Revealed Deliberate Preference Change

Niels Boissonnet, Alexis Ghersengorin, Simon Gleyze

Revealed Deliberate Preference Change*

Niels BOISSONNET† Alexis GHERSENGORIN‡ Simon GLEYZE§

Abstract

We propose a model of rational preference change that is identifiable, empirically testable and founded on two normative principles. First, the decision maker (DM) must be able to *justify* her preference change by making attributes of the alternatives relevant or irrelevant. For instance, an employer who stops discriminating against candidates based on race or gender must make the attribute "race" or "gender" irrelevant for all her future choices. Second, DM must be *consistent* in her preference change. For instance, in the future the employer cannot make relevant again the attribute "race" or "gender", unless she makes (ir)relevant other complementary attributes. We show that two axioms on choice data which reflect these principles are necessary and sufficient to represent preference change by the maximization of a meta-preference. Finally, we apply our model to derive new insights on the polarization of political preferences.

*First draft: May 19, 2020. This draft: May 17, 2022. We thank the editor and the anonymous referee for their valuable comments. We are grateful to Douglas Bernheim, Yves Breitmoser, Simone Cerreia-Vioglio, Franz Dietrich, Marco Mariotti, Pietro Ortoleva, Jean-Marc Tallon, Al Roth, Ariel Rubinstein and seminar participants at Bielefeld, CREST, PSE, Stanford and TUS-VI for helpful conversations and comments. A. Ghersengorin thanks ANR-17-CE26-0003 for its support. S. Gleyze acknowledges the support of the EUR grant ANR-17-EURE-0001.

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

‡Paris School of Economics, Paris 1 Panthéon-Sorbonne. Email: alexis.ghersengorin @pse-mail.eu

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

1 INTRODUCTION

Understanding how individuals change their behavior is critical for social sciences. Economists traditionally argue that decision makers (DMs) are Bayesian; that is, they adapt their behavior by updating their beliefs about the environment. Although this mechanism has proved powerful and normatively appealing, a wide range of phenomena seem better described with preference change because they involve values such as fairness, conservatism, etc. For instance, Barrera et al. (2020) show experimentally that exposure to fake news about the European refugee crisis increases voting intentions toward far-right politicians, even though voters' beliefs may not change in case of fact-checking. Their explanation is that by raising voters' awareness of the migration issue, politicians may alter preferences. Another example is the expansion of abortion rights in western societies—along with its economical and political implications—that is more plausibly due to the diffusion of new values such as women's rights than to changing beliefs on some underlying state of the world.

Modeling preference changes raises two challenges: first, the lack of normative foundations compared to Bayesian updating; second the lack of testability of the model. To fill these gaps, we propose and axiomatize two testable normative principles: a *principle of sufficient reason* and a *principle of deliberation*. To express these normative principles, we use the attribute-based approach. Our primitive is the observation of successive preferences, as well as the attributes of each alternative. This allows us to define the attributes that are *relevant* to DM's choice at each period and, thereby, to reveal DM's reasoning behind preference changes. In doing so, we make progress toward a testable and normatively founded model of preference change.

The principle of sufficient reason states that DM changes her preferences if and only if it can be justified by an attribute of the alternative that is made relevant or irrelevant. For instance, if an employer becomes aware that her hiring decision is based on the attribute "gender", she might make this attribute irrelevant in the future to stop being discriminatory.[1] Formally, this translates into an identification axiom called Restricted Reversals, which guarantees that preference reversals can be explained by changes of relevant attributes alone (proposition 1).

The principle of deliberation states that DM should not make mistakes (from

[1] *Implicit* discrimination would also imply that the attribute "gender" is *relevant*. Therefore, an attribute can be relevant even if DM does not consciously use this attribute.

her perspective) when changing preferences; that is, she cannot change her mind twice regarding an attribute if no additional event occurred meanwhile. Otherwise, this would indicate that she fails to deliberate and lacks internal consistency. Formally, this translates into an Acyclicity axiom, which guarantees that if an attribute is made relevant and then irrelevant it must be explained by *other* attributes becoming (ir)relevant meanwhile.

Our main representation theorem states that Restricted Reversals and Acyclicity hold if and only if (i) preferences are represented by an ordering on the alternatives' attributes—we call this the *attribute ordering*—, and (ii) preference changes are explained by the maximization of an ordering on preferences themselves—we call this the *meta-preference* (theorem 1).

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. The succession of such changes is consistent with the maximization of a meta-preference relation, capturing DM's moral values, motivated reasoning, social objectives, norms, etc. Therefore, the reasoning behind preference changes is revealed through the meta-preference relation and the sequence of awareness. Such a sequence represents DM's constraint regarding which preferences are reachable at each period. The existence of such a constraint follows from the principle of deliberation and the observation of multiple preference reversals. Indeed, should DM be unconstrained in the maximization of her meta-preference she would directly reach her most preferred set of relevant attributes and never change preferences again. Note that the attribute ordering remains stable and that only the set of relevant attributes changes; this implies that if DM deems relevant the same set of attributes from one period to another, she must make exactly the same choices.[2] See Figure 1 for a representation of the model. Models of chosen preferences are receiving renewed attention since Bernheim et al. (2021), and the present paper is the first to investigate its revealed preference implications.

The attribute ordering and the meta-preference are essentially unique.[3] Furthermore, if two sequences of awareness represent DM's constraint on meta-choices, so does their intersection (proposition 2). We, however, stress that the sequence of

[2]We discuss why it would be problematic that DM changes her "taste" towards the attributes in section 2.5.

[3]That is, if two distinct attribute orderings (resp. meta-preferences) rationalize the preference changes, any completion of their intersection do so.

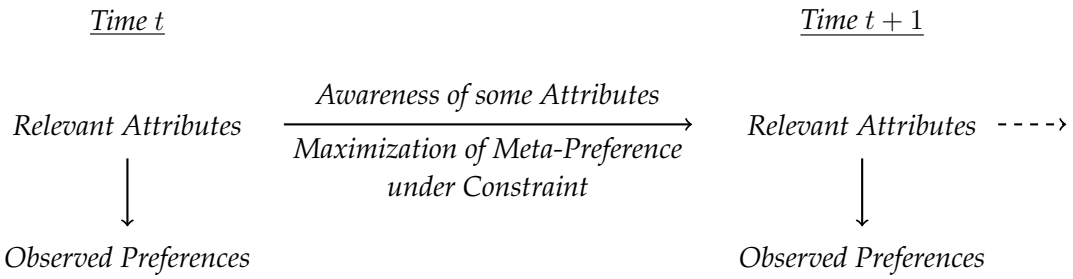

Figure 1: *The Dynamics of Deliberate Preference Change.*

relevant attributes is not uniquely identified in general. Hence, we investigate specific conditions that make this sequence set-identified or point-identified (proposition 4 and theorem 2).

We then investigate a particular type of meta-preference (i.e., a particular type of reasoning) in which DM chooses the preferences that maximize her underlying utility (theorem 3). This captures *motivated preference change* in which DM's evaluation of the attributes is guided by her own-interest alone. We show that motivated preference change provides new insights on the formation of political preferences. For instance, if two voters with identical preferences become aware of the same attributes in a different order, they can end up endorsing antagonistic views. Whether a voter becomes aware that a politician is corrupted before or after learning his political affiliation can lead to very different outcomes: in the latter case, the voter might ignore this attribute because it undermines the view of her preferred candidate. This type of path-dependent motivated reasoning is specific to our model and provides empirically testable implications.

Our contribution is threefold: first, we show that models incorporating preference changes can have empirical content and normative foundations. Second, our model suggests that choice reversals need not be irrational, and may reflect DM aligning her choice behavior with her values. Any deliberate preference change must break (or create) indifference with respect to other pairs of alternatives that share the same attribute, which indicates that this attribute becomes relevant (resp. irrelevant). This is a necessary condition for preference change to be induced by a coherent reasoning from DM. Finally, we illustrate the explanatory power of our model through an application.[4]

[4]All proofs until section 2.6 are in the Appendix. The remaining proofs can be found in the

Related Literature. The idea of representing objects by their attributes goes back to Lancaster (1966). Moreover, we draw on an important literature on reason-based theories of choice, most notably Simonson (1989), Shafir et al. (1993), Tversky and Simonson (1993), and Dietrich and List (2013, 2016). Boissonnet (2019) and Dietrich and List (2011) also use an attribute-based approach to model non-informational preference change. Our paper should be seen as the first counterpart of these models within the revealed preference theory.

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. 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. The main differences with our paper is that they consider deviations between expected behavior and actual behavior which are 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 that she may change preferences in the future. The closest paper in this literature to our own is Nehring (2006) who studies the revealed preference implications of second-order preferences as a self-control mechanism. The main differences with our paper are that he considers preferences over menus whereas we deal with preferences over alternatives, he does not introduce attributes, and the second order preferences act exclusively as a self-control mechanism whereas our meta-preference relation is completely general.

Our work relates to the literature on conflicting motivations—or justifiable choices—as we also obtain a representation with several (more precisely two) orderings. See among other contributions Kalai et al. (2002), Heller (2012), De Clippel and Eliaz (2012), Cherepanov et al. (2013), Dietrich and List (2016) and Ridout (2021). Despite this similarity, these works focus on static choice data that violates the usual rational requirements—namely the Weak Axiom of Revealed Preferences (WARP) or the Independence of Irrelevant Alternatives Axiom (IIA)—whereas in our work, the choice data consists in an ordered sequence of choices on the same collection of menus of options. We explore two distinct situations, one in which within-period choices are represented by not necessarily transitive bi-

Supplement (Boissonnet et al., 2022).

nary relations, one in which within-period choices satisfy WARP. We focus on the irregularities in choices that arise between periods, hence the reversals can happen on the same menus. Furthermore, the time structure is used to rationalize the successive changes as being guided by a meta-maximization.

In the applied theory literature, the closest paper is Bernheim et al. (2021). 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. For the purpose of falsification, our model makes some simplifications: in their model DM anticipates her preference change, and they allow for convex combinations of worldviews. Despite the differences in modelling assumptions, their paper is complementary with ours as we focus on the identification and falsification of deliberate preference changes. Other models of chosen preferences include Becker and Mulligan (1997), Akerlof and Kranton (2000), Palacios-Huerta and Santos (2004).

2 Deliberate Preference Change

2.1 Preliminaries

There is finite set X of **alternatives**, that are defined by their **attributes**. Formally, there are K attributes and an alternative is a vector $\mathbf{x} = (x^1, \dots, x^K)$ in the vector space $\mathbb{R}^K$ whose k^{th}-coordinate describes the value x^k of the attribute k.[5] For any subset $M \subseteq \{1, \dots, K\}$, denote $\mathbf{x}^M = (x^k)_{k \in M}$ and $\mathbf{x}^{-M} = (x^k)_{k \notin M}$.[6] We require that for any attribute k there exist $\mathbf{x}$ and $\mathbf{y}$ such that $x^k \neq y^k$, as otherwise this attribute could be removed.

The analyst observes (i) the value of each attribute for all alternatives, and (ii) choices over options for T periods of time. The latter are represented by a sequence of complete orders $(\succsim_t)_{t=1,\dots,T}$, where $\succ_t$ and $\sim_t$ denote the asymmetric and symmetric parts, respectively. For the first part of the analysis, we do not require each $\succsim_t$ to be transitive. We investigate the implications of transitivity within periods

[5] Attributes can either code different categories (e.g colors, sex, etc.), indicate whether a property is possessed by the alternative (e.g whether a job applicant is a foreigner or not), or measure the intensity of a property (e.g years of experience of a candidate).

[6] If $M = \{k\}$ is a singleton, we simply write $\mathbf{x}^{-k}$ instead of $\mathbf{x}^{-\{k\}}$.

—that is, DM's choices satisfy WARP— in section 2.6.

Example 1: Labor Market Discrimination. *An employer wants to hire a worker. Her decision is based on the resume of each candidate that provides information on three attributes: (1) "education", (2) "experience", and (3) "gender" (1 for female and 2 for male). Therefore, a female college-educated worker entering the labor market is represented by* $\mathbf{x} = (4, 0, 1)$*, while a male non-educated worker with ten years of experience is represented by* $\mathbf{y} = (0, 10, 2)$*.*

2.2 Revealed Relevant Attributes

The attribute-based approach allows us to identify which attributes drive DM's choice behavior. These "relevant attributes" are easy to identify when the choice set X is sufficiently rich: the attribute k is revealed relevant at t if there is a pair of alternatives $\mathbf{x}$ and $\mathbf{y}$ that only differ on the k^{th}-dimension ($\mathbf{x}^{-k} = \mathbf{y}^{-k}$) and such that $\mathbf{x} \not\sim_t \mathbf{y}$. In this case, we are sure that DM uses attribute k in her decision making. This richness assumption—that we can always find two alternatives that differ only on one dimension—would be too restrictive, however. We illustrate the construction of the *revealed relevant* attributes using our running example and then provide a formal definition.

Example 1 (continued): *Suppose that* $\mathbf{z} \succ_t \mathbf{x}$ *for the two candidates* $\mathbf{x} = (4, 0, 1)$*,* $\mathbf{z} = (4, 2, 2)$*. The idea is to identify a set of attributes* $M \subset \{1, 2, 3\}$ *that has to be relevant to explain this strict preference. From* $\mathbf{z} \succ_t \mathbf{x}$*, we can conclude that* $M = \{2, 3\}$ *is revealed relevant because (i) the alternatives differ on* M *and are identical outside of* M*, and (ii) there is no pair of alternatives that differ on a strict subset of* M *and are ranked strictly. The second point captures conservatism in our definition of revealed relevant attributes: if we cannot disentangle which attributes drive DM's behavior exactly, we keep all attributes in* M*. The following definition formalizes points (i) and (ii).*

Definition 1 (Revealed Relevant Attributes). *A set* M *of attributes is* ***revealed relevant*** *at period* t *if:*

(i) there exists $\mathbf{x}, \mathbf{y} \in X$ *with* $\mathbf{x}^{-M} = \mathbf{y}^{-M}$ *and* $x^k \neq y^k$ *for every* $k \in M$*, such that* $\mathbf{x} \not\sim_t \mathbf{y}$*;*

(ii) for every $M' \subsetneq M$ *and every* $\mathbf{w}, \mathbf{z} \in X$ *with* $\mathbf{w}^{-M'} = \mathbf{z}^{-M'}$*,* $\mathbf{w} \sim_t \mathbf{z}$*.*

Remark: if two attributes are systematically revealed relevant together, they might be coded into a single attribute (for instance if colors have been coded into different binary attributes).

Let P_t denote the collection of sets of revealed relevant attributes at period t. We denote $\mathbf{m}_t \in \{0,1\}^K$ the **vector of revealed relevant attributes** such that $m_t^k = 1$ if $k \in \bigcup_{M \in P_t} M$ and $m_t^k = 0$ otherwise.[7]

We emphasize that an attribute can be revealed relevant, yet DM might be unaware that it causes her behavior. For instance, it is well known that implicit discrimination can have a strong impact on job performance (Bertrand et al., 2005; Glover et al., 2017; Bertrand and Duflo, 2017).

2.3 Principle of Sufficient Reason

We impose the following principle of sufficient reason: DM changes preferences if and only if the revealed relevant attributes change. The interpretation is that DM does not "wake up" with different preferences but must be able to justify her new preferences by making some attributes relevant or irrelevant. We view this as a normative principle: unjustified changes would not be normatively compelling.

Formally, the axiom states that if two alternatives $\mathbf{x}$ and $\mathbf{x}'$ have the same relevant attributes between periods t and t'—namely, if $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ where $\circ$ denotes the element-wise (Hadamard) product—DM should rank consistenly $\mathbf{x}$ against the other alternatives in period t and $\mathbf{x}'$ against the other alternatives in period t'.

Restricted Reversals. *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals if for any t, t', and for any $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'}$,*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_{t'} \mathbf{y}'.$$

Remark: although we do not impose restrictions on the values of the attributes, the value 0 has a specific role. If no attribute can take the value 0, this axiom can simply be stated as: if $\mathbf{m}_t = \mathbf{m}_{t'}$, then $\succsim_t = \succsim_{t'}$. This would imply that DM changes her evaluation towards every option each time an attribute is made (ir)relevant. We stress that, although it makes sense to remove 0 for some attributes (e.g a category attribute coding the color of an item),

[7] Our definition of revealed relevant attributes is analogous to the definition of a non-null state in expected utility theory (taking the attributes as states and the alternatives as acts).

it is not necessarily the case for attributes measuring the intensity of a property (e.g years of experience) or binary attributes that indicate whether a property is possessed or not (e.g whether a job applicant is a foreigner or not). This axiom therefore imposes that some alternatives be ranked similarly in different periods, even though the revealed relevant attributes might change.

Example 1 (continued). *Consider four candidates* $\mathbf{x} = (6, 2, 1), \mathbf{x}' = (0, 2, 1)$, $\mathbf{y} = (5, 0, 2)$ *and* $\mathbf{y}' = (0, 0, 1)$. *Suppose that the only strict rankings of* $\succsim_1$ *are* $\mathbf{x} \succ_1 \mathbf{x}' \succ_1 \mathbf{y}'$ *whereas the only strict ranking of* $\succsim_2$ *is* $\mathbf{x}' \succ_2 \mathbf{y}'$. *It is verified that the vectors of revealed relevant attributes are* $\mathbf{m}_1 = (1, 1, 0)$ *and* $\mathbf{m}_2 = (0, 1, 0)$ *respectively. Observe that* $\mathbf{x}' \circ \mathbf{m}_1 = \mathbf{x} \circ \mathbf{m}_2$, *hence* $\mathbf{x}$ *and* $\mathbf{x}'$ *have the same relevant attributes at periods 1 and 2. Similarly,* $\mathbf{y}' \circ \mathbf{m}_1 = \mathbf{y} \circ \mathbf{m}_2$. *Therefore, this sequence of choices violate Restricted Reversals, given that* $\mathbf{x}' \succ_1 \mathbf{y}'$ *whereas* $\mathbf{x} \sim_2 \mathbf{y}$.

A consequence of this axiom is the existence of a bijection between vectors of revealed relevant attributes and preference relations. Namely, this axiom is necessary and sufficient to represent the sequence of preferences $(\succsim_t)_t$ by the sequence of revealed relevant attributes $(\mathbf{m}_t)_t$ together with a time-independent binary relation over vectors of attributes. Formally, for any period t, let $X(\mathbf{m}_t) = \{\mathbf{x} \circ \mathbf{m}_t : \mathbf{x} \in X\}$ be the set of alternatives "filtered" through the revealed relevant attributes $\mathbf{m}_t$, and denote $\bar{X} = \bigcup_t X(\mathbf{m}_t)$.

Proposition 1. *Preferences* $(\succsim_t)_t$ *satisfy Restricted Reversals if and only if there exists a complete binary relation* $\geqslant \subseteq \bar{X}^2$ *(called the* ***attribute ordering****), such that for any period* t *and any* $\mathbf{x}, \mathbf{y} \in X$*:*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t. \tag{1}$$

The interpretation is that DM has a fundamental preference—called the *attribute ordering*—that, unlike her choices $(\succsim_t)_t$, does not change over time. This attribute ordering ranks vectors of attributes and does not depend on the relevant attributes.[8] The main consequence of Proposition 1 is that preference change can only be induced by changes in relevant attributes. Observe that the attribute ordering need not be transitive. We derive necessary and sufficient conditions for a transitive attribute ordering in Section 2.6.

[8] In a slightly different framework, Dietrich and List (2013) provide an equivalence result between this separability condition (their axiom 2) and the existence of an attribute ordering.

2.4 Principle of Deliberation

The second normative principle that guides our analysis is a principle of deliberation: DM must evaluate all possible preferences at time t and consistently choose the best feasible one according to some criterion. This translates into an acyclicity axiom, which states that if DM changes her preference once, every future change should be due to the discovery of some new attributes—i.e that were not involved in the first change—and towards which DM has changed her attitude meanwhile.

ACYCLICITY. *Preferences $(\succsim_t)_t$ satisfy Acyclicity if for any t and any $t' > t+1$, if $\mathbf{m}_{t+1} \neq \mathbf{m}_{t'}$, then there exists k such that $m^k_{t'} \neq m^k_{t+1} = m^k_t$.*

Note that, as soon as several choice reversals are observed, the principle of deliberation implies the existence of a constraint on preference change. Indeed, would preference change be unconstrained, DM would directly reach her most preferred preference once and for all. We interpret this constraint as DM's awareness: she can change only the attributes she is aware of, that is, the ones she is able to question.

Example 1 (continued). *Suppose that $\mathbf{m}_1 = (0,0,1)$ and $\mathbf{m}_2 = (0,1,0)$, namely the recruiter makes gender relevant but experience irrelevant at the second period. This could be because on the market men are more experienced, implying an unfair discrimination. Therefore, she must have been able to modify her relevant attributes (at least) on these two attributes. Acyclicity implies that she could never choose the following relevant attributes in the future: $(0,0,0)$ and $(0,1,1)$ as they were accessible between period 1 and period 2. Since she did not change the relevance of the education attribute, we conclude that she was not aware of this attribute at this point. Assuming for instance that education provides a fair criterion to rank the candidates, she could later on decide to remove again gender only if education is made relevant jointly, reaching $\mathbf{m}_3 = (1,0,0)$.*

2.5 The Representation

The constraint on preference change in the representation is formalized by a sequence of vectors $(\mathbf{a}_t)_{t=1}^{T-1}$, which represents DM's **awareness** between each period t and $t+1$. Namely, $\mathbf{a}_t \subseteq \{0,1\}^K$ for any t and codes as 1 attributes that DM can

modify and as 0 the ones that she cannot modify between t and $t+1$. An awareness vector $\mathbf{a} \in \{0,1\}^K$ together with a vector of relevant attributes $\mathbf{m} \in \{0,1\}^K$ defines a set of **reachable attributes** for the next period $R(\mathbf{m}, \mathbf{a})$:

$$R(\mathbf{m}, \mathbf{a}) = \left\{ \mathbf{m}' \in \{0,1\}^K : \text{for all } k, \mathbf{a}^k = 0 \text{ implies } \mathbf{m}'^k = \mathbf{m}^k \right\}.$$

To state our main result, define for any set A and any linear order $P \subset A^2$, $\max(A, P) = \{a \in A \mid aPb,\ \forall b \in A\}$.

Theorem 1 (Representation)**.** *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals and Acyclicity if and only if there exists a complete binary relation $\geqslant \subseteq \bar{X}^2$, a sequence of awareness $(\mathbf{a}_t)_t$ (with $\mathbf{a}_t \in \{0,1\}^K$), and a linear order $\rhd \subseteq \{0,1\}^K \times \{0,1\}^K$,[9] such that, for any t and any $\mathbf{x}, \mathbf{y} \in X$,*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t, \tag{1}$$

$$\{\mathbf{m}_{t+1}\} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd). \tag{2}$$

The principle of sufficient reason together with the principle of deliberation are necessary and sufficient for what we name a **deliberate preference change model**. If the tuple $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ satisfy the conditions in theorem 1, we say that it **rationalizes** $(\succsim_t)_t$. In this model, DM's behavior is represented by the maximization of two binary relations: a preference relation on alternatives that together with the relevant attributes determine choices in each period (1) and a meta-preference relation on vectors of relevant attributes that determine the change of preference between periods (2). The revealed preference implication of our model is that when we observe choice reversals between alternatives $\mathbf{x}$ and $\mathbf{y}$, we should observe other choice reversals on alternatives that share attributes with $\mathbf{x}$ and $\mathbf{y}$. For instance if an employer stops discriminating at work this should impact her preferences in other contexts, such as her political preferences.

The fact that attributes can only be made relevant or irrelevant—and that DM cannot change her "taste" (attribute ordering) towards an attribute due to the stability of the attribute ordering—might seem arbitrary at first. This is however important for the testability of the model, as otherwise almost any sequence of observed choices could be rationalized by changing DM's tastes. Furthermore, if

[9] It is observationally equivalent to construct a linear order or a complete preorder together with a tie-breaking rule for the meta-choice such that if $\mathbf{m}_t = \mathbf{m}$ and $\mathbf{m}_{t'} \neq \mathbf{m}$ for some $t' > t$, then $\mathbf{m}_\tau \neq \mathbf{m}$ for all $\tau > t'$.

the space of attributes is correctly specified from the beginning, there is no need to change DM's tastes. For instance, if the employer makes "gender" irrelevant to avoid discrimination, but makes it relevant again in the future due to an affirmative action policy, this policy should be thought of an attribute that is complementary with the attribute "gender". Therefore, it is not that DM changes her tastes toward the attribute "gender", but that the combination of "gender" and "affirmative action" is strictly preferred to "gender" alone. This suggests that the specification of the attributes is a crucial step that the researcher should discuss carefully, and commit to before observing choice data to avoid ex-post rationalization.

What can be inferred if one of the two axioms is violated? First, a violation of Restricted Reversal indicates that preference changes do not arise from changes in DM's revealed relevant attributes. Indeed, it is a necessary and sufficient condition for the existence of a time-independent attribute ordering that rationalizes each period's preference together with a set of relevant attributes (proposition 1). Therefore, the analyst's knowledge of what determines DM's preference is incomplete: we may not observe all attributes, or the attribute ordering may change because DM discovers new consequences of an attribute for instance. Second, a violation of Acyclicity suggests that DM does not change her preferences *rationally*, meaning that no linear order can rationalize the sequence of meta choices. Canonical examples of non-deliberate preference changes are nudges, conformism or random utility. Alternatively, a violation of these axioms may suggest that the revealed relevant attributes are not the "truly" relevant attributes for DM, and her behavior could be rationalized by our model with a different sequence $(\mathbf{m}'_t)_t$.[10]

Finally, we emphasize that our model is complementary with Bayesianism to explain preference change. Even though evidence suggests that agents do not always follow Bayes' rule, we do not think that an exhaustive theory of social interactions could do without belief updating. Instead, we argue that preference change and belief updating can occur simultaneously. This thesis receives empirical support in experiments on fake news by Barrera et al. (2020) (cited in the introduction).

Regarding the uniqueness of the ingredients of a deliberate preference change model, without further restrictions, only the attribute ordering is uniquely re-

[10]Note that if one does not want to restrict attention to revealed relevant attributes, it is possible to write axioms on multiple "candidate" sequences of relevant attributes (details available upon request).

vealed and the meta-preference identified up to an arbitrary completion. Furthermore, any intersection of two rationalizing sequences of awareness can also rationalize preference changes.

Proposition 2 (Uniqueness). *Let* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *and* $(\geqslant', \rhd', \mathbf{m}_t, \mathbf{a}'_t)$ *rationalize* $(\succsim_t)_t$. *Then, any completion of* $\geqslant \cap \geqslant'$ *and* $\rhd \cap \rhd'$, *together with* $(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ *also rationalize* $(\succsim_t)_t$.

In Section 2.7, we derive sufficient conditions for the identification of the relevant attributes, in the case where the the preferences at each period are transitive as well as the attribute ordering. Hence we first study in the next section the characterization of deliberate preference changes with transitive attribute ordering.

2.6 Transitive Attribute Ordering

Our main representation theorem does not guarantee that the attribute ordering is transitive and does not require that the observed preferences $(\succsim_t)_t$ are transitive. Indeed Restricted Reversals constraints choices only between pairs of periods which is not enough to guarantee transitivity. For instance, suppose that $\mathbf{x}, \mathbf{y} \in X(\mathbf{m}_t)$, $\mathbf{y}, \mathbf{z} \in X(\mathbf{m}_{t'})$ and $\mathbf{x}, \mathbf{z} \in X(\mathbf{m}_{t''})$ but $\mathbf{z} \notin X(\mathbf{m}_t), \mathbf{x} \notin X(\mathbf{m}_{t'})$ and $\mathbf{y} \notin X(\mathbf{m}_{t''})$. It could be that $\mathbf{x} \succ_t \mathbf{y}, \mathbf{y} \succ_{t'} \mathbf{z}$ and $\mathbf{z} \succ_{t''} \mathbf{x}$ because Restricted Reversals does not constraint choices on triplets of periods. In fact, this problem is more general and may arise with any number of periods strictly greater than two.

Transitivity of preferences is sometimes viewed as a condition for rationality, hence it might be of interest to characterize transitivity of the attribute ordering. The following axiom extends Restricted Reversals to address this problem.

Strong Restricted Reversals. *For any* $\{t_1, \ldots, t_n\}$ *and any* $\{\mathbf{x}_k, \mathbf{x}'_k\}_{k=1,\ldots,n}$ *such that, for* $k = 1, \ldots, n-1$, $\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}}$ *and* $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1}$, *preferences* $(\succsim_t)_t$ *satisfy Strong Restricted Reversals if:*

$$\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k, \text{ for every } k = 1, \ldots, n-1 \implies \mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n.$$

Proposition 3. *Suppose that preferences* $(\succsim_t)_t$ *are transitive. Preferences satisfy Strong Restricted Reversals and Acyclicity if and only if there exists a deliberate preference change model* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *that rationalizes them with* $\geqslant$ *being a complete preorder.*

2.7 Identification of the Revealed Relevant Attributes

The relevant attributes are typically not identified without further restrictions on preferences. This is the case because when we observe an indifference, we cannot always identify whether this is due to an attribute being irrelevant, or whether DM is indifferent towards this attribute in the attribute ordering. Denote $\mathcal{M}(\succsim_t) = \{\mathbf{m} : \exists \text{ a preorder } \geqslant \text{ s.t. } (\mathbf{m}, \geqslant) \text{ rationalizes } \succsim_t\}$ the set of relevant attributes that rationalize preferences at t using a transitive attribute ordering. To explore the structure of $\mathcal{M}(\succsim_t)$ we make the following richness assumption.

RICHNESS ASSUMPTION. *For all $\mathbf{x}, \mathbf{y} \in X$ that differ only on a subset M of n attributes, there is a sequence of alternatives $\mathbf{z}_1, \ldots, \mathbf{z}_n \in X$ such that $\mathbf{z}_1 = \mathbf{x}$, $\mathbf{z}_n = \mathbf{y}$ and $\mathbf{z}_i^{-k} = \mathbf{z}_{i+1}^{-k}$ for some $k \in M$, for all $i = 1, \ldots, n-1$.*[11]

We show that, under the richness assumption and the transitivity of the preferences $\succsim_t$, the set of vectors of relevant attributes $\mathbf{m}$ that can be used to rationalize preferences in the baseline model has a lattice structure. The most parsimonious vector is the vector of revealed relevant attributes $\mathbf{m}_t$,[12] but in principle other vectors could be used to rationalize DM's preferences.

Proposition 4. *Assume richness and suppose that preferences $\succsim_t$ are transitive. If Restricted Reversal is satisfied, $\mathcal{M}(\succsim_t)$ is a lattice ordered by $\geq$. Its minimum is the vector of revealed relevant attributes $\mathbf{m}_t$ and its maximum is $(1, \ldots, 1)$.*

This indeterminacy problem between irrelevant attributes and indifference can be solved if we impose that indifference are *only* caused by an attribute being irrelevant. In this case, an indifference $\mathbf{x} \sim_t \mathbf{y}$ has a clear interpretation in the sense that there is no attribute that motivates DM to choose $\mathbf{x}$ over $\mathbf{y}$. This is the content of the following axiom.

JUSTIFIED INDIFFERENCE. *Preferences $(\succsim_t)_t$ satisfy Justified Indifference if for any t and any alternatives $\mathbf{x}, \mathbf{y} \in X$,*

$$\mathbf{x} \sim_t \mathbf{y} \implies |\mathbf{x} - \mathbf{y}| \circ \mathbf{m}_t = (0, \ldots, 0).$$

[11] For this assumption to be satisfied, it might be necessary to regroup certain attributes. For instance, splitting a category attribute (e.g color) into binary attributes will violate this assumption.

[12] If X is not rich, the vector of revealed relevant attributes $\mathbf{m}_t$ need not be the minimum of the lattice.

When Justified Indifference is satisfied and if we restrict attention to strict attribute ordering, the relevant attributes are uniquely identified by the revealed relevant ones. Formally, let $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m} : \exists \text{ a partial order } > \text{ s.t. } (\mathbf{m}, >) \text{ rationalizes } \succsim_t \}$ be the set of relevant attributes that rationalize preferences at t using a strict attribute ordering. When Justified Indifference is satisfied, we have $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.

Theorem 2. *Assume richness and suppose that preferences $(\succsim_t)_t$ are transitive. Preferences satisfy Strong Restricted Reversal, Acyclicity and Justified Indifference if and only if there exists a deliberate preference change model $(>, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ that rationalizes $(\succsim_t)_t$ with $>$ being a partial order. Furhermore, for any period t, $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.*

3 Motivated Preference Change

Our main representation theorem shows that preference change can be represented by the maximization of a meta-preference. The representation, however, does not provide a straightforward interpretation of the meta-preference. It could be that DM is changing her behavior to make it more aligned with her values, or she may change preferences to serve her own-interests instead of purely disinterested motives—this is referred to as *motivated preference change*. In this section, we investigate the latter idea. We show that motivated preference change admits a tractable functional representation—this proves convenient for applications in the next section.

First, we construct an extension of the attribute ordering which allows us to keep track of (i) preferences over perceived alternatives at period t, and (ii) preferences over perceived alternatives at period t if she were to change her preferences to make good alternatives even better.

Definition 2. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$. Denote $\mathbf{a} \gg_t \mathbf{b}$ if $\mathbf{x} \circ \mathbf{m}_t = \mathbf{a}$ for some $\mathbf{x} \in X$ and*

(i) $\mathbf{y} \circ \mathbf{m}_t = \mathbf{b}$ *for some* $\mathbf{y} \in X$ *and* $\mathbf{x} \succsim_t \mathbf{y}$*; or*

(ii) $\mathbf{y} \circ \mathbf{m} = \mathbf{b}$ *for some* $\mathbf{y} \in X$, $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, *and* $\mathbf{x} \succsim_t \mathbf{z}$ *for all* $\mathbf{z} \in X$.

The following axiom, which extends Strong Restricted Reversals, guarantees that DM makes attributes relevant if and only if these attributes are valued positively—that is, making these attributes (ir)relevant increases DM's utility.

MOTIVATED RESTRICTED REVERSALS. *Preferences $(\succsim_t)_t$ satisfy Motivated Restricted Reversals if for any $\{t_1, \ldots, t_n\}$ and any $(\mathbf{a_k})_{k=1,\ldots,n} \in (\mathbb{R}^K)^n$ such that $\mathbf{a}_{k+1} \gg_{t_k} \mathbf{a}_k$ for $k = 1, \ldots, n-1$,*

$$\mathbf{a}_1 \gg_{t_n} \mathbf{a}_n \implies \mathbf{a}_1 \ll_{t_n} \mathbf{a}_n.$$

The next axiom guarantees that there are no indifference between vectors of relevant attributes when changing preferences. Intuitively, the axiom states that if there is a tie between two vectors $\mathbf{m}$ and $\mathbf{m}'$ that yield identical utility, DM breaks the tie in favor of one vector by virtually increasing her utility for some alternative $x \in X$ so that $\mathbf{m}$ becomes strictly preferred to $\mathbf{m}'$.

MOTIVATED TIE-BREAKING. *Preferences $(\succsim_t)_t$ satisfy Motivated Tie-Breaking if for all t, all $\mathbf{x} \in \max(X, \succsim_t)$, and all $\mathbf{y}, \mathbf{y}' \in X$ such that there exists $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$ with $\mathbf{y}' \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m} \circ \mathbf{m}_t$,*

$$\mathbf{y}' \in \max(X, \succsim_t) \implies \mathbf{m} = \mathbf{m}_t.$$

These two axioms are necessary and sufficient for the motivated preference change representation.

Theorem 3 (Representation)**.** *Suppose that preferences $(\succsim_t)_t$ are transitive. Preferences $(\succsim_t)_t$ satisfy Motivated Restricted Reversals and Motivated Tie-Breaking if and only if there exists a sequence of awareness $(\mathbf{a}_t)_t$ and a function $u : \mathbb{R}^K \longrightarrow \mathbb{R}$ such that for all t and all $\mathbf{x}, \mathbf{x}'$,*

$$\mathbf{x} \succsim_t \mathbf{x}' \iff u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{x}' \circ \mathbf{m}_t),$$

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\operatorname{argmax}} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}).$$

As in the previous representation, DM chooses alternatives to maximize her attribute ordering, which can be represented by a utility function here. The main difference is that preference change must maximize DM's utility. Therefore, all attributes that are "negatively valued" will be made irrelevant as soon as possible, and all attributes that are "positively valued" will be made relevant as soon as possible.

4 An Application

An important feature of the model is path dependence—that is, the order in which DM becomes aware of certain attributes has a strong impact on the path of preference change. We illustrate this aspect in a voting context: ex-ante identical voters deliberately ignore what other voters think is relevant because this would undermine their view of their preferred candidate.[13] Therefore, we show that our model can account for polarization of political preferences among ex-ante identical voters in a simple and intuitive way.

Polarization refers to disagreement on policy issues or distrust of the other party members among politicians and citizens (Iyengar et al., 2019). There is now widespread agreement concerning the growing importance of ideological divisions both among politicized and educated voters as well as non-politicized citizens (Abramowitz and Saunders, 2008). There is no agreement, however, on the causes of polarization.[14]

From a Bayesian perspective, it is surprising that polarization increases as rational agents whose posterior beliefs are common knowledge cannot agree to disagree, even if their posteriors are based on different observed information about the world (Aumann, 1976). Arguing that voters have different priors certainly explains polarization, but it only moves the goalpost: where do differences in prior come from? Instead, our model provides a foundation for the concept of "partisan social identity" introduced in the political science literature (Iyengar and Westwood, 2015). This theory captures the tendency of voters to classify opposing partisans as members of an outgroup and copartisans as members of an ingroup. We show that our model can account for the construction of such opposing groups, and how partisan cues can reinforce division.

We consider a very stylized model with motivated preference change. There are two voters i and j and two candidates: $\mathbf{x}^D = (x^1, x^2, x^3)$ and $\mathbf{x}^R = (\tilde{x}^1, \tilde{x}^2, \tilde{x}^3)$ with $\tilde{x}^1 < 0 < x^1$, $x^2 < 0 < \tilde{x}^2$, $x^3 < 0 < \tilde{x}^3$ and $\tilde{x}^2 - \tilde{x}^1 > x^1 - x^2$. The first attribute captures the candidates' support for social policies (e.g. health care), the second attribute captures how conservative candidates are, and the third attribute

[13] Note that Bayesian updating cannot induce this type of path dependence because it is order invariant (Cripps, 2018).

[14] Recent finding suggests that the emergence of the internet or rising economic inequality are less plausible causes than changes that are specific to the US—e.g., changing party composition, growing racial divisions, or the emergence of partisan cable news (Boxell et al., 2020).

represents corruption. Voters are ex-ante identical: they both value integrity and prefer candidates with strong convictions (represented by a high absolute value of the difference between the first and the second attributes). We can represent their preferences as follows:

$$u(\mathbf{x} \circ \mathbf{m}) = \left(x^1 m^1 - x^2 m^2\right)^2 - x^3 m^3.$$

They both initially start with the vector of relevant attributes $(0, 0, 0)$. Suppose that voter i attends a political debate with both candidates: $\mathbf{a}_1^i = (1, 1, 0)$. She will change her preferences and value more candidate $\mathbf{x}^R$ who has stronger convictions: the meta-maximization writes

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)) = \left(\tilde{x}^2 - \tilde{x}^1\right)^2 &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,1,0)) = \left(\tilde{x}^2\right)^2 \\ &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,0,0)) = \left(x^1\right)^2 \\ &> 0 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,0)).\end{aligned}$$

Later, voter i becomes aware that candidate $\mathbf{x}^R$ is corrupted: $\mathbf{a}_2^i = (0, 0, 1)$. She decides to ignore this information and keep this attribute irrelevant if:

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,1)) &= \max\left\{\left(\tilde{x}^2 - \tilde{x}^1\right)^2 - \tilde{x}^3, \left(x^1 - x^2\right)^2 - x^3\right\} \\ &< \left(\tilde{x}^2 - \tilde{x}^1\right)^2 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)).\end{aligned}$$

i.e. whenever $\left(\tilde{x}^2 - \tilde{x}^1\right)^2 > \left(x^1 - x^2\right)^2 - x^3$. Namely, whenever candidate $\mathbf{x}^R$ has strong convictions that counterbalance her corruption. The intuition is that making "corruption" relevant would undermine her view of candidate $\mathbf{x}^R$. In the end, voter i's most preferred candidate is $\mathbf{x}^R$.

Instead, voter j first becomes aware of a felony committed by candidate $\mathbf{x}^R$: $\mathbf{a}_1^j = (0, 0, 1)$. She will change her preferences to make it relevant: the meta-maximization writes

$$\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,1)) = -x^3 > 0 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,0)).$$

At this point voter j prefers the upstanding candidate $\mathbf{x}^D$.

Later, voter j attends a political debate with both candidates: $\mathbf{a}_2^j = (1, 1, 0)$. She will lean toward the candidate $\mathbf{x}^D$ even though he has less convictions than the

candidate $\mathbf{x}^R$ whenever $(x^1 - x^2)^2 - x^3 > (\tilde{x}^2 - \tilde{x}^1)^2 - \tilde{x}^3$. Namely, whenever the convictions of $\mathbf{x}^R$ does not make up for his felonies. In the end, voter j's most preferred candidate is $\mathbf{x}^D$.

It is quite striking that two identical voters who become aware of the same attributes can become polarized. This arises due to the path dependence of preference change: past justifications can conflict with new justifications leading to rich dynamics.

APPENDIX

Proof of Proposition 1. We say that a pair $(\mathbf{m}, \geqslant)$ **rationalizes** $\succsim$ if for any $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \succsim \mathbf{y} \iff \mathbf{x} \circ \mathbf{m} \geqslant \mathbf{y} \circ \mathbf{m}$. The proof of the necessity is straightforward and therefore omitted.

(*Sufficiency*). Assume that $(\succsim_t)_t$ satisfy Restricted Reversals. First, we fix a period t and show that we can indeed construct an ordering $\geqslant_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$. We define the two following binary relations on $X(\mathbf{m}_t)$:

$$>_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \succ_t \mathbf{y}\},$$
$$\simeq_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \sim_t \mathbf{y}\}.$$

By definition, $\simeq_t$ is reflexive and symmetric. We show that $>_t$ is irreflexive, i.e. for any $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \neq \mathbf{y}$ and $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$, $\mathbf{x} \sim_t \mathbf{y}$. Indeed, $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$ implies that $\mathbf{x}$ and $\mathbf{y}$ do not differ on any attributes k such that $m_t^k = 1$. Hence, if by contradiction we had $\mathbf{x} \succ_t \mathbf{y}$, then there should be a subset of attributes on which $\mathbf{x}$ and $\mathbf{y}$ differ (i.e. with $m_t^k = 0$) and that are revealed relevant. This contradicts the definition of $\mathbf{m}_t$.

Now let $\mathbf{a}, \mathbf{b} \in X(\mathbf{m}_t)$, with $\mathbf{a} \neq \mathbf{b}$, and $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_t = \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_t = \mathbf{b}$. Applying Restricted Reversal with $t = t'$, we obtain $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_t \mathbf{y}'$. Given that $>_t$ is irreflexive, this establishes that it is asymmetric. It also proves that $>_t \cap \simeq_t = \emptyset$.

Therefore, the relation $\geqslant_t := \simeq_t \cup >_t$ is complete on $X(\mathbf{m}_t)$ (by the completeness of $\succsim_t$) ; $\simeq_t$ and $>_t$ being respectively its symmetric and asymmetric parts. Furthermore, $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$.

Second, we show that for any two distinct periods t and t', $\geqslant_t$ does not contradict $\geqslant_{t'}$. Let $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$ be such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'} =: \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'} =: \mathbf{b}$.

Then by Restricted Reversal we have,

$$\mathbf{a} \geqslant_t \mathbf{b} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant_t \mathbf{y} \circ \mathbf{m}_t \iff \mathbf{x} \succsim_t \mathbf{y} \overbrace{\iff}^{\text{Restricted Reversal}} \mathbf{x}' \succsim_{t'} \mathbf{y}'$$
$$\iff \mathbf{x}' \circ \mathbf{m}_{t'} \geqslant_{t'} \mathbf{y}' \circ \mathbf{m}_{t'} \iff \mathbf{a} \geqslant_{t'} \mathbf{b}$$

Finally, define $\geqslant := \bigcup_t \geqslant_t$. By the previous argument, $\geqslant \cap X^2(\mathbf{m}_t) = \geqslant_t$, so for any t, $(\mathbf{m}_t, \geqslant)$ rationalizes $\succsim_t$. Furthermore $\geqslant$ can be innocuously completed on $\bar{X}$. □

Proof of Theorem 1. (Necessity). We only prove the necessity of Acyclicity. Let t and t' such that $t+1 < t'$ and assume $\mathbf{m}_{t'} \neq \mathbf{m}_{t+1}$. By equation (2), $\mathbf{m}_{t'} \rhd \mathbf{m}_{t+1}$ and, thus, $\mathbf{m}_{t'} \notin R(\mathbf{m}_t, \mathbf{a}_t)$. Hence, $m_t^k \neq m_{t'}^k$ and $a_t^k = 0$ for some attribute k. Yet, $a_t^k = 0$ and $\mathbf{m}_{t+1} \in R(\mathbf{m}_t, \mathbf{a}_t)$ imply that $m_t^k = m_{t+1}^k$, and thus $m_{t+1}^k \neq m_{t'}^k$

(Sufficiency). We know from proposition 1 that there exists an attribute ordering $\geqslant \subseteq \bar{X}^2$, such that for any period t, $(\mathbf{m}_t, \geqslant)$ rationalizes $\succsim_t$. Define the sequence of awareness as $\mathbf{a}_t = |\mathbf{m}_{t+1} - \mathbf{m}_t|$ for any t; and the revealed meta-preference relation $\rhd$ as follows: $\mathbf{m} \rhd \mathbf{m}'$ if $\mathbf{m} \neq \mathbf{m}'$ and there exists t, such that $\mathbf{m} = \mathbf{m}_t$ and,

$$\mathbf{m}' \in \bigcup_{t' : t' < t} R(\mathbf{m}_{t'}, \mathbf{a}_{t'}).$$

We verify that $\rhd$ is asymmetric. Suppose that $\mathbf{m} \rhd \mathbf{m}'$ and take $t' < t$, such that $\mathbf{m}' \in R(\mathbf{m}_{t'}, \mathbf{a}_{t'})$ and $\mathbf{m} = \mathbf{m}_t$. First let us show that there cannot be $t'' > t$ such that $\mathbf{m}' = \mathbf{m}_{t''}$. Assume by contradiction that such a t'' exists. Then, we have

$$|\mathbf{m}_{t''} - \mathbf{m}_{t'}| \underbrace{=}_{\text{Def. } \mathbf{m}_{t''}} |\mathbf{m}' - \mathbf{m}_{t'}| \underbrace{\leq}_{\mathbf{m}' \in R(\mathbf{m}_{t'}, |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|)} |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|$$

where $|\cdot|$ is the element-wise absolute value: for any vector $\mathbf{b} \in \mathbb{R}^K$, $|\mathbf{b}| = (|b^1|, \ldots, |b^K|)$. This means that for all k, if $m_{t'}^k = m_{t'+1}^k$, then $m_{t''}^k = m_{t'}^k = m_{t'+1}^k$. Thus Acyclicity implies that $\mathbf{m}_{t'+1} = \mathbf{m}_{t''} = \mathbf{m}'$. But, then we still have that $\mathbf{m}' \in R(\mathbf{m}_{t'+1}, |\mathbf{m}_{t'+2} - \mathbf{m}_{t'+1}|)$ so that, applying the previous reasoning inductively, we obtain $\mathbf{m}' = \mathbf{m}_{t'+2} = \mathbf{m}_{t'+3} = \ldots \mathbf{m}_t = \mathbf{m} \neq \mathbf{m}'$. A contradiction. Second assume by contradiction that $\mathbf{m}' = \mathbf{m}_{t'''}$ and $\mathbf{m} \in R(\mathbf{m}_{t''}, |\mathbf{m}_{t''+1} - \mathbf{m}_{t''}|)$ for some t'', t''' such that $t'' < t''' < t$. By the same argument, Acyclicity would then imply that $\mathbf{m} = \mathbf{m}_{t''+2} = \mathbf{m}_{t''+3} = \ldots \mathbf{m}_{t'''} = \mathbf{m}' \neq \mathbf{m}$. A contradiction.

We now verify that $\rhd$ is transitive. Suppose that $\mathbf{m} \rhd \mathbf{m}'$ and $\mathbf{m}' \rhd \mathbf{m}''$. Then

there exist t, t' with $t > t'$, such that, $\mathbf{m} = \mathbf{m}_t$ and $\mathbf{m}' = \mathbf{m}_{t'}$. Moreover,

$$\mathbf{m}'' \in \bigcup_{t''':t'''<t'} R(\mathbf{m}_{t'''}, \mathbf{a}_{t'''}) \subseteq \bigcup_{t'':t''<t} R(\mathbf{m}_{t''}, \mathbf{a}_{t''})$$

where the inclusion follows from $t > t'$. We conclude that $\mathbf{m} \rhd \mathbf{m}''$, implying the transitivity of $\rhd$.

Moreover, by the definition of $\rhd$, $\mathbf{m}_{t+1} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd)$. By Szpilrajn's theorem, the meta-preference can be completed on $\{0,1\}^K \times \{0,1\}^K$. □

Proof of Proposition 2. For any t, any $\mathbf{a}, \mathbf{b} \in X(\mathbf{m}_t)$, $\mathbf{a} \succcurlyeq \mathbf{b}$ if and only if there exist, $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{a}$, $\mathbf{y} \circ \mathbf{m}_t = \mathbf{b}$ and $\mathbf{x} \succsim_t \mathbf{y}$, which in turn is true if and only if $\mathbf{a} \succcurlyeq' \mathbf{b}$. Which establishes that $\succcurlyeq \cap X^2(\mathbf{m_t}) = \succcurlyeq' \cap X^2(\mathbf{m_t})$ for any t. Therefore, any completion of $\succcurlyeq \cap \succcurlyeq'$ together with $\mathbf{m}_t$ rationalizes $\succsim_t$ for any t.

We next show that by considering $\rhd \cap \rhd'$ and the sequence of awareness $(\mathbf{a}_t \circ \mathbf{a}'_t)_t$, we can rationalize the meta-choices of each period t. Fix a period t, and suppose that being at $\mathbf{m}_t$, DM faces the meta-menu $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$. Note that $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t) = R(\mathbf{m}_t, \mathbf{a}_t) \cap R(\mathbf{m}_t, \mathbf{a}'_t)$. Hence it implies that $(\mathbf{m}_{t+1}, \mathbf{m}) \in \rhd \cap \rhd'$ for any $\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$. This completes the proof than any completion of $\rhd \cap \rhd'$, together with $(\succcurlyeq \cap \succcurlyeq', \mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ rationalize $(\succsim_t)_t$. □

Proof of Proposition 3. (*Necessity*.) Suppose there exists a complete preorder $\succcurlyeq$ such that for every t, $(\mathbf{m}_t, \succcurlyeq)$ represents $\succsim_t$. Take any $\{t_1, \ldots, t_n\}$ and any $\{\mathbf{x_k}, \mathbf{x'_k}\}_{k=1,\ldots,n}$ such that, for $k = 1, \ldots, n-1$: $\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}}$, $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1}$, and for every $k \leq n-1$, $\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k$. The latter implies that $\mathbf{x}_k \circ \mathbf{m}_{t_k} \succcurlyeq \mathbf{x}'_k \circ \mathbf{m}_{t_k}$. Hence by the transitivity of $\succcurlyeq$, we can conclude that $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1} \succcurlyeq \mathbf{x}_{n-1} \circ \mathbf{m}_{t_{n-1}} = \mathbf{x}_n \circ \mathbf{m}_{t_n}$, i.e. $\mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n$. Hence Strong Restricted Reversal is satisfied.

(*Sufficiency*). We fix a period t and show that we can construct a complete preorder $\succcurlyeq_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \succcurlyeq_t)$ rationalizes $\succsim_t$. We define $\succcurlyeq_t$ in the same way as in the proof of proposition 1. Given that Strong Restricted Reversal implies Restricted Reversal, the same arguments apply and we conclude that $(\mathbf{m}_t, \succcurlyeq_t)$ rationalizes $\succsim_t$. Furthermore, the transitivity of $\succcurlyeq_t$ is a direct consequence of the transitivity of $\succsim_t$. We need now to construct a complete preorder $\succcurlyeq$ on $\bar{X}$ that is time-independent.

From the proof of proposition 1, we know that for any two distinct periods t and t', $\succcurlyeq_t$ does not contradict $\succcurlyeq_{t'}$. We define $\succcurlyeq_{1;T} := \bigcup_t \succcurlyeq_t$. We know therefore that for any t, $\succcurlyeq_{1;T} \cap X^2(\mathbf{m}_t) = \succcurlyeq_t$.

We next show that the transitive closure of $\succcurlyeq_{1;T}$, denoted $\succcurlyeq_{1;T}^{C}$, can rationalize the sequence $(\succsim_t)_t$ together with the sequence $(\mathbf{m}_t)_t$. Namely, we show that for any period t, any $\mathbf{a}, \mathbf{b} \in X(\mathbf{m}_t)$, if $\mathbf{b} \succ_t \mathbf{a}$, there cannot be a sequence $(\mathbf{a}_k)_{k=1,\dots,n}$ and $(t_k)_{1\leq k\leq n-1}$ such that $\mathbf{a}_1 = \mathbf{a}$, $\mathbf{a}_n = \mathbf{b}$ and $\mathbf{a}_k \succcurlyeq_{t_k} \mathbf{a}_{k+1}$. Let suppose by contradiction the existence of such a sequence. If $t_1 \neq t$, then complete the sequence with $\mathbf{a}_0 = \mathbf{a}$ and $t_0 = t$; similarly, if $t_{n-1} \neq t$, then complete the sequence with $\mathbf{a}_{n+1} = \mathbf{b}$ and $t_n = t$. Therefore, w.l.o.g we consider the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$.

If $t_k = t_{k'}$ for some $k \neq k'$, we show that we can restrict to a subsequence $(\mathbf{a}_{\tau(k)})_{k=1,\dots,n+1}$ with $\tau(0) = 0, \tau(n+1) = n+1$, such that $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$. Let suppose that $t_k = t_{k'}$ with $k < k'$ and that for any $k \leq i, j < k'$, if $i \neq j$ then $t_i \neq t_j$. Let's consider the sequence $(\mathbf{a}_i)_{k\leq i\leq k'+1}$. There exists a sequence $(\mathbf{x}_i, \mathbf{y}_{i+1})_{k\leq i\leq k'}$ such that $\mathbf{x}_k \circ \mathbf{m}_{t_k} = \mathbf{a}_k$, $\mathbf{y}_{k'+1} \circ \mathbf{m}_{t_{k'}} = \mathbf{a}_{k'+1}$, for any $k \leq i \leq k' - 1$, $\mathbf{y}_{i+1} \circ \mathbf{m}_{t_i} = \mathbf{x}_{i+1} \circ \mathbf{m}_{t_{i+1}} = \mathbf{a}_{i+1}$, and for any $k \leq i \leq k'$, $\mathbf{x}_i \succsim_{t_i} \mathbf{y}_{i+1}$. By applying Strong Restricted Reversal, this must be that $\mathbf{x}_k \succsim_{t_k} \mathbf{y}_{k'+1}$, i.e. $\mathbf{a}_k \succcurlyeq_{t_k} \mathbf{a}_{k'+1}$. Therefore, from the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$, we can construct a subsequence $(\mathbf{a}_{\tau(k)})_{k=0,\dots,n+1}$, with $\tau(0) = 0$, $\tau(n+1) = n+1$, $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$, and such that for any k with $\tau(k) \neq \tau(k+1)$, $\mathbf{a}_{\tau(k)} \succcurlyeq_{\tau(k)} \mathbf{a}_{\tau(k+1)}$. From a similar reasoning, we conclude by Strong Restricted Reversal that $\mathbf{a} \succcurlyeq_t \mathbf{b}$, a contradiction.

By an implication of Szpilrajn's theorem (see Corollary A.1 in Ok (2007)), there exists a complete, transitive and reflexive binary relation that extends $\succcurlyeq_{1;T}^{C}$. We denote it $\succcurlyeq$. We proved that for any t, $X^2(\mathbf{m}_t) \cap \succcurlyeq = \succcurlyeq_t$, hence $(\mathbf{m}_t, \succcurlyeq)$ rationalizes $\succsim_t$. □

REFERENCES

Abramowitz, A. I. and Saunders, K. L. (2008). Is Polarization a Myth? *The Journal of Politics*, 70(2):542–555.

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

Aumann, R. J. (1976). Agreeing to Disagree. *The annals of statistics*, pages 1236–1239.

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics. *Journal of Public Economics*, 182:104123.

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., Martínez-Marquina, A., and Zuckerman, D. (2021). A Theory of Chosen Preferences. *American Economic Review*, 111(2):720–54.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit Discrimination. *American Economic Review*, 95(2):94–98.

Bertrand, M. and Duflo, E. (2017). Field Experiments on Discrimination. *Handbook of Economic Field Experiments*, 1:309–393.

Boissonnet, N. (2019). Rationalizing Preference Formation by Partial Deliberation. *PhD Thesis*.

Boissonnet, N., Ghersengorin, A., and Gleyze, S. (2022). Supplement to "Revealed Deliberate Preference Change". *Working Paper*.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2020). Cross-Country Trends in Affective Polarization. *NBER Working Paper*.

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

Cripps, M. W. (2018). Divisible Updating. *Working Paper*.

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. (2009). Temptation-Driven Preferences. *The Review of Economic Studies*, 76(3):937–971.

Dietrich, F. and List, C. (2011). A model of non-informational preference change. *Journal of Theoretical Politics*, 23(2):145–164.

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

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

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

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.

Heller, Y. (2012). Justifiable choice. *Games and Economic Behavior*, 76(2):375–390.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22:129–146.

Iyengar, S. and Westwood, S. J. (2015). Fear and Loathing Across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, 59(3):690–707.

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

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

Nehring, K. (2006). Self-Control Through Second-Order Preferences. *Working Paper*.

Ok, E. A. (2007). *Real Analysis with Economic Applications*. Princeton University Press.

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.

Ridout, S. (2021). Choosing for the right reasons. *Unpublished manuscript*.

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.