# A Representation of Preference over Preferences and the Act of Choosing

Jun Hyun Ji*

December 24, 2024

**Abstract**

This paper studies how individuals value the act of choosing itself by using the novel concept "preference over preferences" while assuming no preference over outcomes. An agent 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 that of 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 select a *reference option* against which the act of choosing from each menu is assessed. This choice rule captures the individual's menu-dependent paternalistic attitude toward herself (manifested as guilt, pride, the joy of freedom, or the fear of *the act* of making mistakes) which implicitly induces preferences over menus. I discuss connections to prior models that exploit choices over menus, and welfare implications.

*Ph.D. Student, Economics, University of Pittsburgh. Email: juj25@pitt.edu. 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 emotional sensations that the agent immediately experiences when choosing the menu itself (or any higher-level meta-choices). This challenge has profound welfare implications, as policies based solely on observed choices may fail to enhance, or may even diminish, individual well-being.

These prior studies, whether empirical or theoretical, infer preferences over the act of choosing indirectly. Because agents also care about outcomes, both the observed choices and emotional states are the agent's compromises between his preferences for outcomes and preferences over the act of making choices. 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)[1]. I use the phrase "preferring a preference" to mean preferring to *behave* as if one holds that preference[2]. To illustrate, suppose an agent strictly prefers $x$ to $y$. Our standard understanding

[1] 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. It is also possible to simply enjoy fulfilling one's preferences in addition to enjoying the outcome. 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.

[2] To avoid confusion, I distinguish second-order preferences from second-order *desires*. The former pertains to one's observable behavior (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. 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.

is that he will give up $y$ whenever $x$ is available. Now, suppose he prefers "preferring $x$ to $y$" to "preferring $y$ to $x$", 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 interpret a preference over the act of choosing as a second-order preference, and identify its behavioral characteristics by assuming that the agent has no preferences over outcomes (henceforth, 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)[3]. Next, I introduce the key axiom that yields a unique representation. Third, I explore two extreme attitudes toward the act of choosing and investigate their connections to preferences for commitment or freedom of choice. Fourth, I introduce non-extreme menu-dependent attitudes, and discuss menus that enhance the value of the act of choosing even when detached from outcome-driven preferences. 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 $A$"[4]. 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[5]. 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 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.

My first result is a general functional form of the representation (Theorems 1-2). In addition to the standard axioms of expected utility theory, my key axiom yields a unique representation of the form

$$V(x, A) = v(x) - v(r(A))$$

[3] By focusing on the act of choosing, I also rule out the possibility that the agent has a preference over indifference. For example, one might like (or dislike) to be indifferent among some options. In Appendix I, I show that modeling "the act of being indifferent" requires more careful considerations.

[4] 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.

[5] For instance, throwing away a \$10 bill is generally foolish but becomes a very reasonable choice if the only other alternative is to throw away \$100. And choosing $-\$10$ over $-\$100$ might be just as reasonable as choosing \$90 over \$0 even though the outcomes differ.

where $v$ is an von Neumann-Morgenstern (vNM) utility function over lotteries and $\mathbf{r}$ is a choice function that chooses a lottery $r(A)$ from the menu $A$ that serves as a reference against which the act of choosing $x$ from $A$ is assessed[6]. The function $v$ represents the preference that the agent believes he (or someone) should ideally adopt (e.g., ideally, an alcoholic should prefer coffee over beer). According to this ideal ranking, $r(A)$ is either the best or worst option (or even something in between) in $A$[7]. As I explain in more detail in subsequent sections, the uniquely identified $r$ is a function of sets with minimal linear properties such that the function $V$ subsumes most menu-dependent components of menu preference representations in the literature.

Next, I investigate how the choice function $\mathbf{r}$ captures the agent's menu-dependent preferences for freedom of choice. As the first step, I introduce two extreme attitudes toward preferences: paternalism and libertarianism[8]. First, 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 being unable to exhibit any preference at all (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 future freedom of choice, protecting himself from the act of making a mistake[9]. Roughly speaking, a sense of pride is not possible—only guilt is. The opposite holds for the *pure libertarianism*: the freedom of choice is valued above avoiding mistakes, hence feeling guilt is impossible[10]. As a result, a purely paternalistic (libertarian) preference over the act of choosing implies that the reference of each menu is the most (least) ideal option in the menu.

As my main result, Theorem 3 captures an agent who is *locally* purely paternalistic. A sense of pride is highly menu-dependent. 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 as opposed to how they are *expected* to behave. Hence, the agent's paternalistic attitude might weaken when menus present a strong conflict

[6] 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.

[7] The menus are either convex or finite, which does not affect the uniqueness of the choice function $\mathbf{r}$. In Appendix C, I provide the theorems for the finite case.

[8] 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.

[9] 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).

[10] Experiments involving social preferences provide evidence that demonstrates preferences for autonomy (see Bartling et al., 2014).

between his ideal preference and expected preference. A purely libertarian alcoholic—who expects himself to behave as an addict—might prefer the act of choosing coffee over beer because it reflects strong self-control or favorable self-evaluation[11]. Now, a less preferable act of choosing can be made by adding a very favorable option to his menu: for example, spending time with his niece whom he deeply loves more than anything. In this case, he becomes purely paternalistic because choosing his niece over the two types of beverages is neither in conflict with his expectation nor contrary to his ideals. It is an obvious choice he would make, and thus would not evoke a higher sense of pride, despite being a better outcome overall. I identify a condition that generates these behavior, and pins down the agent's unique expected preference. In Section 4.3, I apply my model to a social setting—specifically, a dictator who exhibits locally pure paternalism in a dictator game.

I discuss the connections between my model and prior models of menu preferences by presenting an optimal menu choice problem of an agent who has additively separable first- and second-order preferences. 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)[12].

I also provide welfare implications. The concept of higher-order preferences suggests that the design of welfare policies can be influenced by the social planner's own judgments and values. For example, even when a parent knows precisely what her child will choose and how he will feel—such as experiencing guilt for not completing a chore—the parent's own values determine whether to allow that negative sensations (e.g., to promote the child's personal growth and future welfare). Consequently, having extensive data on the decision-maker's first- and second-order preferences may not resolve the inherent complexities in welfare assessments.

In the next section, I present the model. Section 3 provides the general representation. Section 4 offers the representations for paternalism and libertarianism. In Section 5, I discuss prior menu preference literature and welfare implications. Proofs (if omitted) are collected in Appendices A-D. In Appendix H, I discuss my contributions to the literature specifically on second-order preference.

[11] Frankfurt (1971) noted that second-order desires are a manifestation of the capacity for reflective self-evaluation. Then, we can infer that second-order preferences (over one's own or others' preferences) also arise when one evaluates the preference of the decision-maker from the perspective of a third person, as an objective observer.

[12] 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).

## 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 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[13]. 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 $\alpha$ and $y$ with probability $1-\alpha$.

**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*[14]. 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 $A$. Yet, in the Appendix, I also consider the case where Bob can only face finite menus[15]. My explanations and examples will often feature finite menus because they simplify the narrative

[13] This approach is inspired by philosophical discussions of higher-order volitions and the 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.

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

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

[15] In Appendix C, I show that my results hold for finite menus with a moderate modification of my axioms. In particular, the representation is not affected by whether Bob faces convex or finite menus.

and aid in understanding the intuitions. One technical issue 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). 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 $A \subseteq X$. 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

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

where $\neg(yPx)$ means not $yPx$[16]. 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[17].

[16] 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 set of options ($X$) is not finite. See Laffond et al. (2020) for more detail on "metrizability" of the set $\mathbb{P}(X)$.

[17] 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 in the menu, or revealing one's preference over the menu entirely. Yet, in many cases, we only regard

**The act of choosing.** The primitive of my model is a preference $\succeq$ 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*. In Section 2.2, I show that my axioms restrict Amy's second-order preference to $\mathbb{C}$ instead of $\mathcal{P}$. The precise 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 $B$"*. Henceforth, $(x, A)$ naturally implies $x \in A$. A special notation will be used to indicate vacuous choices—any choices made from singleton menus: let $\phi$ denote a vacuous choice, i.e., $\phi \in \{(x, \{x\}) : x \in X\} \subset \mathbb{C}$. 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 $\lambda$ 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 $(\lambda x + (1 - \lambda) y, \lambda A + (1 - \lambda)B)$.

## 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 that Amy does care about either what Bob could have chosen if he had been presented with a different menu, or the non-chosen options that he was willing to choose. For example, suppose on two separate occasions (e.g., period 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 period 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. If Amy's second-order preference is restricted to the act of choosing, we must have $c_1 \sim c_2$. 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. Again, we must have $c_1 \sim c_2$.

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

**Axiom 1** (Preference for Revealed Preference). *Given $A \in \mathbb{M}$ and $P_1, P_2 \in \mathbb{P}(A)$, $\mathcal{C}_{P_1}(A) = \mathcal{C}_{P_2}(A)$ implies $P_1 \sim P_2$.*

**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, $P_1 \succeq P_2$ implies $P_1 \sim P_3 \succeq P_2$.*

one's favorite as the vital part of his preference. In a presidential election, we do not count a voter's non-favorite candidates. Some voting methods (e.g., Condorcet method) do look at a whole ranking of alternatives.

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 in 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 $P$.

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[18]. 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 $\mathcal{C}_{P_3}(A) = \{x, y\}$[19]. 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 $y$[20]. 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 $\mathcal{C}_P(A) = \{x\}$[21]. Formally, let $\mathcal{P}_s = \bigcup_{A \in \mathbb{M}} \mathbb{P}_s(A)$ be the set of all preferences across all menus

[18] 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. In Appendix I, I discuss preferences for (or against) indifference by relaxing Axiom 2.

[19] Notice that since $A \in \mathbb{M}$ 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 $\mathcal{C}_{P_3}(A)$.

[20] 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 contingent plan $z = \alpha x + (1 - \alpha) y$ for some $\alpha \in (0, 1)$. Yet, if Amy has no preference for indifference, she is only concerned with the behavioral outcome of his plan: either $(x, A)$ or $(y, A)$.

[21] 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.

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 (which can still be described by using $\succeq$ 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 *rankings* as well if $\succeq$ is further restricted to complete contingency plans for each possible binary choice situation.

## 2.3. Standard Axioms

I employ the standard von Neumann-Morgenstern axioms of continuity and independence used in prior literature, and impose the following axioms[22]:

**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[23]. Axiom 6 states that when Bob's menu is $X$—the entire set of 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*. If $\succeq_1$ has a representation $v : X \to \mathbb{R}$, then $v(x) \geq v(y)$ implies that she wants Bob to prefer $x$ to $y$, or equivalently, I say $x$ is ideally preferred to $y$. To explain the

[22] 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.

[23] 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 menus (see Gul and Pesendorfer, 2001; Dekel et al., 2001, 2007)

intuition behind Axiom 6, if Bob's menu is his entire consumption space (i.e., $X$), there is a first-order preference $\succeq_1$ over $X$ that Amy believes Bob should ideally have[24]. 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[25].

### 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 essence of second-order preference[26]. It states that given any two 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. The following ranking directly violates Axiom 8:

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

(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 $B$".

To 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 is considering two potential business partners and evaluating them based on their 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

[24] 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. Also, the fact that $\succeq_1$ is a preference relation implies that I do not consider the case where Amy particularly wants Bob to want to behave irrationally. This does not imply that she particularly favors rational behavior: there is no value added to Bob's rationality itself.

[25] 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: $(x, A) \succeq_2' (y, B)$ if and only if $x \succeq_1' y$ for all $A, B$ containing $x, y$, 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. Axioms 6-7 together imply $(x, A) \succeq (y, A)$ if and only if $x \succeq_1 y$ for all $A$ containing $x, y$, in which case, $(x, A) \succeq (y, B)$ is not a guarantee.

[26] In Appendix C, I present a modified version of Axiom 8 to address preferences restricted to finite menus.

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[27]:

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

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!" fixating on the menu $\{x, y\}$ 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,

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

# 3. Representation

I use the following definitions. 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 $x \succeq_1 y$. $v$ 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

[27] A choice function is any well-defined function $f : \mathbb{M} \to X$ that satisfies $f(A) \in A$.

$\succeq$ if $V(x, A) \geq V(y, B)$ is equivalent to $(x, A) \succeq (y, B)$. A 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 $\lambda \in [0, 1]$.

**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* $\lambda \in [0, 1]$.

Henceforth, the function $\mathbf{r}$ is also referred to as *the reference function*. 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 (reference) 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. □

In Appendix C, I also provide the same theorem when $\succeq$ is restricted to finite menus. 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, 4 and 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 $A$. Let $\mathbf{r}_A := \mathbf{r}(A)$ for brevity. The second step is Lemma 1 in the Appendix which shows that due to Axioms 4-5, the functions $v$ and $V$ have the following relationship[28]:

**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 $\succeq_1$, we have

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

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:

[28] In the Appendix, 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).

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

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 technical implication is that Amy's references are consistent with the properties of her ideal first-order preference $\succeq_1$. 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. (In Section 3.1, I provide behavioral intuitions behind the affinity of $\mathbf{r}$.)

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 Lemma 2 implies that $\succeq_{\mathbf{r}}$ is a complete, transitive, continuous and independent binary relation on $\mathbb{M}$—the necessary and sufficient conditions for the existence of a continuous affine function $K$ representing $\succeq_{\mathbf{r}}$ (see Herstein and Milnor, 1953). My axioms ensures that $K(\cdot) = v(\mathbf{r}(\cdot))$. This completes the proof.

I discuss the relationship between my model and prior models of menu preference in Section 5. In Appendix F, I also provide a detail comparison between my model and Kőszegi and Rabin (2006)'s seminal model of reference-dependent preference.

## 3.1. Behavioral Remarks

I present behavioral intuitions behind three important consequences of my axioms. The first one is Lemma 2. The implication is that Amy's reference is independent of Bob's *personal* contingencies, which is easier to see in the case of finite menus[29]. 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 $\mathbf{r}_{\lambda A+(1-\lambda)A}$[30]. 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

[29] In Appendix C, I show that the reference function $\mathbf{r}$ is unique whether Bob faces convex or finite menus.

[30] When an agent has both first- and second-order preferences, the case where Amy and Bob refer to a single individual, the personal contingency plans are always observable unless we introduce some dynamics of imperfect recall.

change how much Amy would be disappointed at his choice to do $y$ instead of $x$. Thus, she will evaluate $\lambda x + (1-\lambda) y$ based on the reference she has formed for $A$[31].

The second and third noteworthy consequences of my axioms are as follows:

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

(ii) Given any $A$, $(x, A) \succeq \phi \succeq (y, A)$ 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[32]. 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 $v$ is an affine function[33]. 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. Amy's loss or gain from Bob's choice from any non-singleton menu would not occur if it were a singleton.

## 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 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. □

In Appendix C, I also provide the same theorem when $\succeq$ is restricted to finite menus.

[31] 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 contingent 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.

[32] 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 $\mathbf{r}_B \succeq_1 \mathbf{r}_A$.

[33] 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 $\mathbf{r}_{\{y\}} = y$.

# 4. Paternalism and Libertarianism

This section investigates how preferences over the act of choosing are influenced by the extent to which one values freedom of choice, and how preferences for freedom might depend on menus.

Consider parents who want their child to prefer doing homework to watching television: let $x$ be homework and $y$ be television. 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 reference function $\mathbf{r}$ affine with respect to $v$. (Note that we have $v(x) \geq v(\mathbf{r}_{\{x,y\}}) \geq v(y)$ by definition of $\mathbf{r}$.) 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 $y = \mathbf{r}_{\{x,y\}}$.

More formally, I present two extreme attitudes toward the act of choosing (or preferences): paternalism and 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* $(x, A) \in \mathbb{C}$.

Axiom 9 states that a vacuous choice is weakly preferred to any given act of choosing $(x, A)$. This relates to the concept of paternalism—the tendency to restrict someone's options either to ensure his best well-being or prevent any possible mistakes. Suppose Amy's preference satisfies Axiom 9, or *Pure Paternalism*. 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 preferences. 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 in the menu. The purely libertarian parents would satisfy $(z_\alpha, A) \succ (x, \{x\})$ for any $\alpha > 0$.

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 in the menu: for all menu $A$, we have $r(A) \in \arg\max_{y \in A} v(y)$ if $\succeq$ is paternalistic, and $r(A) \in \arg\min_{y \in A} v(y)$ if libertarian.

What the two extreme attitudes toward the act of choosing 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 Relativity: $(\mathbf{r}_A, A) \sim (\mathbf{r}_B, B)$. 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 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, $(c, \{y, c\}) \succeq (c, \{c\})$ is plausible.

**Remark 1.** *Axiom 11b-c are testable conditions even when the DM also has outcome preferences, because they provide a ranking of choices conditional on the same outcome.*

## 4.1. Constant Paternalistic Attitudes

Examining the two polar attitudes toward the act of choosing raise 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 in 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 in the menu[34].) Consider the following representation $V_{v,\alpha}$ where the reference value function $v(\mathbf{r}(\cdot))$ takes the form of the $\alpha$-maxmin utility function of sets of lotteries presented by Olszewski (2007) in his characterization of ambiguity aversion[35]:

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

where the parameter $\alpha \in [0, 1]$ can be interpreted as Amy's paternalistic attitude toward Bob's act of choosing. The two extreme cases discussed above are when $\alpha \in \{0, 1\}$[36]. When $\alpha \in (0, 1)$, 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. □

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

[35] 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.

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

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 |

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 generally feel little to no pride in avoiding an obviously bad outcome (e.g., choosing life over committing suicide), or in choosing an obviously good one. 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, it emerges when a person acts according to his ideal preference as opposed to how he is *expected* to behave. In other words, the paternalistic attitude might weaken when menus present a strong conflict between ideal preference and the expected preference. Intuitively, when a menu contains a stronger temptation, Amy is 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 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 reference function and 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[37]. 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. $\square$

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. $\square$

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 referred to as

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

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.

[37] 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.

- *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 $\{\text{coffee, non-alcoholic beer, wine}\}$). 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 in 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 $u$:

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

## 4.3. Application: A Dictator's Menu Preference

Consider a dictator game context where the dictator chooses an option $x \in X$ (or a degenerate lottery) that refers to an allocation $(x_1, x_2)$ of wealth between himself (who gets $x_1$) and a recipient (who gets $x_2$). Consider three allocations: a fair allocation $f = (5, 5)$, a selfish allocation $s = (6, 4)$, and let $p_w = (6 + w, 6 + w)$ be called a Pareto optimal allocation with an increment $w > 0$. Before the dictator chooses an allocation, he is allowed to privately choose one of the following menus of allocations: $A = \{f, s\}$ and $B = \{f, s, p_w\}$. This set up—where the dictator privately chooses the set of allocations assuming the recipient does not know that the menu is chosen by the dictator—has been employed previously (see Dillenberger and Sadowski, 2012; Saito, 2015).

A standard dictator only has a first-order preference for greater wealth, and thus would prefer the choice $(p_w, \{f, s, p_w\})$ to any other possible act of choosing. His menu choice would be $B$ trivially for any increment $w > 0$. Alternatively, suppose the dictator only has a preference over the act of choosing, and prefers "preferring being fair to being selfish". In particular, let $v$ and $u$ represent his ideal preference and expected preference over allocations, respectively. Consider the following:

| Allocations | $v$ | $u$ |
|---|---|---|
| $f$ | 5 | 5 |
| $s$ | 4 | 6 |
| $p_w$ | $6+w$ | $6+w$ |

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

$$V_{v,u}(f, \{f, s\}) = 1 > 0 = V_{v,u}(p_w, \{f, s, p_w\}).$$

That is, the dictator prefers the act of being fair over being selfish to the act of choosing the Pareto optimal allocation $p_w$ over the two Pareto inferior allocations. Thus, the menu choice would be $A$. This is true because the dictator is purely libertarian when the menu is $A = \{f, s\}$, but becomes purely paternalistic when the menu is $B = \{f, s, p_w\}$. We can easily verify that

$$R_{v,u}(A) = v(s) = \min_{x \in A} v(x) \quad \text{and} \quad R_{v,u}(B) = v(p_w) = \max_{x \in B} v(x).$$

Intuitively, when the menu is $A = \{f, s\}$, the dictator feels a sense of pride in making the hard choice—sacrificing his own wealth to willingly pursue fairness—while feeling no pride at all in choosing $p_w$ because it is both the easiest and the best choice to make.

The result suggests that even when the dictator also cares about wealth itself, if he deems the act of choosing extremely more important than the wealth outcome (e.g., when $w$ is very small), he might deliberately remove the best outcome from the menu to pursue the best act of choosing.

**Remark 2.** *Prior models of menu preference—where the DM also cares about outcome—do not allow the menu $A$ to be more preferable than $B$ in the above example. Specifically, the below ranking was not a possibility:*

$$(f, \{f, s\}) \succ (p_w, \{f, s, p_w\}) \succ (f, \{f, s, p_w\}) \succ (s, \{f, s, p_w\})$$

*where $p_w$ is both the most tempting and normatively superior option among $\{f, s, p_w\}$.*

# 5. Discussion

## 5.1. Menu Preference

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

$$U_{u,f}(A) := \max_{x \in A} u(x) + f(x, A) \tag{3}$$

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). Theorems 1-2 imply that (i) the uniquely identified affine choice function $\mathbf{r}$ is a function of sets and thus, the study of the reference function technically lies within the study of preferences over sets; and (ii) the representation $V_{v,\mathbf{r}}$ subsumes any affine menu preference representation $U_{u,f}$ assuming $u$ is constant and $V_{v,\mathbf{r}} = f$.

One way to reproduce the prior menu preference representations is to model an optimization problem where the DM has an additively separable first- and second-order preferences represented by $u$ and $f$, respectively, and has a utility function $U_{u,f}$ as in (3). By finding the optimal menus, we can uniquely reproduce the same menu-choice patterns generated 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 $\succeq_1$ (represented by $u$) and a purely paternalistic preference $\succeq$ (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)[38].

[38] 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).

## 5.2. Welfare and Non-comparability Problem

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 DM'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 emotional sensations that the agent immediately experiences when choosing the menu itself (or any higher-level meta-choices) (see also Kőszegi and Rabin, 2008). To address this, Bernheim et al. (2024) proposed an empirical strategy to estimate welfare measures by combining choice data with the DM's self-reported emotional states. The welfare measures estimated in their experiments suggest the non-comparability problem exists, and that conditional on the same choice, two acts of choosing may yield different welfare outcomes.

However, the concept of higher-order preferences suggests that the design of welfare policies can be influenced by the social planner's own judgments and values. Consequently, having extensive data on the decision-maker's first- and second-order preferences (or choice data and self-reported emotional states) may not resolve the inherent complexities in welfare assessments. For example, suppose a parent is deciding whether or not to instruct her child to clean his room. Suppose further that the parent has enough data on the child's choices and emotional states so that she knows that (i) the child will succumb to the temptation of playing with his smartphone instead, and (ii) he will experience feelings of guilt by choosing play over responsibility. According to Bernheim et al. (2024)'s welfare measures, the parent should clean the room herself so that the child can enjoy playing with his phone without any sense of guilt or shame. Yet, some parents might intentionally instruct the child to clean because they believe feeling guilty is an important experience, and want to promote the child's personal growth and future welfare. The example suggests a higher-order non-comparability problem: if the social planner has a preference over the DM's preference over the act of choosing, then a unique welfare measure may not be inferred either from (i) the DM's choice data and self-reported emotional states, or (ii) the DM's choice data and the social planner's choice data. Suppose a policy-maker retires, leaving behind the DM *and* his own choice data for the newly hired policy-maker. From the available choice data, the newly hired cannot infer whether his predecessor's objective was to promote the DM's present or future welfare.

# 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,

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

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

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

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

We say the function $\mathbf{r} : \mathbb{M} \to X$ is affine with respect to a binary relation $\succeq_1$ on $X$ if it satisfies Lemma 2. When Bob faces the menu $\lambda A + (1-\lambda)\,B$ and chooses a contingent plan $\lambda x + (1-\lambda)\,y$, Amy observes and evaluates this choice based on her subjective reference of $\lambda A + (1-\lambda)B$ denoted by $\mathbf{r}_{\lambda A+(1-\lambda)B}$. 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 technical implication is that Amy's references are consistent with the properties of her ideal first-order preference $\succeq_1$. 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.

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, $K(\{x\}) = v(x)$. 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

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

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 $(\mathbf{r}_A, A) \sim (z, \{z\}) \sim (\mathbf{r}'_A, A)$ for any vacuous choice $(z, \{z\})$, which implies $\mathbf{r}_A \sim_1 \mathbf{r}'_A$. $\square$

## C. Theorems 1-2 for Finite Menus

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

**Axiom 15.** *For any $A, B \in \mathbb{M}_f$, there are $\mathbf{r}_A \in conv(A)$ and $\mathbf{r}_B \in conv(B)$ such that*

*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.* $(\mathbf{r}_A, A \cup \{\mathbf{r}_A\}) \sim (\mathbf{r}_B, B \cup \{\mathbf{r}_B\})$

**Theorem 4** (Finite Menu). *$\succeq$ restricted to the set $\{(x, A) : x \in A \in \mathbb{M}_f\}$ satisfies Axioms 1-7, and Axiom 15 if and only if $\succeq$ has a unique representation as in Theorems 1-2.*

*Proof of Theorem 4.* 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

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

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* $A \in \mathbb{M}_f$.

*Proof of Lemma 4.* Let $(x, A_n) = \sum_{s=1}^{n} \lambda_s (x, A)$ where $A_n = \sum_{s=1}^{n} \lambda_s A$ and $\lambda_s = \frac{1}{n}$ for all $s = 1, ..., n$. Then, $(x, A_n)$ converges to $(x, conv(A))$[39]. 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 $\frac{1}{2}x + \frac{1}{2}\mathbf{r}_A \sim_1 \frac{1}{2}\mathbf{r}_{conv(A)} + \frac{1}{2}x$. Since $\succeq_1$ is independent, this gives us $\mathbf{r}_{conv(A)} \sim_1 \mathbf{r}_A$. For the second indifference relation, note that $\sum_{s=1}^{n} \lambda_s A \cup \{\mathbf{r}_A\}$ converges to $conv(A \cup \{\mathbf{r}_A\}) = conv(A)$ 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 $\mathbf{r}_{conv(A)} \sim_1 \mathbf{r}_{A\cup\{\mathbf{r}_A\}}$. □

For Lemma 2, Note that by Axiom 15, we have $(\mathbf{r}_A, A \cup \{\mathbf{r}_A\}) \sim (\mathbf{r}_B, B \cup \{\mathbf{r}_B\})$. By Lemma 4, we have $(\mathbf{r}_A, conv(A)) \sim (\mathbf{r}_B, conv(B))$, and thus

$$\begin{aligned}(\mathbf{r}_A, A \cup \{\mathbf{r}_A\}) &\sim \lambda(\mathbf{r}_A, conv(A)) + (1-\lambda)(\mathbf{r}_B, conv(B)) \\ &= (\lambda \mathbf{r}_A + (1-\lambda)\mathbf{r}_B, \lambda conv(A) + (1-\lambda) conv(B)).\end{aligned}$$

Axiom 15 and the Shapley-Folkman theorem also give us

$$\begin{aligned}(\mathbf{r}_A, A \cup \{\mathbf{r}_A\}) &\sim \left(\mathbf{r}_{\lambda A+(1-\lambda)B}, \lambda A + (1-\lambda) B \cup \{\mathbf{r}_{\lambda A+(1-\lambda)B}\}\right) \\ &\sim \lim_{n\to\infty} \sum_{s=1}^{n} \lambda_s \left(\mathbf{r}_{\lambda A+(1-\lambda)B}, \lambda A + (1-\lambda) B \cup \{\mathbf{r}_{\lambda A+(1-\lambda)B}\}\right) \\ &= \left(\mathbf{r}_{\lambda A+(1-\lambda)B}, conv(\lambda A + (1-\lambda) B)\right) \\ &= \left(\mathbf{r}_{\lambda A+(1-\lambda)B}, \lambda conv(A) + (1-\lambda) conv(B)\right).\end{aligned}$$

By Axioms 6-7, we can conclude $\mathbf{r}_{\lambda A+(1-\lambda)B} \sim_1 \lambda \mathbf{r}_A + (1-\lambda)\mathbf{r}_B$. 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 $A \in \mathbb{M}_f$. 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, both $(\mathbf{r}_A, conv(A))$ and $(\mathbf{r}'_A, conv(A))$ are indifferent from a vacuous choice, say $(z, \{z\})$. That is,

$$(\mathbf{r}_A, conv(A)) \sim (\mathbf{r}'_A, conv(A)) \sim (z, \{z\}) \tag{4}$$

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

[39] To be precise, $A_n$ converges to $conv(A)$ in the Hausdorff metric—see Emerson and Greenleaf (1969) and Starr (1969) for this result, also known as the Shapley-Folkman theorem.

## D. Proofs of Corollaries

Define the best and worst lotteries in 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

$$\tfrac{1}{2}b_B + \tfrac{1}{2}b_{A\cup C} \succ_1 \tfrac{1}{2}b_A + \tfrac{1}{2}b_{B\cup C}. \tag{5}$$

If $b_A \succeq_1 b_C \succeq_1 b_B$, then (5) 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 (5) 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 (5) 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 $\alpha \in (0, 1)$,

$$v(x) - v(y) \geq \alpha\Big[v(b_{\{x\}\cup C}) - v(b_{\{y\}\cup C})\Big] + (1-\alpha)\Big[v(w_{\{x\}\cup C}) - v(w_{\{y\}\cup C})\Big] \geq 0 \tag{6}$$

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, (6) 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

$$(1-\alpha)\Big[v(w_B) - v(w_A)\Big] < \alpha\Big[v(b_A) - v(b_B)\Big]. \tag{7}$$

Suppose $v(b_C) \leq v(b_B) < v(b_A)$ and $v(w_C) \leq v(w_A) < v(w_B)$. Then, assuming (7), the following holds for all $\alpha \in (0,1)$,

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

which violates Axiom 11a.

For Axiom 11b, suppose $v(b_B) < v(b_A) \le v(b_C)$ and $v(w_A) < \min\{v(w_B), v(w_C)\}$. Then, assuming (7), 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 $\alpha(v(b_A) - v(b_B)) \ge (1-\alpha)(v(w_B) - v(w_A))$. Then, we have $\mathbf{r}_A \succeq_1 \mathbf{r}_{A\cup B}$ since

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

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

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

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[40]. In this case, Amy's reference takes every option into account equally when evaluating Bob's preference, and thus, subjectively expects the choice of 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. Conversely, 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 $A = \{x, y\}$

[40] Many axiomatic models of menu preferences put nonzero utility weights on very few non-chosen options in a menu, which inhibit the agent's ability or willingness to consider every option in 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.

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 $B_\alpha$ converges to $A$ in the Hausdorff metric as $\alpha \to 1$[41]. 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 $B_\alpha$ should be indifferent. However, we have $V_{v,avg}(x, A) \neq V_{v,avg}(x, B_\alpha)$ for $\alpha \neq 0.5$.

In a general non-finite setting, we can assume that Amy has a probability measure $\mu$ on $X$ such that an expected option conditional on $A \subseteq X$ is Amy's subjective expectation of Bob's choice from $A$. 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 $V_\mu$ deviates from previously discussed properties, and its technical friendliness relies heavily on the design of topology and the set of menus[42].

$V_{v,avg}$ can still partially satisfy a number of properties discussed in my paper. In particular, it satisfies the following weaker version of Axiom 13.

**Axiom 16** (Weak VCB)**.** *For any disjoint $A, B$,*

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

**Corollary 5.** *The preference represented by $V_{v,avg}$ satisfies Axiom 12 and Axiom 16.*

*Proof.* To prove, I will assume that

$$\mathbf{r}_A \sim_1 \sum_{x \in A} \frac{1}{|A|} x$$

for all $A \in \mathbb{M}$ where $|A|$ is the number of options in $A$. Let $V_{v,\mathbf{r}}$ represent $\succeq$. By Lemma 1, this condition is equivalent to $V_{v,\mathbf{r}} = V_{v,avg}$.

For Axiom 12, suppose $x \succeq_1 y$. By Lemma 1, it is sufficient to show that

$$v(x) - v(y) \geq \frac{1}{1+|C|}\left[v(x) + \sum_{y \in C} v(y)\right] - \frac{1}{1+|C|}\left[v(y) + \sum_{y \in C} v(y)\right] \geq 0.$$

It is easy to show that the first inequality holds whenever $1 + |C| \geq 1$ and the second one holds whenever $v(x) \geq v(y)$.

[41] 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.

[42] As a model of attitudes towards ambiguity, 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.

For Axiom 16, note that $\mathbf{r}_A \succeq_1 \mathbf{r}_B$ implies $\frac{1}{|A|}\sum_A v \geq \frac{1}{|B|}\sum_B v$. Supppose $\mathbf{r}_{A\cup B} \succ_1 \mathbf{r}_A$. That is,

$$\frac{1}{|A|+|B|}\left(\sum_A v + \sum_B v\right) > \frac{1}{|A|}\sum_A v$$

which implies $\frac{1}{|B|}\sum_B v > \frac{1}{|A|}\sum_A v$, a contradiction. Suppose $\mathbf{r}_B \succ_1 \mathbf{r}_{A\cup B}$. That is,

$$\frac{1}{|B|}\sum_B v > \frac{1}{|A|+|B|}\left(\sum_A v + \sum_B v\right)$$

which implies $\frac{1}{|B|}\sum_B > \frac{1}{|A|}\sum_A v$, a contradiction. □

## 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

$$U(x|\mathbf{r}) := u(x) + \ell\Big(V_{v,\mathbf{r}}(x, A)\Big) \tag{8}$$

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 $\ell$ 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[43]. 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 unanticipated menu and loss aversion (*i.e.* when $\ell$ is linear), the second term of (8) can affect her behavior.

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

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[44]. 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[45].

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

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

[44] 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.

[45] 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 $\mu$ on $u$. Then, her second-order preference induces a menu preference characterized by $(v, \mathbf{r}, \mu)$ and represented by a form:

$$E_\mu \overline{V}(A) = E_\mu \left[ \max_{x \in B(A;u)} v(x) \right] - v(\mathbf{r}_A). \tag{9}$$

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

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 1 demonstrates Axiom 1, where the payoff vectors are of Amy's and Bob's utilities:

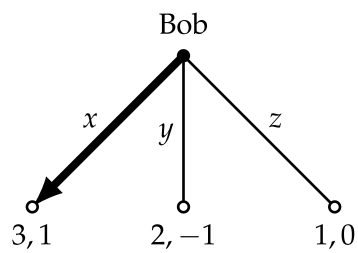


Figure 1: $(x, Z)$ induced by both $xPyPz$ and $xQzQy$.

The bold arrow in Figure 1 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 2-3, where *Nature* decides with equal probability which menu he will face.

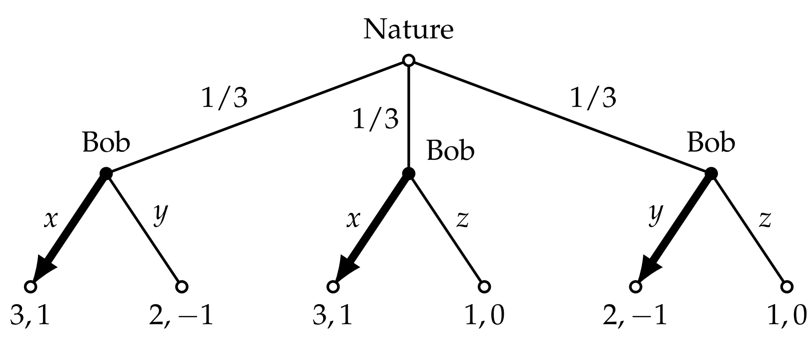


Figure 2: $\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A\right)$ induced by $xPyPz$.

The contingency plan (indicated by the bold arrows) in Figure 2 represents the choice $(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A)$ induced by $P$ while the one in Figure 3 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

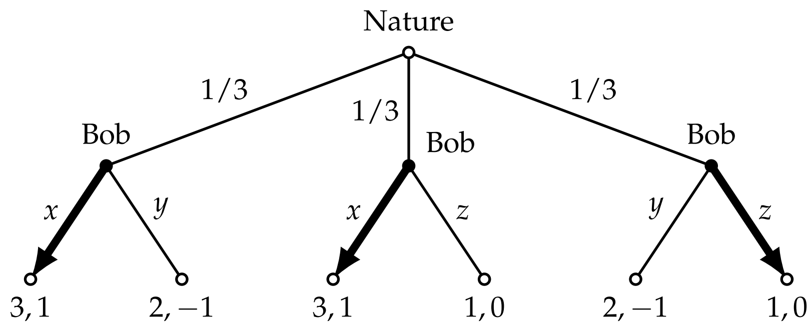

Figure 3: $\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}z, A\right)$ induced by $xQzQy$.

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 2. Hence, from any choice from $A$, Amy can precisely infer which preference Bob has among those in $\mathbb{P}(Z)$[46].

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 2 to the one in Figure 3. 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

$$V_{v,\mathbf{r}}\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}y, A\right) \geq V_{v,\mathbf{r}}\left(\frac{1}{3}x + \frac{1}{3}x + \frac{1}{3}z, A\right). \tag{10}$$

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 (10) since Bob's menu is fixed at $A$. Indeed, (10) 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, independent 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

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

[46] 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.

which, essentially, is the trees in Figure 2-3 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 2** (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 $\succeq$ 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).

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 which a second-order preference is formally integrated into a microeconomic framework. This paper is the first to contribute 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.

More importantly, 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 Axiom 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 $\{\text{smoke}, \text{abstain}\}$). 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,

$$v_H(x_1) - v_H(y_1) > v_H(x_2) - v_H(y_2). \tag{11}$$

Halldén (1980) used the fact that without imposing any cardinal property on utility functions, (11) is equivalent to stating that an expected utility maximizer prefers a coin toss $\frac{1}{2}x_1 + \frac{1}{2}y_2$ to $\frac{1}{2}x_2 + \frac{1}{2}y_1$[47].

I identify three key limitations of Halldén's axiom. First, the decision-maker's menus are restricted to contain exactly two options[48]. 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.

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 (11), 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

[47] 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.

[48] 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.

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

Recall that the model 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 indifference. 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 $P_1 \sim P_2 \succ P_3$[49].

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[50]. 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 an announcement may not characterize Bob's 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.

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.


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

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

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.

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¡sup¿p¡/sup¿ = C + + 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.

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.

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.

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.

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.

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