A Representation of Preference over Preferences and the Act of Choosing

Jun Hyun Ji*

January 27, 2025

Abstract

This paper studies how individuals value the act of choosing itself by using the concept "preference over preferences". The agent in my model has no preference over outcomes. Instead, she prefers some preference relation if she prefers *behaving* as if she holds that preference (e.g., "preferring x to y" is preferred to "preferring y to x" if she values the act of willingly giving up y for x more than giving up x for y). My axioms yield a unique representation that identifies (i) the individual's *ideal* preference over outcomes, and (ii) a choice rule that selects a *reference option* against which the act of choosing from each menu is assessed. This choice rule captures the individual's paternalistic attitude toward willful choices—manifested as guilt, pride, the joy of freedom, or the fear of *the act* of making mistakes—which is inherently menu-dependent. Without relying on menu choices or econometric models of welfare measures, I provide a revealed-preference approach to testing my model even when the agent cares about outcomes. Welfare implications are discussed.

*Ph.D. Student, Economics, University of Pittsburgh. Email: juj25@pitt.edu. I am indebted to Luca Rigotti for his guidance and encouragement throughout the project, and to Sven Neth and Kevin Zollman for many useful discussions. I have also benefited from comments from In-Koo Cho, Antonio Penta, Michael Woodford, and participants at the 2024 Asian School of Economic Theory, NYU Abu Dhabi. I gratefully acknowledge the financial support by 2023 Tamara Horowitz Memorial Fund.

Contents

1. Introduction

People experience different emotional sensations when making a choice depending not only on the consequence of the choice but also on what they are willingly giving up for it. For example, the existing choice-theoretic models of temptation have significantly advanced our understanding of how individuals design their future behavior by anticipating the psychological experiences associated with opportunities they may forgo. When choosing a menu of future options, the agent might remove tempting options from his menu, anticipating that the act of willingly giving them up to achieve a long-term goal requires costly self-control (Gul and Pesendorfer, 2001) or perhaps fearing that they might succumb to temptations and feel a sense of guilt or shame (Kopylov, 2012; Dillenberger and Sadowski, 2012; Saito, 2015).

From a social planner's perspective, recognizing that the act of making a choice is more than a means to an end raises questions about engaging in paternalistic interventions and restricting a rational decision-maker's options. The non-comparability problem, formally characterized by Bernheim et al. (2024), posits that observing choice data alone is insufficient for deriving valid welfare policies because choices do not uniquely reveal the emotional sensations that the agent immediately experiences when choosing an option or the menu itself (including any higher-level meta-choices)[1]. This challenge has profound welfare implications, as policies based solely on observed choices may fail to enhance, or may even diminish, individual well-being. Bernheim et al. (2024) addressed this by proposing an econometric method to estimate the decision-maker's welfare by combining choice data with self-reported well-being methods.

However, these prior studies infer preferences over the act of choosing indirectly, relying either on econometric models or menu choices that are strongly affected by the agent's outcome preferences. Consequently, a crucial aspect of economic behavior remains insufficiently understood—the standalone nature of preferences over the act of choosing.

In this paper, I provide a theoretical framework for preferences over the act of choosing by using a novel concept of preference over preferences (henceforth, *second-order preference*)[2]. I use the phrase "preferring a preference" to mean preferring to *behave* as if one holds that preference[3]. To illustrate, suppose an agent strictly prefers x to y. Our standard understanding is that he is willing to give up y for x. A *first-order preference* is described in this manner. Now,

[1] I explain the non-comparability problem in more detail in Section 5.

[2] Philosophers have long discussed that human beings can have preferences over their own preferences (Frankfurt, 1971; Jeffrey, 1974). For example, some may wish to become a person who prefers exercising to indulging in eating, being altruistic to being selfish, or wish that they find drinking coffee more enjoyable than alcohol. Others might be concerned with someone else's preferences (e.g., parents wishing that their child prefers doing homework to watching television, or wanting one's romantic partner to prefer marriage to otherwise). See Appendix H for prior literature on second-order preference.

[3] To avoid confusion, I distinguish second-order preferences from second-order *desires*. The former pertains to one's observable behavior—manifestations of preferences (e.g., a killer might prefer preferring not killing to killing). The latter pertains to one's state of mind (e.g., he might desire not to have the desire to kill even after he decided not to kill). I focus on the study of the former. See Appendix H for more detail.

suppose he prefers "preferring x to y" to "preferring y to x". Then, it must be that he values the act of willingly giving up y for x more than that of giving up x for y. Hence, we can study and interpret an act of choosing as a manifestation of certain preferences, and examine how an individual might prefer one manifestation to another, under the assumption that he has no first-order preferences. That is, the agent in my model cares only about the act of choosing, not about the consequences that follow.

I first introduce the model foundation of second-order preference and axiomatize the agent's attention to the act of choosing (a single option from a menu)[4]. Next, I introduce the key axiom that yields a unique representation in a general functional form. Third, I explore special cases starting with two extreme attitudes toward the act of choosing, and investigate their limitations to understanding how individuals value the act of choosing. Fourth, I introduce a non-extreme attitude, and discuss how emotional experiences such as pride and guilt are inherently menu-dependent. Fifth, I provide a direct revealed-preference approach to testing my model even when the agent cares about outcomes, without relying on menu choices or econometric models. Lastly, I discuss the relationship between my model and prior models of menu preferences, and welfare implications.

I define an act of choosing as a pair (x, A) where A is the menu and x is the chosen option, representing "preferring x to all else in A"[5]. In my model, the agent has a preference $\succeq$ over all possible act of choosing. My key axiom states that given any two menus A and B, we can find an option from each menu—say, x and y—such that the two acts of choosing (x, A) and (y, B) are indifferent. I call this axiom *Relativity*. This is based on the idea that the quality of a choice is relative to constraints: one can always make a good (bad) choice from a bad (good) menu[6]. The essential role of this axiom is to remove all utility variations possibly attributed to the design of the menus. If the axiom is false, there must be two menus A^* and B^* such that "preferring anything in A^*" is preferred to "preferring anything in B^*" which implies that preferences do not matter: the agent merely wants some outcomes in A^* more than the ones in B^*. Consequently, if the axiom holds, any two acts of choosing an option from a singleton menu—henceforth, *vacuous choices*—are indifferent since preferences do not influence the choices in those cases.

Section 3 offers my first result: a general functional form of the representation (Theorems 1-

[4] In the next section, I show that a preference over the act of choosing belongs to a special class of second-order preferences in general. In particular, my axioms rule out the possibility that the agent has a preference over indifference (e.g., one might like (or dislike) to be indifferent among some options). Appendix I briefly discusses extending my model to allow for "the act of being indifferent".

[5] While my model uses the outcome-menu pairs as its primitives, some prior models use higher-order menus (e.g., menus of menus of outcomes, and so on) as their primitives (see Noor, 2011; Noor and Ren, 2023). However, my model's applicability is not restricted to the lowest-level menu-dependent preferences, as the act of choosing from any higher-order menu can encompass the entire series of choices involved in that action, including choices at each level down to the final outcome selection.

[6] For instance, a poor person's act of donating \$100 to charity might be more praiseworthy than a rich person's act of donating the same amount. He probably needs to donate much more to be equally praiseworthy.

2). In addition to the standard axioms of expected utility theory, my key axiom yields a unique representation of the form

$$V(x, A) = v(x) - v(\mathbf{r}(A))$$

where v is an von Neumann-Morgenstern (vNM) utility function over lotteries and $\mathbf{r}$ is a choice function that selects a lottery $\mathbf{r}(A)$ from the menu A that serves as a reference against which the act of choosing x from A is assessed[7]. The function v represents the preference that the agent believes he (or someone) should ideally adopt (e.g., an alcoholic thinks that ideally, he should prefer coffee to beer; a parent thinks that her child should ideally prefer doing homework to watching television). According to this ideal ranking, $\mathbf{r}(A)$ is either the best or worst option (or even something in between) in A[8]. As I explain in more detail, the significance of my theorems lies in translating the choice function $\mathbf{r}(\cdot)$ into an induced binary relation $\succeq_{\mathbf{r}}$ on sets, enabling the use of standard results on menu preferences to explicitly derive the function $v(\mathbf{r}(\cdot))$ without finding the function $\mathbf{r}(\cdot)$ itself.

Section 4 introduces several representations of special forms to investigate how the function $\mathbf{r}(\cdot)$ relates to preferences for exercising freedom of choice. I start with two extreme attitudes toward the act of choosing called paternalism and libertarianism[9]. I say the second-order preference exhibits *pure paternalism* if a vacuous choice is weakly preferred to any given act of choosing. This implies that preferring the most ideal option from any menu (e.g., choosing coffee over beer) is indifferent from the state of not being able to give up anything (e.g., vacuously choosing either of them). Consequently, if there is even a slight chance that the best option will not be chosen, the agent would rather abandon his freedom of choice, preventing himself from *the act* of making mistakes[10]. Roughly speaking, any non-singleton menu can potentially evoke feelings of guilt, but not pride. The opposite holds for *pure libertarianism*: the freedom of choice is valued above avoiding wrong choices[11]. As a result, the representation of a purely paternalistic (libertarian) preference is such that the choice function $\mathbf{r}(\cdot)$ selects the most (least) ideal option on the menu.

[7] Our tendency to assess an outcome of a choice in contrast with a reference has been discussed previously. Kőszegi and Rabin (2006)'s reference-dependent preference captured a loss-averse agent's tendencies to assess an outcome of a choice in contrast with his expectation of the outcome, which arises from uncertainty. Yet, my model stays within expected utility theory and the reference stems from the agent's preference over the act of choosing. See Appendix F for a more detail comparison.

[8] The menus are convex sets. Appendix C provides the same results that allow for non-convex menus, including finite sets.

[9] While paternalism usually refers to one's willingness to intervene in others' autonomy to enhance their welfare, the agent with a paternalistic second-order preference adopts a paternalistic stance toward his own preference, not his outcomes.

[10] Although my model does not explicitly present the chances of making mistakes, the paternalistic tendency can also be applied to the setting where the agent has a paternalistic preference over others' preferences (e.g., a parent who wants the child to prefer doing homework to watching television might decide whether to offer a choice or not given her expectation of what the child will choose).

[11] See Bartling et al. (2014) for experimental evidence that individuals value "decision rights" beyond their instrumental benefit.

As my main result, Theorem 3 captures an agent who is *locally* purely paternalistic. The idea is that a sense of pride emerges from making *hard choices*, which are not always available on a menu. People generally feel little to no pride in avoiding an obviously bad outcome (e.g., choosing life over committing suicide) while they are proud when their choices are aligned with their ideality that are distinctively better than how they are *expected* to behave. Hence, the agent's paternalistic attitude might weaken when menus present a strong conflict between his ideal preference and expected preference.

Suppose a purely libertarian dictator—tasked with allocating resources between himself and a passive recipient—expects himself to be selfish, but prefers "preferring fairness to selfishness". So, he might be proud of himself for choosing a fair allocation over an unfair one that disproportionately benefits him. Now, suppose an allocation that Pareto-dominates both the fair and unfair allocations is added to his menu. He might become purely paternalistic, and the sense of pride might completely vanish because the act of giving up the two Pareto-inferior allocations is neither giving up being fair nor giving up being selfish. Despite being the best outcome overall, it is simply an obvious choice he would make, neither in conflict with his expectation nor contrary to his ideals. In this sense, the dictator is strictly better off without the Pareto-dominant allocation on his menu. This result is generated by my axiom called *locally pure paternalism* motivated by Gul and Pesendorfer (2001)'s *Set-betweenness* axiom, which allows the choice function $\mathbf{r}(\cdot)$ to be determined by two conflicting preferences that are uniquely identified: the ideal and expected preferences.

Second-order preferences are fundamentally difficult to observe because in most observable cases, the agent *does* care about the outcomes. Section 5 provides three approaches to addressing this problem. The first one relies on menu choices, analogous to the standard social settings designed to test the models of menu preference. These models emphasize the influence of outcome preferences, and thus do not rationalize the tendency to willingly remove the best outcome from the menu (e.g., the dictator's Pareto-dominant allocation). Hence, observing such tendency would clearly verify the existence of preference over the act of choosing that exhibits locally pure paternalism. To address the potential non-comparability problem in menu-choice environments, the second approach uses my model to fit Bernheim et al. (2024)'s econometric estimation of the DM's welfare. In the dictator game context as above, my model predicts that a social planner can strictly improve an altruistic dictator's welfare by removing the Pareto-dominant allocation from his menu if the differences in outcomes are small enough.

The third method is a direct revealed-preference approach through a recipient's decision problem prior to passively playing a dictator game. The recipient, who cares only about his wealth, must appoint a person as the dictator from a pool of non-strategic candidates. Before making this choice, the recipient observes the choice data of each candidate, who previously played a game with other players. In this setting, he is interested in what choices the candidates made previously, not the outcomes they achieved. In other words, he is choosing *the best*

preference. By observing his ranking of the candidates, this setup transforms his standard preference for wealth into a preference over the dictators' preferences, allowing all my axioms to be tested without relying on menu choices or econometric models. This problem applies to many other social settings to some extent (e.g., dating, recruiting, courtrooms and elections) where an agent ranks potential decision-makers based on their past choices (hence, preferences) before they directly affect his outcome.

Section 6 discusses (i) how my model is conceptually related to prior models of menu preferences, and (ii) welfare implications. In particular, the concept of higher-order preferences suggests that the design of welfare policies can be influenced by the social planner's own higher-order preferences. As I explain more in detail, this leads to a higher-order non-comparability problem, which cannot be resolved even when the decision-maker's preferences over outcomes and the act of choosing are fully known. Thereby, the inherent complexities in welfare assessments persist.

Proofs (if omitted) are collected in Appendices A-D. I also discuss my contributions to the literature specifically on second-order preference in Appendix H.

2. Model

This section provides the foundations for second-order preference and in turn, preference over the act of choosing. Mainly, I consider a decision-maker (DM) who does not have preferences over standard objects such as actions, money, or any other forms of outcomes, but has a preference over preference relations. To capture this idea, I represent the DM as composed of two conceptual entities: Bob, who corresponds to the conventional decision problem—choosing an option (a lottery) from an exogenously given menu of options—and Amy, who forms a "second-order" judgment over Bob's preferences from an observer's perspective[12]. It is immediately intuitive that this formulation applies equally well to situations in which a person (e.g., a parent) has a preference over someone else's preferences (e.g., a child's).

I first characterize Bob's choice environment.

Options (lotteries). Let Z be the finite set of alternatives other than preference relations, and X be the set of lotteries on Z, endowed with a metric d generating the standard weak topology. X is Bob's entire consumption space where any elements $x, y, z \in X$ are called lotteries or *options*. For $\alpha \in [0,1]$, let $\alpha x + (1-\alpha) y$ denote the mixture of lotteries x and y that yields x with probability α and y with probability $1-\alpha$.

[12] This is inspired by the philosophical discussions of a person's capacity to reflect upon one's own tastes and dispositions (see Frankfurt, 1971). The observer's point of view generalizes the notion of "meta-preference" while abstracting away from any particular first-order preference structure, and helps focus on how one might evaluate different preference relations.

Menus. Let $\mathbb{M}$ denote the set of nonempty compact *convex* subsets of X whose elements $A, B, C \in \mathbb{M}$ are called *menus*[13]. And let $conv(A)$ denote the convex hull of A. I define convex combinations of menus as follows: $\lambda A + (1-\lambda)B := \{\lambda x + (1-\lambda)y : x \in A,\ y \in B\}$ for $\lambda \in [0,1]$. There is a reason why I expose Bob only to the convex menus. I allow Bob to randomize and announce a personal state-contingent plan whenever he has a non-convex menu. For example, when a menu $\{x, y\}$ is given, we can think of Bob declaring a state-contingent plan $z = \alpha x + (1-\alpha)\, y$ for some $\alpha \in (0,1)$. Indeed, this particular lottery z is not available in $A = \{x, y\}$, but if Amy cannot stop him from forming probabilities or tossing an imaginary coin in his head, then we can say that he is actually facing the menu $conv(A)$ instead of A. Appendix C provides the same results that allow for non-convex menus, including finite sets[14]. My explanations and examples will often feature finite menus because they simplify the narrative and aid in understanding the intuitions. One technical issue with convex menus is that the elements in $\mathbb{M}$ are not closed under standard set operations (e.g., the union of two nonempty convex disjoint sets is not convex in general). I use alternative set operations, defined as

$$A \cup^* B = conv(A \cup B); \quad A \cap^* B = conv(A \cap B); \quad A \setminus^* B = conv(A \setminus B).$$

Henceforth, I use these alternative operations but maintain the use of the standard notations $\cup, \cap, \setminus$.

2.1. Defining Second-order Preference

I now characterize Amy's second-order preference in general.

First-order Preference. Define $\mathbb{P}(A)$ as the set of all strict preference relations over the subset $A \subseteq X$[15]. The elements $P, Q \in \mathbb{P}(A)$ for any $A \in \mathbb{M}$ are called *first-order preferences*. For example, if $A = \{x, y\}$, then $\mathbb{P}(A) = \{P_1, P_2, P_3\}$ such that

$$xP_1y; \quad yP_2x; \quad \neg(xP_3y) \quad \text{and} \quad \neg(yP_3x)$$

[13] I endow $\mathbb{M}$ with the Hausdorff metric

$$d_H(A,B) := \max\left\{\max_{x\in A}\min_{y\in B} d(x,y), \max_{y\in B}\min_{x\in A} d(x,y)\right\}.$$

[14] Appendix C shows that my results hold for nonempty compact menus with a moderate modification of my key axiom *Relativity*. In particular, the representation is not affected by whether or not Bob faces convex menus.

[15] Notice that the cardinality of the set $\mathbb{P}(A)$ explodes as the menu A becomes larger. In fact, a menu with n elements gives us $n!$ different strict preference relations without considering indifference. This explosion creates mathematical challenges since our consumption space (X) is not finite. See Laffond et al. (2020) for more detail on "metrizability" of the set $\mathbb{P}(X)$.

where $\neg(yPx)$ means not yPx. For each $P \in \mathbb{P}(A)$, define $\mathcal{C}_P(A) := \{x \in A : \neg(yPx) \ \ \forall y \in A\}$ as the set of choices in A induced by P.

Second-order Preference. Generally, a second-order preference $\succeq$ is a binary relation on the set

$$\mathcal{P} := \bigcup_{A \in \mathbb{M}} \mathbb{P}(A)$$

which is the collection of all preference relations defined across all possible menus. This comprehensive set $\mathcal{P}$ contains all preferences Bob could potentially exhibit, each corresponding to a different choice situation or menu he may encounter. Let $P_A, Q_B \in \mathcal{P}$ where $P_A \in \mathbb{P}(A)$ and $Q_B \in \mathbb{P}(B)$. I say Amy prefers P_A to Q_B if she prefers "the action induced by P_A given the menu A" to "the action induced by Q_B given B". The nature of the model can vary widely depending on how we define what "an action induced by P" refers to. My analysis focuses on the case where the action of Amy's interest pertains solely to Bob's act of choosing a single option from a menu—thereby, willingly giving up all feasible others on the menu[16].

The act of choosing. The primitive of my model is a binary relation $\succeq$ over the set

$$\mathbb{C} := \bigcup_{A \in \mathbb{M}} \{(\mathcal{C}_P(A), A) : P \in \mathbb{P}_s(A)\}$$

where $\mathbb{P}_s(A) = \{P \in \mathbb{P}(A) : |\mathcal{C}_P(A)| = 1\}$ is the set of preferences inducing a single option in each menu. I abuse notations and let $\mathbb{C} = \{(x, A) : x \in A \in \mathbb{M}\}$. A pair (x, A) refers to *the act of choosing x over everything else in A*, but for brevity, each element in $\mathbb{C}$ will also be called *a choice*[17]. A more accurate interpretation of the relation $(x, A) \succeq (y, B)$ is that Amy prefers *"preferring x to everything else in A"* to *"preferring y to everything else in B"*. I define convex combinations of choices as follows: for $\lambda \in [0, 1]$,

$$\lambda(x, A) + (1 - \lambda)(y, B) := (\lambda x + (1 - \lambda)y, \lambda A + (1 - \lambda)B).$$

The interpretation of $\lambda A + (1 - \lambda)B$ is that Bob faces the menu A with probability λ and B with probability $1 - \lambda$. Before this uncertainty is resolved, he chooses a contingency plan $\lambda x + (1 - \lambda)\, y$ which constitutes the act of choosing $(\lambda x + (1 - \lambda)\, y, \lambda A + (1 - \lambda)B)$.

[16] The action induced by a preference in general can refer to many different behaviors: the act of consuming n options from a menu where $n \in \{1, 2, ...\}$, declaring indifference among some options, declaring the least favorite option on the menu, or revealing one's preference over the menu entirely. Yet, in many cases, we only regard one's favorite as the vital part of his preference. In a presidential election, we do not count a voter's non-favorite candidates while some voting methods (e.g., Condorcet method) do look at a whole ranking of alternatives.

[17] Henceforth, (x, A) naturally implies $x \in A$.

Vacuous choices. A special notation will be used to indicate vacuous choices—any choices made from singleton menus: let ϕ denote a vacuous choice, i.e., $\phi \in \{(x, \{x\}) : x \in X\} \subset \mathbb{C}$.

2.2. Axiomatizating Preference over the Act of Choosing

I now provide two axioms that restrict Amy's attention to the act of choosing. The idea is twofold. First, Amy does not care about what Bob might have chosen if he had been presented with a different menu. Second, she does not care about any other option Bob was willing to choose, apart from the one he actually did. For example, suppose on two separate occasions (e.g., periods 1 and 2), Bob chose x from the menu $\{x, y, z\}$. Let $c_1 = (x, \{x, y, z\})$ and $c_2 = (x, \{x, y, z\})$ denote his act of choosing in periods 1 and 2, respectively. Suppose Amy found out that Bob's second favorite option in $\{x, y, z\}$ was y in period 1 and z in period 2. The first axiom presented below will dictate that Amy is indifferent—i.e., $c_1 \sim c_2$. Consider another scenario: Amy found out that Bob was indifferent between x and y in period 1, but became indifferent between x and z in period 2. The second axiom below still requires $c_1 \sim c_2$.

Formally, suppose Bob's menu A is fixed. The two axioms are as follows:

Axiom 1 (Preference for Revealed Preference)**.** *Given $A \in \mathbb{M}$ and $P_1, P_2 \in \mathbb{P}(A)$,*

$$\mathcal{C}_{P_1}(A) = \mathcal{C}_{P_2}(A) \quad \Longrightarrow \quad P_1 \sim P_2.$$

Axiom 2 (No Preference for Indifference)**.** *Given $A \in \mathbb{M}$, suppose $\mathcal{C}_{P_1}(A)$ and $\mathcal{C}_{P_2}(A)$ form a partition of $\mathcal{C}_{P_3}(A)$ for some $P_1, P_2, P_3 \in \mathbb{P}(A)$. Then,*

$$P_1 \succeq P_2 \quad \Longrightarrow \quad P_1 \sim P_3 \succeq P_2.$$

By Axiom 1, Amy associates a preference relation only with its contribution to Bob's willingness to choose (or give up) certain options. I refer to her second-order preference as a *preference for revealed preference* if she does not care about how Bob orders the non-favorite options on a menu. Consider $A = \{x, y, z\}$ and $P_1, ..., P_4 \in \mathbb{P}(A)$ where

$$xP_1yP_1z; \quad xP_2zP_2y; \quad yP_3xP_3z; \quad yP_4zP_4x.$$

Then, Axiom 1 requires $P_1 \sim P_2$ and $P_3 \sim P_4$ since

$$\{x\} = \mathcal{C}_{P_1}(A) = \mathcal{C}_{P_2}(A) \neq \mathcal{C}_{P_3}(A) = \mathcal{C}_{P_4}(A) = \{y\}.$$

Yet, if $\mathcal{C}_P(A)$ is not a singleton, then P does not directly induce the act of willingly choosing a *single* option. Hence, under Axiom 1 alone, Amy also regards the act of announcing indifference as a valid external behavior that corresponds to a preference $P \in \mathbb{P}(A)$.

I say Amy has *no preference for indifference* if Axiom 2 holds. In other words, she does not particularly favor or disfavor Bob's indifference among some options. Consequently, we can focus on $P \in \mathbb{P}(A)$ such that $\mathcal{C}_P(A)$ is a singleton. To elaborate, consider $A = \{x, y\}$ and $\mathbb{P}(A) = \{P_1, P_2, P_3\}$ where $\mathcal{C}_{P_1}(A) = \{x\}$ and $\mathcal{C}_{P_2}(A) = \{y\}$ form a partition of $\mathcal{C}_{P_3}(A) = \{x, y\}$[18]. Suppose Amy wants Bob to *want* to choose x from A (i.e., $P_1 \succ P_2$). Then, by Axiom 2, we have $P_1 \sim P_3$ which implies that she does not care whether he gave up y for x because he is indifferent or because he strictly prefers x to y[19]. This brings $P_1 \sim P_3 \succ P_2$. When $P_1 \sim P_2$, she simply does not care whether Bob wants to choose x or y in which case, we have $P_1 \sim P_3 \sim P_2$.

Assuming that Bob's menu A is fixed and not subject to change, Axioms 1-2 allow us to jettison irrelevant information inferred from some $P \in \mathbb{P}(A)$ and restrict Amy's attention to the act of choosing. In other words, if $U : \mathbb{P}(A) \to \mathbb{R}$ is her utility function, then we can preserve all variations with a function $\overline{U} : A \to \mathbb{R}$ defined by $\overline{U}(x) = U(P)$ for all P such that $\mathcal{C}_P(A) = \{x\}$[20]. Formally, let $\mathcal{P}_s = \bigcup_{A \in \mathbb{M}} \mathbb{P}_s(A)$ be the set of all preferences across all menus inducing a single choice.

Observation 1. *$\succeq$ defined on $\mathcal{P}$ satisfies Axioms 1-2 if and only if the equivalence classes of $\mathcal{P}$ under $\succeq$ can be mapped onto $\mathcal{P}_s$ which can be further mapped onto $\mathbb{C}$.*

The above observation implies that I can define $\succeq$ on $\mathbb{C}$ instead of $\mathcal{P}$, and that preferences over the act of choosing are a special class of second-order preferences. (The preferences over the act of choosing can still be characterized by $\succeq$ on $\mathcal{P}$ with unnecessary complexity.) However, in Appendix G, I show that a second-order preference $\succeq$ restricted to the act of choosing, mainly due to Axiom 1, allows for a ranking of *complete rankings* as well if $\succeq$ is further restricted to complete contingency plans for each possible binary choice situation.

In Appendix I, I discuss preferences for (or against) indifference by relaxing Axiom 2. Amy might particularly like or dislike indifference. When we are making a decision as a group (for example, what to eat for lunch), we often witness people who claim to be indifferent among all alternatives. Sometimes, this benefits the group because they allow others with strong preferences to make decisions according to their needs. However, some may not appreciate the presence of indifferent individuals if they interpret indifference as a lack of interest or engagement.

[18] Notice that since $A = \{x, y\}$ is nonempty and finite, $\mathcal{C}_P(A)$ is nonempty for all $P \in \mathbb{P}(A)$. Thus, if $\mathcal{C}_{P_1}(A)$ and $\mathcal{C}_{P_2}(A)$ form a partition of $\mathcal{C}_{P_3}(A)$, then both $\mathcal{C}_{P_1}(A)$ and $\mathcal{C}_{P_2}(A)$ are always proper subsets of $\mathcal{C}_{P_3}(A)$.

[19] Yet, if Bob's preference is P_3, then we need an additional context in which his indifference is mapped into consumption. One possible context is that after Bob truthfully announces his indifference between x and y, Amy—who learns that he is willing to consume x—chooses x for him. We can also think of Bob choosing a contingency plan $z = \alpha x + (1 - \alpha) y$ for some $\alpha \in (0, 1)$. If this is possible, then Bob is facing the convex menu $conv(A)$ instead of $A = \{x, y\}$.

[20] The fact that we can define $\succeq$ on A instead of $\mathbb{P}(A)$ implies that under Axioms 1-2, a preference relation $\succeq$ defined on $\mathbb{P}(X)$ is behaviorally indistinguishable from first-order preferences over X. In other words, if the menu is fixed, it limits our understanding of second-order preferences themselves. Yet, the impact of different choice sets had not been explored by the past literature on second-order preferences. See Appendix H for more detail.

2.3. Standard Axioms

I employ the standard vNM axioms of continuity and independence used in prior literature, and impose the following axioms[21]:

Axiom 3 (Weak Order). $\succeq$ *is complete and transitive.*

Axiom 4 (Independence). *For all* $\lambda \in (0,1)$,

$$(x,A) \succ (y,B) \text{ implies } \lambda(x,A) + (1-\lambda)(z,C) \succ \lambda(y,B) + (1-\lambda)(z,C).$$

Axiom 5 (Continuity). $\{(x,A) : (x,A) \succeq (y,B)\}$ *and* $\{(x,A) : (y,B) \succeq (x,A)\}$ *are closed.*

Axiom 6 (EU Ideality). *There is a continuous and independent relation* $\succeq_1$ *in* $\mathbb{P}(X)$ *such that*

$$(x,X) \succeq (y,X) \iff x \succeq_1 y.$$

Axiom 7 (Menu-independent Ideality).

$$(x,A) \succeq (y,A) \iff (x,B) \succeq (y,B).$$

Axioms 3-5 are in alignment with the standard axioms of the expected utility theory[22]. I provide the motivation for *independence* (Axiom 4) in detail in Section 3.1. Axiom 6 states that when Bob's menu is X—the set of all lotteries—there is a first-order preference over X denoted by $\succeq_1$ that determines the ranking of preferences. $\succeq_1$ is called Amy's *ideal first-order preference*. $x \succeq_1 y$ implies that she wants Bob to prefer x to y, or equivalently, I say x is ideally preferred to y. To explain the intuition behind Axiom 6, $\succeq_1$ is the first-order preference that Amy believes Bob should ideally have[23]. I further impose continuity and independence on $\succeq_1$ to follow the underlying framework of the standard expected utility theory. Axiom 7 states that Amy adheres to the ideal preference even when Bob faces menus other than X: her ideal preference is menu-independent.

Note that any first-order preference $\succeq_1'$ can be defined in terms of a preference $\succeq_2'$ over the act of choosing, as follows: $x \succeq_1' y$ if and only if $(x,A) \succeq_2' (y,B)$ for all A, B containing x, y,

[21] A first-order preference $\succeq_1$ over X is *independent* if $x \succ_1 y$ and $\alpha \in (0,1)$ imply $\alpha x + (1-\alpha)z \succ_1 \alpha y + (1-\alpha)z$. It is *continuous* if $\{x \in X : x \succeq_1 y\}$ and $\{x \in X : y \succeq_1 x\}$ are closed.

[22] In particular, Axiom 4 is consistent with the assumption that the decision-maker remains impartial concerning the timing of uncertainty resolution, as implied by the *independence* axiom imposed on menu preferences (see Gul and Pesendorfer, 2001; Dekel et al., 2001, 2007)

[23] Note that Amy's ideal preference $\succeq_1$ does not necessarily reflect a sense of morality or better judgements. The philosopher Mele (1992) pointed out that self-control is not always exercised to motivate moral actions via 2nd-order desires. He presented a story of a young man Bruce who agreed to participate in a crime, but 'chickened out' and left the scene before the crime began. Although Bruce's inaction agrees with his sense of morality, it can also be a sign of his lack of self-control against fear and anxiety.

in which case, the menus are merely means to an end: the first-order ranking $\succeq_1'$ completely determines the ranking of the act of choosing. In contrast, Axioms 6-7 together imply $x \succeq_1 y$ if and only if $(x, A) \succeq (y, A)$ for all A containing x, y, in which case, $(x, A) \succeq (y, B)$ does not hold in general.

2.4. Key Axiom

Axiom 8 (Relativity). *For any $A, B \in \mathbb{M}$, there are $x \in A$ and $y \in B$ such that*

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

Axiom 8 is the essential property of second-order preference. It states that given any two convex menus A, B, we can find an option from each menu—say, x in A and y in B—such that "preferring x to all else in A" is just as good as "preferring y to all else in B". The key role of this axiom is to remove any utility variations possibly attributed to the *design* of Bob's menu, thereby eliminating any preference for Bob's outcomes. To see this, suppose Axiom 8 is violated. Then, we can immediately find two menus A, B such that

$$(x, A) \succ (y, B) \quad \forall x \in A,\ \forall y \in B. \tag{1}$$

(1) essentially implies that Amy prefers the menu A to B regardless of Bob's preferences over the two menus: she prefers "preferring anything in A" to "preferring anything in B". This naturally implies that Amy cares only about Bob's outcomes, not preferences.

I discuss the motivation for Axiom 8 in more detail. First, the name "relativity" suggests that the quality—not the consequence—of a choice is relative to constraints: one can always make a good (bad) choice from a bad (good) menu. To illustrate, suppose Amy considers investing in two potential businesses by evaluating them based on their owners' past choices. One candidate chose \$10 from the menu $\{\$10, \$0\}$, while another candidate chose \$20 from the menu $\{\$20, \$30\}$. In absolute terms, the second candidate's choice yielded more profit. However, if Amy is looking for a partner who prioritizes profit, she would prefer the first candidate whose choice clearly revealed a preference for money while the other's did not.

Second, the axiom implies that Amy is indifferent between any two vacuous choices:

$$(x, \{x\}) \sim (y, \{y\}) \quad \forall x, y \in X.$$

This clearly shows that the axiom entirely eliminates all variations in Amy's preference attributed to consumption utilities. If Bob chooses y from $\{y\}$, Amy has no room for judgment because his preference had no impact on his choice $(y, \{y\})$: he did not *willingly* choose or give up anything. Suppose she wants him to prefer x to y. Is she happier if Bob is given $\{x\}$ instead,

so that he ends up with the choice $(x, \{x\})$? If yes, her satisfaction must come from appreciating the *design* of the menu $\{x\}$. Yet, she has no reason to appreciate Bob's preference which had no contribution to the design.

Third, the axiom implies that vacuous choices serve as reference points against which Bob's preference is evaluated given any menu. By the axiom, we can find an option from any menu (say, $\mathbf{r}(A)$ from A) such that a vacuous choice is indifferent from the act of choosing $\mathbf{r}(A)$ from A. This means we can define a choice function $\mathbf{r} : \mathbb{M} \to X$ by the following indifference relation[24]:

$$(\mathbf{r}(A), A) \sim (\mathbf{r}(B), B) \quad \forall A, B \in \mathbb{M}.$$

When B is a singleton menu, we have $(\mathbf{r}(A), A) \sim \phi$ for all A. To see why each option $\mathbf{r}(A)$ in menu A serves as a reference point, consider a classic family question "Who do you like better, mom (x) or dad (y)?". Let $A = conv(\{x, y\})$ be the child's menu. Each parent wants to be their child's favorite. Suppose the child wants to flip a coin to decide it in front of his parents who would probably say "Flipping a coin doesn't count. You have to choose!" because they would not regard the coin flip as a choice. In other words, it is the option that gives neither gain nor loss for both mom and dad, but serves as a reference point when evaluating the child's preference over A. Suppose the parents have the power to force a desired answer from the child: either $(x, \{x\})$ or $(y, \{y\})$. Yet, the value of these vacuous choices would be commensurate to that of willfully choosing the coin toss: that is,

$$(\mathbf{r}(A), A) = \left(\tfrac{1}{2}x + \tfrac{1}{2}y, A\right) \sim (x, \{x\}) \sim (y, \{y\}).$$

Since Axiom 8 is only compatible with convex menus, I introduce a discrete version, called *Discrete Relativity*, in Appendix C to allow for non-convex menus (including finite menus).

3. Representation

I use the standard definitions of preference representations and *affine* functions for both first- and second-order preferences[25]. The following is a non-standard definition I use.

Definition 1 (Affine Choice Function)**.** *A choice function $\mathbf{r} : \mathbb{M} \to X$ is affine with respect to a binary relation $\succeq_1$ on X if $\mathbf{r}(\lambda A + (1-\lambda) B) \sim_1 \lambda \mathbf{r}(A) + (1-\lambda)\, \mathbf{r}(B)$ for $\lambda \in [0, 1]$.*

[24] A choice function is any well-defined function $f : \mathbb{M} \to X$ that satisfies $f(A) \in A$. If the set of menus include non-convex sets, I instead use a *stochastic choice function* g that satisfies $g(A) \in conv(A)$ for any A. See Appendix C for more detail.

[25] Given a first-order preference $\succeq_1$ over X, I say the function v represents $\succeq_1$ when $v(x) \geq v(y)$ if and only if $x \succeq_1 y$. The function v is *affine* if $v(\alpha x + (1-\alpha)\, y) = \alpha v(x) + (1-\alpha)\, v(y)$ for all $x, y \in X$ and $\alpha \in [0, 1]$. I say the function $V : \mathbb{C} \to \mathbb{R}$ represents $\succeq$ if $V(x, A) \geq V(y, B)$ is equivalent to $(x, A) \succeq (y, B)$. The function $V : \mathbb{C} \to \mathbb{R}$ is affine if $V(\lambda(x, A) + (1-\lambda)\,(y, B)) = \lambda V(x, A) + (1-\lambda)\, V(y, B)$ for all $(x, A), (y, B) \in \mathbb{C}$ and $\lambda \in [0, 1]$.

I also refer the function $\mathbf{r}$ to as *the reference function*, and $\mathbf{r}(A)$ is called *the reference of* A. My axioms yield the following result:

Theorem 1. *$\succeq$ satisfies Axioms 1-8 if and only if there is a pair $(v, \mathbf{r})$ where $v : X \to \mathbb{R}$ is a continuous affine function of lotteries and $\mathbf{r} : \mathbb{M} \to X$ is an affine choice function with respect to $\succeq_1$ such that $\succeq_1$ is represented by v, and $\succeq$ is represented by a continuous affine function $V_{v,\mathbf{r}}$ of the form*

$$V_{v,\mathbf{r}}(x, A) := v(x) - v(\mathbf{r}(A)).$$

Proof. See Appendix A. □

The "if" part is straightforward. I provide a sketch of proof for the "only if" part. First, note that Axiom 6 grants the existence and uniqueness of the continuous affine function v representing $\succeq_1$ due to the standard expected utility theory. Moreover, by the result of Herstein and Milnor (1953), Axioms 3-5 are equivalent to the existence of a continuous affine function $V : \mathbb{C} \to \mathbb{R}$ representing $\succeq$.

Let $\mathbf{r}$ be the choice function defined by $(\mathbf{r}(A), A) \sim \phi$ for all A. (It is well-defined, thanks to Axiom 8.) The second step is Lemma 1 in the Appendix which shows that due to Axioms 4-5, the relations $\succeq_1$ and $\succeq$ have the following relationship[26]:

Lemma 1. $(x, A) \succeq (y, B) \iff \frac{1}{2}x + \frac{1}{2}\mathbf{r}(B) \succeq_1 \frac{1}{2}\mathbf{r}(A) + \frac{1}{2}y.$

Since v is an affine function representing $\succeq_1$, Lemma 1 implies

$$(x, A) \succeq (y, B) \iff v(x) - v(\mathbf{r}(A)) \geq v(y) - v(\mathbf{r}(B)).$$

As the third step, define $V_{v,\mathbf{r}} : \mathbb{C} \to \mathbb{R}$ by $V_{v,\mathbf{r}}(x, A) := v(x) - v(\mathbf{r}(A))$. The goal is to show that $V_{v,\mathbf{r}}$ is also a continuous affine function and thus, $V_{v,\mathbf{r}} = V$. That is, we need to show that $v(\mathbf{r}(\cdot))$ is a continuous affine function of sets. To accomplish this, I first show that the reference function $\mathbf{r}$ responds to state-contingent menus in a linear manner. In the Appendix, I prove the following lemma, which is a consequence mainly of Axiom 4 and Axiom 8:

Lemma 2. $\mathbf{r}(\lambda A + (1 - \lambda) B) \sim_1 \lambda \mathbf{r}(A) + (1 - \lambda)\, \mathbf{r}(B)$ *for* $\lambda \in [0, 1]$.

Lemma 2 states that $\mathbf{r}(\lambda A + (1 - \lambda)B)$ is ideally indifferent to $\lambda \mathbf{r}(A) + (1 - \lambda)\, \mathbf{r}(B)$, the convex combination of the two separate references.

The last step of the proof involves defining a binary relation $\succeq_{\mathbf{r}}$ on $\mathbb{M}$ as $A \succeq_{\mathbf{r}} B$ if and only if $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$. I show in the Appendix that we can use Lemma 2 to conclude that $\succeq_{\mathbf{r}}$ is a complete, transitive, continuous and independent binary relation on $\mathbb{M}$—the necessary and

[26] In Appendix H, I show that Lemma 1 is a generalized version of the axiom of second-order preference originally introduced in the book *The foundations of decision logic* by the philosopher Halldén (1980).

sufficient conditions for the existence of a continuous affine function K representing $\succeq_{\mathbf{r}}$[27]. My axioms ensure that $K(\cdot) = v(\mathbf{r}(\cdot))$. This completes the proof.

3.1. Behavioral Remarks

I present several behavioral intuitions behind the representation in Theorem 1.

Signs of Utilities. First, what does it mean to have a strictly positive utility of the act of choosing? A natural interpretation of the signs of utilities is that positive utilities are associated with a sense of pride in willingly making choices while negative utilities are tied to negative psychological experiences such as disappointment or guilt. To see this, suppose $V_{v,\mathbf{r}}(x, A) > 0$. This implies that x should ideally be preferred to the reference of A—i.e., $x \succ_1 \mathbf{r}(A)$. Also, it means the act of choosing x from A is preferred to "the inability to give up anything"—i.e., $(x, A) \succ \phi$. Therefore, conditional on x being chosen, the freedom of having the menu A is preferred to any exogenous outcome. Naturally, Amy finds value in Bob's willingness to choose x from A, which I interpret as feelings of pride.

Values Relative to Constraints. Second, I offer motivations for two noteworthy consequences of *Relativity* (Axiom 8):

(i) $(x, A) \succeq (x, B) \iff \mathbf{r}(B) \succeq_1 \mathbf{r}(A)$.

(ii) $(x, A) \succeq \phi \succeq (y, A)$ for some $x, y \in A$.

(i) states that the same consumption is preferred less whenever the reference of the menu from which it was chosen has greater value[28]. This reflects the relative nature of choice evaluations. $(x, A) \succeq (x, B)$ implies that while the two choices have the common chosen option x, Amy's reference of B has a higher value than that of A: $\mathbf{r}(B) \succeq_1 \mathbf{r}(A)$. (ii) is due to the fact that $\succeq_1$ satisfies *independence* and thus, v is an affine function[29]. Since $\mathbf{r}(A) \in A$, the affinity of v requires that

$$\min_{x \in A} v(x) \leq v(\mathbf{r}(A)) \leq \max_{x \in A} v(x) \quad \forall A \in \mathbb{M}.$$

Thus, a menu always offers a choice that is better or worse than vacuous choices.

[27] I say a binary relation $\succeq_{\mathbf{r}}$ on $\mathbb{M}$ is *independent* if $A \succ_{\mathbf{r}} B$ implies $\lambda A + (1-\lambda)C \succ_{\mathbf{r}} \lambda B + (1-\lambda)C$ for all $\lambda \in (0, 1)$. $\succeq_{\mathbf{r}}$ is *continuous* if $\{A : A \succeq_{\mathbf{r}} B\}$ and $\{A : B \succeq_{\mathbf{r}} A\}$ are closed.

[28] Proof of (i). Note that $(x, A) \succeq (x, B)$ is equivalent to $\frac{1}{2}x + \frac{1}{2}\mathbf{r}(B) \succeq_1 \frac{1}{2}x + \frac{1}{2}\mathbf{r}(A)$ by Lemma 1. Since $\succeq_1$ is independent, this implies $\mathbf{r}(B) \succeq_1 \mathbf{r}(A)$.

[29] Proof of (ii). Given any A, let $x \in \{a \in A : a \succeq_1 b \ \forall b \in A\}$ and $y \in \{a \in A : b \succeq_1 a \ \forall b \in A\}$. Since $\mathbf{r}(A) \in A$ and $\succeq_1$ is independent, $x \succeq_1 \mathbf{r}(A) \succeq_1 y$ holds. By Lemma 1, we have $(x, A) \succeq \phi \succeq (y, A)$ if and only if $\frac{1}{2}x + \frac{1}{2}\mathbf{r}(\{x\}) \succeq_1 \frac{1}{2}x + \frac{1}{2}\mathbf{r}(A)$ and $\frac{1}{2}y + \frac{1}{2}\mathbf{r}(A) \succeq_1 \frac{1}{2}y + \frac{1}{2}\mathbf{r}(\{y\})$ which are true since $\mathbf{r}(\{x\}) = x$ and $\mathbf{r}(\{y\}) = y$.

Independence. Third, what is the intuition behind Lemma 2? The answer lies in *independence* (Axiom 4). The axiom requires that when Bob chooses a contingency plan $\lambda x + (1-\lambda)\, y$ from the menu $\lambda A + (1-\lambda)B$ for some $\lambda \in (0,1)$, which constitutes the act of choosing $(\lambda x + (1-\lambda)\, y, \lambda A + (1-\lambda)B)$, Amy's reference of the menu $\lambda A + (1-\lambda)B$ is formed by weighing her references of A and B proportionally with the probability measure $(\lambda, 1-\lambda)$. This is in alignment with her ideal first-order preference which also satisfies *independence*. If she believes Bob should ideally be an expected utility maximizer, then it is reasonable to assume that she evaluates his expected choice from his expected menu accordingly in a linear manner.

More importantly, *independence* implies that Amy's reference is independent of Bob's *personal* contingencies. Suppose Bob faces a non-singleton finite menu A with certainty. Even though his choices are limited to A, Amy cannot stop him from considering various scenarios in his head, rolling an imaginary die and creating multiple states or personal contingencies in which he chooses a different option in A. Notice that whenever a non-singleton A is finite, we have $\lambda A + (1-\lambda)\, A \neq A$ for $\lambda \in (0,1)$. For example, if $A = \{x, y\}$, then $\lambda A + (1-\lambda)\, A$ offers the state-contingent plans $\lambda x + (1-\lambda)\, y$ and $\lambda y + (1-\lambda)\, x$ which are not in A. Of course, Amy only observes either (x, A) or (y, A) if Bob does not inform her of his personal plans. However, if the plan is announced or observable, then she begins to perceive $\lambda A + (1-\lambda)\, A$ and updates her reference to $\mathbf{r}(\lambda A + (1-\lambda)\, A)$. By Lemma 2, her reference is unchanged: $\mathbf{r}(\lambda A + (1-\lambda)\, A) \sim_1 \mathbf{r}(A)$. To illustrate, suppose Amy has a 9-year-old child named Bob. For the upcoming weekend, Bob wants to play soccer (y), while Amy believes he should prefer studying (x) to y. He claims that he will study if it rains during the weekend. That is, his choice is $(\lambda x + (1-\lambda)\, y, \lambda A + (1-\lambda)\, A)$ where the probability of rain is $1-\lambda$. According to Lemma 2, his plan conditional on the weather forecasts cannot change how much Amy would be disappointed at his choice to do y instead of x. Thus, she will continue to evaluate $\lambda x + (1-\lambda)\, y$ based on the reference she has formed for A[30].

Reference-depedence. Lastly, since the idea of the representation in Theorem 1 lies in reference-dependence, I also provide a detail comparison between my model and Kőszegi and Rabin (2006)'s seminal model of reference-dependent preference in Appendix F.

3.2. Uniqueness

Analogous to the standard expected utility theory, the representation $V_{v,\mathbf{r}}$ is unique up to positive affine transformations. When two affine functions $v : X \to \mathbb{R}$ and $\mathbf{r} : \mathbb{M} \to X$, put

[30] Lemma 2 also implies Amy is indifferent between Bob's choice of a compound lottery and a simple lottery, a condition known as *the reduction of compound lotteries axiom*. See Samuelson (1952). Notice that the menu $\frac{1}{2}A + \frac{1}{2}A$ contains two compound lotteries $\frac{1}{2}x + \frac{1}{2}y$ and $\frac{1}{2}y + \frac{1}{2}x$, which may be two different contingency plans from Bob's perspective. However, Amy would not distinguish them since they both yield x with probability 0.5 and y with probability 0.5.

together as a pair $(v, \mathbf{r})$, represent $\succeq$ as in Theorem 1, then $(v', \mathbf{r}')$ also represents $\succeq$ if and only if $v' = \alpha v + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$, and $v(\mathbf{r}(A)) = v(\mathbf{r}'(A))$ for each $A \in \mathbb{M}$.

Theorem 2 (Uniqueness)**.** *Suppose $(v, \mathbf{r})$ represents $\succeq$ as in Theorem 1. Then, $(v', \mathbf{r}')$ represents $\succeq$ if and only if $v' = \alpha v + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$, and $v(\mathbf{r}(A)) = v(\mathbf{r}'(A))$ for each $A \in \mathbb{M}$.*

Proof. See Appendix B. □

Appendix C shows that Theorems 1-2 still hold even when menus are nonempty compact subsets of X—including finite menus. In particular, whether or not the menus are convex does not affect the uniqueness of the choice function $\mathbf{r}$. Roughly speaking, I show that $\mathbf{r}(A) = \mathbf{r}(conv(A))$ for any menu A.

4. Paternalism and Libertarianism

The key component that distinguishes the second-order preference representation in Theorem 1 from any first-order preference representation is the choice function $\mathbf{r}$. This section investigates what kinds of preferences are captured by this function. In short, the function $\mathbf{r}$ reflects the DM's *paternalistic attitude toward the act of choosing*. Conversely, $\mathbf{r}$ is influenced by the extent to which one values autonomy, or freedom of choice.

Consider parents who want their child to prefer doing homework (x) to watching television (y). The parents believe that ideally, the child should strictly prefer x to y, which is identified with $(x, \{x, y\}) \succ (y, \{x, y\})$. This is when $v(x) > v(y)$. The parents would be disappointed at the child's choice not to do homework, which can be identified with $(y, \{y\}) \succeq (y, \{x, y\})$. This can be shown by $v(y) - v(\mathbf{r}(\{y\})) = v(y) - v(y) = 0 \geq v(y) - v(\mathbf{r}(\{x, y\}))$, which holds for all choice function $\mathbf{r}$ affine with respect to v. A disciplinarian would prefer enforcing homework time to granting freedom, and thus, would satisfy $(x, \{x\}) \succ (y, \{x, y\})$. This is when the parents' reference of $\{x, y\}$ is valued more than y: i.e., $v(\mathbf{r}(\{x, y\})) > v(y)$. Paternalistic parents—who would remove television from his choice set even when the child is willing to engage in schoolwork—can be described by $(x, \{x\}) \sim (x, \{x, y\})$. This is true when $x = \mathbf{r}(\{x, y\})$. Yet, parents who are libertarian might grant leeway and allow the child to choose from $\{x, y\}$ even when they know he will not choose to do homework, which is identified with $(x, \{x\}) \sim (y, \{x, y\})$. This is true when $y = \mathbf{r}(\{x, y\})$. This example shows that given the same ideal preference, the parents can have different attitudes toward the child's act of choosing, which is captured by the choice function $\mathbf{r}$.

More formally, I now present two extreme attitudes called *pure paternalism* and *pure libertarianism*.

Axiom 9 (Pure Paternalism)**.** *$\phi \succeq (x, A)$ for all $(x, A) \in \mathbb{C}$.*

Axiom 10 (Pure Libertarianism). $(x, A) \succeq \phi$ *for all* $(x, A) \in \mathbb{C}$.

In general, paternalism refers to one's willingness to intervene in the DM's autonomy, restricting his options to promote his best welfare. The prior experimental studies on paternalistic preferences focus on the case where a social planner (alternatively referred to as a policy maker or a choice architect) is concerned only with the DM's outcome, not with the act of choosing (see Ambuehl et al., 2021; Bartling et al., 2023). I say the planner has a *paternalistic attitude toward outcomes* if she has the standard motivation for paternalistic interventions discussed in the prior literature, which is to prevent the DM's mistakes—choosing the wrong outcomes. In contrast, the paternalistic attitude toward the act of choosing refers to the motivation to prevent *the act* of making mistakes itself.

Axiom 9 states that a vacuous choice is weakly preferred to any given act of choosing (x, A). To see how this relates to the concept of paternalism, suppose Amy's preference satisfies Axiom 9. Then, her most preferable act of choosing is a vacuous choice—the state of not being able to willingly make any choice at all. Given any menu A, if there is even a slight chance that Bob will not choose the most ideal option, then Amy would abandon his freedom of choice and enforce a vacuous choice, preventing the act of making a mistake. The paternalistic parents mentioned above has a purely paternalistic attitude toward the child's act of choosing. Suppose $A = conv(\{x, y\})$. The most ideal option in this example is x (homework). If the parents believe the child will choose $z_\alpha = \alpha x + (1 - \alpha)\, y$ given any $\alpha < 1$, then we have $(x, \{x\}) \succ (z_\alpha, A)$. That is, they do not allow even a small chance of the act of choosing y over x. Rough speaking, any non-singleton menu given to Bob is a potential loss (e.g., a sense of guilt or disappointment) for Amy.

The opposite is true for Axiom 10, or *Pure Libertarianism*, which states that any given act of choosing is weakly preferred to a vacuous choice. In this case, the least preferable act of choosing is the vacuous choice. For a libertarian who values freedom of choice, any non-singleton menu is a potential gain (e.g., a sense of pride or the joy of exercising autonomy). Consequently, willingly making a choice is strictly preferred to a vacuous choice if there is a even a slight chance of avoiding the least ideal option on the menu. The purely libertarian parents would satisfy $(z_\alpha, A) \succ (x, \{x\})$ for any $\alpha > 0$.

The value of autonomy has been widely documented by philosophers and psychologists (see Mill, 1859; Deci and Ryan, 1985). In economics, Bartling et al. (2014) provided experimental evidence that individuals value "decision rights" beyond their instrumental benefit. Yet, there has not been an axiomatic approach to modeling preferences for freedom of choice independent of outcomes.

The two extreme cases have the following representations.

Corollary 1 (Representations of Paternalism and Libertarianism). *Suppose* $\succeq$ *satisfies Axioms 1-8*

whose representation is $V_{v,\mathbf{r}}$ as in Theorem 1. Then, Axiom 9 and Axiom 10 are equivalent to

$$V_{v,\mathbf{r}}(x,A) = v(x) - \max_{y\in A} v(y) \quad \textit{and} \quad V_{v,\mathbf{r}}(x,A) = v(x) - \min_{y\in A} v(y),$$

respectively. The former (latter) is referred to as the representation of a purely paternalistic (libertarian) preference over the act of choosing.

Proof. See Appendix D.1. □

Corollary 1 shows that a purely paternalistic (libertarian) preference over the act of choosing implies that the reference of each menu is the most (least) ideal option on the menu: for all menu A, we have $r(A) \in \arg\max_{y\in A} v(y)$ if $\succeq$ is purely paternalistic, and $r(A) \in \arg\min_{y\in A} v(y)$ if purely libertarian.

What the two extreme attitudes have in common is that when there is a common set of opportunities, the ranking of two choices is determined solely by the ideal ranking $\succeq_1$; and the ranking of the act of *giving up* two different sets are determined by their reference values. The following axiom is called *Independence of Common Alternatives* (ICA).

Axiom 11 (ICA). $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$ *implies for any* $C \in \mathbb{M}$ *disjoint from* $A \cup B$,

a. $(\mathbf{r}(A), A\cup C) \succeq (\mathbf{r}(B), B\cup C)$, *and*

b. $(c, B\cup C) \succeq (c, A\cup C)$ *for all* $c \in C$.

c. $(c, B\cup C) \succeq (c, C) \succeq (c, A\cup C)$ *for all* $c \in C$ *if* $\mathbf{r}(A) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B)$.

Corollary 2. *Axioms 1-8, and either Axiom 9 or Axiom 10 imply Axiom 11.*

Proof. See Appendix D.2. □

Roughly speaking, Axiom 11a-b state that, with every other opportunities equal, the act of choosing (giving up) ideally superior (inferior) options is preferred to the act of choosing (giving up) ideally inferior (superior) ones. Formally, Axiom 11a states that if the reference value of the menu A (i.e., $\mathbf{r}(A)$) is greater than that of B (i.e., $\mathbf{r}(B)$), then the act of choosing $\mathbf{r}(A)$ from A is weakly preferred to the act of choosing $\mathbf{r}(B)$ from B once any new set of options C is commonly added to each menu. Note that if C is not added, then choosing $\mathbf{r}(A)$ from A can never be strictly preferred to choosing $\mathbf{r}(B)$ from B due to Axiom 8: $(\mathbf{r}(A), A) \sim (\mathbf{r}(B), B)$. Intuitively, while the forgone opportunities in A and B are equal in relative value, the addition of C offers a new context in which giving up the common opportunities in C for the ideally superior option $\mathbf{r}(A)$ is better than giving them up for the inferior option $\mathbf{r}(B)$. The idea is that the common *forgone* opportunities are ignored when comparing two choices.

Axiom 11b states that given any commonly chosen option c and non-chosen options in C, giving up B is preferred to giving up A. Note that the two choices $(c, B\cup C)$ and $(c, A\cup C)$

differ only in the forgone sets B and A. According to the result, Amy determines the value of giving up a set of options by its reference value. Intuitively, willingly giving up the bad options—i.e., a menu of options with a smaller reference value (in this case, $\mathbf{r}(B)$)—is preferred to willingly giving up the good options—i.e., a menu with a greater reference value (in this case, $\mathbf{r}(A)$).

Axiom 11c states that if the reference value of the commonly added set C is between the reference values of A and B, then having B as a part of the forgone set of alternatives is better than when B is not available; and forgoing A is worse than when A is not available. Intuitively, the act of "not choosing a bad option" is preferred to "being unable to choose it", and "being unable to choose a good option" is preferred to "not choosing it".

To illustrate, let $A = \{x\}$, $B = \{y\}$, and $C = \{c\}$ for simplicity. If a parent wants a child to prefer doing homework (x) to playing with friends outside (y), then by Axiom 11a, regardless of what c is, the parent prefers the act of choosing homework over c to the act of choosing the social activity over c: i.e., $(x, \{x, c\}) \succeq (y, \{y, c\})$ for all $c \neq x, y$. Now instead suppose the child chose c. Then, by Axiom 11b, the parent prefers the act of giving up playing with friends for c to the act of giving up homework for c: i.e., $(c, \{y, c\}) \succeq (c, \{x, c\})$ for all $c \neq x, y$. For Axiom 11c, suppose c is watching an educational television show. For an academically-focused parent, it is reasonable to presume $x \succeq_1 c \succeq_1 y$. Then, from $(c, \{c\}) \succeq (c, \{x, c\})$, we can infer that if the parent witnesses the child watching the educational show, then she would wish that he does not have any homework to do. $(c, \{y, c\})$ implies that the child prefers the educational show to playing outside with friends, which is good news that the parent would not have inferred from $(c, \{c\})$. Thus, $(c, \{y, c\}) \succeq (c, \{c\})$.

4.1. Constant Paternalistic Attitudes

Examining the two polar attitudes toward the act of choosing raises the question: What lies between them? Can the attitude change with menus? The two extremes focus exclusively on a single aspect—under pure paternalism (libertarianism), the reference of each menu is always the best (least) ideal option on the menu. This limitation is formally captured in Corollary 2, which demonstrates that the common addition of any disjoint set C has no impact on how a purely paternalistic (libertarian) person ranks the two choices.

One simple extension that can address this limitation would be imposing a constant measure that captures a non-extreme paternalistic attitude—the one that weighs both the best and worst options with a fixed ratio. (In Appendix E, I present an alternative utility function where the paternalistic attitude depends on *every* option on the menu[31].) Consider the following representation $V_{v,\alpha}$ where the reference value function $v(\mathbf{r}(\cdot))$ takes the form of the α-maxmin utility function of sets of lotteries presented by Olszewski (2007) in his characterization of am-

[31] To be specific, each reference value is the average value of the options on the menu.

biguity aversion[32]:

$$V_{v,\alpha}(x,A) = v(x) - \underbrace{\left[\alpha \max_{y\in A} v(y) + (1-\alpha)\min_{y\in A} v(y)\right]}_{=v(\mathbf{r}(A))}$$

where the parameter $\alpha \in [0,1]$ can be interpreted as the constant paternalistic attitude toward the act of choosing. The two extreme cases discussed above are when $\alpha \in \{0,1\}$[33]. When $\alpha \in (0,1)$, the DM's attitude deviates from being purely paternalistic (libertarian) if the value of the best (least) ideal option increases (decreases). It is easy to verify that the function $V_{v,\alpha}$ is an affine function, and thus it is a special case of Theorem 1.

I show in the Appendix that when $\alpha \in (0,1)$, $V_{v,\alpha}$ satisfies the following two weaker versions of Axiom 11.

Axiom 12 (Weak ICA). *$x \succeq_1 y$ implies for any $C \in \mathbb{M}$ disjoint from $\{x\}$ and $\{y\}$,*

a. $(x, \{x\} \cup C) \succeq (y, \{y\} \cup C)$, *and*

b. $(c, \{y\} \cup C) \succeq (c, \{x\} \cup C)$ *for all $c \in C$.*

Specifically, Axiom 11 is partially satisfied when the sets A and B are singletons, as stated in Axiom 12. Formally, define the preference $\succeq_{v,\alpha}$ on $\mathbb{C}$ by $(x,A) \succeq_{v,\alpha} (y,B)$ if and only if $V_{v,\alpha}(x,A) \geq V_{v,\alpha}(y,B)$.

Corollary 3. *For all $\alpha \in (0,1)$, $\succeq_{v,\alpha}$ satisfies Axiom 12, but not Axiom 11.*

Proof. See Appendix D.3. □

As an example that violates Axiom 11, consider the above parent-child example, but alternatively, assume the menu A contains the option to watch television (w) as well as doing homework (x)—i.e., $A = conv(\{x,w\})$—and let c be watching a movie. It seems natural for the parent to have the following ideal ranking v:

The child's options	v
Doing homework (x)	10
Playing outside with friends (y)	2
Watching television (w)	0
Watching a movie (c)	0

[32] Olszewski (2007) described a preference of an agent who chooses a menu of lotteries from which Nature ambiguously chooses a lottery for the agent to consume. In his model, $\alpha \in (0,1)$ represents the agent's optimism toward ambiguity. My construct shares a common feature with this setup: Amy is not the one making a choice.

[33] Olszewski (2007)'s representation is only defined for $\alpha \in (0,1)$.

Suppose the parent's constant paternalistic attitude is $\alpha = 0.5$. Then, the above ranking yields the following:

$$v\left(\tfrac{1}{2}x + \tfrac{1}{2}w\right) > v(y), \tag{2a}$$

$$V_{v,0.5}\left(\tfrac{1}{2}x + \tfrac{1}{2}w, A\right) = V_{v,0.5}(y, \{y\}) = 0, \tag{2b}$$

$$V_{v,0.5}\left(y, \{y\} \cup \{c\}\right) = 1 > 0 = V_{v,0.5}\left(\tfrac{1}{2}x + \tfrac{1}{2}w, A \cup \{c\}\right). \tag{2c}$$

As shown in (2a), the parent believes preferring homework is so important that a coin toss between homework and television is ideally preferred to playing outside with friends. As shown in (2b), due to her paternalistic attitude fixed at $\alpha = 0.5$, the parent thinks that the act of choosing the coin toss from the menu A is just as impressive as vacuously choosing to play with his friends. When watching a movie is added to the menu A, the parent still believes the coin toss $\frac{1}{2}x + \frac{1}{2}w$ corresponds to a vacuous choice. However, when the child chooses between playing with friends and watching a movie, the parent's reference changes: an outdoor social activity is now considered a good option compared to watching a movie alone at home. Consequently, as shown in (2c), the parent prefers the act of choosing y over c to choosing the coin toss over homework, television, and a movie. (2a)-(2c) violate Axiom 11.

4.2. Locally Pure Paternalism

The constant paternalistic attitude still has its flaws. In particular, it does not allow the DM to be purely paternalistic or libertarian. Yet, people do not feel pride whenever they make good choices. For instance, they generally feel little to no pride in avoiding an obviously bad outcome (e.g., choosing life over committing suicide). A sense of pride—corresponding to a positive utility of the act of choosing—usually comes from making a hard choice which often involves a trade-off between competing values, goals, or desires.

More specifically, pride emerges when a person acts according to his ideal preference as opposed to how he is *expected* to behave. Psychological theories and evidence suggest that emotional experiences—such as pride or guilt—are intrinsically associated with discrepancies between an individual's actual self and his ideal self (see Markus and Nurius, 1986; Higgins, 1987; Tracy and Robins, 2004; Gilchrist et al., 2019). In other words, the paternalistic attitude might be menu-dependent, and it might weaken when menus present a strong conflict between ideal preference and the expected preference. In the context of the tension between temptation and normative goals, when a menu contains a stronger temptation, Amy might be prouder of Bob for resisting it or less disappointed when he succumbs to it.

For instance, an alcoholic—who expects himself to behave as an addict—might not feel particularly proud of choosing filtered water over tap water. Even though the former is ideally

preferable, he would not prefer the latter regardless of his alcohol addiction. However, when presented with a menu offering a cup of coffee, a non-alcoholic beer, and a glass of his favorite wine, he may experience a great sense of achievement and a positive self-evaluation by choosing a non-alcoholic beer as a compromise—and an even greater sense of pride if he chooses the coffee, fully aligning with his ideal preference against his expected preference for alcohol.

Formally, consider the following representation $V_{v,u}$ where the reference value function $v(\mathbf{r}(\cdot))$ takes the form of Gul and Pesendorfer (2001)'s representation of preference for commitment:

$$V_{v,u}(x, A) := v(x) - \left[\max_{y \in A} \left\{v(y) + u(y)\right\} - \max_{z \in A} u(z)\right]$$

where the function $u : X \to \mathbb{R}$ is Bob's first-order preference that Amy expects. The function $v + u$ is the ranking that reflects the expected choice when Bob reaches a compromise between the ideal ranking v and u. Using Gul and Pesendorfer (2001)'s terms, u can be referred to as Amy's expectation of Bob's temptation ranking.

The following axiom identifies Bob's first-order preference u that Amy expects such that her preference over the act of choosing is represented by $V_{v,u}$.

Axiom 13 (Vacuous Choice Betweenness (VCB))**.**

$$\mathbf{r}(A) \succeq_1 \mathbf{r}(B) \quad \Longrightarrow \quad (\mathbf{r}(A), A \cup B) \succeq \phi \succeq (\mathbf{r}(B), A \cup B).$$

According to Axiom 13, when the reference value of A surpasses that of B, the reference value of $A \cup B$ (i.e., $\mathbf{r}_{A \cup B}$) falls in between. Intuitively, Amy weighs the two references when formulating the reference of $A \cup B$ rather than interpreting $A \cup B$ in a fresh perspective. This idea resembles the *set betweenness* axiom of Gul and Pesendorfer (2001) which states that if the agent prefers a menu A to B, then A is preferred to $A \cup B$, which is preferred to B.

Theorem 3 (Locally Pure Paternalism)**.** *Suppose $\succeq$ satisfies Axioms 1-8 whose representation is $V_{v,\mathbf{r}}$ as in Theorem 1. Then, Axiom 13 holds if and only if there exists a continuous affine function $u : X \to \mathbb{R}$ such that*

$$V_{v,\mathbf{r}} = V_{v,u}.$$

Proof. I provide the proof here to emphasize the technical link between the choice function $\mathbf{r}$ and the study of preferences over sets. For the "only if" part, define a binary relation $\succeq_{\mathbf{r}}$ on $\mathbb{M}$ as $A \succeq_{\mathbf{r}} B$ if and only if $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$. Lemma 3 in the Appendix shows that $\succeq_{\mathbf{r}}$ is complete, transitive, continuous and independent. By Axiom 13, $A \succeq_{\mathbf{r}} B$ implies $A \succeq_{\mathbf{r}} A \cup B \succeq_{\mathbf{r}} B$, which is the *set betweenness* axiom of Gul and Pesendorfer (2001). By their Theorem 1, there exists continuous affine functions K, v_0, u such that $K(A) = \max_{y \in A} v_0(y) + u(y) - \max_{z \in A} u(z)$ represents $\succeq_{\mathbf{r}}$. By definition of $\succeq_{\mathbf{r}}$, the ranking of singleton sets follows $\succeq_1$ and thus, $K(\{x\}) = v_0(x) = v(x)$. It follows that $v(\mathbf{r}(A)) = K(A)$. Then, it is clear that the "if" part is straightforward, which completes the proof. □

The significance of this proof lies not in novel technical machinery but rather in the ability to bypass the explicit construction of the choice function $\mathbf{r}$. By translating $\mathbf{r}(\cdot)$ into an induced binary relation $\succeq_{\mathbf{r}}$ on sets, we can invoke standard results on menu preferences (e.g., Theorem 1 in Gul and Pesendorfer (2001)) to derive the explicit form of the real-valued function $v(\mathbf{r}(\cdot))$ directly as a representation of $\succeq_{\mathbf{r}}$. Notice that $\succeq_{\mathbf{r}}$ satisfies

$$A \succeq_{\mathbf{r}} B \quad \Longleftrightarrow \quad v(\mathbf{r}(A)) \geq v(\mathbf{r}(B))$$

where $v(\mathbf{r}(\cdot))$ is a continuous affine function of sets. Thus, any continuous affine representation commonly seen in the menu preference literature is a special form of $v(\mathbf{r}(\cdot))$. The axioms imposed in the prior framework can be directly translated in terms of the choice function $\mathbf{r}(\cdot)$. By doing so, it becomes straightforward to derive an explicit functional form for $v(\mathbf{r}(\cdot))$, without having to characterize $\mathbf{r}(\cdot)$ itself.

The preference represented by $V_{v,u}$ satisfies Axiom 11 partially when the common disjoint set C has a reference value between the reference values of A and B. Formally, it satisfies the following.

Axiom 14 (Local ICA). $\mathbf{r}(A) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B)$ *implies for any* $C \in \mathbb{M}$ *disjoint from* $A \cup B$,

a. $(\mathbf{r}(A), A \cup C) \succeq (\mathbf{r}(B), B \cup C)$,

b. $(c, B \cup C) \succeq (c, C) \succeq (c, A \cup C)$ *for all* $c \in C$

Corollary 4. *Axioms 1-8, and Axiom 13 imply Axiom 14, but not Axiom 11 or Axiom 12.*

Proof. See Appendix D.4. □

When Amy's preference is represented by $V_{v,u}$, her paternalistic attitude can become anything from purely paternalistic to purely libertarian, depending on the menu. We can mainly consider three cases. Let the compromise—the option aligned with $v + u$—be denoted by

$$y_{v+u} \in \arg\max_{y \in A} v(y) + u(y).$$

For brevity, let $R_{v,u}(A) := \max_{y \in A}\{v(y) + u(y)\} - \max_{z \in A} u(z)$ be the reference value function of the representation $V_{v,u}$.

- *Case 1 (Pure paternalism: when the reference is the most ideal option).*

  When u is aligned with v (e.g., when the alcoholic's menu is {filtered water, tap water}), Amy becomes purely paternalistic: i.e., if $u = v$, then $R_{v,u}(A) = \max_{y \in A} v(y)$. In this case, there actually is no need to call y_{v+u} the compromise since there are no conflicting preferences.

- *Case 2 (Pure libertarianism: when the reference is the least ideal option).*

  When u is perfectly misaligned with v, Amy becomes purely libertarian: i.e., if $v = -u$, then $R_{v,u}(A) = \min_{y \in A} v(y)$.

- *Case 3 (Weakening paternalism).*

  The word "compromise" is actually meaningful when y_{v+u} is not aligned with v (e.g., y_{v+u} is not the coffee from the menu $\{$coffee, non-alcoholic beer, wine$\}$). In this case, Amy's paternalistic attitude weakens: i.e., $y_{v+u} \notin \arg\max_{y \in A} v(y)$ implies $\min_{y \in A} v(y) \leq R_{v,u}(A) < \max_{y \in A} v(y)$.

  - *Case 3-1 (The reference is the compromise).*

    When the expected preference u is so strong compared to the ideal preference v that the compromise is aligned with u (e.g., y_{v+u} is the wine), the compromise becomes the reference of the menu: i.e., $y_{v+u} \notin \arg\max_{y \in A} v(y)$ and $y_{v+u} \in \arg\max_{y \in A} u(y)$ imply $R_{v,u}(A) = v(y_{v+u})$. In particular, Amy becomes purely libertarian if the compromise is the least ideal option on the menu (i.e., when $y_{v+u} \in \min_{y \in A} v(y)$).

  - *Case 3-2 (The reference is less ideal than the compromise).*

    Lastly, when the compromise is neither aligned with v nor u (e.g., y_{v+u} is the non-alcoholic beer), the reference is less ideal than the compromise: i.e.,

    $$y_{v+u} \notin \left(\arg\max_{y \in A} v(y) \right) \cup \left(\arg\max_{y \in A} u(y) \right)$$

    implies $R_{v,u}(A) < v(y_{v+u})$. In this case, Amy is proud of the compromise: she strictly prefers the compromise to a vacuous choice. The utility of the act of choosing the compromise is the utility distance between the compromise and the choice aligned with u:

    $$V_{v,u}(y_{v+u}, A) = \max_{z \in A} u(z) - u(y_{v+u}) > 0.$$

5. Three Approaches to Model Testing

This section provides three different approaches to testing my model even when the DM cares about the consequences (outcomes) of his choices. The first one is a revealed-preference approach analogous to the standard social settings in which the models of menu preference are tested. The second one employs Bernheim et al. (2024)'s empirical estimation of the DM's welfare. These two methods rely either on the DM's menu choices or on econometric assumptions, and thus are considered indirect approaches in the sense that the DM's ranking of the act of choosing cannot be directly observed. The last approach allows all my axioms to be tested directly.

5.1. Menu-choice approach: a dictator's menu choice

Consider the two-period dictator game context adopted by Dillenberger and Sadowski (2012) and Saito (2015). In period 1, the dictator publicly chooses an option (or a degenerate lottery) from a menu A. The option refers to an allocation $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ of wealth between himself (who gets x_1) and a passive recipient (who gets x_2). When the allocation is chosen, the wealth is distributed accordingly which ends the game. The prior studies assumed that the recipient believes the menu A is exogenously given; however, there is an ex ante period, say period 0, in which the dictator is allowed to privately choose the menu A from an exogenously given set $\mathcal{A}$ that contains menus of allocations (see Figure 1). Hence, in fact, A is not exogenous.

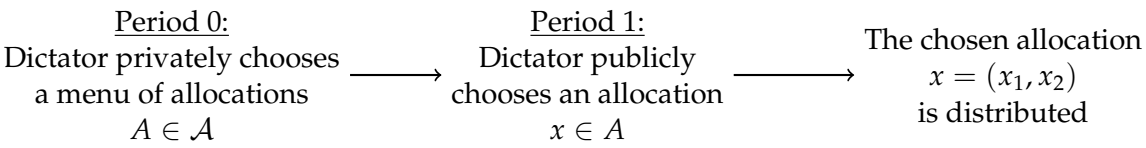

Figure 1: Timeline of the dictator game

Assume the dictator chooses a menu in period 0 to maximize his utility of the act of choosing in period 1. That is, his menu choice is

$$\arg\max_{A \in \mathcal{A}} \left(\max_{x \in A} V_{v,\mathbf{r}}(x, A) \right).$$

Consider three allocations: a fair allocation $x_f = (5, 5)$, a selfish allocation $x_s = (6, 4)$, and let $x_p^w = (6 + w, 6 + w)$ be called a Pareto optimal allocation with an increment $w > 0$. Suppose the set of menus of allocations is $\mathcal{A} = \{A, B^w\}$ where $A = \{x_f, x_s\}$ and $B^w = \{x_f, x_s, x_p^w\}$. A standard dictator—who only has a first-order preference for his own wealth—would prefer the choice $(x_p^w, \{x_f, x_s, x_p^w\})$ to any other possible act of choosing. His menu choice would be B^w trivially for any increment $w > 0$. Alternatively, suppose the dictator only has a preference over the act of choosing, and prefers "preferring being altruistic to being selfish". In particular, let v and u represent his ideal preference and expected preference over allocations, respectively. Consider the following[34]:

Allocations	v	u
$x_f = (5, 5)$	5	5
$x_s = (6, 4)$	4	6
$x_p^w = (6 + w, 6 + w)$	$6 + w$	$6 + w$

[34] We can assume the following functions: $v(x) = x_2$ and $u(x) = x_1$. The dictator's expected ranking u is driven by selfishness, and thus is determined by his own wealth outcome (x_1). In contrast, his ideal ranking v is driven by altruism, and thus determined by the recipient's wealth outcome (x_2).

If the dictator's preference exhibits locally pure paternalism, represented by $V_{v,u}$ as in Theorem 3, the following holds: for all $w > 0$,

$$V_{v,u}(x_f, \{x_f, x_s\}) = 1 > 0 = V_{v,u}(x_p^w, \{x_f, x_s, x_p^w\}).$$

That is, the dictator prefers the act of being altruistic over being selfish to the act of choosing the Pareto optimal allocation x_p^w over the two Pareto inferior allocations. Thus, the menu choice would be A. The dictator is purely libertarian when the menu is A, but becomes purely paternalistic when the menu is B^w. We can easily verify that the value of reference of each menu is

$$R_{v,u}(A) = v(x_s) = \min_{x \in A} v(x) \quad \text{and} \quad R_{v,u}(B^w) = v(x_p^w) = \max_{x \in B^w} v(x).$$

Intuitively, when the menu is A, the dictator feels a sense of pride in making the hard choice—sacrificing his own wealth to willingly pursue altruism. In contrast, he feels no pride at all in choosing x_p^w from B^w because he does not feel like willingly giving up anything: he is neither giving up being altruistic nor giving up being selfish. $(x_p^w, \{x_f, x_s, x_p^w\})$ is both the easiest and the best choice to make.

The result suggests that even when the dictator also cares about outcomes (i.e., his own wealth), the choice of A over B^w would be a clear sign that he has a (locally paternalistic) preference over the act of choosing as shown above. That is, he might deliberately remove the best outcome from the menu to pursue the best act of choosing if the sense of pride in willingly giving up being selfish is significant enough (or alternatively, when the wealth increment w is very small).

We can also use this dictator's choice environment to capture the monetary value of the sense of pride in choosing x_f over x_s. When $w > 0$ is very large, the dictator's preference for greater wealth would obviously outweigh his preference over the act of choosing, and thus B^w would be chosen over A. However, given that $(x_f, \{x_f, x_s\})$ invokes a significant mental benefit, we can look for a threshold $\overline{w} > 0$ small enough such that the dictator is indifferent between A and $B^{\overline{w}}$. Given that such $\overline{w}$ exists, he is willing to pay $6 + \overline{w} - 5$ amount of wealth to pursue the sense of pride[35].

The models of temptation using menu preferences in the literature—where the DM also cares about outcome—do not allow the menu A to be preferred to B^w in the above example.

[35] More formally, we could consider a more general setting where $x_f = (a - b, a - b)$, $x_s = (a, a - c)$, and $x_p^w = (a + w, a + w)$ for some constants $a, b, c > 0$ with $c > b$. Suppose the dictator's menu choice is generated by the function $U : \mathbb{M} \to \mathbb{R}$ of the form $U(A) := \max_{x \in A} W(x) + V_{v,u}(x, A)$ where $W : X \to \mathbb{R}$ represents the dictator's first-order preference over wealth. The experimental success will rely on finding the right parameters (a, b, c) such that $U(A) = U(B^{\overline{w}})$ for some $\overline{w}$. Then, assuming that the dictator would choose x_f over x_s—i.e., $\{x_f\} = \arg\max_{x \in A} W(x) + V_{v,u}(x, A)$—we have $W(x_f) + V_{v,u}(x_f, A) = W(x_p^{\overline{w}}) + V_{v,u}(x_p^{\overline{w}}, B^{\overline{w}})$. Since $V_{v,u}(x_p^w, B^w) = 0$ for all $w > 0$, the equality becomes $V_{v,u}(x_f, A) = W(x_p^{\overline{w}}) - W(x_f)$. That is, $V_{v,u}(x_f, A)$ is equal to the difference in wealth utilities between the Pareto optimal allocation with increment $\overline{w}$ and the fair allocation.

Specifically, the below ranking was deemed irrational:

$$(x_f, \{x_f, x_s\}) \succ (x_p^w, \{x_f, x_s, x_p^w\}) \succ (x_f, \{x_f, x_s, x_p^w\}) \succ (x_s, \{x_f, x_s, x_p^w\}) \tag{3}$$

The ranking (3) represents the tendency to remove an option (x_p^w) from a menu that the DM would otherwise have chosen. The standard models can only rationalize (3) by identifying x_p^w as temptation that is normatively inferior—e.g., against the DM's long-term goal (see Gul and Pesendorfer, 2001). However, in this example, since x_p^w is Pareto optimal, it is both the most tempting and normatively superior option. Alternatively, the tendency to remove a normatively superior option from a menu had been rationalized by the DM's motivation to avoid a sense of guilt that stems from not choosing it (see Kopylov, 2012). This can only make sense if the DM would give up x_p^w for either x_f or x_s, thereby validating his anticipated guilt. Yet, clearly, the DM would choose x_p^w if it is available in my example: there is no negative emotions such as guilt involved in choosing the Pareto optimal allocation.

The model most closely related to this dictator's problem is Saito (2015)'s menu preference representation. They adopt the same problem, and capture what they refer to as *impure altruism*, which is exhibited when an intrinsically selfish dictator behaves altruistically in order to feel pride and to avoid the shame of acting selfishly. However, even in their model, the menu A is never preferred to B^w. That is, the value of pride cannot outweigh the value of better outcomes[36].

5.2. Welfare-measure approach: the dictator's welfare

The menu-choice approach runs into a crucial identification problem if the dictator's menu choice itself reflects his preference over the act of choosing. Note that the dictator prefers "preferring x_p^w to x_f and x_s", which means when x_p^w is available, he wants the act of giving up x_f and x_s for x_p^w. Consider the following scenario. In period 0, the dictator sees the two available menus of allocations A and B^w, and notices that the Pareto optimal allocation x_p^w is available in B^w. He thinks that if he chooses the menu A and the fair allocation x_f afterwards, then he is consequentially giving up x_p^w by giving up the menu B^w, feeling responsible for leaving himself

[36] If I remove the parameter for the cost of shame, and impose the maximal parameter for the sense of pride in Saito (2015)'s model, so that the dictator's pride of acting altruistically is maximized, his representation becomes V_S of the form:

$$V_S(A) := \max_{x \in A} \alpha W(x_1) + W(x_2) + \alpha \left[\max_{y \in A} W(y_1) - W(x_1) \right]$$

for some $\alpha > 0$ where $W : \mathbb{R} \to \mathbb{R}$ is a ranking of wealth outcomes. Here, the ex post choice of allocation is the most altruistic allocation—i.e., $\arg\max_{x \in A} W(x_2)$. The sense of pride is captured by the term $\max_{y \in A} W(y_1) - W(x_1)$ which is the dictator's maximal wealth outcome available on the menu A minus his chosen wealth outcome $W(x_1)$. When the menu is $A = \{x_f, x_s\}$, we have $V_S(A) = W(5) + \alpha W(6)$. By choosing x_f over x_s, the dictator gains the sense of pride measured by $\alpha W(6)$. When the menu is B^w, we have $V_S(B^w) = W(6 + w) + \alpha W(6 + w)$. Assuming that W is an increasing function, we have $V_S(B^w) \geq V_S(A)$.

and the recipient with the Pareto inferior allocation x_f. Thus, even if he is a person who would feel a great sense of pride in choosing x_f from A, he might think that his menu in period 0 is not $\mathcal{A} = \{A, B^w\}$ but essentially $A \cup B^w = B^w$ because he has control over his own menu in the future[37]. He wishes that the menu A would be exogenously given to him without his own influence, but as long as he has control, his choice is B^w in period 0, and x_p^w in period 1 because

$$(x_p^w, A \cup B^w) \succ (x_f, A \cup B^w) \succ (x_s, A \cup B^w).$$

As a result, the menu-choice approach in the previous section may not identify preferences over the act of choosing that exhibit locally pure paternalism as in (3).

The above problem arises because it is crucial in my model that the menus are exogenous when the DM engages in the act of choosing. Given any choice situation, the DM with a second-order preference might only be concerned about the question "What would the person with the ideal first-order preference choose?" This is regardless of whether or not the DM is making a meta choice from higher-order menus. Indeed, a person who only has a first-order preference $\succeq_1$ satisfying $x_p^w \succ_1 x_f \succ_1 x_s$ would obviously look for a menu that contains x_p^w.

This identification problem is analogous to the non-comparability problem formally illustrated by Bernheim et al. (2024). The problem arises when the DM cares about the act of choosing, in which case, his choices and welfare are not necessarily aligned[38]. As a result, the DM's welfare cannot be uniquely recovered from standard choice data. To illustrate using the dictator's example above, suppose a social planner (she) wants to figure out which menu among A and B^w would make the dictator (he) better off. She observes that his choices are B^w in period 0 and x_p^w in period 1. Based on this observation alone, she is unsure whether or not he would feel proud of himself for choosing x_f from A, and thus be happier than when B^w is given.

To address this problem, Bernheim et al. (2024) developed an econometric method to measure welfare by combining the DM's choice data and self-reported well-being data. Briefly, since the mental benefits (or costs) of the DM's act of choosing is not revealed by choices, they not only observe the DM's choices but also request the DM to report his mental states in multiple categories (such as pride and guilt) that he associates with each possible choice from exogenously given menus. Then, using the standard discrete choice techniques, they use

[37] The dictator's preference is consistent with the constraint-set dependent preference illustrated by Bernheim et al. (2024) (see also Kőszegi and Rabin, 2008). The constraint set refers to the DM's outcome constraints. Hence, regardless of whether or not the DM is making a meta choice from higher-order menus, his object of interest is the pair (x, C) where C is the set of outcomes that the DM can ultimately choose (e.g., $C = A \cup B^w$ for the dictator).

[38] Bernheim et al. (2024) showed that even when the DM cares about the act of choosing *menus* (or any other higher-order menus), the non-comparability problem persists. Suppose the DM faces a n-stage decision problem. An outcome is chosen from a first-order menu (i.e., a menu of outcomes) in the final (nth) stage; the first-order menu is chosen from a second-order menu (i.e., a menu of first-order menus) in the $n-1$th stage, and so on. Even from the choice data in this setting, the social planner cannot recover the DM's preference over the act of choosing in the first stage when he chooses the nth-order menu from an exogenously given $n+1$th-order menu, due to the same reason why she cannot recover the dictator's preference from his choice of B^w over A.

the choice data to estimate preferences over the mental states, and obtain the DM's welfare function $\mathcal{W} : \mathbb{C} \to \mathbb{R}$ that escapes from the non-comparability problem.

Bernheim et al. (2024)'s welfare measure $\mathcal{W}$ and their experimental design can be implemented in my dictator's problem. In period 0, the dictator will be given K exogenous menus $(A_k \in \mathcal{A}, k = 1, ..., k)$ of allocations randomly chosen by a computer. In period 1, the dictator will sequentially choose an allocation $x_k \in A_k$ for each $k = 1, ..., K$. Then, the experimenter will gather the dictator's self-reported well-being data—Bernheim et al. (2024) uses proxies for multiple composites of mental states, which they call *Categorical Subjective Assessments* (CSAs). Lastly, using their econometric method, we obtain the welfare function $\mathcal{W}$ (see Figure 2).

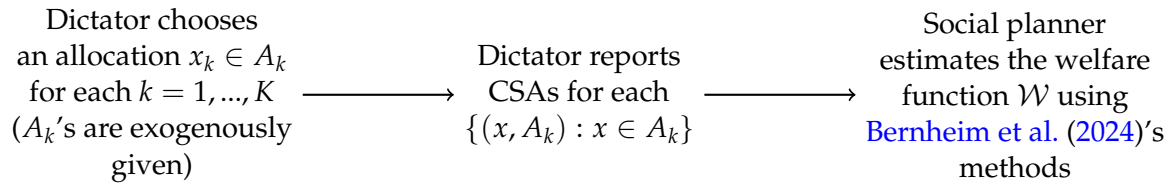

Figure 2: Timeline of experiment using Bernheim et al. (2024)'s design

While their experiments were focused on verifying the existence of the non-comparability problem, the focus here is to verify the existence of preferences over the act of choosing—more specifically, verifying that the following inequality is possible:

$$\mathcal{W}(x_f, \{x_f, x_s\}) > \mathcal{W}(x_p^w, \{x_f, x_s, x_p^w\}).$$

5.3. Choice based on choice data: choosing the right dictator

Second-order preferences are fundamentally difficult to observe because in most imaginable cases, the DM cares about the outcomes of his choices as well. The two approaches in the previous sections partially address this concern, but they still seems to lack observational power in the sense that the axioms in this paper cannot be directly tested. For example, the indifference of the vacuous choices holds only if we shut down the DM's preferences for outcomes.

I address this problem by considering a decision problem of the recipient in the dictator game who cares only about his wealth outcomes. Suppose two dictators (Dictators 1 and 2) played a game: each Dictator $i \in \{1, 2\}$ was exogenously given a menu A_i of allocations, and chose $x_i \in A_i$. (In general, there can be $K \geq 2$ dictators.) The DM who was not involved in the two previous games is about to play a dictator game as a recipient. His decision problem is as follows: before the game begins, he must choose whether Dictator 1 or Dictator 2 will be the dictator in his own game. Before making this decision, the DM observes the choice data of the past games. That is, he observes the two choices (x_1, A_1) and (x_2, A_2) made by the two dictators, knowing that the menus were exogenously given. Also, the DM privately knows the

menu A^* of allocations that will be given to his dictator. When the game begins, the chosen dictator will choose an allocation from A^*. See Figure 3.

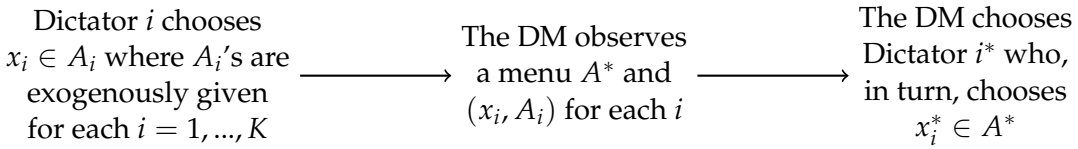

Figure 3: Timeline of the game of choosing the dictator

There are several important assumptions. First, it is common knowledge that all players have no control over the menus. Second, to avoid any strategic motivations, the two dictators are not aware of the possibility of playing another game, or that their choices are being observed by the DM. Third, the dictators do not make mistakes (i.e., there are no cognitive imprecision or trembling hands involved). Also, their preferences over the allocations are stable (i.e., no changing tastes). Fourth, the DM knows that these assumptions hold. Lastly, I allow the dictators to choose more than one allocation if they are indifferent among some options, in which case, the recipient decides the final allocation[39].

In this setting, even though the DM cares only about his wealth outcomes, he must care about the two dictators' preferences because the chosen dictator's preference over the allocations will determine his outcome. The outcomes of the past games do not affect the DM's outcome because he was not involved. Consequently, this decision problem turns the DM's standard preference for wealth into a preference over (the dictator's) preferences. This means all of my axioms (including Axioms 1-2) can be tested directly by a revealed-preference approach without relying on menu choices or econometric models using the self-reported well-being data.

Suppose the DM knows that the menu for his chosen dictator will be $A^* = \{x_f, x_s\}$, and observes five dictators' choices as below:

$$\begin{aligned}
&\text{Dictator 1's choice:} \quad (x_1, A_1) = (x_f, \{x_f, x_s\});\\
&\text{Dictator 2's choice:} \quad (x_2, A_2) = (x_p^w, \{x_f, x_s, x_p^w\});\\
&\text{Dictator 3's choice:} \quad (x_3, A_3) = (x_f, \{x_f\});\\
&\text{Dictator 4's choice:} \quad (x_4, A_4) = (x_s, \{x_s\});\\
&\text{Dictator 5's choice:} \quad (x_5, A_5) = (x_s, \{x_f, x_s\}).
\end{aligned}$$

Then, it is reasonable for the DM to choose Dictator 1 to be his dictator. While the choice $(x_f, \{x_f, x_s\})$ suggests that Dictator 1 cares about fairness, the choice $(x_p^w, \{x_f, x_s, x_p^w\})$ does not

[39] For example, suppose the menu is $\{x, y, z\}$, and the dictator chooses x and y because he is indifferent. Then, the recipient chooses from $\{x, y\}$. If this happened in the previous game, the DM observes $(\{x, y\}, \{x, y, z\})$.

suggest anything about Dictator 2's preference over $\{x_f, x_s\}$. Similarly, Dictator 5 would be the last person the DM wants because the choice $(x_s, \{x_f, x_s\})$ reveals selfishness. If the DM's preference exhibits the locally pure paternalism as in (3), we should observe the following ranking:

$$\text{Dictator 1} \succ \text{Dictator 2} \sim \text{Dictator 3} \sim \text{Dictator 4} \succ \text{Dictator 5}$$

where the indifference relations are due to Axiom 8, or the indifference of vacuous choices.

To test Axiom 1, we need to let each dictator play two games prior to the DM's decision problem. Suppose the DM knows that the menu for his chosen dictator will be $A^* = \{x_f, x_s, x_p^w\}$, and observes two dictators' choices as below:

$$\begin{array}{lllll} \text{Dictator 1's choices:} & (x_p^w, \{x_f, x_s, x_p^w\}) & \text{and} & (x_f, \{x_f, x_s\}); \\ \text{Dictator 2's choices:} & (x_p^w, \{x_f, x_s, x_p^w\}) & \text{and} & (x_s, \{x_f, x_s\}). \end{array}$$

Based on this choice data, Dictator 1 cares about fairness while Dictator 2 does not. However, because both dictators chose the Pareto optimal allocation from $\{x_f, x_s, x_p^w\}$, the DM knows that regardless of who the dictator will be in his game, x_p^w will be chosen from A^*. For this reason, Axiom 1 dictates that the DM is indifferent between the two dictators.

6. Discussion

6.1. Models of Menu Preference

Most menu preference representations in the prior literature on temptation and self-control are a function $U_{u,f}$ of the form:

$$U_{u,f}(A) := \max_{x \in A} u(x) + f(x, A) \tag{4}$$

where $u : X \to \mathbb{R}$ is an outcome ranking that determines the ranking of singleton sets, and $f : \mathbb{C} \to \mathbb{R}$ is the menu-dependent component that takes a special form in each prior model (e.g., $f(x, A) = v(x) - \max_{y \in A} v(y)$ in the seminal model by Gul and Pesendorfer (2001)). While the prior work mainly focused on identifying the outcome ranking u based on menu-choice data, my analysis focuses on studying the conceptual understanding of f alone, investigating the class of functions that f can be, by assuming no outcome preferences (i.e., u is a constant function).

The representation $V_{v,\mathbf{r}}$ in Theorem 1 subsumes any affine menu preference representation $U_{u,f}$ assuming u is constant and $V_{v,\mathbf{r}} = f$. Of course, (4) is not a theorem in this paper. It is simply an optimization problem. Yet, we can think of an agent who maximizes additively separable first- and second-order preferences represented by the form (4). In fact, by observing the menu choices, we can use (4) to uniquely reproduce the same menu-choice patterns gen-

erated by the prior work. For instance, given any Gul and Pesendorfer (2001)'s representation $U_{GP}(A) := \max_{x\in A} u(x) + v(x) - \max_{y\in A} v(y)$, there is a unique optimization problem of additively separable first-order preference representation u and a purely paternalistic preference (represented by $v(x) - \max_{y\in A} v(y)$) over the act of choosing that yields the same behavior.

The optimization approach suggests that one's paternalistic stance toward the act of choosing yields preferences for smaller menus implying costly self-control (i.e., preferences for commitment) as well as guilt-avoidance behavior. Also, the libertarian attitudes yield preferences for larger menus, implying pride-seeking behavior (i.e., preferences for menus that require self-control)[40].

6.2. Welfare Implications: a higher-order non-comparability problem

The concept of higher-order preferences suggests that the design of welfare policies is influenced by the social planner's own preferences. Consequently, even with extensive data on the DM's first- and second-order preferences—such as choice data and data on mental states—the inherent complexities in welfare assessments may not be resolved. This challenge arises because the social planner must interpret these preferences in light of her own higher-order goals, which might prioritize the DM's immediate well-being, long-term welfare, or some combination of both.

For example, consider a mother deciding whether or not to instruct her child to clean his room. Suppose the mother has sufficient data to know that (i) the child will surely succumb to the temptation of playing with his smartphone instead, and (ii) he will feel guilty for choosing play over fulfilling the parent's request[41]. Based on Bernheim et al. (2024)'s welfare measures, the mother should silently clean the room herself, allowing the child to play without any feelings of guilt or shame. However, some parent might intentionally instruct the child to clean, not to have the room cleaned, but because she believes experiencing guilt is crucial for the child's personal growth and long-term welfare. In this case, the parent's decision reflects her own higher-order preference to prioritize the child's future development over immediate well-being.

This example highlights how the non-comparability problem presented by Bernheim et al. (2024) extends to a higher-order level. The social planner (the parent, in this case) faces a meta-preference challenge, balancing the DM's immediate pleasure (e.g., the child's joy in playing with the smartphone) against his future welfare (e.g., cultivating responsibility through guilt). This parallels the standard tension in the temptation literature, where a DM may struggle be-

[40] Guilt-avoidance behavior has been observed in several experiments in the social preference literature (e.g., avoiding the opportunity to act prosocially; Dana et al. (2006)). Non-axiomatic models as well as other empirical studies suggest that people sometimes prefer facing temptation because self-control improves self-image and willpower (Prelec and Bodner, 2003; Bénabou and Tirole, 2004; Dunning, 2007; Dhar and Wertenbroch, 2012).

[41] Say, the mother is choosing the child's menu: $A = \{\text{clean, smartphone}\}$ vs. $B = \{\text{smartphone}\}$.

tween short-term indulgence and long-term goals. Evidently, this tension is not limited to the DM alone; it also exists in the social planner's interpretation of what welfare entails.

The higher-order non-comparability problem can be illustrated more clearly when we introduce another observer, such as the father, who observes both the mother's welfare policy (instructing the child to clean) and the child's choice (playing over cleaning). Suppose the father knows that the mother had sufficient data on the child's choices and mental states when she implemented her policy. Additionally, if he knows that the mother's goal was to promote the child's immediate well-being, then he may conclude that the child feels pride in willingly resisting the mother's request. However, if the father assumes the mother's intent is to foster long-term welfare, he might conclude that the child feels guilty for not cleaning, a completely opposite interpretation. Thus, without knowing the social planner's higher-order preferences, even extensive data on the DM's choices and mental states fails to resolve the ambiguity in welfare judgments.

This example suggests that the combination of choice and policy data alone cannot recover the DM's preferences if the previous policy maker's goals remain unclear. This higher-order non-comparability problem necessitates a more transparent framework for articulating the social planner's goals in welfare policy design. Welfare analysis thus requires not only an understanding of the DM's first- and second-order preferences but also explicit knowledge of the social planner's preference over the DM's second-order preferences.

Appendix

A. Proof of Theorem 1

The "if" part is straightforward. To prove the "only if" part, note that Axiom 6 grants the existence and uniqueness of the continuous affine function v representing $\succeq_1$ due to the standard expected utility theory. Moreover, note that $\mathbb{C}$ is a mixture space. Then, by the result of Herstein and Milnor (1953), Axioms 3, 4 and 5 ensure the existence of a continuous affine function $V : \mathbb{C} \to \mathbb{R}$ representing $\succeq$. That is,

$$\begin{aligned} &V(x, A) \geq V(y, B) \iff (x, A) \succeq (y, B); \\ &V(\lambda(x, A) + (1 - \lambda)(y, B)) = \lambda V(x, A) + (1 - \lambda) V(y, B) \quad \forall \lambda \in [0, 1]. \end{aligned}$$

Then, we have the following result:

Lemma 1. $(x, A) \succeq (y, B) \iff \frac{1}{2}x + \frac{1}{2}\mathbf{r}(B) \succeq_1 \frac{1}{2}\mathbf{r}(A) + \frac{1}{2}y.$

Proof of Lemma 1. Note that

$$\begin{aligned}(x,A) \succeq (y,B) &\iff \tfrac{1}{2}(x,A) + \tfrac{1}{2}(\mathbf{r}(B),B) \succeq \tfrac{1}{2}(y,B) + \tfrac{1}{2}(\mathbf{r}(A),A) \\ &\iff \left(\tfrac{1}{2}x + \tfrac{1}{2}\mathbf{r}(B), \tfrac{1}{2}A + \tfrac{1}{2}B\right) \succeq \left(\tfrac{1}{2}\mathbf{r}(A) + \tfrac{1}{2}y, \tfrac{1}{2}A + \tfrac{1}{2}B\right) \\ &\iff \tfrac{1}{2}x + \tfrac{1}{2}\mathbf{r}(B) \succeq_1 \tfrac{1}{2}\mathbf{r}(A) + \tfrac{1}{2}y \text{ by Axioms 6-7}\end{aligned}$$

which completes the proof of Lemma 1. □

Since v is an affine function representing $\succeq_1$, we have

$$(x,A) \succeq (y,B) \iff v(x) - v(\mathbf{r}(A)) \geq v(y) - v(\mathbf{r}(B)).$$

Define $V_{v,\mathbf{r}} : \mathbb{C} \to \mathbb{R}$ by $V_{v,\mathbf{r}}(x,A) \coloneqq v(x) - v(\mathbf{r}(A))$. The goal is to let $V = V_{v,\mathbf{r}}$. To show that $V_{v,\mathbf{r}}$ is also a continuous affine function, we need to show that $K(A) \coloneqq v(\mathbf{r}(A))$ is a continuous affine function of sets. I first derive the following lemma:

Lemma 2 (Reference Affinity)**.** $\mathbf{r}(\lambda A + (1-\lambda)\,B) \sim_1 \lambda\mathbf{r}(A) + (1-\lambda)\,\mathbf{r}(B)$ *for* $\lambda \in [0,1]$.

Proof of Lemma 2. By Axiom 8, we have $(\mathbf{r}(A),A) \sim (\mathbf{r}(B),B)$. Then, by Axiom 4, we have

$$(\mathbf{r}(A),A) \sim \lambda(\mathbf{r}(A),A) + (1-\lambda)\,(\mathbf{r}(B),B) \sim (\lambda\mathbf{r}(A) + (1-\lambda)\,\mathbf{r}(B), \lambda A + (1-\lambda)\,B)\,.$$

Axiom 8 also gives us

$$(\mathbf{r}(A),A) \sim (\mathbf{r}(\lambda A + (1-\lambda)\,B), \lambda A + (1-\lambda)\,B)\,.$$

By Axioms 6-7, we can conclude $\mathbf{r}(\lambda A + (1-\lambda)\,B) \sim_1 \lambda\mathbf{r}(A) + (1-\lambda)\,\mathbf{r}_B$. □

Next, I define a binary relation $\succeq_{\mathbf{r}}$ on $\mathbb{M}$ as $A \succeq_{\mathbf{r}} B$ if and only if $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$. I say $\succeq_{\mathbf{r}}$ is *independent* if $A \succ_{\mathbf{r}} B$ implies $\lambda A + (1-\lambda)C \succ_{\mathbf{r}} \lambda B + (1-\lambda)C$ for all $\lambda \in (0,1)$. $\succeq_{\mathbf{r}}$ is *continuous* if $\{A : A \succeq_{\mathbf{r}} B\}$ and $\{A : B \succeq_{\mathbf{r}} A\}$ are closed. The next lemma is a useful consequence of Lemma 2:

Lemma 3. $\succeq_{\mathbf{r}}$ *is complete, transitive, continuous and independent.*

Proof of Lemma 3. Since $\succeq_1$ is complete and transitive, $\succeq_{\mathbf{r}}$ is as well. For continuity, since $\mathbb{M}$ is a topological space, it is sufficient to show that $A \succ_{\mathbf{r}} C \succ_{\mathbf{r}} B$ implies that there are $\alpha, \beta \in (0,1)$ such that

$$\alpha A + (1-\alpha)\,B \succ_{\mathbf{r}} C \succ_{\mathbf{r}} \beta A + (1-\beta)\,B.$$

Since $\succeq_1$ is continuous, there are $\alpha, \beta \in (0,1)$ such that

$$\alpha\mathbf{r}(A) + (1-\alpha)\,\mathbf{r}(B) \succ_1 \mathbf{r}(C) \succ_1 \beta\mathbf{r}(A) + (1-\beta)\,\mathbf{r}(B).$$

By Lemma 2, we have $\mathbf{r}(\alpha A + (1-\alpha)\,B) \succ_1 \mathbf{r}(C) \succ_1 \mathbf{r}(\beta A + (1-\beta)\,B)$ which is equivalent to our desired result by definition of $\succeq_{\mathbf{r}}$. For *independence*, suppose $A \succ_{\mathbf{r}} B$ or equivalently, $\mathbf{r}(A) \succ_1 \mathbf{r}(B)$. Since $\succeq_1$ is *independent*, $\lambda \in (0,1)$ implies $\lambda\mathbf{r}(A) + (1-\lambda)\mathbf{r}(C) \succ_1 \lambda\mathbf{r}(B) + (1-\lambda)\mathbf{r}(C)$. By Lemma 2, it implies $\mathbf{r}(\lambda A + (1-\lambda)C) \succ_1 \mathbf{r}(\lambda B + (1-\lambda)C)$. By definition, we have $\lambda A + (1-\lambda)C \succ_{\mathbf{r}} \lambda B + (1-\lambda)C$. □

By the result of Herstein and Milnor (1953), Lemma 3 holds if and only if there is a continuous affine representation $K : \mathbb{M} \to \mathbb{R}$ of $\succeq_{\mathbf{r}}$. By definition of $\succeq_{\mathbf{r}}$, the ranking of singleton sets follows $\succeq_1$ and thus, $K(\{x\}) = v(x)$ for all $x \in X$. Since $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$ is equivalent to $A \succeq_{\mathbf{r}} B$ which is represented by $K(A) \geq K(B)$, we conclude $K(A) = v(\mathbf{r}(A))$ for all $A \in \mathbb{M}$.

As the final step, $V_{v,\mathbf{r}} = v - K$ is a continuous function since both v and K are continuous. And $V_{v,\mathbf{r}}$ is affine since

$$\begin{aligned} V_{v,\mathbf{r}}(\lambda(x,A) + (1-\lambda)(y,B)) &= v(\lambda x + (1-\lambda)y) - v(\mathbf{r}(\lambda A + (1-\lambda)B)) \\ &= \lambda v(x) + (1-\lambda)v(y) - v(\lambda \mathbf{r}(A) + (1-\lambda)\mathbf{r}(B)) \\ &= \lambda v(x) + (1-\lambda)v(y) - [\lambda v(\mathbf{r}(A)) + (1-\lambda)v(\mathbf{r}(B))] \\ &= \lambda[v(x) - v(\mathbf{r}(A))] + (1-\lambda)[v(y) - v(\mathbf{r}(B))] \\ &= \lambda V_{v,\mathbf{r}}(x,A) + (1-\lambda)V_{v,\mathbf{r}}(y,B) \end{aligned}$$

which completes the proof of Theorem 1. □

B. Proof of Theorem 2

For the "if" part. Suppose $v' = \alpha v + \beta$ and $\mathbf{r}'(A) \sim_1 \mathbf{r}(A)$ for all $A \in \mathbb{M}$. Then

$$\begin{aligned} (x,A) \succeq (y,B) &\iff v(x) - v(\mathbf{r}(A)) \geq v(y) - v(\mathbf{r}(B)) \\ &\iff \left[\alpha v(x) + \beta\right] - \left[\alpha v(\mathbf{r}(A)) + \beta\right] \geq \left[\alpha v(y) + \beta\right] - \left[\alpha v(\mathbf{r}(B)) + \beta\right] \\ &\iff v'(x) - v'(\mathbf{r}(A)) \geq v'(y) - v'(\mathbf{r}(B)) \\ &\iff v'(x) - v'(\mathbf{r}'(A)) \geq v'(y) - v'(\mathbf{r}'(B)) \end{aligned}$$

where the last equivalence is due to $\mathbf{r}'(A) \sim_1 \mathbf{r}(A)$.

To prove the "only if" part, suppose $(v, \mathbf{r})$ and $(v', \mathbf{r}')$ represent $\succeq$. I need to show that (i) there exists $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $v' = \alpha v + \beta$ and that (ii) $\mathbf{r}'(A) \sim_1 \mathbf{r}(A)$ for all $A \in \mathbb{M}$. (i) is the result of the standard expected utility theory. For (ii), note that since $\mathbb{M}$ is the set of convex subsets, the proof is trivial: we always have $\mathbf{r}(A), \mathbf{r}'(A) \in A$ for all $A \in \mathbb{M}$, and thus we have $(\mathbf{r}(A), A) \sim \phi \sim (\mathbf{r}'(A), A)$, which implies $\mathbf{r}(A) \sim_1 \mathbf{r}'(A)$. □

C. Non-convex Menus and Finite menus

Let $\mathbb{M}^*$ be the set of nonempty compact subsets of X. (Note that $\mathbb{M}^*$ includes all finite menus.) I now use the standard set operations instead of the alternative ones $(\cup^*, \cap^*, \setminus^*)$ defined in Section 2. I replace Axiom 8 with the following axiom:

Axiom 15 (Discrete Relativity). *For any $A, B \in \mathbb{M}^*$, there are $x' \in conv(A)$ and $y' \in conv(B)$ such that*

a. $(x, A \cup \{x'\}) \sim (x, A)$ and $(y, B \cup \{y'\}) \sim (y, B)$ for all $x \in A$, $y \in B$, and

b. $(x', A \cup \{x'\}) \sim (y', B \cup \{y'\})$.

I redefine the choice function as follows:

Definition 2. *A function* $\mathbf{r} : \mathbb{M}^* \to X$ *is called a stochastic choice function if* $\mathbf{r}(A) \in conv(A)$ *for all* $A \in \mathbb{M}^*$.

I now have the following result:

Theorem 4. $\succeq$ *on* $\mathbb{M}^*$ *satisfies Axioms 1-7, and Axiom 15 if and only if* $\succeq$ *has a unique representation as in Theorems 1-2 where the choice function* $\mathbf{r}(\cdot)$ *is a stochastic choice function.*

Proof of Theorem 4. Let $\mathbf{r} : \mathbb{M}^* \to X$ be the choice function defined according to Axiom 15. That is, given any $A, B \in \mathbb{M}^*$,

a. $(x, A \cup \{\mathbf{r}(A)\}) \sim (x, A)$ and $(y, B \cup \{\mathbf{r}(B)\}) \sim (y, B)$ for all $x \in A$, $y \in B$, and

b. $(\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) \sim (\mathbf{r}(B), B \cup \{\mathbf{r}(B)\})$.

For Theorem 1, it is sufficient to show that Lemmas 1-3 hold. For Lemma 1, the "if" part is again straightforward. For the "only if" part, note that $(x, A) \succeq (y, B)$ is equivalent to

$$
\begin{aligned}
&(x, A \cup \{\mathbf{r}(A)\}) \succeq (y, B \cup \{\mathbf{r}(B)\}) \text{ by Axiom 15a} \\
&\iff \tfrac{1}{2}(x, A \cup \{\mathbf{r}(A)\}) + \tfrac{1}{2}(\mathbf{r}(B), B \cup \{\mathbf{r}(B)\}) \succeq \tfrac{1}{2}(y, B \cup \{\mathbf{r}(B)\}) + \tfrac{1}{2}(\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) \\
&\qquad \text{by Axiom 4 and Axiom 15b} \\
&\iff \left(\tfrac{1}{2}x + \tfrac{1}{2}\mathbf{r}(B), \tfrac{1}{2}A \cup \{\mathbf{r}(A)\} + \tfrac{1}{2}B \cup \{\mathbf{r}(B)\}\right) \succeq \left(\tfrac{1}{2}\mathbf{r}(A) + \tfrac{1}{2}y, \tfrac{1}{2}A \cup \{\mathbf{r}(A)\} + \tfrac{1}{2}B \cup \{\mathbf{r}(B)\}\right) \\
&\iff \tfrac{1}{2}x + \tfrac{1}{2}\mathbf{r}(B) \succeq_1 \tfrac{1}{2}\mathbf{r}(A) + \tfrac{1}{2}y \text{ by Axioms 6-7} \\
&\iff v(x) - v(\mathbf{r}(A)) \geq v(y) - v(\mathbf{r}(B)).
\end{aligned}
$$

Hence, Lemma 1 holds. To show that Lemmas 2-3 hold, the following result will be used:

Lemma 4. $\mathbf{r}(conv(A)) \sim_1 \mathbf{r}(A) \sim_1 \mathbf{r}(A \cup \{\mathbf{r}(A)\})$ *for all* $A \in \mathbb{M}^*$.

Proof of Lemma 4. Consider the following choice

$$(x, A_n) = \sum_{s=1}^{n} \lambda_s (x, A)$$

where $A_n = \sum_{s=1}^{n} \lambda_s A$ and $\lambda_s = \frac{1}{n}$ for all $s = 1, ..., n$. Then, by the result known as the Shapley-Folkman theorem (see Emerson and Greenleaf, 1969; Starr, 1969), A_n converges to $conv(A)$ in the Hausdorff metric, and thus, (x, A_n) converges to $(x, conv(A))$. Because $\succeq$ has an affine representation, we have $(x, A) \sim (x, A_n)$ for all $n \in \mathbb{N}$. Then, by Axiom 5, we have $(x, conv(A)) \sim (x, A)$, which, by Lemma 1, is equivalent to $\frac{1}{2}x + \frac{1}{2}\mathbf{r}(A) \sim_1 \frac{1}{2}\mathbf{r}(conv(A)) + \frac{1}{2}x$ for all $n \in \mathbb{N}$. Since $\succeq_1$ is independent, this gives us $\mathbf{r}(conv(A)) \sim_1 \mathbf{r}(A)$. For the second indifference relation, I use the Shapley-Folkman theorem again to conclude $\sum_{s=1}^{n} \lambda_s A \cup \{\mathbf{r}(A)\}$ converges to $conv(A)$. Note that $conv(A \cup \{\mathbf{r}(A)\}) = conv(A)$ since $\mathbf{r}(A) \in conv(A)$. This implies $(\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) \sim (\mathbf{r}(A), conv(A))$ which, by Lemma 1, means $\mathbf{r}(conv(A)) \sim_1 \mathbf{r}(A \cup \{\mathbf{r}(A)\})$. This completes the proof of Lemma 4. □

Using Lemma 4, we can now prove Lemma 2. We now know that

$$\phi \sim (\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) \sim (\mathbf{r}(B), B \cup \{\mathbf{r}(B)\}) \sim (\mathbf{r}(A), conv(A)) \sim (\mathbf{r}(B), conv(B)).$$

The first two indifference relations are due to Axiom 15 and the last two relations are due to Lemma 4. Then, for any $\lambda \in [0,1]$, we have

$$\begin{aligned}(\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) &\sim \lambda(\mathbf{r}(A), conv(A)) + (1-\lambda)(\mathbf{r}(B), conv(B)) \quad \text{by Axiom 4} \\ &= (\lambda \mathbf{r}(A) + (1-\lambda)\mathbf{r}(B), \lambda conv(A) + (1-\lambda) conv(B)). \end{aligned} \tag{5}$$

Moreover, we also have

$$\begin{aligned}&(\mathbf{r}(A), A \cup \{\mathbf{r}(A)\}) \qquad (6)\\ &\sim (\mathbf{r}(\lambda A + (1-\lambda)B), \lambda A + (1-\lambda)B \cup \{\mathbf{r}(\lambda A + (1-\lambda)B)\}) \quad \text{by Axiom 15} \\ &\sim \lim_{n\to\infty} \sum_{s=1}^{n} \lambda_s (\mathbf{r}(\lambda A + (1-\lambda)B), \lambda A + (1-\lambda)B \cup \{\mathbf{r}(\lambda A + (1-\lambda)B)\}) \quad \text{by Axiom 5} \\ &= (\mathbf{r}(\lambda A + (1-\lambda)B), conv(\lambda A + (1-\lambda)B)) \quad \text{by the Shapley-Folkman theorem} \\ &= (\mathbf{r}(\lambda A + (1-\lambda)B), \lambda conv(A) + (1-\lambda) conv(B)). \qquad (7)\end{aligned}$$

By Axioms 6-7, the two results (5) and (6) imply $\mathbf{r}(\lambda A + (1-\lambda)B) \sim_1 \lambda \mathbf{r}(A) + (1-\lambda)\mathbf{r}(B)$, which proves Lemma 2. Then, Lemma 3 and Theorem 1 follow.

For Theorem 2, I only prove the "only if" part. Suppose $(v, \mathbf{r})$ and $(v', \mathbf{r}')$ represent $\succeq$. I need to show that (i) there exists $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $v' = \alpha v + \beta$ and that (ii) $\mathbf{r}'(A) \sim_1 \mathbf{r}(A)$ for all $A \in \mathbb{M}^*$. Since v and v' both represent $\succeq_1$, Lemma 1 ensures (i). For (ii), note that since $\mathbf{r}(A_n) \sim_1 \mathbf{r}(A)$ and $\mathbf{r}'(A_n) \sim_1 \mathbf{r}'(A)$ for all $n \in \mathbb{N}$, we can choose two sequences (x_n, A_n) and (x'_n, A_n) converging to $(\mathbf{r}(A), conv(A))$ and $(\mathbf{r}'(A), conv(A))$, respectively. Since $\mathbf{r}(A) \sim_1 \mathbf{r}(conv(A))$ and $\mathbf{r}'(A) \sim_1 \mathbf{r}'(conv(A))$ by Lemma 4, we have

$$(\mathbf{r}(A), conv(A)) \sim (\mathbf{r}'(A), conv(A)) \sim \phi \tag{8}$$

by Axiom 15 and Lemma 1. If $\mathbf{r}(A) \not\sim_1 \mathbf{r}'(A)$, then (8) contradicts Axiom 6. This completes the proof of Theorem 2, and thus the proof of Theorem 4. □

D. Proofs of Corollaries

Define the best and worst lotteries on a menu A by $b_A \in \{x \in A : x \succeq_1 y \ \forall y \in A\}$ and $w_A \in \{x \in A : y \succeq_1 x \ \forall y \in A\}$.

D.1. Proof of Corollary 1

I first claim that if Axiom 9 holds, then $\mathbf{r}(A) \sim_1 b_A$ for all A. That is, $(b_A, A) \sim \phi$ for all A. For the sake of contradiction, suppose $\phi \succ (b_A, A)$ for some A. Then, by definition of b_A, we have $\phi \succ (b_A, A) \succeq (x, A)$

for all $x \in A$ which violates Axiom 8. Similarly, we can show that if Axiom 10 holds, then $\mathbf{r}(A) \sim_1 w_A$ for all A. Then, the desired result follows by Lemma 1.

D.2. Proof of Corollary 2

Suppose Axiom 9 holds, $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$, and $C \cap (A \cup B) = \emptyset$. Suppose Corollary 2a does not hold: i.e., $(\mathbf{r}(B), B \cup C) \succ (\mathbf{r}(A), A \cup C)$. By Lemma 1, this implies

$$\tfrac{1}{2}b_B + \tfrac{1}{2}b_{A\cup C} \succ_1 \tfrac{1}{2}b_A + \tfrac{1}{2}b_{B\cup C}. \tag{9}$$

If $b_A \succeq_1 b_C \succeq_1 b_B$, then (9) becomes $\frac{1}{2}b_B + \frac{1}{2}b_A \succ_1 \frac{1}{2}b_A + \frac{1}{2}b_C$ which is a contradiction since $\succeq_1$ is independent. If $b_A \succeq_1 b_B \succ_1 b_C$, then (9) becomes $\frac{1}{2}b_B + \frac{1}{2}b_A \succ_1 \frac{1}{2}b_A + \frac{1}{2}b_B$ which is also a contradiction. If $b_C \succ_1 b_A \succeq_1 b_B$, then (9) becomes $\frac{1}{2}b_B + \frac{1}{2}b_C \succ_1 \frac{1}{2}b_A + \frac{1}{2}b_C$, a contradiction. Hence, Corollary 2a holds. For Corollary 2b, it is sufficient to show that $\mathbf{r}(A \cup C) \succeq_1 \mathbf{r}(B \cup C)$. This is immediately true because $b_{A\cup C} \succeq_1 b_{B\cup C}$ for any C. For Corollary 2c, it is sufficient to show that $\mathbf{r}(A \cup C) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B \cup C)$ holds if $\mathbf{r}(A) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B)$. Note that $\mathbf{r}(A) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B)$ immediately implies $b_{A\cup C} \succeq_1 b_C \succeq_1 b_{B\cup C}$. Thus, the proof is done. (We can similarly prove for Axiom 10.)

D.3. Proof of Corollary 3

I first show that Axiom 12 holds. Suppose $x \succeq_1 y$ and $x, y \notin C$. By Lemma 1, it is sufficient to show that for all $\alpha \in (0,1)$,

$$v(x) - v(y) \geq \alpha\Big[v(b_{\{x\}\cup C}) - v(b_{\{y\}\cup C})\Big] + (1-\alpha)\Big[v(w_{\{x\}\cup C}) - v(w_{\{y\}\cup C})\Big] \geq 0 \tag{10}$$

where the first and second inequalities imply Axiom 12a and Axiom 12b, respectively. Notice that $v(x) - v(y) \geq v(b_{\{x\}\cup C}) - v(b_{\{y\}\cup C}) \geq 0$ and $v(x) - v(y) \geq v(w_{\{x\}\cup C}) - v(w_{\{y\}\cup C}) \geq 0$ regardless of the value of $v(b_C)$ and $v(w_C)$. Hence, (10) holds for all $\alpha \in (0,1)$.

I now show that $\succeq_{v,\alpha}$ does not satisfy Axiom 11 when $\alpha \in (0,1)$. To be specific, Axiom 11a-b are not satisfied. For Axiom 11a, suppose $v(\mathbf{r}(A)) > v(\mathbf{r}(B))$ which implies

$$(1-\alpha)\Big[v(w_B) - v(w_A)\Big] < \alpha\Big[v(b_A) - v(b_B)\Big]. \tag{11}$$

Suppose $v(b_C) \leq v(b_B) < v(b_A)$ and $v(w_C) \leq v(w_A) < v(w_B)$. Then, assuming (11), the following holds

$$\begin{aligned}
V(\mathbf{r}(A), A\cup C) - V(\mathbf{r}(B), B\cup C) &= v(\mathbf{r}(A)) - v(\mathbf{r}(B)) + v(\alpha b_{B\cup C} + (1-\alpha)w_{B\cup C}) - v(\alpha b_{A\cup C} + (1-\alpha)w_{A\cup C}) \\
&< (1-\alpha)\Big[v(w_A) - v(w_B) + v(w_{B\cup C}) - v(w_{A\cup C})\Big] \\
&= (1-\alpha)\Big[v(w_A) - v(w_C) + v(w_C) - v(w_B)\Big] \\
&= (1-\alpha)\Big[v(w_A) - v(w_B)\Big] \\
&< 0
\end{aligned}$$

for all $\alpha \in (0,1)$, which violates Axiom 11a.

For Axiom 11b, suppose $v(b_B) < v(b_A) \leq v(b_C)$ and $v(w_A) < \min\{v(w_B), v(w_C)\}$. Then, assuming (11), the following holds for all $\alpha \in (0,1)$,

$$\begin{aligned} V(c, A \cup C) - V(c, B \cup C) &= v(\alpha b_{B\cup C} + (1-\alpha) w_{B\cup C}) - v(\alpha b_{A\cup C} + (1-\alpha) w_{A\cup C}) \\ &= (1-\alpha)\Big[v(w_{B\cup C}) - v(w_{A\cup C})\Big] \\ &= (1-\alpha)\Big[v(w_{B\cup C}) - v(w_A)\Big] \\ &> 0 \end{aligned}$$

which violates Axiom 11b.

I show that $\succeq_{v,\alpha}$ satisfies Axiom 11c. In the proof of Corollary 4 in Section D.4, I show that Axiom 13 implies Axiom 11c. Hence, it is sufficient to show that $\succeq_{v,\alpha}$ satisfies Axiom 13. Suppose $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$, which means $\alpha(v(b_A) - v(b_B)) \geq (1-\alpha)(v(w_B) - v(w_A))$. Then, we have $\mathbf{r}(A) \succeq_1 \mathbf{r}(A \cup B)$ since

$$v(\mathbf{r}(A)) - v(\mathbf{r}(A \cup B)) = \alpha\Big[v(b_A) - v(b_{A\cup B})\Big] + (1-\alpha)\Big[v(w_A) - v(w_{A\cup B})\Big] \geq 0.$$

This is true since $v(\mathbf{r}(A)) - v(\mathbf{r}(A \cup B))$ is zero if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_A$; it is greater than zero if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_B$ or if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_B$. Note that $\mathbf{r}(A) \succeq_1 \mathbf{r}(B)$ does not allow $b_{A\cup B} = b_B$ and $w_{A\cup B} = w_A$. Similarly, we have $\mathbf{r}(A \cup B) \succeq_1 \mathbf{r}(B)$ since

$$v(\mathbf{r}(A \cup B)) - v(\mathbf{r}(B)) = \alpha\Big[v(b_{A\cup B}) - v(b_B)\Big] + (1-\alpha)\Big[v(w_{A\cup B}) - v(w_B)\Big] \geq 0.$$

This is true since $v(\mathbf{r}(A \cup B)) - v(\mathbf{r}(B))$ is greater than zero if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_A$ or if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_B$; it is zero if $b_{A\cup B} = b_A$ and $w_{A\cup B} = w_B$.

D.4. Proof of Corollary 4

I first show that Axiom 14 is satisfied. Suppose $\mathbf{r}(A) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B)$. By Axiom 13, this implies

$$\mathbf{r}(A) \succeq_1 \mathbf{r}(A \cup C) \succeq_1 \mathbf{r}(C) \succeq_1 \mathbf{r}(B \cup C) \succeq_1 \mathbf{r}(B).$$

I can use Lemma 1 to conclude (a) $v(\mathbf{r}(A)) - v(\mathbf{r}(A \cup C)) \geq 0 \geq v(\mathbf{r}(B)) - v(\mathbf{r}(B \cup C))$, which implies Axiom 14a, and (b) $(c, B \cup C) \succeq (c, C) \succeq (c, A \cup C)$ if and only if $\frac{1}{2}c + \frac{1}{2}\mathbf{r}(C) \succeq_1 \frac{1}{2}c + \frac{1}{2}\mathbf{r}(B \cup C)$ and $\frac{1}{2}c + \frac{1}{2}\mathbf{r}(A \cup C) \succeq_1 \frac{1}{2}c + \frac{1}{2}\mathbf{r}(C)$ which are true since $\succeq_1$ is independent—(b) is Axiom 14b.

Next, I show that Axiom 12, and in turn Axiom 11a-b, are not satisfied. Consider $A = \{x\}$, $B = \{y\}$, $C = \{c\}$, and

Options	v	u_0	u_1
x	3	2	5
y	2	5	4
c	1	4	5

Then, it is easy to verify that $v(x) > v(y)$, but $V_{v,u_0}(c, \{y\} \cup C) = -1 < 0 = V_{v,u_0}(c, \{x\} \cup C)$, which violates Axiom 12b. Also, $V_{v,u_1}(y, \{y\} \cup C) = 1 > 0 = V_{v,u_1}(x, \{x\} \cup C)$, which violates Axiom 12a. Since Axiom 12 is not satisfied, the stronger version Axiom 11 cannot hold.

Note that Axiom 14b is equivalent to Axiom 11c. Hence, Axiom 13 implies Axiom 11c.

E. Reference that depends on the number of options

Consider the preference $\succeq$ on $\mathbb{C}$ restricted to finite menus, and the utility function $V_{v,avg}$ of the form:

$$V_{v,avg}(x, A) = v(x) - \frac{1}{|A|} \sum_{y \in A} v(y)$$

where $|A|$ is the number of options in A and the reference value function $v(\mathbf{r}(\cdot))$ is the average value of v within A. That is, the size and relative values of options affect Amy's utility directly[42]. In this case, Amy's reference takes every option into account equally when evaluating Bob's preference, and thus, the choice function $\mathbf{r}(\cdot)$ selects the average point.

To illustrate, suppose Amy wants Bob to overcome his alcoholism. If Bob orders coffee at a wine bar where a variety of tempting options are served, Amy would be extremely proud of his choice since her reference would lean toward the choice of wine. Intuitively, abstaining from alcohol at a bar is regarded as a significant achievement for an addict. Yet, she may not be as impressed if he chose coffee at a morning buffet that serves many healthy alternatives to alcohol, but offers a small collection of wine. Thus, failing to resist alcohol at the buffet might raise a concern more serious about alcoholism than at a bar.

However, the preference represented by $V_{v,avg}$ does not satisfy Lemma 2 since $V_{v,avg}$ violates the continuity of $\succeq$ under the Hausdorff metric which does not allow a utility jump to be caused by a sudden change in the number of options. Suppose $1 = v(x) > v(y) = 0$ given two menus $A = \{x, y\}$ and $B_\alpha = \{x, \alpha x + (1-\alpha)\, y, y\}$. We have $\frac{1}{|A|} \sum_A v = \frac{1}{2}$ and $\lim_{\alpha \to 1} \frac{1}{|B_\alpha|} \sum_{B_\alpha} v = \frac{2}{3}$ although B_α converges to A in the Hausdorff metric as $\alpha \to 1$[43]. Notice that $V_{v,avg}$ also reacts to an option to randomize. The independence axiom commonly imposed on menu preferences in the literature implies that A and B_α should be indifferent. However, we have $V_{v,avg}(x, A) \neq V_{v,avg}(x, B_\alpha)$ for $\alpha \neq 0.5$.

[42] Many axiomatic models of menu preferences put nonzero utility weights on very few non-chosen options on a menu, which inhibit the agent's ability or willingness to consider every option on the menu. The representation in the seminal model of temptation by Gul and Pesendorfer (2001) only depends on at most two options: the most tempting and/or the most normatively superior options. Some representations (see Dekel et al., 2009; Dekel and Lipman, 2012; Stovall, 2010) can have many influential non-chosen options, which, however, often rely on the presence of uncertain temptations, not on the agent's willingness to consider all options.

[43] Since a menu $A \in \mathbb{M}$ is Amy's information rather than a consumption space, an alternative to the Hausdorff metric can be implemented to reflect how she topologically perceives $\mathbb{M}$. Consider a distance between the centroids of two sets A, B defined as

$$d_c(A, B) = d\left(\sum_{x \in A} \frac{1}{|A|} x \,,\ \sum_{y \in B} \frac{1}{|B|} y \right)$$

which is a pseudometric. If $\mathbb{M}$ is endowed with d_c, then $V_{v,avg}$ is continuous.

In a general non-finite setting, we can assume that Amy has a probability measure μ on X such that the reference $\mathbf{r}(A)$ is the corresponding expected option conditional on $A \subseteq X$. Consider the following economic utility function of menus defined for A with $\mu(A) > 0$ as

$$V_\mu(x, A) = v(x) - \frac{\int_A v d\mu}{\mu(A)}.$$

However, the function V_μ deviates from previously discussed properties, and its technical friendliness relies heavily on the design of topology and the set of menus[44].

F. Reference-dependence and Subjective Expectations

Note that the tuple $(\succeq_1, \mathbf{r})$ characterizes Amy's taste in a deterministic setting. The function $\mathbf{r}$ can not only reflect her belief, but also nest her subjective point of view on Bob's possible choice situations. Hence, two imperative conceptual departures from the standard reference-dependence model by Kőszegi and Rabin (2006) lie in the origin of the references and how the agent perceives the menu.

Consider an agent who has both first and second-order preferences represented by u and $V_{v,\mathbf{r}}$, respectively. Given a menu A, assume that her utility U of choosing an option $x \in A$ is in an additively separable form of

$$U(x|\mathbf{r}) := u(x) + \ell\Big(V_{v,\mathbf{r}}(x, A)\Big) \tag{12}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is the universal gain-loss function of Kőszegi and Rabin (2006)'s model. That is, the agent gains consumption utility from x via u and economic utility of her choice (x, A) via $V_{v,\mathbf{r}}$ while ℓ reflects her loss-averse attitude toward her economic utility. $U(x|\mathbf{r})$ is functionally identical to their model under five conditions: (i) $X = \triangle(\mathbb{R}^N)$ for some $N \in \mathbb{N}$, (ii) the utilities u, v of sure outcomes are additively separable across dimensions, (iii) Lemma 2 holds, (iv) $\mathbf{r}$ reflects objective information (the expectation of her choice), and (v) $u = v$, a special case where the agent's first-order preference is identical to her ideal ranking and thus, whenever a choice is made, she not only enjoys consumption, but also her will to make the choice.

In Kőszegi and Rabin (2006)'s personal equilibrium, the agent endogenously forms her reference point to be equal to her expectation of the outcome. Hence, interesting behavior arise only when the true menu is "ex ante" unanticipated (*i.e.* an "out-of-equilibrium") and the agent does not "ex post" update her reference[45]. However, if her reference stems from the second-order preference, it is formed for every possible menu and thus, as long as she is able to observe her present menu, she updates her reference in the event of an unanticipated menu. The matter at hand would rather be whether or not the menu can be observed correctly.

More importantly, the reference is not necessarily her expectation and thus, even without an unan-

[44] As a model of ambiguity attitudes, Ahn (2008) presented a utility function U of sets similar to the form $U(A) = \frac{\int_A v d\mu}{\mu(A)}$. Yet, he replaced the Hausdorff continuity with what he referred to as *Lebesgue continuity* in the topology generated by the symmetric difference metric, focusing his attention to a class of menus called regular sets.

[45] Note that if $\mathbf{r}(A) = x$, then $\ell(V_{v,\mathbf{r}}(x, A))$ is always zero.

ticipated menu and loss aversion (*i.e.* when ℓ is linear), the second term of (12) can affect her behavior. If Amy's reference is her expectation of what Bob will do for each menu, then she expects a vacuous choice from any menu, implying that her second-order preference is trivial[46]. Yet, what Amy wants Bob to want to do may not be what she thinks he will do. In economics, we often overlook the subtle nuances of the word "expectation", misinterpreting it solely as an indication of likelihood. However, it can also imply one's desire or hope and thus, disappointment can arise from anticipated outcomes. In this sense, Amy's reference can be regarded as her personal wish, or *subjective expectation* of Bob's choice[47].

Consider parents whose child, a habitual video gamer prioritizing leisure over academics, continues his trend. When they tell him that they *expect* him to do homework, are they announcing their belief or preference? Some parents who are highly committed to their child's academic success tend to set the bar high, perhaps influenced by observing a neighbor's children who own even more video games yet diligently engage in their schoolwork. In turn, they might still experience profound disappointment at their child's choice to indulge in games, despite the predictability, due to the disparity in the quality of his choice subjectively compared to a few others in their interest.

G. Preference over Rankings

In this section, I show that a second-order preference $\succeq$ restricted to the act of choosing, mainly due to Axiom 1, allows for a ranking of *rankings* as well. Since $\mathbb{M}$ contains menus that are essentially lotteries over deterministic menus, some acts of choosing are a complete contingency plan that determines what will be chosen in each possible binary choice situation. Consequently, a ranking of such choices corresponds to a ranking of rankings of alternatives, which is inherently implied by the second-order preference in my model. Hence, as the title "Preference over Preferences" suggests, Theorem 1 fundamentally addresses preferences over *preference relations* (and even complete binary relations), even though it appeared to focus on preferences over simpler objects.

To illustrate, let the set of alternatives be $Z = \{x, y, z\}$ and $\succeq$ be the second-order preference repre-

[46] Note that when Y is a random variable, we have $E(Y - E(Y)) = 0$. Intuitively, if a person wants to want to do what, she believes, she wants to do, she will simply do what she wants. Assuming an out-of-equilibrium, if it turns out that her belief is wrong, then she will simply do what, she now believes, she wants.

[47] Suppose Amy is uncertain about Bob's preference and chooses his menu to discover it. Let u be Bob's utility function while Amy has some belief μ on u. Then, her second-order preference induces a menu preference characterized by $(v, \mathbf{r}, \mu)$ and represented by a form:

$$E_\mu \overline{V}(A) = E_\mu \left[\max_{x \in B(A;u)} v(x) \right] - v(\mathbf{r}(A)) \tag{13}$$

where $B(A;u)$ is the set of Bob's favorite options on a menu A. The first term reflects the expectation of Bob's choice based on μ while the second term is her subjective expectation representing what she wants him to do which is unaffected by μ.

sented by the pair $(v, \mathbf{r})$ as in Theorem 1, satisfying

$$V_{v,\mathbf{r}}(x, Z) = v(x) - v(\mathbf{r}(Z)) = 3;$$
$$V_{v,\mathbf{r}}(y, Z) = v(y) - v(\mathbf{r}(Z)) = 2;$$
$$V_{v,\mathbf{r}}(z, Z) = v(z) - v(\mathbf{r}(Z)) = 1.$$

I use x, y, z and Z to denote the degenerate lotteries with prizes x, y, z, and the menu containing them, respectively. Consider two strict preferences P and Q in $\mathbb{P}(Z)$ satisfying

$$xPyPz; \quad xQzQy.$$

By Axiom 1, when Bob faces the menu Z, Amy cares only about his favorite option, and thus P and Q are indifferent. Suppose Bob's preference is Q represented by a utility function u satisfying

$$u(x) = 1; \quad u(y) = -1; \quad u(z) = 0.$$

The tree in Figure 4 demonstrates Axiom 1, where the payoff vectors are of Amy's and Bob's utilities:

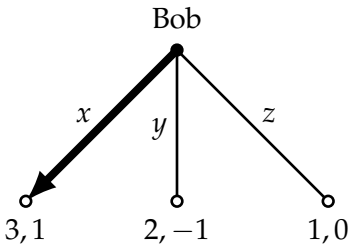

Figure 4: (x, Z) induced by both $xPyPz$ and $xQzQy$.

The bold arrow in Figure 4 represents Bob's choice (x, Z) induced by either P or Q. Amy's utilities are equally $v(x) = 3$ regardless of whether Bob's preference is P or Q because they both induce (x, Z).

Now, instead of a deterministic menu such as Z, suppose Bob's menu is either $\{x, y\}$, $\{x, z\}$, or $\{y, z\}$, each with equal probability. That is, his menu is

$$A = \frac{1}{3}\{x, y\} + \frac{1}{3}\{x, z\} + \frac{1}{3}\{y, z\}$$

which can be illustrated as the trees in Figures 5-6, where *Nature* decides with equal probability which menu he will face.

The contingency plan (indicated by the bold arrows) in Figure 5 represents the choice $(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A)$ induced by P while the one in Figure 6 represents $(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}z, A)$ induced by Q. Notice that when Bob faces the menu A, whether his preference is P or Q critically influences his contingency plan that determines his willingness to choose. While P induces the choice of y from $\{y, z\}$, Q induces z from it. From the game theory perspective, each of Bob's possible choices represents a (pure) strategy in the game tree depicted in Figure 5. Hence, from any choice from A, Amy can precisely infer which

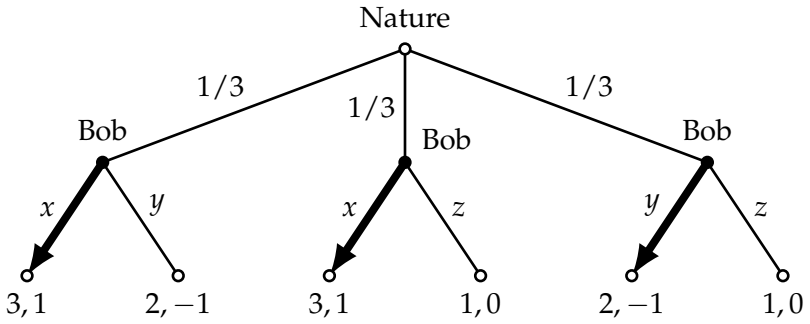

Figure 5: $\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A\right)$ induced by $xPyPz$.

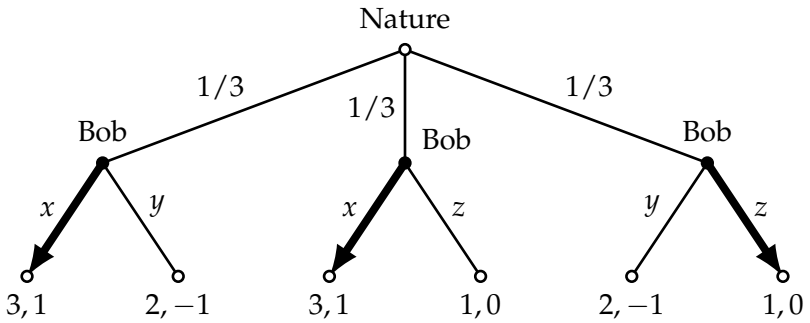

Figure 6: $\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}z, A\right)$ induced by $xQzQy$.

preference Bob has among those in $\mathbb{P}(Z)$[48].

Suppose $\succeq_0$ is Amy's preference over the rankings of Z induced by her second-order preference representation $(v, \mathbf{r})$ as in Theorem 1. It is reasonable to presume that $P \succeq_0 Q$ whenever Amy prefers the choice in Figure 5 to the one in Figure 6. To eliminate any influence of the probability of specific menus on Amy's preference over rankings, I (or Nature) assign equal probability to each menu. Intuitively, when Amy is interested in how Bob ranks the options, his plan for some menu does not particularly concern Amy more than the ones for other menus. Then, $P \succeq_0 Q$ is true whenever

$$V_{v,\mathbf{r}}\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A\right) \geq V_{v,\mathbf{r}}\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}z, A\right). \tag{14}$$

Notice that Amy's reference values $v(\mathbf{r}(\{x,y\}))$, $v(\mathbf{r}(\{x,z\}))$ and $v(\mathbf{r}(\{y,z\}))$ do not play a role in the inequality (14) since Bob's menu is fixed at A. Indeed, (14) is equivalent to

$$v(x) + v(x) + v(y) \geq v(x) + v(x) + v(z)$$

which true since we assumed $v(y) > v(z)$.

The above example implies that (i) Amy's preference over the act of choosing from the non-deterministic menu A corresponds to her preference over the rankings of Z, or even any complete binary relation on Z; and (ii) this corresponding preference is induced solely by Amy's ideal first-order preference, inde-

[48] Here, I adhere to the context I assumed along with Axiom 2 that if Bob is indifferent among two or more options, then he truthfully announces his indifference and Amy will choose the one that she thinks is the most ideal for him.

pendent of the reference function $\mathbf{r}$.

I will now formalize the idea. Let $|Z| < \infty$ denote the number of alternatives in Z. Then, $n = \binom{|Z|}{2}$ is the number of two-element subsets of Z. Let $\mathbb{D}(Z) = \{D_1, D_2, ..., D_n\}$ be the set of all two-element subsets of Z. Define the non-deterministic menu

$$\mathbb{D}_n(Z) = \frac{1}{n}D_1 + \cdots + \frac{1}{n}D_n$$

which, essentially, is the trees in Figures 5-6 if $n = 3$. The following set includes all acts of choosing, which are contingency plans specifying the option to be chosen from each possible menu in $\mathbb{D}(Z)$:

$$\overline{\mathbb{D}} = \left\{ \left(\frac{1}{n}x + \cdots + \frac{1}{n}z_n, \mathbb{D}_n(Z) \right) : z_i \in D_i \ \forall i \in \{1, ..., n\} \right\}.$$

When Amy is interested in Bob's ranking of options, her second-order preference is restricted to $\overline{\mathbb{D}} \subset \mathbb{C}$. Let $\mathbb{B}(Z)$ be the set of all complete binary relations on Z. Given any $P \in \mathbb{B}(Z)$ and a preference $\succeq$ over the act of choosing on $\mathbb{C}$, define the choice correspondence $\mathcal{C}^* : Z^2 \times \mathbb{B}(Z) \times \mathbb{P}(\mathbb{C}) \to 2^Z$ by

$$\mathcal{C}^*(D; P, \succeq) := \{x \in \mathcal{C}_P(D) : (x, D) \succeq (y, D) \ \forall y \in \mathcal{C}_P(D)\}.$$

The correspondence $\mathcal{C}^*$ essentially breaks the indifference induced by some $P \in \mathbb{B}(Z)$. For example, if x and y are indifferent according to P, then $\mathcal{C}_P(\{x, y\}) = \{x, y\}$; and if $(x, \{x, y\})$ is strictly preferred to $(y, \{x, y\})$ according to $\succeq$, then we have $\mathcal{C}^*(\{x, y\}; P, \succeq) = \{x\}$. This is consistent with the context that if Bob is indifferent among some options, then he will choose the one that aligns with Amy's ideal preference.

I define Amy's preference $\succeq_0$ over $\mathbb{B}(Z)$, who has a second-order preference representation as in Theorem 1, as follows:

Definition 3 (Induced Preference over Rankings)**.** *A preference relation $\succeq_0$ on $\mathbb{B}(Z)$ is a preference over rankings of Z induced by a second-order preference $\succeq$ represented as in Theorem 1, if for all $P, Q \in \mathbb{B}(Z)$, $z_i \in \mathcal{C}^*(D_i; P, \succeq)$ and $z_i' \in \mathcal{C}^*(D_i; Q, \succeq)$ for each $i \in \{1, ..., n\}$,*

$$P \succeq_0 Q \iff \left(\frac{1}{n}x + \cdots + \frac{1}{n}z_n, \mathbb{D}_n(Z) \right) \succeq \left(\frac{1}{n}x' + \cdots + \frac{1}{n}z_n', \mathbb{D}_n(Z) \right).$$

The following result holds:

Theorem 5. *Suppose $\succeq_0$ on $\mathbb{B}(Z)$ is a preference over rankings of Z induced by a second-order preference $\succeq$ represented as in Theorem 1. Then, $\succeq_0$ is represented by $V_0 : \mathbb{B}(Z) \to \mathbb{R}$ of the form:*

$$V_0(P) := \sum_{D \in \mathbb{D}(Z)} \left(\max_{x \in \mathcal{C}_P(D)} v(x) \right).$$

Proof. Notice that $\mathcal{C}^*(D; P, \succeq) = \arg\max_{x \in \mathcal{C}_P(D)} v(x)$ for any $P \in \mathbb{B}(Z)$ and $D \in \mathbb{D}(Z)$. Suppose $P \succeq_0 Q$. Choose any

$$z_i \in \arg\max_{x \in \mathcal{C}_P(D_i)} v(x) \ \text{ and } \ z_i' \in \arg\max_{x \in \mathcal{C}_Q(D_i)} v(x)$$

for each $i \in \{1, ..., n\}$. By construction, we have

$$\left(\frac{1}{n}x + \cdots + \frac{1}{n}z_n, \mathbb{D}_n(Z)\right) \succeq \left(\frac{1}{n}x' + \cdots + \frac{1}{n}z'_n, \mathbb{D}_n(Z)\right)$$

which is equivalent to

$$\sum_{i=1}^{n} v(z_i) \geq \sum_{i=1}^{n} v(z'_i).$$

Hence, we have

$$\sum_{D \in \mathbb{D}(Z)} \left(\max_{x \in \mathcal{C}_P(D)} v(x)\right) \geq \sum_{D \in \mathbb{D}(Z)} \left(\max_{x \in \mathcal{C}_{P'}(D)} v(x)\right).$$

□

H. Prior Literature on Second-order Preference

This paper also contributes to the prolonged philosophical studies on the relationship among higher-order preferences and self-control. Frankfurt (1971) first introduced the concept of "second-order desires". In his account, "first-order desires" are desires directed toward actions or states of affairs in the world (e.g., I want to eat a piece of cake), while second-order desire are desires to have certain first-order desires (e.g., I want to want to eat vegetables). In my paper, I use the phrase "preferring a preference" to mean preferring to *behave* as if one holds that preference, thereby essentially differentiating from second-order *desires* which pertains to one's state of mind (e.g., a killer might desire not to have the desire to kill even after he decided not to kill).

This distinction between one's inner desires and the motives that lead to actions is also acknowledged by Frankfurt (1971) who discussed the case where a person desires certain desires without ever wanting them to lead to action. Watson (1975) similarly noted that the strength of one's desires does not solely determine their impact on action. Jeffrey (1974) provided the first formal illustration of a heavy smoker who prefers smoking to abstaining but prefers "preferring abstaining to smoking" to "preferring smoking to abstaining" (see also McPherson, 1982; Carballo, 2018; González de Prado, 2020). Yet, it is hard to find a narrative in the existing literature where second-order preferences, formally defined *and* distinguished from second-order desires, are integrated into a microeconomic framework. This paper is the first to make a contribution in this regard.

Moreover, while preferences over one's own preferences dominated the philosophical discussion, my model captures the distinctive characteristics of second-order preferences in the absence of first-order preferences. Thus, it can also reflect a preference over others' preferences in general. Social relationships such as romantic partners, trainer-trainee, parent-child, judge-defendant and voter-politician can to some extent be subsumed under the Amy-Bob paradigm.

Economic theories have persistently adhered to the use of first-order preferences—binary relations defined on practically any set that is not composed of preferences themselves. Sen (1977) and Hirschman (1984) characterized a second-order preference by a preference over "a sense of morality". However, it was technically a first-order preference over real numbers such that higher numbers were assumed to

indicate greater moral outcomes. Bolle (1983)'s utility function portrayed a state-dependent moral ranking while Dowell et al. (1998) presented morality-dependent budget constraints, assuming that moral actions might have a negative impact on one's wealth. Notably, this paper is the first to formally capture the menu-dependent nature of second-order preferences. My model shows that when the decision-maker's menu is fixed, a second-order preference under Axioms 1-2 is behaviorally indistinguishable from a first-order preference. Notice that the representation in Theorem 1 is entirely captured by the function v of lotteries since the term $v(\mathbf{r}(A))$ is constant unless the menu A is subject to change. In other words, if the choice situation remains unchanged, second-order preferences are practically absent.

This revelation underscores significant limitations in previous studies, which often concentrated on a binary choice problem (e.g., the heavy smoker of Jeffrey (1974) chooses only from $\{\text{smoke, abstain}\}$). Subsequent studies encountered skepticism regarding the validity of investigating second-order preferences. Hirschman (1984) discussed practical challenges in observing the existence of second-order preferences through individual choices, and Bruckner (2011) posited that second-order preferences should be integrated into the analysis of first-order preferences. Philosopher Mele (1992) argued that second-order desires are not necessarily present when deciding between a continent and an incontinent action. He illustrated that a person might resist the desire to eat a piece of cake not because of a higher-order desire to lose weight, but as a compromise between two conflicting first-order desires. My model suggests that such skepticism arises from the narrow focus on outcome rankings while the distinctive nature of second-order preferences lies in individuals' subjective perceptions of different choice situations.

H.1. Generalization of Halldén's Axiom

I present the philosophical rationale behind Lemma 1. In particular, Lemma 1 is a generalized version of the axiom of second-order preference originally introduced in the book *The foundations of decision logic* by the philosopher Halldén (1980) who proposed that the value of discriminating between two options is determined by the extent to which these options differ in value. Specifically, the utility of "preferring x to y" is the difference between the utility of consuming x and that of consuming y. Similarly, the rationale behind my representation is that the utility of "preferring x to all else in A" is the difference in utility between x and $\mathbf{r}(A)$. While this idea seems to assume "cardinal utilities" of x and $\mathbf{r}(A)$, Ramsey (1926) showed that the difference in expected (ordinal) utilities still possesses ordinal information. To see this, suppose an expected utility maximizer prefers a coin toss between x and $\mathbf{r}(B)$ to the one between y and $\mathbf{r}(A)$. Then, for any expected utility function representing the ranking of the two coin tosses, the difference in utility between x and $\mathbf{r}(A)$ is larger than that between y and $\mathbf{r}(B)$. This allowed Halldén (1980) to equate the ordinal ranking of two coin tosses to that of preferences.

Halldén (1980)'s axiom can be translated formally as follows:

Halldén's Axiom (1980). *Let $\succeq_H$ be a preference over X represented by an affine function. If $x_1 \succeq_H y_1$ and $x_2 \succeq_H y_2$ for some $x_1, x_2, y_1, y_2 \in X$, then "preferring x_1 to y_1" is preferred to "preferring x_2 to y_2" if and only if $\frac{1}{2}x_1 + \frac{1}{2}y_2 \succeq_H \frac{1}{2}x_2 + \frac{1}{2}y_1$.*

He regarded a preference over one's own preferences as a preference over one's abilities to distinguish each option from another. This is motivated by his thought experiment. Consider a cup of water (x_1),

gasoline (y_1), orange juice (x_2), and grape juice (y_2). Suppose an agent has a preference $\succeq_H$ that satisfies

$$x_2 \succ_H y_2 \succ_H x_1 \succ_H y_1.$$

He is about to have severe brain surgery after which he will inevitably lose the ability to distinguish either x_1 from y_1 or x_2 from y_2. That is, his preference will no longer satisfy either $x_1 \succ_H y_1$ or $x_2 \succ_H y_2$. The surgeon asks which ranking he would prefer to maintain after the surgery. The agent would obviously choose to keep preferring x_1 to y_1 because he would not want to risk being a person who is indifferent between the taste of gasoline and that of water, while being a little picky about types of juice is not a vital part of his life.

Halldén (1980) proposed the consistent rule that "preferring x_1 to y_1" is preferred to "preferring x_2 to y_2" if and only if the value difference between x_1 and y_1 is larger than that between x_2 and y_2. In other words, x_1 is *more* preferred to y_1 than x_2 is preferred to y_2, and thus, maintaining the ranking $x_1 \succ_H y_1$ is more valuable than keeping $x_2 \succ_H y_2$. If v_H is a cardinal utility function representing $\succeq_H$, then this would imply that the utility difference between x_1, y_1 is larger than the that of the other pair. That is,

$$v_H(x_1) - v_H(y_1) > v_H(x_2) - v_H(y_2). \tag{15}$$

Halldén (1980) used the fact that without imposing any cardinal property on utility functions, (15) is equivalent to stating that an expected utility maximizer prefers a coin toss $\frac{1}{2}x_1 + \frac{1}{2}y_2$ to $\frac{1}{2}x_2 + \frac{1}{2}y_1$[49].

I identify three key limitations of Halldén's axiom. First, the decision-maker's menus are restricted to contain exactly two options[50]. Axioms 1-2 allow for any larger menus. Second, there is no notion of ideal preferences in Halldén's axiom. Notice that in terms of my model, the "if and only if" condition in his axiom can be rewritten as

$$(x_1, \{x_1, y_1\}) \succeq (x_2, \{x_2, y_2\}) \iff \tfrac{1}{2}x_1 + \tfrac{1}{2}y_2 \succeq_H \tfrac{1}{2}x_2 + \tfrac{1}{2}y_1.$$

This holds only when $x_1 \succeq_H y_1$ and $x_2 \succeq_H y_2$, which means in Halldén's axiom, the second-order preference $\succeq$ is defined on the agent's first-order preference—the set $\succeq_H \subseteq X \times X$ itself. Defining $\succeq$ on $\succeq_H$ implies that the agent's preference is already ideal because he does not consider the value of acting against his own preference. In contrast, I identify the first-order preference that is desired by the agent (Amy) but not necessarily his (Bob's). By separating the owners of first- and second-order preferences, I establish the behavioral dichotomy between first- and second-order preferences and capture the distinctive characteristics of the latter—Axiom 8. Furthermore, the presence of the ideal preference also allows a person to have a preference over others' preferences as well as characterizing a conflict between one's own ideal and non-ideal desires.

[49] Halldén (1980) referred to Ramsey (1926) who first showed that if the utility function is affine, comparing utility differences between two pairs of options is equal to comparing the two coin tosses as shown. Sahlin (1981) conducted an empirical study that supported the theoretical link between the second-order preference and the comparison of two coin tosses.

[50] If $x_1 = y_1$, then his axiom allows singleton menus. However, Halldén (1980) neither provided any implication for this case nor the notion of vacuous choices.

The third limitation relates to a detail part of his thought experiment. Suppose the agent chooses to maintain the ranking $x_1 \succ_H y_1$. It implies that he will no longer be able to discriminate between x_2 and y_2. Since indifference between x_2 and y_2 can lead to any choice within $conv(\{x_2, y_2\})$, it remains ambiguous which of x_2 or y_2 the agent would expect to choose when presented with the menu $\{x_2, y_2\}$. The reference function $\mathbf{r}$ in my model directly addresses this ambiguity. Based on (15), we can see that Halldén (1980) implicitly assumed a purely libertarian preference over the act of choosing—Axiom 10. Consequently, the agent in Halldén (1980)'s thought experiment preferred "preferring water (x_1) to gasoline (y_1)" to "preferring orange juice (x_2) to grape juice (y_2)". However, a person with the locally pure paternalistic preference over the act of choosing as in Theorem 3 would prefer the otherwise. Given that the expected preference will also put water over gasoline, the DM will become purely paternalistic and would regard the act of choosing water over gasoline as a vacuous choice.

H.2. Beyond EU Theory

The seminal identification of the relationship between 'value distance' and 'the ranking of two coin tosses' illustrated in Lemma 1 was initially made by Ramsey (1926), affirming its role as a foundational aspect of Expected Utility (EU) theory. It is noteworthy that the concept of second-order preference is derived from Halldén (1980)'s interpretation of this relationship. He posited that second-order preference quantifies the extent to which one option is preferred over another, more than a third is over a fourth, through a systematic ranking of value distances. He posed the critical inquiry: "How can we meaningfully measure those distances?" Given that a utility function, u, essentially represents a ranking rather than a quantifiable level of satisfaction, the difference $u(x) - u(y)$ ostensibly lacks inherent significance. Halldén (1980)'s resolution was predicated on the validity of the EU theory, suggesting that these differences acquire significance within its framework. This rationale underscores the adherence of my model to the EU theory principles.

Looking ahead, my research will explore modifications to the axioms of second-order preference to encompass theories beyond the EU theory. This exploration will address a pivotal question: "How can meaningful value distances be quantified within non-EU theoretical frameworks?" For instance, given two Anscombe-Aumann acts f, g, and a function $M(f)$ representing the maximin EU function, the difference $M(f) - M(g)$ diverges from its interpretation under the traditional EU functions, thereby breaking the linkage to coin toss rankings previously established in Lemma 1. Addressing these questions will broaden the understanding and the scope of value distance measurements in decision theory.

I. Preference over Indifference

As I mentioned in Section 2, the model of second-order preference varies widely depending on how we define what "the action induced by P_A given the menu A" refers to. By relaxing Axiom 2, I can allow Amy to regard "declaring indifference" as a valid action.

I present two examples to demonstrate that Amy can particularly favor or disfavor Bob's indiffer-

ence. Let $A = \{x, y\}$. Then, Bob's possible strict preferences are $P_1, P_2, P_3 \in \mathbb{P}(A)$ such that

$$xP_1y, \quad yP_2x, \quad \neg(xP_3y) \quad \text{and} \quad \neg(yP_3x).$$

First, Amy might strictly prefer P_1 and P_2 to P_3. That is, she disfavors being indifferent between x and y. Suppose she is a wine expert and Bob is her student. x is a bottle of red wine from Chile and y is from Italy. As a beginner, Bob's preference is P_3 who is not yet trained to feel the subtle difference in tastes between x and y. All Amy wants to accomplish as a teacher is to see Bob refining his own tastes for wine so that he strictly prefers either one of the two bottles. In this case, $\succeq$ would satisfy $P_1 \sim P_2 \succ P_3$[51].

An example of positive values added to indifference can be found in people who try not to discriminate against certain aspects of others or objects. Suppose Amy is a parent with two children x and y. She recently won two traveling tickets to Paris and plans to take one of her children for the summer. She personally prefers taking her firstborn x who was always her favorite. However, if she chooses x, she knows she will suffer from overwhelming guilt and shame as a parent for discriminating among her children and reinforcing y's prolonged belief that he is always her second choice[52]. Hence, she might start thinking that a parent should ideally be indifferent between taking x and y. Her second-order preference would satisfy $P_3 \succ P_1 \sim P_2$.

To technically approach the two examples above, notice that once Axiom 2 is relaxed, Amy can no longer rank choices in $\mathbb{C}$. Instead, she also considers Bob's possible preferences that induce more than one choice. Then, $\succeq$ needs to be defined on the set

$$\overline{\mathbb{C}} := \bigcup_{A \in \mathbb{M}} \left\{ (\mathcal{C}_P(A), A) \in \mathbb{M}^2 : P \in \mathbb{P}(A) \right\} = \{(\mathbf{x}, A) : \mathbf{x} \subseteq A \subseteq X\}$$

where the pair $(\mathbf{x}, A)$ is referred to as *the act of choosing a set* $\mathbf{x}$ *of favorite options among* A. This construction implies that I do not specify how Bob maps his indifference into consumption. Suppose his announcement is $(\{x, y\}, A)$. While many theories would require him to choose an option from $\{\alpha x + (1-\alpha)\, y \ : \ \alpha \in [0,1]\}$, I believe it distorts the essence of indifference. Although the choice $(\{x, y\}, A)$ does not characterize Bob's final consumption, it allows Amy to clearly process his indifference as it is—the set of options that he is willing to consume.

References

Ahn, D. S. (2008). Ambiguity Without a State Space. *The Review of Economic Studies*, 75 (1): 3–28.

Ambuehl, S., Bernheim, B. D., and Ockenfels, A. (2021). What Motivates Paternalism? An Experimental Study†. *American Economic Review*, 111 (3).

[51] I thank Kevin Zollman for inspiring this example.

[52] The example resembles the "Machina's mom" story in Machina (1989) who characterized the mother's preference using a non-expected utility theory.

Bartling, B., Cappelen, A. W., Hermes, H., Skivenes, M., and Tungodden, B. (2023). Free to Fail? Paternalistic Preferences in the United States. *SSRN Electronic Journal*.

Bartling, B., Fehr, E., and Herz, H. (2014). The Intrinsic Value of Decision Rights. *Econometrica*, 82 (6): 2005–2039.

Bénabou, R. and Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, 112 (4): 848–886.

Bernheim, B. D., Kim, K., and Taubinsky, D. (2024). Welfare and the Act of Choosing. *National Bureau of Economic Research Working Paper Series*, No. 32200.

Bolle, F. (1983). On Sen's Second-Order Preferences, Morals, and Decision Theory. *Erkenntnis (1975-)*, 20 (2): 195–205.

Bruckner, D. W. (2011). Second-Order Preferences and Instrumental Rationality. *Acta Analytica*, 26 (4): 367–385.

Carballo, A. (2018). Rationality & Second-Order Preferences. *Noûs*, 52 (1): 196–215.

Dana, J., Cain, D. M., and Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100 (2): 193–201.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*.

Dekel, E. and Lipman, B. L. (2012). Costly Self-Control and Random Self-Indulgence. *Econometrica*, 80 (3): 1271–1302.

Dekel, E., Lipman, B. L., and Rustichini, A. (2001). Representing Preferences with a Unique Subjective State Space. *Econometrica*, 69 (4): 891–934.

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

Dekel, E., Lipman, B. L., Rustichini, A., and Sarver, T. (2007). Representing preferences with a unique subjective state space: A corrigendum. *Econometrica*, 75 (2): 591–600.

Dhar, R. and Wertenbroch, K. (2012). Self-signaling and the costs and benefits of temptation in consumer choice. *Journal of Marketing Research*, 49 (1): 15–25.

Dillenberger, D. and Sadowski, P. (2012). Ashamed to be selfish. *Theoretical Economics*, 7 (1): 99–124.

Dowell, R. S., Goldfarb, R. S., and Griffith, W. B. (1998). Economic Man As A Moral Individual. *Economic Inquiry*, 36 (4): 645–653.

Dunning, D. (2007). Self-Image Motives and Consumer Behavior: How Sacrosanct Self-Beliefs Sway Preferences in the Marketplace. *Journal of Consumer Psychology*, 17 (4): 237–249.

Emerson, W. R. and Greenleaf, F. P. (1969). Asymptotic Behavior of Products $C^p = C + \cdots + C$ in Locally Compact Abelian Groups. *Transactions of the American Mathematical Society*, 145: 171–204.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68 (1): 5–20.

Gilchrist, J. D., Sabiston, C. M., and Kowalski, K. C. (2019). Associations between actual and ideal self-perceptions and anticipated pride among young adults. *Journal of Theoretical Social Psychology*, 3 (2): 127–134.

González de Prado, J. (2020). Akrasia and the Desire to Become Someone Else: Venturinha on Moral Matters. *Philosophia*, 48 (5): 1705–1711.

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

Halldén, S. (1980). *The foundations of decision logic*. (Library of Theoria, 14.) Lund: CWK Gleerup.

Herstein, I. N. and Milnor, J. (1953). An Axiomatic Approach to Measurable Utility. *Econometrica*, 21 (2): 291.

Higgins, E. T. (1987). Self-Discrepancy: A Theory Relating Self and Affect. *Psychological Review*, 94 (3).

Hirschman, A. O. (1984). Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse. *Bulletin of the American Academy of Arts and Sciences*, 37 (8): 11.

Jeffrey, R. C. (1974). Preference Among Preferences. *The Journal of Philosophy*, 71 (13): 377.

Kopylov, I. (2012). Perfectionism and Choice. *Econometrica*, 80 (5): 1819–1843.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences*. *The Quarterly Journal of Economics*, 121 (4): 1133–1165.

Kőszegi, B. and Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92 (8-9): 1821–1832.

Laffond, G., Lainé, J., and Sanver, M. R. (2020). Metrizable preferences over preferences. *Social Choice and Welfare*, 55 (1): 177–191.

Machina, M. J. (1989). Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty. *Journal of Economic Literature*, 27 (4): 1622–1668.

Markus, H. and Nurius, P. (1986). Possible Selves. *American Psychologist*, 41 (9).

McPherson, M. S. (1982). Mill's Moral Theory and the Problem of Preference Change. *Ethics*, 92 (2): 252–273.

Mele, A. R. (1992). Akrasia, Self-Control, and Second-Order Desires. *Noûs*, 26 (3): 281.

Mill, J. S. (1859). *On Liberty*, volume 55. Broadview Press.

Noor, J. (2011). Temptation and Revealed Preference. *Econometrica*, 79 (2): 601–644.

Noor, J. and Ren, L. (2023). Temptation and guilt. *Games and Economic Behavior*, 140: 272–295.

Olszewski, W. (2007). Preferences Over Sets of Lotteries. *The Review of Economic Studies*, 74 (2): 567–595.

Prelec, D. and Bodner, R. (2003). Self-signaling and self-control. In *Time and decision: Economic and psychological perspectives on intertemporal choice.*, pages 277–298. Russell Sage Foundation, New York, NY, US.

Ramsey, F. P. (1926). Truth and Probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, number 7, pages 156–198. McMaster University Archive for the History of Economic Thought.

Sahlin, N. 1981). Preference among preferences as a method for obtaining a higher-ordered metric scale. *British Journal of Mathematical and Statistical Psychology*, 34 (1): 62–75.

Saito, K. (2015). Impure altruism and impure selfishness. *Journal of Economic Theory*, 158: 336–370.

Samuelson, P. A. (1952). Probability, Utility, and the Independence Axiom. *Econometrica*, 20 (4): 670.

Sen, A. K. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, 6 (4): 317–344.

Starr, R. M. (1969). Quasi-Equilibria in Markets with Non-Convex Preferences. *Econometrica*, 37 (1): 25–38.

Stovall, J. E. (2010). Multiple Temptations. *Econometrica*, 78 (1): 349–376.

Tracy, J. L. and Robins, R. W. (2004). TARGET ARTICLE: "Putting the Self Into Self-Conscious Emotions: A Theoretical Model". *Psychological Inquiry*, 15 (2): 103–125.

Watson, G. (1975). Free Agency. *The Journal of Philosophy*, 72 (8): 205–220.