<t} R(M_{t'''}^*, A_{t'''})$$

From the previous argument, note that $t' < t$, hence,

$$M'' \in \bigcup_{t'':t''<t} R(M_{t''}^*, A_{t''}) \subseteq \bigcup_{t''':t'''<t'} R(M_{t'''}^*, A_{t'''})$$

We conclude that $M'' \lhd M$ which proves the transitivity of $\rhd$. We can finally complete it in any way on subsets that are not ranked yet.

□

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Apesteguia, J. and Ballester, M. A. (2015). A Measure of Rationality and Welfare. *Journal of Political Economy*, 123(6):1278–1310.

Arneson, R. J. (1989). Equality and Equal Opportunity for Welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 56(1):77–93.

Becker, G. S. and Mulligan, C. B. (1997). The Endogenous Determination of Time Preference. *The Quarterly Journal of Economics*, 112(3):729–758.

Bernheim, B. D., Braghieri, L., Martinez-Marquina, A., and Zuckerman, D. (2019). A Theory of Chosen Preferences. *Working Paper*.

Bernheim, B. D. and Rangel, A. (2009). Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *The Quarterly Journal of Economics*, 124(1):51–104.

Boissonnet, N. (2019). Rationalizing Preference Formation by Partial Deliberation. *Working Paper*.

Bouacida, E. (2019). Eliciting Choice Correspondences A General Method and an Experimental Implementation. *Working Paper*.

Chambers, C. P. and Hayashi, T. (2012). Choice and Individual Welfare. *Journal of Economic Theory*, 147(5):1818–1849.

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

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and Taxation: Theory and Evidence. *American economic review*, 99(4):1145–77.

De Clippel, G. and Eliaz, K. (2012). Reason-Based Choice: A Bargaining Rationale for the Attraction and Compromise Effects. *Theoretical Economics*, 7(1):125–162.

Dekel, E., Lipman, B. L., and Rustichini, A. (1998). Recent Developments in Modeling Unforeseen Contingencies. *European Economic Review*, 42(3-5):523–542.

Dekel, E., Lipman, B. L., and Rustichini, A. (2009). Temptation-Driven Preferences. *The Review of Economic Studies*, 76(3):937–971.

Dietrich, F. (2018). Savage's Theorem under Changing Awareness. *Journal of Economic Theory*, 176:1–54.

Dietrich, F. and List, C. (2013a). A Reason-Based Theory of Rational Choice. *Nous*, 47(1):104–134.

Dietrich, F. and List, C. (2013b). Where Do Preferences Come From? *International Journal of Game Theory*, 42(3):613–637.

Dietrich, F. and List, C. (2016). Reason-Based Choice and Context-Dependence: An Explanatory Framework. *Economics & Philosophy*, 32(2):175–229.

Dutta, P., Murgai, R., Ravallion, M., and van de Walle, D. (2014). *Right to Work?: Assessing India's Employment Guarantee Scheme in Bihar*. The World Bank.

Exley, C. L. (2016). Excusing Selfishness in Charitable Giving: The Role of Risk. *The Review of Economic Studies*, 83(2):587–628.

Grune-Yanoff, T. and Hansson, S. (2009). Preference change: an introduction. In Till Grune-Yanoff and Sven Ove Hansson, editor, *Preference Change: Approaches from Philosophy, Economics and Psychology*, pages 1–26.

Gul, F. and Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, 69(6):1403–1435.

Gul, F. and Pesendorfer, W. (2005). The Revealed Preference Theory of Changing Tastes. *The Review of Economic Studies*, 72(2):429–448.

Kalai, G., Rubinstein, A., and Spiegler, R. (2002). Rationalizing Choice Functions by Multiple Rationales. *Econometrica*, 70(6):2481–2488.

Karni, E. and Vierø, M.-L. (2013). "Reverse Bayesianism": A Choice-Based Theory of Growing Awareness. *American Economic Review*, 103(7):2790–2810.

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2):132–157.

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

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

Nielsen, K. and Rehbeck, J. (2020). When Choices are Mistakes. *Working Paper*.

Nishimura, H. (2018). The Transitive Core: Inference of Welfare from Nontransitive Preference Relations. *Theoretical Economics*, 13(2):579–606.

Ok, E. A., Ortoleva, P., and Riella, G. (2015). Revealed (P)reference Theory. *American Economic Review*, 105(1):299–321.

Palacios-Huerta, I. and Santos, T. J. (2004). A Theory of Markets, Institutions, and Endogenous Preferences. *Journal of Public Economics*, 88(3-4):601–627.

Peleg, B. and Yaari, M. E. (1973). On the Existence of a Consistent Course of Action when Tastes are Changing. *The Review of Economic Studies*, 40(3):391–401.

Salant, Y. and Rubinstein, A. (2008). (A,f): Choice with Frames. *The Review of Economic Studies*, 75(4):1287–1296.

Sarver, T. (2008). Anticipating Regret: Why Fewer Options may be Better. *Econometrica*, 76(2):263–305.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-Based Choice. *Cognition*, 49(1-2):11–36.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, 16(2):158–174.

Strotz, R. H. (1955). Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies*, 23(3):165–180.

Tversky, A. and Simonson, I. (1993). Context-Dependent Preferences. *Management Science*, 39(10):1179–1189.