A Representation of Preference over Preferences and the Act of Choosing

Jun Hyun Ji*

July 6, 2025

Abstract

This paper studies how individuals value the act of choosing itself by using the concept "preference over preferences." In addition to preferring outcomes, the agent likes a preference if she prefers *behaving* as if she holds that preference: "preferring x to y" is preferred to "preferring y to x" if she values the act of willingly giving up y for x more than giving up x for y. I show that a preference over the acts of choosing has a unique representation that identifies the agent's second-order preference, which *induces* all standard menu preferences in the literature on self-control. Unlike how outcomes are ranked, the value of preferences is relative to contextual factors—counterintuitively, *the best act of choosing* may only be feasible when *the best outcome* is unavailable. My model also accommodates preferences over others' preferences, analyzing choices that condition on others' choices and the social signaling value of choice.

Keywords: Preference over preferences; second-order preference; meta preference; pride; guilt; menus; paternalism

*Ph.D. student, Economics, University of Pittsburgh. Email: juj25@pitt.edu. I am deeply indebted to Luca Rigotti for his guidance, support, and encouragement throughout this project. I am also grateful to Sven Neth and Kevin Zollman for many insightful discussions. I received valuable comments from In-Koo Cho, Antonio Penta, Michael Woodford, and participants at the 2024 Asian School of Economic Theory at NYU Abu Dhabi. I thank Jawwad Noor, Evan Piermont, and all participants at the 2025 BSE Summer Forum (Choice and Decision) for helpful feedback and engaging discussion. Additional thanks go to David Ahn, David Huffman, Colin Sullivan, Gerelt Tserenjigmid, and my colleague Bruno Kömel for their helpful suggestions. I gratefully acknowledge the financial support by 2023 Tamara Horowitz Memorial Fund.

1. Introduction

When individuals make a choice, they often care not only about the consequence of that choice, but also about what they are willingly giving up for it. The existing choice-theoretic models of temptation provide the foundational insight that people have preferences for designing their future choice sets (or menus) because they care about the psychological experiences associated with opportunities they may forgo. For example, commitment is valuable: one might wish to eliminate unwanted temptations, anticipating that the act of willingly giving them up requires costly self-control (Gul and Pesendorfer, 2001).

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 that restrict a rational decision-maker's (DM's) options. Recent work by Bernheim et al. (2024) shows that choices over outcomes—and even menus—do not uniquely reveal welfare-relevant experiences involved in choosing (e.g., guilt, pride, exertion of self-control), a limitation they refer to as *the non-comparability problem*, implying that policies based solely on observed choices may fail to enhance, or may even diminish, individual well-being.

Understanding how individuals value—more precisely, *rank*—the acts of choosing has therefore become increasingly important. However, prior studies—relying on either menu choices or welfare estimations via econometric methods—do not formally offer such theoretical insights.[1] In this paper, I provide a theoretical framework for preferences over the acts of choosing. An act of choosing is a manifestation of certain preferences; therefore, I introduce a novel concept of preference over preferences (henceforth, *second-order preference*). I use the phrase "preferring a preference" to mean preferring to *behave* as if one holds that preference. To illustrate, suppose an agent strictly prefers x to y. Our standard understanding is that she is willing to give up y for x—this describes her *first-order preference*. Suppose she prefers "preferring x to y" to "preferring y to x." Then, it must be that there is some intrinsic value in the act of willingly giving up y for x greater than in that of giving up x for y.[2] My representation

[1] Since choices and welfare may be misaligned when the DM cares about the acts of choosing, Bernheim et al. (2024) proposed an econometric method to estimate the DM's welfare associated with each choice by combining choice data with self-reported well-being methods.

[2] Philosophers have long discussed that human beings can have preferences over their own preferences (Frankfurt, 1971; Jeffrey, 1974). For example, some may wish to become a person who prefers exercising to indulging in eating. 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).

identifies these first- and second-order preferences separately.

My model yields several important implications. First, preferences over acts of choosing induce all standard menu preferences found in the literature, offering a clearer conceptual understanding of the motivations underlying menu choice behavior. Second, and crucially, my model is not limited to the self-control problems that dominate the menu preference literature. Rather, it addresses the broader question: How does an individual assess the quality of choices—made either by herself or by others—separately from their outcomes? My axioms capture the insight that the quality of a choice is relative to contextual factors such as constraints and conflicts among the DM's expected behavior, ideal behavior, and temptations. Consequently, I demonstrate counter-intuitive cases in which the best act of choosing is feasible only when the best outcome is unavailable. Third, by exploring the broader question, my model accommodates preferences over others' preferences, providing a natural framework for analyzing choices that condition on others' choices (e.g., a planner's paternalistic interventions based on observing the DM's choices; voters assessing politicians based on their previous policy decisions; hiring a CEO from a pool of applicants based on their prior business choices). Finally, my approach captures the social signaling value inherent in the act of choosing, without assuming that individuals explicitly care about others' beliefs. Instead, signaling emerges naturally from how choices reveal underlying preferences that the agent values.

Section 2 presents the model. I let the pair (x, A) denote the act of choosing an option x from a menu A.[3] The agent in my model has a preference $\succeq$ over all possible acts of choosing, caring not only about outcomes but also about how choices manifest preferences. Hence, her preference is not necessarily a second-order preference: (x, A) might be preferred to (y, B), simply because the outcome of x is greater than that of y. I say $\succeq$ is a second-order preference if it satisfies two axioms, called *the axioms of second-order preference*. In this case, the relation $(x, A) \succeq (y, B)$ is interpreted as *"preferring x to all else in A"* is preferred to *"preferring y to all else in B."* The first axiom is that any two *vacuous choices*—any two acts of choosing from a singleton menu—are indifferent. I call

[3] 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.

this axiom *Indifference of Vacuous Choices* (IVC). Since a vacuous choice represents exogenous consumption, IVC implies that the agent does not care about outcomes (i.e., no first-order preferences).

The second axiom, called *Relativity*, states that we can find a mixture of lotteries on any given menu, such that the agent finds no value in the act of either giving it up or choosing it from that menu. As an intuitive example, consider the classic family question: "Who do you like better, Mom or Dad?" Suppose the child decides to flip a coin to settle the matter. The parents might object, saying, "Flipping a coin doesn't count—you have to choose!" because choosing to flip the coin is as meaningless as if the child were told to flip it, and giving up the coin toss brings no additional gain or loss for either parent. *Relativity* captures 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.[4] If the two axioms are violated, there must be two menus A^* and B^* such that "preferring anything in A^*" is strictly preferred to "preferring anything in B^*" which implies that preferences do not matter: the agent merely wants some outcomes in A^* more than the ones in B^*.

Section 3 presents my main theorems (Theorems 1-2): the preference over the acts of choosing—exhibiting both first- and second-order preferences—is uniquely represented by the function $U_{u,v,r}$ of the form:

$$U_{u,v,r}(x, A) = u(x) + v(x) - v(r(A))$$

where u, v are von Neumann-Morgenstern (vNM) utility functions over lotteries, and r is a choice function that selects a mixture $r(A)$ of options on the menu A. The function u is a ranking of vacuous choices, and thus represents the first-order preference. If the agent does not care about outcomes (i.e., if IVC holds), then u is constant. I show that given a preference $\succeq$ over the acts of choosing, we can uniquely identify a second-order preference $\succcurlyeq_2$ (i.e., $\succcurlyeq_2$ satisfies the axioms of second-order preference) that is represented by the term $v(x) - v(r(A))$. The function v represents the preference that the agent believes the DM (herself or another) should ideally adopt (e.g., an alcoholic thinks that ideally, he should prefer coffee to beer; a parent thinks that her child should ide-

[4] For example, a poor person's act of donating $100 to charity can be more praiseworthy than a billionaire doing the same. The billionaire would likely have to give a much larger sum to earn equal praise. It's not just the raw dollar amount that impresses people, but by the amount relative to one's financial constraints which reveals how strongly someone values helping others.

ally prefer doing homework to watching television). The lottery $r(A)$ is called *the reference of* A that serves as a reference against which the act of choosing x from A is assessed.[5]

My model suggests that menu preferences arise when the agent chooses menus for herself anticipating the outcome as well as the value of her own preference manifestations. Let $\succeq_M$ be the induced menu preference defined by $A \succeq_M B$ if and only if there exists $x \in A$ such that $(x, A) \succeq (y, B)$ for all $y \in B$. Then, any standard menu preference in the literature is a special case of $\succeq_M$ where the function $r(\cdot)$ takes a special form in each prior model.[6] For instance, $\succeq_M$ is exactly Gul and Pesendorfer (2001)'s preference for commitment if $r(\cdot)$ selects the most ideal option on each menu—i.e., $v(r(A)) = \max_{z \in A} v(z)$ for all A.[7]

Section 4 introduces several special forms that illustrate how the function $r(\cdot)$ relates to the agent's paternalistic attitude toward the acts of choosing. I start with two extreme attitudes. I say the second-order preference is purely paternalistic (purely libertarian) if the reference is the most (least) ideal option on each menu. This means a paternalistic perspective assesses choices by how much they fall short of the most ideal outcome, whereas a libertarian perspective evaluates them by how much they exceed the least ideal option.[8] For example, if a child does homework half the time and chooses leisure the other half, a purely paternalistic parent would enforce homework time because she is disappointed by the 50 percent of homework he willingly gave up. In contrast, a purely libertarian parent would allow the child to choose freely since she takes

[5] 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 about the outcome, which arises from uncertainty. Yet, my model stays within expected utility theory and the reference stems from the agent's second-order preference. See Section 5.2 for a more detail comparison.

[6] A standard menu preference has affine representations; it is not subject to uncertainty about consumption preferences; and it yields ex post consumption choices satisfying the weak axiom of revealed preference. (See, for example, Sarver, 2008; Kopylov, 2012; Dillenberger and Sadowski, 2012; Saito, 2015; Kopylov and Noor, 2018; Noor and Ren, 2023)

[7] In Gul and Pesendorfer (2001)'s context, u is the ranking of normative goals, v is the temptation ranking, and the term $v(x) - \max_{y \in A} v(y)$ captures the cost of self-control. This context is carried into the present paper as follows: the agent evaluates each outcome based on its normative value, but believes she should *ideally* prefer the most enjoyable option available. Consequently, the act of willingly giving up the most enjoyable option incurs a mental cost such as self-control.

[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 takes a similar stance toward preferences; therefore, she is concerned with the design of the acts of choosing, rather than outcomes.

pride in the 50 percent of leisure the child willingly gave up. Ultimately, the attitude is a matter of perspective—as in the familiar question, "Is the glass half full or half empty?" These attitudes shape preferences over menus across various contexts. A paternalistic attitude favors smaller menus, implying a desire for commitment or avoidance of guilt. In contrast, a libertarian attitude favors larger menus, implying a desire for a sense of pride, freedom, or even the embrace of guilty pleasures.[9]

As my main result, the representation of the *temptation-adjusted preference* in Theorem 3 captures paternalistic attitudes that flexibly respond to the DM's temptations. The idea is that simply choosing a good option over a clearly bad one does not necessarily reflect having *good preferences*. People generally feel little to no pride in avoiding an obviously bad outcome (e.g., choosing delicious vegetable juice over a cup of gasoline), but an alcoholic might feel extremely proud of himself for making a "hard choice" such as choosing coffee over beer. To capture this insight, I allow an act of choosing to have greater value the more it exceeds the agent's expectation and the more she perceives the menu as *difficult*, where difficulty is defined by the gap between the expected choice and the strongest temptation on the menu.

For instance, suppose a menu contains an option highly tempting yet misaligned with the ideal preference. The stronger this temptation, the more likely the agent anticipates that the DM will yield to it. If the DM is nevertheless expected to resist, the agent anticipates greater cognitive effort required to process this challenging choice situation. In such cases, she would feel greater pride if the temptation is actually resisted and be less disappointed if otherwise. Formally, I impose an additional axiom called *Reference-betweenness* that states if the reference value of A is ideally preferred to that of B, then the reference value of $A \cup B$ falls in between. Motivated by Gul and Pesendorfer (2001)'s *Set-betweenness* axiom, *Reference-betweenness* uniquely identifies the agent's expectations about the DM's choices, temptations, and the difficulty of each menu.

The novelty of the temptation-adjusted preference is that the best act of choosing may only be feasible when the best outcome is unavailable. Suppose a dictator—tasked with allocating resources between himself and a passive recipient—prefers "preferring being altruistic to being selfish." He is proud of

[9] 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).

himself for choosing a fair allocation over an unfair one that disproportionately benefits him. Now, suppose an allocation that Pareto-dominates both the fair and unfair allocations is added to his menu. In this case, the sense of pride might completely vanish because the act of giving up the two Pareto-inferior allocations is neither giving up being altruistic nor giving up being selfish. Despite being the best outcome overall, the dictator might be strictly better off without the Pareto-dominant allocation on his menu if the sense of pride is significantly valuable. In contrast, menu preferences in the literature emphasize the influence of outcome preferences, and thus do not rationalize the tendency to willingly remove the best outcome from the menu.

Section 5 discusses related literature in detail. In Section 6, I consider broader implications of the model, including welfare assessment, the signaling value of choice, and the relationship between the model and a more general notion of second-order preference. Proofs (if omitted) are collected in Appendices A-B. The Supplemental Appendix presents three methodological approaches for testing my model: a menu-choice approach, a welfare-measure approach, and a revealed-preference approach based on *"choices over DMs."* Appendix C, which is available upon request, discusses my contributions to the prolonged philosophical studies on the relationship among higher-order preferences, desires, and self-control.

2. Model

I consider an agent who not only has a preference over standard objects (e.g., consumption goods, money, or any other forms of consequential outcomes), but also has a preference over preference relations on those objects. To capture this idea, I represent the agent as composed of two conceptual entities: a decision-maker (DM), who corresponds to the conventional decision problem—choosing an option (a lottery) from an exogenously given menu of options—and an *evaluator*, who forms a "second-order" judgment over the DM's choices from an observer's perspective. Namely, the evaluator is assessing which choice is superior to one another by considering both the outcomes and the DM's preferences inducing those choices. This formulation applies equally well to situations in which the DM and the evaluator are a single individual.[10]

[10] This is inspired by the philosophical discussions of a person's capacity to reflect upon one's own tastes and dispositions (see Frankfurt, 1971)—essentially, the concept of "meta-preference."

I characterize the DM's choice environment, as follows.

Options (lotteries) and Menus. Let Z be the finite set of outcomes, and X be the set of lotteries on Z, endowed with a metric d generating the standard weak topology. X is the DM's entire consumption space where any elements $x, y, z \in X$ are called lotteries or *options*. For $\alpha \in [0,1]$, let $\alpha x + (1-\alpha)\, y$ denote the mixture of lotteries x and y that yields x with probability α and y with probability $1-\alpha$. Let $\mathbb{M}$ denote the set of nonempty compact subsets of X whose elements $A, B, C \in \mathbb{M}$ are called *menus*.[11] And let $\text{conv}(A)$ 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 $\lambda \in [0,1]$.

The act of choosing. The primitive of my model is a binary relation $\succeq$ over the set $\mathbb{C} = \{(x, A) : x \in A \in \mathbb{M}\}$. 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*.[12] 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 the DM faces the menu A with probability λ and B with probability $1-\lambda$. Before this uncertainty is resolved, he chooses a contingency plan $\lambda x + (1-\lambda)\, y$ which constitutes the act of choosing $(\lambda x + (1-\lambda)\, y, \lambda A + (1-\lambda)B)$. A special notation ϕ will be used to indicate an arbitrary vacuous choice—a choice made from a singleton menu: i.e., $\phi \in \{(x, \{x\}) : x \in X\} \subset \mathbb{C}$.

First-order Preferences. I use the symbol $\succcurlyeq_1$ (along with $\succ_1$ and $\sim_1$) to denote a complete and transitive binary relation (henceforth, a first-order preference) on X. In terms of the preference over the acts of choosing, I use $\succcurlyeq_1^e$ to denote *the evaluator's first-order preference*, which I define as the ranking of vacuous choices:

Definition 1 (The Evaluator's First-order Preference)**.** A binary relation $\succcurlyeq_1^e$ on X is *the evaluator's first-order preference* if $x \succcurlyeq_1^e y$ whenever $(x, \{x\}) \succeq (y, \{y\})$.

The definition of second-order preference will be presented after I impose the standard axioms below.

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

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

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

2.1. Standard Axioms

I employ the standard vNM axioms of continuity and Independence used in prior literature, and impose the following axioms:[13]

Axiom 1 (Weak Order). $\succeq$ *is complete and transitive.*

Axiom 2 (Independence). *For all* $\lambda \in (0,1)$, $(x,A) \succ (y,B)$ *implies* $\lambda(x,A) + (1-\lambda)(z,C) \succ \lambda(y,B) + (1-\lambda)(z,C)$.

Axiom 3 (Continuity). $\{(x,A) : (x,A) \succeq (y,B)\}$ *and* $\{(x,A) : (y,B) \succeq (x,A)\}$ *are closed.*

Axiom 4 (Consistency). *There is a continuous and independent preference relation* $\succcurlyeq_1^c$ *on* X *such that* $x \succcurlyeq_1^c y$ *whenever* $(x,A) \succeq (y,A)$ *for all* A *with* $x, y \in A$.

Axioms 1-3 are in alignment with the standard axioms of the expected utility (EU) theory.[14] I provide the motivation for the Independence axiom in detail in Section 3.1.1. Consistency states that there is a vNM preference over lotteries—denoted by $\succcurlyeq_1^c$—that governs $\succeq$ *within* a menu. In other words, the evaluator's preference is consistent: if she prefers the act of choosing x from A to choosing y from A, then choosing x from any other menu B is preferred to choosing y from B.[15] Note that Consistency does not determine the ranking *across* menus: i.e., $(x,A) \succeq (y,A)$ does not guarantee $(x,A) \succeq (y,B)$. This fundamentally differs from any individuals who care only about outcomes because they would prefer (x,A) to (y,B) whenever x is a better outcome than y.

2.2. Axioms of Second-order Preference

The two axioms below are called *the axioms of second-order preference*:

Axiom 5 (Relativity). *For any* $A \in \mathbb{M}$, *there exists* $x' \in \text{conv}(A)$ *such that*

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

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

[13] A binary relation $\succcurlyeq_1$ on 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 $\{x \in X : x \succcurlyeq_1 y\}$ and $\{x \in X : y \succcurlyeq_1 x\}$ are closed.

[14] In particular, Axiom 2 is consistent with the assumption that the decision-maker remains impartial concerning the timing of uncertainty resolution, as implied by the Independence axiom imposed on menu preferences (see Gul and Pesendorfer, 2001; Dekel et al., 2001, 2007)

[15] Technically speaking, Consistency implies that an outcome-choice rule induced by $\succeq$ satisfies the weak axiom of revealed preference (WARP). The preference $\succeq$ induces an outcome-choice rule defined by $c(A;\succeq) := \{x \in A : (x,A) \succeq (y,A)\ \forall y \in A\}$.

Axiom 6 (Indifference of Vacuous Choices)**.** $(x, \{x\}) \sim (y, \{y\})$ *for all* $x, y \in X$.

I say the evaluator's preference $\succeq$ over the acts of choosing *is* her second-order preference if $\succeq$ satisfies Relativity and IVC. In this case, $(x, A) \succeq (y, B)$ is interpreted as a ranking of preferences, not outcomes: she intrinsically prefers *"preferring x to everything else in A"* to *"preferring y to everything else in B."* This motivates the following definition.

Definition 2 (Second-order Preference)**.** The preference $\succeq$ is called a *consistent EU second-order preference over the acts of choosing a single option* (for brevity, simply referred to as a second-order preference) if $\succeq$ satisfies Axioms 1, 2, 3, Consistency, Relativity and IVC.

Note that the second-order preference defined above focuses on how the evaluator ranks the acts of choosing *a single option from a menu*, without regard to the underlying ranking that may have led to the choice. In Section 6.3, I discuss a more general notion of second-order preference—one that can depend on the full structure of the DM's underlying preference relation (e.g., caring about which option was his second favorite)—and identify two assumptions that narrow this broader preference to the class of preferences analyzed in this paper.

I now provide the motivations for the two axioms. IVC explicitly states that the evaluator does not have a first-order preference (i.e., $x \sim_1^e y$ for all x, y), and thus her preference over the acts of choosing should find any two vacuous choices indifferent. Suppose the evaluator wants the DM to prefer x to y, while the DM chose y from $\{y\}$. Then, the DM did not *willingly* choose or give up anything. Is the evaluator happier if the DM is given $\{x\}$ instead, so that he ends up with the choice $(x, \{x\})$? If yes, her satisfaction must come from appreciating the outcome of x. Yet, she has no reason to intrinsically appreciate the DM's preference which had no contribution to the outcome.

Relativity states that given any menu A, we can find a mixture of options in A, say x', such that (a) adding x' to A does not change how each choice from A is evaluated, and (b) the act of choosing x' over A is indifferent to the exogenous consumption of x'. Intuitively, each menu offers an option such that there is no intrinsic value in the act of either (a) giving it up or (b) choosing it. (Nor is there any outcome value if IVC holds.) The intuitive reason why (a) and (b) should hold together is that if there is no value in the act of giving up some option on a menu, then there must be no value in the act of choosing it, and vice versa.

I next discuss the motivation for Relativity 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. Suppose the evaluator considers investing in two potential businesses by evaluating them based on their owners' past choices. One candidate chose \$10 from the menu $\{\$10, \$0\}$, while another candidate chose \$20 from the menu $\{\$20, \$30\}$. In absolute terms, the second candidate's choice yielded more money. However, if the evaluator is looking for a partner who prioritizes money, she would prefer the first candidate whose choice clearly revealed a preference for money while the other's did not. Alternatively, we can naturally think of a middle school physical education teacher's grading system, where her first-order preference pertains to performance-based assessments and her second-order preference relates to effort-based grading. If the effort-based grading prevails, then it must be possible to award the same grade to both the most athletic and the least athletic student if the former does not try harder than the latter given their distinct capabilities.

Furthermore, Relativity and IVC together imply that vacuous choices serve as reference points against which the DM's preference is evaluated given any menu. Note that by Relativity and IVC, we can define a choice function $r : \mathbb{M} \to X$ by the following indifference conditions:

(a) $(x, A \cup \{r(A)\}) \sim (x, A)$ for all $x \in A$,

(b) $(r(A), A \cup \{r(A)\}) \sim (r(B), B \cup \{r(B)\})$ for all $A, B \in \mathbb{M}$.[16]

Since (b) holds even when B is a singleton menu, it follows that $(r(A), A) \sim \phi$ for all $A \in \mathbb{M}$ and any vacuous choice ϕ. This means each $r(A)$ itself serves as the reference point matching the value of vacuous choices. To motivate the existence of such $r(A)$ for any A, recall the classic family question "Who do you like better, Mom (x) or Dad (y)?" Let $A = \{x, y\}$ be the child's menu. Each parent wants to be their child's favorite. The coin flip $\frac{1}{2}x + \frac{1}{2}y$ 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, $r(A) = \frac{1}{2}x + \frac{1}{2}y$ so that

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

[16] Technically, r is a *stochastic* choice function which satisfies $r(A) \in \text{conv}(A)$.

What happens if Relativity and IVC are violated? For simplicity, assume that all menus are convex sets, and thus Relativity(a) holds trivially. Then, we can find two menus A, B such that $(x, A) \succ (y, B)$ for all $x \in A$ and $y \in B$ which implies that the evaluator prefers the menu A to B regardless of the DM's preferences: she prefers "preferring *anything* in A" to "preferring *anything* in B". We can naturally infer that the evaluator cares only about outcomes, not preferences. Therefore, if $\succeq$ is a second-order preference, then for any A, B, there must be $x' \in A$ 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."

3. Representations

I use the standard definitions of preference representations and *affine* functions. Given a first-order preference $\succcurlyeq_1$ on X, I say the function v represents $\succcurlyeq_1$ when $v(x) \geq v(y)$ if and only if $x \succcurlyeq_1 y$. The function v is *affine* if $v(\alpha x + (1-\alpha)\, y) = \alpha v(x) + (1-\alpha)\, v(y)$ for all $x, y \in X$ and $\alpha \in [0,1]$. I say the function $U : \mathbb{C} \to \mathbb{R}$ represents $\succeq$ if $U(x, A) \geq U(y, B)$ is equivalent to $(x, A) \succeq (y, B)$, and it is affine if $U(\lambda(x, A) + (1-\lambda)\,(y, B)) = \lambda U(x, A) + (1-\lambda)\, U(y, B)$ for all $(x, A), (y, B) \in \mathbb{C}$ and $\lambda \in [0,1]$. A function $r : \mathbb{M} \to X$ is called a (stochastic) *choice function* if $r(A) \in \text{conv}(A)$ for all $A \in \mathbb{M}$. In all representations in this paper, the chosen element $r(A)$ is called *the reference of* A. Given a function $v : X \to \mathbb{R}$, let $\succcurlyeq_1^v$ denote the corresponding binary relation: that is, define $\succcurlyeq_1^v$ by $x \succcurlyeq_1^v y$ if and only if $v(x) \geq v(y)$. I define the affinity of a choice function as follows:

Definition 3 (Affine Choice Function)**.** A choice function $c : \mathbb{M} \to X$ is *affine with respect to* $\succcurlyeq_1$ if $c(\lambda A + (1-\lambda)\, B) \sim_1 \lambda c(A) + (1-\lambda)\, c(B)$ for $\lambda \in [0,1]$.

3.1. Second-order Preference Representation

I first provide the special case where $\succeq$ is a second-order preference. Although this is a special case of a more general theorem presented later, I state it separately because it highlights an especially insightful case and uses a different proof that helps clarify the core ideas.

Definition 4 (Second-order Preference Representation)**.** The EU representation of a *second-order preference over the acts of choosing* (SPA) is a pair (v, r) where v is an affine function of lotteries and r is an affine choice function with respect to

$\succcurlyeq_1^v$ such that $\succeq$ is represented by

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

The SPA representation has two components: a ranking v of outcomes and a choice function $r(\cdot)$ such that the utility of "preferring x to all else in A" is the difference in value between the option x chosen by the DM and the evaluator's reference of A. The function v is interpreted as the first-order preference that the evaluator believes the DM should ideally have. In this sense, v is referred to as the representation of the *ideal first-order preference*, which I formally define as follows:

Definition 5 (Ideal Preference)**.** When $\succeq$ is a second-order preference, the binary relation $\succcurlyeq_1^*$ on X is called *the evaluator's ideal first-order preference* if $x \succcurlyeq_1^* y$ whenever $(x, \{x, y\}) \succeq (y, \{x, y\})$.

Since $(x, \{x, y\}) \succeq (y, \{x, y\})$ means "preferring x to y" is preferred to "preferring y to x," I denote this relation by $\succcurlyeq_1^*$, written as $x \succcurlyeq_1^* y$, and read it as "x is *ideally (weakly) preferred* to y," meaning that the evaluator wants the DM to prefer x to y whenever he needs to make the binary choice between them. It is straightforward to see that $\succcurlyeq_1^*$ is $\succcurlyeq_1^c$—the consistent ranking of choices within menus, stated in Consistency axiom. Since $x \succcurlyeq_1^c y$ holds if and only if $(x, A) \succeq (y, A)$ for any A containing x, y, it also holds when $A = \{x, y\}$, in which case, we have $x \succcurlyeq_1^* y$.

I now state the theorem:

Theorem 1. *$\succeq$ is a second-order preference if and only if $\succeq$ has a SPA representation. Furthermore, the representation (v, r) is unique: (v', r') also represents $\succeq$ if and only if $v' = \alpha v + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$, and $v(r(A)) = v(r'(A))$ for each $A \in \mathbb{M}$.*

The "if" part of the existence result is straightforward. The complete proof of the "only if" part is in Appendix A. I provide a sketch below.

Proof Sketch. Since $\succcurlyeq_1^c$ and $\succcurlyeq_1^*$ are equivalent, I replace $\succcurlyeq_1^c$ with $\succcurlyeq_1^*$. First, note that Consistency grants the existence and uniqueness of the continuous affine function v representing $\succcurlyeq_1^*$ due to the standard EU theory. Moreover, by the result of Herstein and Milnor (1953), Axioms 1-3 are equivalent to the existence and uniqueness of a continuous affine function $V : \mathbb{C} \to \mathbb{R}$ representing $\succeq$.[17]

[17] Note that $\mathbb{C}$ is a mixture space.

As the second step, let $r(\cdot)$ be the choice function defined by Relativity: i.e., if $(x, A \cup \{x'_A\}) \sim (x, A)$ for all $x \in A$ and $(x'_A, A \cup \{x'_A\}) \sim (x'_A, \{x'_A\})$ for some $x'_A \in \text{conv}(A)$, then let $r(A) = x'_A$.[18] Then, $\succeq$, $\succcurlyeq_1^*$, and $r(\cdot)$ have the following relationship:

Lemma 1. $(x, A) \succeq (y, B) \iff \frac{1}{2}x + \frac{1}{2}r(B) \succcurlyeq_1^* \frac{1}{2}r(A) + \frac{1}{2}y$.

Proof of Lemma 1. Note that IVC and Relativity(b) implies $(r(A), A \cup \{r(A)\}) \sim (r(B), B \cup \{r(B)\})$. Then, $(x, A) \succeq (y, B)$ is equivalent to $(x, A \cup \{r(A)\}) \succeq (y, B \cup \{r(B)\})$ by Relativity(a). It follows that by Axiom 2, a coin toss between $(x, A \cup \{r(A)\})$ and $(r(B), B \cup \{r(B)\})$ is preferred to a coin toss between $(y, B \cup \{r(B)\})$ and $(r(A), A \cup \{r(A)\})$. That is, $\frac{1}{2}(x, A \cup \{r(A)\}) + \frac{1}{2}(r(B), B \cup \{r(B)\}) \succeq \frac{1}{2}(y, B \cup \{r(B)\}) + \frac{1}{2}(r(A), A \cup \{r(A)\})$. In other words,

$$\left(\tfrac{1}{2}x + \tfrac{1}{2}r(B), \tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\right) \succeq \left(\tfrac{1}{2}r(A) + \tfrac{1}{2}y, \tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\right)$$

which holds if and only if $\frac{1}{2}x + \frac{1}{2}r(B) \succcurlyeq_1^* \frac{1}{2}r(A) + \frac{1}{2}y$ by Consistency. This proves Lemma 1. *Q.E.D.*

Since v is an affine function representing $\succcurlyeq_1^*$, Lemma 1 implies $(x, A) \succeq (y, B)$ holds if and only if $v(x) - v(r(A)) \geq v(y) - v(r(B))$.[19]

The third step involves Lemma 3 in the Appendix, which shows that $r(\cdot)$ is affine with respect to $\succcurlyeq_1^*$. To see this briefly, assume that A and B are convex sets. Notice that due to Relativity, we have $(r(A), A) \sim (r(B), B) \sim (r(\lambda A + (1-\lambda)B), \lambda A + (1-\lambda)B)$. Then, the affinity of $r(\cdot)$ follows by Axiom 2. Lemma 2 in the Appendix shows that $r(\text{conv}(A)) \sim_1^* r(A) \sim_1^* r(A \cup \{r(A)\})$ for any compact menu A, which implies the affinity of $r(\cdot)$ is preserved even with any pair of non-convex menus A, B.

Next, I define a binary relation $\succeq_r$ on $\mathbb{M}$ as $A \succeq_r B$ if and only if $r(A) \succcurlyeq_1^* r(B)$. We can use the affinity of $r(\cdot)$ to conclude that $\succeq_r$ 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 f representing $\succeq_r$ (see Lemma 4 in the Appendix).[20] My axioms ensure that $f(\cdot) = v(r(\cdot))$.

[18] If x'_A is not unique, then choose any of them.

[19] In Appendix C, which is available upon request, I show that Lemma 1 is a generalized version of the axiom of *secondary preference* introduced in the book *The foundations of decision logic* by the philosopher Halldén (1980).

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

Lastly, define a function $V_{v,r} : \mathbb{C} \to \mathbb{R}$ by $V_{v,r}(x, A) := v(x) - v(r(A))$. Since I have shown that $v(r(\cdot))$ is a continuous affine function of sets, $V_{v,r}$ is also a continuous affine function and thus, $V_{v,r} = V$. This completes the proof of the existence.

The uniqueness of the SPA representation is analogous to the standard EU theory: it is unique up to positive affine transformations. When the pair (v, r) constitutes the SPA representation of $\succeq$, then (v', r') also represents $\succeq$ if and only if $v' = \alpha v + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$, and $v(r(A)) = v(r'(A))$ for each $A \in \mathbb{M}$. I leave the proof of uniqueness to the Appendix.

Q.E.D.

3.1.1. Behavioral Remarks

I present several behavioral intuitions behind the SPA representation.

Signs of Utilities. What does it mean to have a strictly positive utility of the act of choosing in the absence of outcome preferences? When the ideal preference is a ranking of normative values (e.g., long-term goals or moral values), a natural interpretation is that positive utilities are associated with a sense of pride in willingly making choices while negative utilities are tied to negative psychological experiences such as disappointment or guilt. To see this, suppose $V_{v,r}(x, A) > 0$. This implies that x should ideally be preferred to the reference of A—i.e., $x \succ_1^* r(A)$. Also, it means the act of choosing x from A is preferred to "the inability to give up anything"—i.e., $(x, A) \succ \phi$ for any vacuous choice ϕ. Therefore, conditional on x being chosen, the freedom of having the menu A is preferred to any exogenous outcome. Naturally, the evaluator finds value in the DM's willingness to choose x from A, which I interpret as feelings of pride. When the ideal preference is a ranking of immediate pleasure (e.g., the evaluator wants the DM to prefer what he enjoys), negative utilities are associated with mental costs such as self-control, while positive utilities capture the mental reward of indulgence—what we might call a guilty pleasure—such as when the DM finds psychological relief or enjoyment in choosing cake over salad.

Consistent Relativity. Theorem 1 also captures the idea that the value of a choice is relative to constraints. Note that we have $(x, A) \succeq (x, B)$ if and only if $r(B) \succcurlyeq_1^* r(A)$.[21] That is, conditional on preferences for the same outcome,

[21] To prove this, note that $(x, A) \succeq (x, B)$ is equivalent to $\frac{1}{2}x + \frac{1}{2}r(B) \succcurlyeq_1^* \frac{1}{2}x + \frac{1}{2}r(A)$ by Lemma 1. Since $\succcurlyeq_1^*$ is independent, this implies $r(B) \succcurlyeq_1^* r(A)$.

the evaluator prefers that the DM chooses it from the menu A rather than B if and only if the evaluator's reference of B has a higher value than that of A. Moreover, as a consequence of Consistency, we have $(x, A) \succeq (x, B)$ if and only if $(y, A) \succeq (y, B)$.[22] That is, Consistency not only implies a consistent ranking of choices within a menu, but also imposes consistency in the ranking of menus conditional on outcomes. Roughly speaking, if "the act of giving up A for x" is preferred to "giving up B for x", then "giving up A" is preferred to "giving up B" for any other common option.

Independence. Lastly, what is the intuition behind the affinity of $r(\cdot)$? The answer lies in Independence (Axiom 2). The axiom requires that when the DM chooses a contingency plan $\lambda x + (1-\lambda)\, y$ from the menu $\lambda A + (1-\lambda)B$ for some $\lambda \in (0,1)$, which constitutes the act of choosing $(\lambda x + (1-\lambda)\, y, \lambda A + (1-\lambda)B)$, the evaluator's reference of the menu $\lambda A + (1-\lambda)B$ is formed by weighing her references of A and B proportionally with the probability measure $(\lambda, 1-\lambda)$. This is in alignment with her ideal first-order preference which also satisfies Independence. If she believes the DM should ideally be an EU maximizer, then it is reasonable to assume that she evaluates his expected choice from his expected menu accordingly in a linear manner.[23]

More importantly, Independence implies that the evaluator's reference is independent of the DM's *personal* contingencies. Suppose the DM faces a non-singleton finite menu A with certainty. Even though his choices are limited to A, the evaluator 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 $\lambda \in (0,1)$—e.g., 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, the evaluator only observes

[22] I provide the proof here. For the sake of contradiction, suppose there are two menus A, B such that $(x, A) \succeq (x, B)$ and $(y, B) \succ (y, A)$ for some $x, y \in A \cap B$. There are two cases to consider: (i) $x \succcurlyeq_1^c y$ and (ii) $y \succ_1^c x$. For (i), we have $(x, A) \succeq (x, B) \succeq (y, B) \succ (y, A)$ due to Consistency. By Axiom 2, we must have $\frac{1}{2}(x, A) + \frac{1}{2}(y, B) \succ \frac{1}{2}(x, B) + \frac{1}{2}(y, A)$, which is equivalent to $(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}B) \succ (\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}B + \frac{1}{2}A)$. This contradicts Consistency. A similar contradiction is reached in the case of (ii).

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

either (x, A) or (y, A) if the DM 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 $r(\lambda A + (1-\lambda) A)$. By the affinity of $r(\cdot)$, her reference is unchanged: $r(\lambda A + (1-\lambda) A) \sim_1^* r(A)$.[24]

3.2. Representation of Preference over the Acts of Choosing

I now relax IVC so that the evaluator cares about both outcomes and the DM's preferences. In this case, $\succeq$ is no longer the evaluator's second-order preference, nor is $\succcurlyeq_1^c$ the ideal preference. Accordingly, $(x, A) \succeq (y, B)$ is no longer interpreted as a ranking of preferences.

Definition 6 (Representation of Preference over the Acts of Choosing)**.** The EU representation of a *preference over the acts of choosing* (PA) is a tuple (u, v, r) where u is an affine function of lotteries and (v, r) is a SPA representation such that $\succeq$ is represented by

$$U_{u,v,r}(x, A) := u(x) + v(x) - v(r(A)).$$

There are several preferences that can be inferred from a PA representation. First, note that $U_{u,v,r}(x, \{x\}) = u(x)$ for all $x \in X$. This means the function u represents the evaluator's first-order preference $\succcurlyeq_1^e$, which is additively separated from her SPA representation $V_{v,r}(x, A) = v(x) - v(r(A))$. This also implies that the function v represents her ideal preference $\succcurlyeq_1^*$, and the choice function $r(\cdot)$ is affine with respect to $\succcurlyeq_1^*$. Lastly, the ranking of choices within menus (i.e., $\succcurlyeq_1^c$) is no longer equal to $\succcurlyeq_1^*$. Rather, it is represented by the function $u + v$, which is the evaluator's linear *compromise* between her outcome preference and ideal preference.

The following is my main result:

Theorem 2. *$\succeq$ satisfies Axioms 1, 2, 3, Consistency and Relativity if and only if $\succeq$ has a PA representation. Furthermore, the representation (u, v, r) is unique: (u', v', r') also represents $\succeq$ if and only if $u' = \alpha u + \beta_1$ and $v' = \alpha v + \beta_2$ for some $\alpha > 0$ and $\beta_1, \beta_2 \in \mathbb{R}$, and $v(r(A)) = v(r'(A))$ for each $A \in \mathbb{M}$.*

[24] For instance, suppose Amy (the evaluator) has a 9-year-old child named Bob (the DM). For the upcoming weekend, Bob wants to play soccer (y), while Amy believes that he should prefer studying (x) to y. He claims that he will study if it rains during the weekend. That is, his choice is $(\lambda x + (1-\lambda) y, \lambda\{x, y\} + (1-\lambda)\{x, y\})$ where the probability of rain is $1-\lambda$. By the affinity of $r(\cdot)$, his personal plan contingent on the weather does not change how much Amy would be disappointed at his choice to do y instead of x. Thus, she will continue to evaluate $\lambda x + (1-\lambda) y$ based on the reference she has formed for the menu $\{x, y\}$.

Proof. See Appendix B. *Q.E.D.*

The sketch of proof is outlined in the following steps. First, I show that $\succeq$ is represented by a function U that additively separates the value of options from that of menus: i.e., $U(x, A) = h(x) - f(A)$ for some unknown affine functions h, f. This holds because even in the presence of outcome preferences, the consistent relativity is preserved: i.e., $(x, A) \succeq (x, B)$ if and only if $(y, A) \succeq (y, B)$. (See Lemma 5 in the Appendix.) Second, I show that the function h is a sum of two affine functions u, v. By the standard axioms (Axioms 1, 2, 3), we can define the affine function u by $u(x) := U(x, \{x\})$. Then, let $v(x) := U(x, X) - u(x) + \beta$ for some constant $\beta \in \mathbb{R}$ so that the function $u + v$ is an affine transformation of h. Lastly, I use Relativity to capture the function f. Since Relativity(a) implies $U(x, A \cup \{r(A)\}) = U(x, A)$, it follows that $f(A) = f(A \cup \{r(A)\})$. Then, since Relativity(b) implies $U(r(A), \{r(A)\}) = U(r(A), A \cup \{r(A)\})$, we have $f(A) = v(r(A))$ for all A. (See Appendix B for the complete proof.)

Let $\succcurlyeq_2^e$ denote the evaluator's second-order preference. To elaborate on how $\succcurlyeq_2^e$ is identified, I present the following result.

Corollary 1. *Suppose $\succeq$ has a PA representation (u, v, r). Define $\succcurlyeq_2^e$ by*

$$(x, A) \succcurlyeq_2^e (y, B) \iff (\tfrac{1}{2}x + \tfrac{1}{2}y, \tfrac{1}{2}A + \tfrac{1}{2}\{y\}) \succeq (\tfrac{1}{2}y + \tfrac{1}{2}x, \tfrac{1}{2}B + \tfrac{1}{2}\{x\}).$$

Then, (v, r) is the SPA representation of $\succcurlyeq_2^e$.

The corollary implies that $\succcurlyeq_2^e$ defined as above is a second-order preference (i.e., $\succcurlyeq_2^e$ satisfies Axioms 1, 2, 3, Consistency, Relativity and IVC). It is straightforward to verify this result using Theorem 2. By definition, we have

$$\begin{aligned}(x, A) \succcurlyeq_2^e (y, A) &\iff U_{u,v,r}(\tfrac{1}{2}x + \tfrac{1}{2}y, \tfrac{1}{2}A + \tfrac{1}{2}\{y\}) \geq U_{u,v,r}(\tfrac{1}{2}y + \tfrac{1}{2}x, \tfrac{1}{2}A + \tfrac{1}{2}\{x\}) \\ &\iff v(x) - v(r(A)) \geq v(y) - v(r(B)).\end{aligned}$$

Intuitively, "preferring x to all else in A" is preferred to "preferring y to all else in B" if and only if the act of choosing x from A is preferred to the act of choosing y from B, *conditional on achieving the same expected outcome*. Note that the expected outcomes of the two choices $(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}\{y\})$ and $(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}B + \frac{1}{2}\{x\})$ are the same. They differ only in their manifestations of the DM's preferences. The choice of $\frac{1}{2}x + \frac{1}{2}y$ from the menu $\frac{1}{2}A + \frac{1}{2}\{y\}$ implies that the DM chooses x from A willingly with probability 0.5; otherwise, y is

exogenously given. The opposite is true for $(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}B + \frac{1}{2}\{x\})$ which means he chooses y from B willingly.

Once we have a second-order preference $\succcurlyeq_2^e$, the evaluator's ideal first-order preference $\succcurlyeq_1^*$ can be naturally induced, as follows.

Corollary 2. *Suppose $\succeq$ has a PA representation (u, v, r), and $\succcurlyeq_2^e$ is defined as in Corollary 1. Define $\succcurlyeq_1^*$ by $x \succcurlyeq_1^* y$ if and only if $(x, \{x, y\}) \succcurlyeq_2^e (y, \{x, y\})$. Then, the following conditions are equivalent:*

(a) $x \succcurlyeq_1^* y$.

(b) $\left(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}\{y\}\right) \succeq \left(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}A + \frac{1}{2}\{x\}\right)$ *for all A with $x, y \in A$.*

(c) $v(x) \geq v(y)$.

It is also easy to verify this using Theorem 2.[25] By definition, x is ideally preferred to y if and only if "preferring x to y" is preferred to "preferring y to x", which is true when the act of choosing x over y is preferred to the act of choosing y over x, conditional on achieving the same expected outcome. Since $\succcurlyeq_2^e$ satisfies Consistency, this holds even when x and y are chosen from any other common menu A.

Lastly, the reference value $v(r(A))$ of any menu A can also be induced.

Corollary 3. *Suppose $\succeq$ has a PA representation (u, v, r). Then, given any $x \in A$, $y \in B$ and C with $x, y \in C$,*

$$v(r(A)) \geq v(r(B)) \iff (\tfrac{1}{2}x + \tfrac{1}{2}y, \tfrac{1}{2}C + \tfrac{1}{2}B) \succeq (\tfrac{1}{2}y + \tfrac{1}{2}x, \tfrac{1}{2}C + \tfrac{1}{2}A).$$

It states that the reference value of A is higher than that of B if and only if the act of choosing *something* from B is preferred to the act of *something* from A, *conditional on the act of choosing the same expected outcome*. Notice that both choices $(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}C + \frac{1}{2}B)$ and $(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}C + \frac{1}{2}A)$ constitute the act of choosing the same outcome $\frac{1}{2}x + \frac{1}{2}y$. The only difference is that B is given up by

[25] Formally, for any A with $x, y \in A$,

$$\begin{aligned} x \succcurlyeq_1^* y &\iff (x, \{x, y\}) \succcurlyeq_2^e (y, \{x, y\}) \\ &\iff U_{u,v,r}(\tfrac{1}{2}x + \tfrac{1}{2}y, \tfrac{1}{2}\{x, y\} + \tfrac{1}{2}\{y\}) \geq U_{u,v,r}(\tfrac{1}{2}y + \tfrac{1}{2}x, \tfrac{1}{2}\{x, y\} + \tfrac{1}{2}\{x\}) \\ &\iff v(x) - v(r(\{x, y\})) \geq v(y) - v(r(\{x, y\})) \\ &\iff v(x) - v(r(A)) \geq v(y) - v(r(A)) \\ &\iff (x, A) \succcurlyeq_2^e (y, A). \end{aligned}$$

the former while A is given up by the latter, both with probability 0.5. Hence, if the reference value of A is greater than that of B, then giving up B must be preferred to giving up A, and vice versa.

4. Paternalistic Attitude

The distinctive feature of the SPA representation, unlike any first-order preference representation, lies in the reference choice function $r(\cdot)$. This section explores the kinds of preferences this function captures, especially when the ideal preference is a ranking of normative values (e.g., long-term goals or moral values). In short, $r(\cdot)$ reflects the evaluator's *paternalistic attitude toward the act of choosing*: a paternalistic perspective assesses choices by how much they fall short of an ideal outcome, whereas a libertarian perspective evaluates them by how much they exceed the worst possible option. This contrast is analogous to the common metaphor: "Is the glass half full or half empty?"

Example 1. *Let h denote doing homework and t watching television. Consider parents (the evaluator) who not only want their child to do homework rather than watch television, but also want him to willingly make the right choice (i.e., they want their child to strictly prefer h to t). Formally, they exhibit*

$$(h, \{h,t\}) \succeq (h, \{h\}) \succ (t, \{t\}) \succeq (t, \{h,t\}).$$

Suppose they observed that the child chose homework a fraction $p \in [0,1]$ of the time and ended up choosing television $1-p$ of the time. That is, his choice from the menu $A = \text{conv}(\{h,t\})$ was $(ph + (1-p)\,t, A)$. At what threshold $\tilde{p}$ would the parents switch from enforcing homework (by banning television) to allowing him to keep deciding freely? The most paternalistic parents would be disappointed by any $p < 1$, and will enforce homework unless $p = 1$ (i.e., $\tilde{p} = 1$). Analogous to someone who sees the glass as "half empty", they focus on the shortfall $1-p$, and wants to prevent it even when the child never had willingly chosen t. This is identified with $(h, \{h\}) \sim (h, \{h,t\})$, which holds when $r(\{h,t\}) = h$. In contrast, the least paternalistic parents—like someone who sees the glass as "half full"—would be proud of any positive p. This is identified with $(t, \{t\}) \sim (t, \{h,t\})$ which is true when $t = r(\{h,t\})$. In this case, they would have the lowest threshold $\tilde{p}$ such that $(ph + (1-p)\,t, A) \succeq (h, \{h\})$ for all $p \geq \tilde{p}$. Thus, given the same outcome preference and ideal preference, a different choice function $r(\cdot)$ yields distinct paternalistic decisions, driven by fundamentally different perspectives on how the value of each

choice is assessed.

In general, paternalism refers to one's willingness to intervene in the DM's autonomy, restricting his options to promote his welfare. Prior experimental studies on paternalistic preferences focus on the case where a social planner (alternatively referred to as a policy maker or a choice architect) is concerned only with the DM's outcome, not with the act of choosing (see Ambuehl et al., 2021; Bartling et al., 2023). The parents in Example 1 are concerned with both the outcome and the quality of their child's choice, who face the trade-off: granting freedom inherently comes at the expense of opening the possibility of bad outcomes, while preemptively preventing bad outcomes restricts the DM's opportunity to make good choices.

4.1. Two Extreme Attitudes

More formally, I present two extreme attitudes called *pure paternalism* and *pure libertarianism*. Let $\succcurlyeq_2^e$ be the second-order preference induced by $\succeq$ as in Corollary 1, and let ϕ denote an arbitrary vacuous choice.

Axiom 7 (Pure Paternalism)**.** $\phi \succcurlyeq_2^e (x, A)$ *for all* $(x, A) \in \mathbb{C}$.

Axiom 8 (Pure Libertarianism)**.** $(x, A) \succcurlyeq_2^e \phi$ *for all* $(x, A) \in \mathbb{C}$.

To focus on the nature of second-order preferences, assume that the evaluator does not have a first-order preference, i.e., $\succeq = \succcurlyeq_2^e$. Axiom 7 states that a vacuous choice is weakly preferred to any act of choosing.[26] To see how this relates to the concept of paternalism, suppose Axiom 7 holds. Then, the evaluator's most preferable act of choosing is a vacuous choice—the state of not being able to willingly make any choice at all. Consequently, given any menu A, if there is even a slight chance that the DM will not prefer the most ideal option, then the evaluator would abandon his freedom of choice and enforce a vacuous choice, preventing the *act* of making a mistake. Hence, any non-singleton menu given to the DM is a potential loss for the evaluator. The opposite is true for Axiom 8: 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. Consequently, willingly making a choice is strictly preferred to a vacuous choice if there is even a slight chance of avoiding the least ideal option

[26] In terms of the preference $\succeq$ over the acts of choosing, Axiom 7 can be alternatively written as follows: $(\frac{1}{2}r(A) + \frac{1}{2}x, \frac{1}{2}A + \frac{1}{2}\{x\}) \succeq (\frac{1}{2}x + \frac{1}{2}r(A), \frac{1}{2}A + \frac{1}{2}\{r(A)\})\ \forall x \in A$. We can also write: $r(A) \succcurlyeq_1^* r(\{x\})$ for all $x \in A$.

on the menu.[27] The paternalistic parents in Example 1 has a purely paternalistic attitude toward the child's act of choosing. If the child's choice constitutes $ph + (1-p)\,t$ given any $p < 1$, then we have $(h, \{h\}) \succ_2^e (ph + (1-p)\,t, A)$. In contrast, the purely libertarian parents satisfy $(ph + (1-p)\,t, A) \succ_2^e (h, \{h\})$ for any $p > 0$.

The two extreme cases have the following representations.

Definition 7. An EU representation of *a purely paternalistic preference over the acts of choosing* (PPA) is a pair (u, v) of affine functions of lotteries such that $\succeq$ is represented by

$$U_{u,v}^{Pat}(x, A) = u(x) + v(x) - \max_{y \in A} v(y).$$

Definition 8. An EU representation of *a purely libertarian preference over the acts of choosing* (LPA) is a pair (u, v) of affine functions of lotteries such that $\succeq$ is represented by

$$U_{u,v}^{Lib}(x, A) = u(x) + v(x) - \min_{y \in A} v(y).$$

Corollary 4. *Suppose $\succeq$ has a PA representation. Then, Axiom 7 holds if and only if $\succeq$ has a PPA representation. Axiom 8 holds if and only if $\succeq$ has a LPA representation.*

Proof. Let $\succcurlyeq_1^*$ be the induced ideal preference defined as in Corollary 2. Define the best and worst lotteries on a menu A by $b_A \in \{x \in A : x \succcurlyeq_1^* y \ \forall y \in A\}$ and $w_A \in \{x \in A : y \succcurlyeq_1^* x \ \forall y \in A\}$. I first claim that if Axiom 7 holds, then $r(A) \sim_1^* b_A$ for all A. That is, $(b_A, A) \sim_2^e \phi$ for all A. For the sake of contradiction, suppose $\phi \succ_2^e (b_A, A)$ for some A. Then, by definition of b_A, we have $\phi \succ (b_A, A) \succcurlyeq_2^e (x, A)$ for all $x \in A$ which violates Relativity. Similarly, we can show that if Axiom 8 holds, then $r(A) \sim_1^* w_A$ for all A. Then, the desired result follows by Lemma 1. *Q.E.D.*

According to Corollary 4, a PPA (LPA) representation implies that the reference of each menu is the most (least) ideal option on the menu: for all menu A, we have $r(A) \in \arg\max_{y \in A} v(y)$ if $\succeq$ is purely paternalistic, and $r(A) \in \arg\min_{y \in A} v(y)$ if purely libertarian. In other words, a paternalistic perspective tends to evaluate a choice by its shortfall from the ideal outcome while a libertarian stance focuses on the improvement over the worst option.

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

4.2. Constant Paternalistic Attitude

Examining the two polar attitudes toward the act of choosing raises the question: What lies between them? One simple extension would be imposing a constant measure that captures a non-extreme paternalistic attitude—namely, one that weighs both the best and worst options with a fixed ratio.

Example 2. *Suppose the parents' preference in Example 1 is represented by a function* $U^{\delta}_{v,\alpha}$ *defined by*

$$U^{\delta}_{v,\alpha}(x,A) := \delta v(x) + v(x) - \underbrace{\left[\alpha \max_{y\in A} v(y) + (1-\alpha)\min_{y\in A} v(y)\right]}_{=v(r(A))} \tag{1}$$

where v *represents both their outcome preference and ideal preference,* $\delta > 0$ *is the relative weight on the child's outcome, and the value of the reference* $v(r(A))$ *takes the form of the* α*-maxmin utility function of sets of lotteries presented by Olszewski (2007) in his characterization of attitudes toward ambiguity.*[28] *The parameter* $\alpha \in [0,1]$ *can be interpreted as the constant paternalistic attitude toward the act of choosing. The two extreme cases in Corollary 4 are when* $\alpha \in \{0,1\}$.[29]

Let $A = \text{conv}(\{h,t\})$. *After observing the child's choices that constitute* $(ph + (1-p)\,t, A)$ *for some* $p \in [0,1]$, *the parents prefer granting freedom to enforcing homework time if* p *satisfies* $U^{\delta}_{v,\alpha}(ph + (1-p)\,t, A) \geq U^{\delta}_{v,\alpha}(h,\{h\})$. *Assuming* $v(h) > v(t)$, p *needs to be greater than the following threshold:*

$$\tilde{p} = \frac{\alpha + \delta}{1 + \delta}$$

which increases in both δ *and* α. *This means the parents are more reluctant to allow the child to choose from the menu* $\{h,t\}$ *either because the child's outcome is more important than his act of choosing, or because they are more paternalistic toward his act of choosing.*

[28] The α-maxmin utility function, also known as the "α-MEU" decision rule, can be traced back to the Hurwicz's criterion (Hurwicz, 1951), which has inspired a substantial body of research on attitudes toward ambiguity (see also Arrow and Hurwicz, 1972; Gilboa and Schmeidler, 1989). Olszewski (2007) developed a framework in which the basic object of choice is a set of lotteries, with Nature ultimately choosing the lottery from the chosen set. In his framework, the parameter $\alpha \in (0,1)$ is the agent's degree of optimism toward ambiguity. My model parallels this environment by featuring the evaluator as an entity that does not make the lottery choice.

[29] The function $U^{\delta}_{v,\alpha}$ is an affine function, and thus it is a special case of Theorem 2.

4.3. Menu-dependent Paternalistic Attitude and Temptation

The constant paternalistic attitude still has its flaws. In particular, it does not allow the evaluator to flexibly change its attitude depending on the choice situation. This limitation fails to reflect the fact that people generally feel little to no pride in avoiding an obviously bad outcome (e.g., choosing safe drinking water over a cup of gasoline), but an alcoholic might feel extremely proud of himself for choosing coffee over beer. That is, merely choosing a good option is not necessarily a manifestation of *good preferences*. A sense of pride—corresponding to a positive utility of preference manifestations—usually comes from making a "hard choice" which often involves a trade-off between competing values, goals, or desires. Psychological theories and evidence suggest that emotional experiences—such as pride or guilt—are intrinsically associated with discrepancies between an individual's actual self and his ideal self (see Markus and Nurius, 1986; Higgins, 1987; Tracy and Robins, 2004; Gilchrist et al., 2019).

To capture these psychological insights, I let the evaluator's paternalistic attitude to vary with two factors: (i) her expectation about the DM's choice, and (ii) how difficult she judges the menu to be. More specifically, the DM's choice invokes greater pride the more it exceeds the evaluator's expectation and the more she deems the menu difficult. For instance, suppose the menu contains a tempting option against the evaluator's ideal preference. As temptation grows stronger, she anticipates that the DM will endure a higher mental cost of processing this choice situation. In such cases, the evaluator would feel greater pride if the DM resists the temptation and be less disappointed if he gives in. This means her paternalistic attitude weakens when the menu presents a strong conflict between her ideal preference and the DM's temptation.

To model the evaluator's belief about the "difficulty" of the menu, I employ the definitions that resemble how temptations and self-control were defined by Gul and Pesendorfer (2001).[30] I say the evaluator thinks y tempts the DM more than x if $r(\{x\}) \succ_1^* r(\{x,y\})$; and she thinks the menu $\{x,y\}$ is difficult (or requires the DM's self-control) if $r(\{x\}) \succ_1^* r(\{x,y\}) \succ_1^* r(\{y\})$. If x is water and y is a cup of gasoline, then the evaluator would be purely paternalistic: i.e., $r(\{x\}) \sim_1^* r(\{x,y\})$. If the DM is an alcoholic and x, y are coffee and beer,

[30] In their setting, the agent chooses a menu of options, one of which will be chosen and consumed later. The temptation and self-control are identified as follows: (i) y is more tempting than x if the agent strictly prefers committing to x (i.e., $\{x\}$ is strictly preferred to $\{x,y\}$); and (ii) the menu $\{x,y\}$ requires self-control if the agent strictly prefers having the tempting option on his menu to exogenously consuming it (i.e., $\{x,y\}$ is strictly preferred to $\{y\}$).

respectively, then the evaluator's reference value of $\{x, y\}$ would be less than coffee, implying that the DM's act of choosing x over y evokes pride.

The following representation captures the above idea.

Definition 9 (Representation of Temptation-adjusted Preference over the Acts of Choosing)**.** The EU representation of a *temptation-adjusted preference over the acts of choosing* (TPA) is a tuple (u, v, τ) of affine functions of lotteries such that $\succeq$ is represented by

$$U^{Temp}_{u,v,\tau}(x, A) := u(x) + v(x) - \left[\max_{y \in A} \left\{v(y) + \tau(y)\right\} - \max_{z \in A} \tau(z)\right].$$

In the TPA representation, the function $V^{Temp}_{v,\tau}(x, A) := U^{Temp}_{u,v,\tau}(x, A) - u(x)$ is a SPA representation where the reference value function $v(r(\cdot))$ takes the form of Gul and Pesendorfer (2001)'s representation of preference for commitment.[31] The interpretation slightly differs. The functions τ and $v + \tau$ represent the evaluator's expectations about the DM's temptation and the DM's preference, respectively. Let y^*_A denote the maximizer of $v + \tau$. Then, the function $U^{Temp}_{u,v,\tau}(x, A)$ can be rewritten as

$$U^{Temp}_{u,v,\tau}(x, A) = u(x) + \underbrace{\left[v(x) - v(y^*_A)\right]}_{\text{Expectation gap}} + \underbrace{\left[\max_{z \in A} \tau(z) - \tau(y^*_A)\right]}_{\text{Anticipated mental cost}}.$$

The value difference $v(x) - v(y^*_A)$ between the DM's actual choice x and the expected choice y^*_A is called *the expectation gap*; the larger this gap, the more preferable the act of choosing (x, A) becomes. I call the term $\max_{z \in A} \tau(z) - \tau(y^*_A)$ *anticipated mental cost*, which measures how far the expected choice deviates from the menu's strongest temptation; the greater this deviation, the higher the mental cost the evaluator expects to incur when the DM faces this menu.

A few brief remarks follow. First, the TPA representation shows that the evaluator's pride in the DM's act of choosing (x, A) can arise (i.e., $V^{Temp}_{v,\tau}(x, A) > 0$) in two ways. The DM can make a choice that exceeds the evaluator's expectation. Yet, even when the expectation gap is zero, a difficult menu—one that incurs high anticipated mental cost—can still evoke pride. This is possible whenever the expected choice y^*_A is neither the most ideal nor the most tempting option.[32] By contrast, if the evaluator expects zero mental cost (because she

[31] More precisely, the pair (v, r^τ) is the SPA representation where the choice rule r^τ is defined by $r^\tau(A) = \{x \in \text{conv}(A) : v(x) = \max_{y \in A}\{v(y) + \tau(y)\} - \max_{z \in A} \tau(z)\}$.

[32] To see this, I describe the reference value as a function $R_{v,\tau}(A) := \max_{y \in A}\{v(y) + \tau(y)\} -$

believes the DM will pick the most tempting option), then a choice that meets that expectation produces neither pride nor disappointment.[33]

Second, the evaluator becomes purely paternalistic whenever she expects the DM to choose the most ideal option which happens to be the most tempting one. Suppose $v = \tau$. Then, it follows that $R_{v,\tau}(A) = \max_{y \in A} v(y)$. Conversely, she becomes purely libertarian whenever she expects the DM to choose the most tempting option which happens to be the least ideal option: e.g., when $v = -\tau$, we have $R_{v,\tau}(A) = \min_{y \in A} v(y)$. Otherwise, she adopts a non-extreme attitude.

To derive the TPA representation formally, I impose the following axiom which resembles the *set betweenness* axiom of Gul and Pesendorfer (2001).

Axiom 9 (Reference-betweenness). *$r(A) \succcurlyeq_1^* r(B)$ implies $r(A) \succcurlyeq_1^* r(A \cup B) \succcurlyeq_1^* r(B)$.*

The axiom states that when the reference value of A surpasses that of B, the reference value of $A \cup B$—i.e., $r(A \cup B)$—falls in between. Intuitively, the evaluator subjectively weighs the two references when formulating the reference of $A \cup B$ rather than interpreting $A \cup B$ in a fresh perspective.

I now state the result.

Theorem 3. *Suppose $\succeq$ has a PA representation. Then, Axiom 9 holds if and only if $\succeq$ has a TPA representation.*

Proof. The "if" part is straightforward. The proof of the "only if" part is a direct application of Gul and Pesendorfer (2001)'s Theorem 1. Let $\succcurlyeq_1^*$ be the induced ideal preference as in Corollary 3. Define a binary relation $\succeq_r^*$ on $\mathbb{M}$ as $A \succeq_r^* B$ if and only if $r(A) \succcurlyeq_1^* r(B)$. Then, we can show that $\succeq_r^*$ is complete, transitive, continuous and independent (see Lemma 4 in Appendix A). By Axiom 9, $A \succeq_r^* B$ implies $A \succeq_r^* A \cup B \succeq_r^* B$, which is the *set betweenness* axiom of Gul and Pesendorfer (2001). By their Theorem 1, there exists continuous affine functions $\bar{f}, \bar{v}, \bar{\tau}$ such that $\bar{f}(A) = \max_{y \in A} \bar{v}(y) + \bar{\tau}(y) - \max_{z \in A} \bar{\tau}(z)$ represents $\succeq_r^*$. By definition of $\succeq_r^*$, the ranking of singleton sets follows $\succcurlyeq_1^*$ and thus, $\bar{f}(\{x\}) = \bar{v}(x) = v(x)$ for all $x \in X$. It follows that $v(r(A)) = \bar{f}(A)$ for all $A \in \mathbb{M}$, which completes the proof. *Q.E.D.*

$\max_{z \in A} \tau(z)$. Then, $y_A^* \notin (\arg\max_{y \in A} v(y)) \cup (\arg\max_{y \in A} \tau(y))$ implies $R_{v,\tau}(A) < v(y_A^*)$ which yields $V_{v,\tau}^{Temp}(y_A^*, A) > 0$.

[33] Formally, $y_A^* \in \arg\max_{z \in A} \tau(z)$ implies $R_{v,\tau}(A) = v(y_A^*)$.

The identification of the ranking τ in the TPA representation also follows the way Gul and Pesendorfer (2001) identified their agent's temptation ranking. Choose any two options $a, b \in X$ such that $v(r(\{a\})) > v(r(\{a,b\})) > v(r(\{b\}))$. This means that the evaluator thinks that b tempts the DM more than a, and that the menu $\{a,b\}$ incurs the DM's mental cost. Then, choose any $\lambda \in (0,1)$ small enough so that $v(r(\{a\})) > v(r(\{a, \lambda x + (1-\lambda)\, b\})) > v(r(\{\lambda x + (1-\lambda)\, b\}))$ for all x. The function τ can be defined as follows:

$$\tau(x) := \frac{v(r(\{a,b\})) - v(r(\{a, \lambda x + (1-\lambda)\, b\}))}{\lambda}.$$

As a result,

$$\tau(x) \geq \tau(y) \iff v(r(\{a, \lambda y + (1-\lambda)\, b\})) \geq v(r(\{a, \lambda x + (1-\lambda)\, b\})).$$

That is, x is expected to tempt the DM more than y if and only if the evaluator's reference value of $\{a,b\}$ is greater when the tempting option b shifts toward y than when it shifts toward x.

In the parents-child context, the TPA representation can explain the parents' preference in the following example:

Example 3. *Suppose the parents in Example 1 have two children—Bob and Cindy. As before, at each decision point, each child had homework to complete but chose to do it only a fraction p of the time, opting instead to watch television the remaining $1-p$. Suppose Bob watched a regular television show denoted by t, whereas Cindy watched her all-time favorite program—say t^*. The parents do not differentiate between the outcomes of t and t^*, nor do they think that one should be ideally preferred to another. That is, $(t,\{t\}) \sim (t^*,\{t^*\})$ and $(t,\{t,t^*\}) \sim_2^e (t^*,\{t,t^*\})$. This means if they have a constant paternalistic attitude, the same threshold $\tilde{p}$ applies to both children: if $p < \tilde{p}$, the parents ban television for both; otherwise, each child chooses freely from $\{h,t\}$ and $\{h,t^*\}$, respectively. However, because the parents recognize that Cindy faced a stronger temptation, they believe that her decision to give up t^* entailed stronger self-control than giving up t. To account for this difference, they instead set separate thresholds: $\tilde{p}_{Bob} > \tilde{p}_{Cindy}$. As a result, for some p, they would enforce Bob's homework time but allow Cindy to decide on her own.*

4.3.1. The Best Act of Choosing vs. The Best Outcome

The novelty of the TPA representation is that the best act of choosing may only be feasible when the best outcome is unavailable. I say x is the best outcome

in A if (i) the outcome of x is preferred to every element in A, and (ii) the preference for x over all alternatives in A is itself preferred to any other preference over A. The best act of choosing is simply the most preferable act of choosing available. Consider the following ranking:

$$(y,\{y,z\}) \succ (x,\{x,y,z\}) \succ (y,\{x,y,z\}) \succ (z,\{x,y,z\}) \tag{2}$$

where x is the best outcome among the three alternatives, and yet $(y,\{y,z\})$ is the best act of choosing. Given a PA representation (u,v,r), the ranking (2) implies that (i) x maximizes both u and v, (ii) $u(y)+v(y) > u(z)+v(z)$, and (iii) $r(\cdot)$ satisfies

$$u(x)-u(y) < \underbrace{[v(y)-v(r(\{y,z\}))]}_{V_{v,r}(y,\{y,z\})} - \underbrace{[v(x)-v(r(\{x,y,z\}))]}_{V_{v,r}(x,\{x,y,z\})}. \tag{3}$$

(3) implies that the outcome value difference between x and y is smaller than the preference value difference between choosing y over z and choosing x over y and z.[34]

The constant paternalistic attitude discussed in Section 4.2—including the two extreme cases—cannot satisfy such conditions.[35] However, the TPA representation (u,v,τ) can rationalize it. Suppose $\tau(x) > \tau(z) > \tau(y)$. Then, we have $v(r(\{x,y,z\})) = v(x)$, implying that $(x,\{x,y,z\})$ is equivalent to a vacuous choice that yields only consumption value. In contrast, $v(r(\{y,z\}))$ is either $v(y)-v(z)$ or $\tau(z)-\tau(y)$, which captures the strictly positive value of preferring y to z. Intuitively, the value $v(r(\{x,y,z\}))$ moves closer to $v(z)$, the more z tempts the DM relative to y; by comparison, choosing x over y,z is an obvious decision since x is both tempting and ideal. Therefore, if the outcome difference between x and y is sufficiently small, then the value of preferring y to z can outweigh the value of consuming x.

In the menu preference framework, the ranking (2) implies that the menu $\{y,z\}$ is preferred to $\{x,y,z\}$ even though x is preferred to y and z. The models of temptation using menu preferences in the literature cannot rationalize the tendency to remove an option from a menu that the DM would otherwise have chosen. The standard models can only rationalize (2) by identifying x as temptation that is normatively inferior—e.g., against the agent's long-term goal

[34] (3) is derived directly from the inequality $U_{u,v,r}(y,\{y,z\}) > U_{u,v,r}(x,\{x,y,z\})$.

[35] Suppose $v(r(A)) = \alpha \max_{y\in A} v(y) + (1-\alpha)\min_{y\in A} v(y)$ for all A given $\alpha \in [0,1]$. Then, (3) implies $u(x)-u(y) < (1-\alpha)(v(y)-v(x))$ which contradicts that x maximizes both u,v.

(see Gul and Pesendorfer, 2001). However, in my example, x is both the most tempting and ideal option. Alternatively, the tendency to remove a normatively superior option from a menu had been rationalized by the agent's motivation to avoid a sense of guilt that stems from not choosing it (see Kopylov, 2012). This can only make sense here if either y or z would be chosen over x, thereby validating the anticipated guilt. Yet, clearly, x is the best outcome to choose.

In the Supplemental Appendix, I provide a detailed example in which an altruistic dictator—tasked with allocating resources between himself and a passive recipient—is strictly better off without a Pareto-optimal allocation added to his choice set because its presence prevents the manifestation of a preference for altruism.

5. Related Literature

5.1. Menu Preference

This paper is closely related to the choice-theoretic models of temptation and self-control that exploit choices over menus, initiated by Gul and Pesendorfer (2001). These models capture a rational agent's self-control concerns by assuming that he anticipates future temptation and therefore prefer committing to normative goals. The literature typically adopts a two-period structure: in the first period, the agent selects a menu from which he will make a consumption choice in the next period. A standard menu preference representation is of the form:

$$U^{Menu}_{u,v,r}(A) := \max_{x \in A} u(x) + v(x) - v(r(A))$$

where the function $v(r(\cdot))$ takes a special form in each prior model.

Clearly, the function $U^{Menu}_{u,v,r}$ represents the menu preference that is naturally induced by a SP representation $U_{u,v,r}$ in Theorem 2. Let $\succeq_M$ be the induced menu preference defined by $A \succeq_M B$ if and only if there exists $x \in A$ such that $U_{u,v,r}(x, A) \geq U_{u,v,r}(y, B)$ for all $y \in B$. Then, $U^{Menu}_{u,v,r}$ represents $\succeq_M$. In fact, any menu preference representation in the literature—that (i) is affine, (ii) is not subject to uncertainty about preferences, and (iii) yields consistent ex post outcome-choice behavior (i.e., WARP is satisfied)—is a special case of $U^{Menu}_{u,v,r}$.[36]

[36] Several studies on menu preferences presented potential reasons for violations of WARP: Dekel et al. (2009) and Stovall (2010) introduced uncertainty about temptation which motivated the agent to make inconsistent choices, whereas Noor and Takeoka (2015) presented menu-dependent costs of self-control. In the presence of ambiguity, the Independence axiom was

For instance, $\succeq_M$ induced by a PPA representation $U^{Pat}_{u,v}$ is exactly the preference for commitment introduced by Gul and Pesendorfer (2001). (See also Sarver, 2008; Kopylov, 2012; Dillenberger and Sadowski, 2012; Saito, 2015; Kopylov and Noor, 2018; Noor and Ren, 2023)

Hence, the model of second-order preference applied to menu preferences suggests that the agent chooses menus for himself anticipating the outcome as well as the value of his own preference manifestations. In this sense, one's paternalistic stance toward his own acts 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).[37]

Why adopt preferences over the acts of choosing, particularly when menu preferences seem more accessible in empirical or experimental settings? The non-comparability problem raised by Bernheim et al. (2024) underscores the importance of understanding how people rank the acts of choosing. While existing studies on menu preferences have been foundational in capturing self-control and commitment behavior, they do not formally offer such insights. Yet, individuals value menus because they care about the acts of choosing, not the other way around. In this light, this paper offers a conceptual advantage, and fills the theoretical gap by introducing the framework for second-order preferences within standard economic theory—formalizing an idea that existing approaches have so far treated only implicitly.

Importantly, my model is not limited to the conflict between normative goals and temptation that dominates the menu preference literature. While the standard models of menu preferences interpret the function u as *the normative ranking*, v as *the temptation ranking*, and the term $v(x) - v(r(A))$ as the self-control cost function, my framework treats u more generally as the agent's first-order outcome preference, and the term $v(x) - v(r(A))$ as the second-order preference. Depending on the context, v may correspond to the temptation ranking, a normative standard, or a social ideal.[38] This generalization allows

often relaxed and thus, the representations were not affine (see Ahn, 2008).

[37] 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).

[38] Note that the ideal preference does not necessarily reflect a sense of morality or better judgments. The philosopher Mele (1992) pointed out that self-control is not always exercised to

the model to accommodate a broader range of internal conflicts and evaluative motivations.

Furthermore, unlike existing models that focus on the value of self-restrictions, my framework can just as naturally represent the value of restricting others' choices. For example, a parent may evaluate a child's past choices and, in the next period, select a menu that encourages certain patterns of behavior—reversing the standard dynamics of menu choices. This flexibility broadens the model's applicability beyond personal self-control to include social contexts like paternalistic decision-making, institutional design, or even political choice architecture.

5.2. Reference-dependence

The standard reference-dependence model by Kőszegi and Rabin (2006) captures individuals' tendency to assess an outcome of a choice in contrast with a reference. The PA representation (u, v, r) has two imperative conceptual departures from the standard model, which lie in the origin of the references and how the agent perceives the menu.

Consider the following modification of the PA representation:

$$U^{\ell}_{u,v,r}(x, A) := u(x) + \ell\Big(v(x) - v(r(A))\Big) \tag{4}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is some continuous real-valued function. If ℓ is linear, then $U^{\ell}_{u,v,r}$ is exactly the PA representation. $U^{\ell}_{u,v,r}$ is exactly Kőszegi and Rabin (2006)'s model if ℓ is the universal gain-loss function in their model, which reflects the loss-averse attitude.[39] In their *personal equilibrium*, the agent's reference of each menu is her expectation about 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. In contrast, in my model, the reference stems from the second-order preference, and is formed for every possible menu. Therefore, as long as she is able to observe her present menu, she updates her reference in the event of an unanticipated

motivate moral actions. 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.

[39] To precisely reproduce Kőszegi and Rabin (2006)'s model, I need to further assume that (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) the choice function r is affine with respect to v, and it reflects the expected choice; and (iv) $u = v$.

menu. The matter at hand would rather be whether or not the menu can be observed correctly.

More importantly, the reference is not necessarily the agent's expectation and thus, even without an unanticipated menu and loss aversion (i.e., when ℓ is linear), the second term of (4) can affect her behavior. If the evaluator's reference is her correct expectation about what the DM will do for each menu, then she expects a vacuous choice from any menu, implying that her second-order preference is trivial.[40] Yet, what the evaluator wants the DM 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, the evalautor's reference can be regarded as her personal wish, or *subjective expectation* of the DM's choice.

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.

6. Discussion

This section discusses broader implications of the model. I address challenges in welfare assessment, the signaling value of observed choices, and the relationship between my model and a more general notion of second-order preference.

6.1. Higher-order Non-comparability Problem

The concept of higher-order preferences suggests that the design of welfare policies is influenced by the social planner's own preferences. Consequently,

[40] 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.

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

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

This example highlights how the non-comparability problem presented by Bernheim et al. (2024) extends to a higher-order level. The social planner (the parent, in this case) faces a meta-preference challenge, balancing the DM's immediate pleasure (e.g., the child's joy in playing with the smartphone) against his future welfare (e.g., cultivating responsibility through guilt). This parallels the standard tension in the temptation literature, where a DM may struggle between short-term indulgence and long-term goals. Evidently, this tension is not limited to the DM alone; it also exists in the social planner's interpretation of what welfare entails.

Furthermore, the combination of choice and policy data alone cannot recover the DM's preferences if the previous policy maker's goals remain unclear. Suppose the father observes both the mother's welfare policy (instructing the child to clean) and the child's choice (playing over cleaning). Assume he knows that the mother had sufficient data on the child's choices and mental states when she implemented her policy. Additionally, if he knows that her goal was to promote the child's immediate well-being, then he may conclude that the child feels pride in willingly resisting the mother's request. However,

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

if the father thinks the mother intended to foster long-term welfare, he might conclude that the child feels guilty for not cleaning—a completely opposite interpretation.

This higher-order non-comparability problem suggests that welfare analysis thus requires not only an understanding of the DM's first- and second-order preferences but also explicit knowledge of the social planner's preference over the DM's second-order preferences.

6.2. Signaling Motives

My framework also sheds light on the social signaling value of choice. When choices are publicly observed, the act of choosing from a given menu becomes informative of the DM's underlying preferences—as shaped and constrained by the menu itself. Thus, individuals may derive utility from aligning their choices with their ideal preferences because of the reputational implications such choices signal to others. For example, consider a dictator who chooses an allocation $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ of wealth between himself (who gets x_1) and a passive recipient (who gets x_2). Let $x_f = (5,5)$ be a fair allocation, $x_s = (6,4)$ a selfish one, and $x_p = (7,7)$ a Pareto-improvement over both. Suppose he wishes to signal that he values altruism. Then, he must choose x_f from $\{x_f, x_s\}$ so that the act of giving up the selfish option is observable. This choice credibly demonstrates a preference for fairness or altruism. In contrast, if the menu is $\{x_f, x_s, x_p\}$, choosing x_f no longer signals altruism. Since x_f is Pareto-dominated by x_p, such a choice appears irrational. And if he chooses x_p, the observer cannot tell whether he prefers being altruistic to being selfish because he is neither giving up being altruistic nor giving up being selfish. Thus, the signaling value of a choice depends on the opportunity cost it implies. (I analytically demonstrate this dictator's problem in more detail in the Supplemental Appendix.)

My model captures this signaling motive *without requiring that individuals explicitly value others' beliefs*; rather, signaling emerges naturally when the value of an act of choosing depends on how it reflects preferences under given constraints.

6.3. General Second-order Preference

This paper focused on a specific kind of second-order preference: one that ranks the acts of choosing a single option from a menu. This section outlines a more

general notion, and shows how two assumptions restrict attention to the narrower class studied in this paper.

I first describe the first-order preferences in more detail. Given any $A \in \mathbb{M}$, let $\mathbb{P}(A)$ denote the set of all complete and transitive binary relations (i.e., the standard weak preference relations) on A. The elements $P, Q \in \mathbb{P}(A)$ are subsets of $A \times A$, and are called first-order preferences. For any $P \in \mathbb{P}(A)$ and $x, y \in A$, the notation xPy is used to indicate that $(x, y) \in P$, which is read as "x is weakly preferred to y." I use the standard definitions of strict preferences and indifference associated with each P.[42] For example, if $A = \{x, y\}$, then $\mathbb{P}(A) = \{P, Q, R\}$ such that

$$\begin{aligned} P: &\quad x \text{ is strictly preferred to } y; \\ Q: &\quad y \text{ is strictly preferred to } x; \\ R: &\quad x \text{ is indifferent to } y. \end{aligned}$$

A second-order preference in general is a binary relation $\succcurlyeq_2$ 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 the DM could potentially exhibit, each corresponding to a different choice situation he may encounter. Given $P \in \mathbb{P}(A)$ and $Q \in \mathbb{P}(B)$, I say the evaluator prefers P to Q, independent of outcomes, if she prefers "the action induced by P given the menu A" to "the action induced by Q given B." The nature of the model can vary widely depending on how we define what "an action induced by P" refers to. This paper focuses on the case where the action of the evaluator's interest pertains solely to the DM's acts of choosing a single option from a menu.[43] In this sense, the second-order preference in this paper can be referred to as the *intrinsic* preference for the acts of choosing. Yet, it is generally not feasible to infer a unique $\succcurlyeq_2$ from the model primitive $\succeq$ since $\succeq$ is on $\mathbb{C}$ while $\succcurlyeq_2$ is on $\mathcal{P}$.

[42] I say x is *strictly preferred* to y if xPy and not yPx; x is *indifferent* to y if xPy and yPx. Since $\mathbb{P}(A)$ does not contain incomplete binary relations, I rule out incomplete preferences or incomparable outcomes.

[43] The action induced by a preference in general can refer to many different behaviors: the act of consuming n options from a menu where $n \in \{1, 2, ...\}$, declaring indifference among some options, declaring the least favorite option on the menu, or revealing one's preference over the menu entirely.

There are two underlying assumptions that allow the second-order preference in this paper to be identified by choices over the acts of choosing a single option. First, the evaluator does not care about what the DM might have chosen if he had been presented with a different menu. Second, she does not care about any other option the DM was willing to choose, apart from the one he actually chose. For example, suppose the DM chose x from the menu $\{x, y, z\}$ on two separate occasions (e.g., periods 1 and 2). Let $c_1 = (x, \{x, y, z\})$ and $c_2 = (x, \{x, y, z\})$ denote his act of choosing in periods 1 and 2, respectively. Suppose the evaluator found out that the DM's second favorite option in $\{x, y, z\}$ was y in period 1 and z in period 2. The first assumption is that the evaluator is indifferent—i.e., $c_1 \sim c_2$. Consider another scenario: the evaluator found out that the DM was indifferent between x and y in period 1, but became indifferent between x and z in period 2. The second assumption still requires $c_1 \sim c_2$.

Formally, consider extending $\succeq$ to $\mathcal{P}$. Suppose the DM's menu A is fixed. For each $P \in \mathbb{P}(A)$, define the choice rule $\mathcal{C}_P(A) := \{x \in A : xPy \quad \forall y \in A\}$. The two assumptions are as follows:

Assumption 1 (Preference for Revealed Preference)**.** *If $\mathcal{C}_P(A) = \mathcal{C}_Q(A)$ for some $P, Q \in \mathbb{P}(A)$, then $P \sim Q$.*

Assumption 2 (No Preference for Indifference)**.** *Suppose $\mathcal{C}_P(A)$ and $\mathcal{C}_Q(A)$ form a partition of $\mathcal{C}_R(A)$ for some $P, Q, R \in \mathbb{P}(A)$. Then, $P \succeq Q$ implies $P \sim R \succeq Q$.*

By Assumption 1, the evaluator associates a preference relation only with its contribution to the DM'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 the DM orders the non-favorite options on a menu. Consider $A = \{x, y, z\}$ and $P, P', Q, Q' \in \mathbb{P}(A)$ where

$$\begin{aligned} P: &\quad x \text{ is strictly preferred to } y, \text{ and } y \text{ is strictly preferred to } z; \\ P': &\quad x \text{ is strictly preferred to } y, \text{ and } z \text{ is strictly preferred to } y; \\ Q: &\quad y \text{ is strictly preferred to } x, \text{ and } x \text{ is strictly preferred to } z; \\ Q': &\quad y \text{ is strictly preferred to } x, \text{ and } z \text{ is strictly preferred to } x. \end{aligned}$$

Then, Assumption 1 requires $P \sim P'$ and $Q \sim Q'$ since

$$\{x\} = \mathcal{C}_P(A) = \mathcal{C}_{P'}(A) \neq \mathcal{C}_Q(A) = \mathcal{C}_{Q'}(A) = \{y\}.$$

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 Assumption 1 alone, the evaluator also regards the act of announcing indifference as a valid external behavior that corresponds to a preference $P \in \mathbb{P}(A)$.

I say the evaluator has *no preference for indifference* if Assumption 2 holds. In other words, she does not intrinsically favor or disfavor the DM's indifference among some options. I assume that the DM can choose multiple options when he finds two or more options indifferent, in which case, the evaluator makes the final choice. 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 $\mathbb{P}(A) = \{P, Q, R\}$ where $\mathcal{C}_P(A) = \{x\}$ and $\mathcal{C}_Q(A) = \{y\}$ form a partition of $\mathcal{C}_R(A) = \{x, y\}$.[44] Suppose the evaluator wants the DM to *want* to choose x from A (i.e., $P \succ Q$). Then, by Assumption 2, we have $P \sim R$ 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. This brings $P \sim R \succ Q$. When $P \sim Q$, she simply does not care whether the DM wants to choose x or y in which case, we have $P \sim R \sim Q$.[45]

Assumptions 1-2 imply that the evaluator's preference $\succeq$ extended to $\mathcal{P}$ can be characterized by a ranking of the acts of choosing, and thus inferring a second-order preference $\succcurlyeq_2$ on $\mathcal{P}$ from $\succeq$ on $\mathbb{C}$ is possible. In other words, if a real-valued function $V : \mathbb{P}(A) \to \mathbb{R}$ represents $\succcurlyeq_2$, then we can preserve all variations with a function $\overline{V} : A \to \mathbb{R}$ defined by $\overline{V}(x) = V(P)$ for all P such that $\mathcal{C}_P(A) = \{x\}$.[46] Formally, the set $\mathbb{P}_s(A) = \{P \in \mathbb{P}(A) : |\mathcal{C}_P(A)| = 1\}$ is the set of preferences inducing a single option on each menu. Then, $\mathcal{P}_s = \bigcup_{A \in \mathbb{M}} \mathbb{P}_s(A)$ is the set of all preferences across all menus inducing a single choice.

[44] Since A is nonempty and compact, $\mathcal{C}_P(A)$ is nonempty for all $P \in \mathbb{P}(A)$. Thus, if $\mathcal{C}_P(A)$ and $\mathcal{C}_Q(A)$ form a partition of $\mathcal{C}_R(A)$, then both $\mathcal{C}_P(A)$ and $\mathcal{C}_Q(A)$ are always proper subsets of $\mathcal{C}_R(A)$.

[45] In Appendix D, which is available upon request, I discuss how my model can be extended to allow for preferences for (or against) indifference, by relaxing Assumption 2. The evaluator might intrinsically like or dislike indifference. When we are making a decision as a group (e.g., 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 (e.g., picky eaters) 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.

[46] The fact that we can define $\succeq$ on A instead of $\mathbb{P}(A)$ implies that under Assumptions 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 in philosophy on second-order preferences. I discuss this in more detail in Appendix C, which is available upon request.

Observation 1. *$\succeq$ defined on $\mathcal{P}$ satisfies Assumptions 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 observation is true because we can define $\mathbb{C}$ alternatively by

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

which is essentially equal to the original definition of $\mathbb{C}$.

7. Conclusion

I show that second-order preference—a concept long confined to philosophical debate—can be identified through choice. This framework bridges philosophical insights and economic theory within the standard EU paradigm, relying exclusively on observable decisions rather than self-reported well-being data or assuming belief-dependent utilities. It thereby opens the door to direct empirical testing and policy analysis in contexts where the quality of choosing itself matters. Future research can leverage this approach to study autonomy, signaling motives, and menu-dependent paternalism.

Appendix

A. Proof of Theorem 1

To complete the proof of the "only if" part for the existence result, I first derive the following lemma:

Lemma 2. *$r(\text{conv}(A)) \sim_1^c r(A) \sim_1^c r(A \cup \{r(A)\})$ for all $A \in \mathbb{M}$.*

Proof of Lemma 2. Consider the following choice

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

where $A_n = \sum_{s=1}^{n} \lambda_s A$ and $\lambda_s = \frac{1}{n}$ for all $s = 1, ..., n$. Then, by the result known as the Shapley-Folkman theorem (see Emerson and Greenleaf, 1969; Starr, 1969), A_n converges to $\text{conv}(A)$ in the Hausdorff metric, and thus, (x, A_n) converges to $(x, \text{conv}(A))$. Because $\succeq$ has an affine representation, we have

$(x, A) \sim (x, A_n)$ for all $n \in \mathbb{N}$. Then, by Axiom 3, we have $(x, \text{conv}(A)) \sim (x, A)$, which, by Lemma 1, is equivalent to $\frac{1}{2}x + \frac{1}{2}r(A) \sim_1^c \frac{1}{2}r(\text{conv}(A)) + \frac{1}{2}x$ for all $n \in \mathbb{N}$. Since $\succcurlyeq_1^c$ is independent, this gives us $r(\text{conv}(A)) \sim_1^c r(A)$. To show $r(A) \sim_1^c r(A \cup \{r(A)\})$, I use the Shapley-Folkman theorem again to conclude $\sum_{s=1}^n \lambda_s A \cup \{r(A)\}$ converges to $\text{conv}(A)$. Note that $\text{conv}(A \cup \{r(A)\}) = \text{conv}(A)$ since $r(A) \in \text{conv}(A)$. This implies $(r(A), A \cup \{r(A)\}) \sim (r(A), \text{conv}(A))$ which, by Lemma 1, means $r(\text{conv}(A)) \sim_1^c r(A \cup \{r(A)\})$. This completes the proof of Lemma 2. *Q.E.D.*

Using Lemma 2, I prove that $r : \mathbb{M} \to X$ is affine with respect to $\succcurlyeq_1^c$:

Lemma 3. *$r(\lambda A + (1-\lambda) B) \sim_1^c \lambda r(A) + (1-\lambda) r(B)$ for $\lambda \in [0,1]$.*

Proof of Lemma 3. I first consider convex menus. For any convex menus A and B, we have $(r(A), A) \sim (r(B), B)$ by Relativity. Then, by Axiom 2, we have $(r(A), A) \sim (\lambda r(A) + (1-\lambda) r(B), \lambda A + (1-\lambda) B)$. Relativity also gives us $(r(A), A) \sim (r(\lambda A + (1-\lambda) B), \lambda A + (1-\lambda) B)$. By Consistency, we can conclude $r(\lambda A + (1-\lambda) B) \sim_1^c \lambda r(A) + (1-\lambda) r(B)$.

For any nonempty compact menus $A, B \in \mathbb{M}$, Lemma 2 allows us to conclude that, for any vacuous choice ϕ, we have $\phi \sim (r(A), A \cup \{r(A)\}) \sim (r(B), B \cup \{r(B)\}) \sim (r(A), \text{conv}(A)) \sim (r(B), \text{conv}(B))$ where the first two indifference relations are due to Relativity and the last two relations are due to Lemma 2. Then, by Axiom 2, it follows that $(r(A), A \cup \{r(A)\}) \sim (\lambda r(A) + (1-\lambda) r(B), \lambda \text{conv}(A) + (1-\lambda) \text{conv}(B))$ for any $\lambda \in [0,1]$. Moreover, by Relativity, $(r(A), A \cup \{r(A)\})$ is also indifferent to $(r(\lambda A + (1-\lambda) B), \lambda A + (1-\lambda) B \cup \{r(\lambda A + (1-\lambda) B)\})$, which, by Axiom 3, is indifferent to

$$\lim_{n\to\infty} \sum_{s=1}^{n} \lambda_s \left(r(\lambda A + (1-\lambda) B), \lambda A + (1-\lambda) B \cup \{r(\lambda A + (1-\lambda) B)\}\right).$$

By the Shapley-Folkman theorem, the limit is $(r(\lambda A + (1-\lambda) B), \text{conv}(\lambda A + (1-\lambda) B))$. Since $\text{conv}(\lambda A + (1-\lambda) B) = \lambda \text{conv}(A) + (1-\lambda) \text{conv}(B)$, it follows that $r(\lambda A + (1-\lambda) B) \sim_1^c \lambda r(A) + (1-\lambda) r(B)$ by Consistency. *Q.E.D.*

Next, I define a binary relation $\succeq_r$ on $\mathbb{M}$ as $A \succeq_r B$ if and only if $r(A) \succcurlyeq_1^c r(B)$.

Lemma 4. *$\succeq_r$ is complete, transitive, continuous and independent.*

Proof of Lemma 4. Since $\succcurlyeq_1^c$ is complete and transitive, so is $\succeq_r$. For continuity, since $\mathbb{M}$ is a topological space, it is sufficient to show that $A \succ_r C \succ_r B$ implies

that there are $\alpha, \beta \in (0,1)$ such that $\alpha A + (1-\alpha) B \succ_r C \succ_r \beta A + (1-\beta) B$. Since $\succcurlyeq_1^c$ is continuous, there are $\alpha, \beta \in (0,1)$ such that $\alpha r(A) + (1-\alpha) r(B) \succ_1^c r(C) \succ_1^c \beta r(A) + (1-\beta) r(B)$. By Lemma 3, we have $r(\alpha A + (1-\alpha) B) \succ_1^c r(C) \succ_1^c r(\beta A + (1-\beta) B)$ which is equivalent to our desired result. For Independence, suppose $A \succ_r B$ or equivalently, $r(A) \succ_1^c r(B)$. Since $\succcurlyeq_1^c$ is *independent*, $\lambda \in (0,1)$ implies $\lambda r(A) + (1-\lambda) r(C) \succ_1^c \lambda r(B) + (1-\lambda) r(C)$. By Lemma 3, it implies $r(\lambda A + (1-\lambda) C) \succ_1^c r(\lambda B + (1-\lambda) C)$. By definition, we have $\lambda A + (1-\lambda) C \succ_r \lambda B + (1-\lambda) C$. *Q.E.D.*

By the result of Herstein and Milnor (1953), Lemma 4 holds if and only if $\succeq_r$ has a unique continuous affine representation $f : \mathbb{M} \to \mathbb{R}$. By construction, the ranking of singleton sets follows $\succcurlyeq_1^c$ and thus, $f(\{x\}) = v(x)$ for all $x \in X$. Since $r(A) \succcurlyeq_1^c r(B)$ is equivalent to $A \succeq_r B$ which is represented by $f(A) \geq f(B)$, we conclude $f(A) = v(r(A))$ for all $A \in \mathbb{M}$.

As the final step, $V_{v,r}(x, A) = v(x) - f(A)$ is a continuous function since both v and f are continuous. It is easy to verify that $V_{v,r}$ is affine. This completes the proof of the "only if" part of the existence result in Theorem 1.

For the uniqueness result, let ϕ denote any vacuous choice. For the "if" part, suppose $v' = \alpha v + \beta$ and $r'(A) \sim_1^c r(A)$ for all $A \in \mathbb{M}$. Then $(x, A) \succeq (y, B)$ is equivalent to

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

where the last equivalence is due to $r'(A) \sim_1^c r(A)$.

To prove the "only if" part, suppose (v, r) and (v', 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) $r'(A) \sim_1^c r(A)$ for all $A \in \mathbb{M}$. (i) is the result of the standard EU theory. For (ii), I first consider a convex menu A. In this case, the proof is trivial: we always have $r(A), r'(A) \in A$, and thus we have $(r(A), A) \sim \phi \sim (r'(A), A)$, which implies $r(A) \sim_1^c r'(A)$. Now suppose A is any nonempty compact set. Since Lemma 2 implies $r(A) \sim_1^c r(\mathrm{conv}(A))$ and $r'(A) \sim_1^c r'(\mathrm{conv}(A))$, we have $(r(A), \mathrm{conv}(A)) \sim (r'(A), \mathrm{conv}(A)) \sim \phi$ by Relativity and Lemma 1. Then, $r(A) \not\sim_1^c r'(A)$ contradicts Consistency. This completes the proof of Theorem 1. *Q.E.D.*

B. Proof of Theorem 2

I first prove the following lemma.

Lemma 5. *Axioms 1, 2, 3, and Consistency imply* $(x, A) \succeq (x, B)$ *if and only if* $(y, A) \succeq (y, B)$.

Proof of Lemma 5. For the sake of contradiction, suppose there are two menus A, B such that $(x, A) \succeq (x, B)$ and $(y, B) \succ (y, A)$ for some $x, y \in A \cap B$. There are two cases to consider: (i) $x \succcurlyeq_1^c y$ and (ii) $y \succ_1^c x$. For (i), we have $(x, A) \succeq (x, B) \succeq (y, B) \succ (y, A)$ due to Consistency. By Axiom 2, we must have $\frac{1}{2}(x, A) + \frac{1}{2}(y, B) \succ \frac{1}{2}(x, B) + \frac{1}{2}(y, A)$, which is equivalent to $(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}B) \succ (\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}B + \frac{1}{2}A)$. This contradicts Consistency. A similar contradiction is reached in the case of (ii). *Q.E.D.*

Axioms 1, 2, 3 allow us to find a unique affine function U representing $\succeq$, while Consistency grants a unique affine function h representing $\succcurlyeq_1^c$. By Consistency and Lemma 5, we know that U is of the additively separable form $U(x, A) = h(x) - f(A)$ for some unknown affine function $f : \mathbb{M} \to \mathbb{R}$. Choose any lottery—say, the worst lottery x_w that satisfies $(x, X) \succeq (x_w, X)$ for all $x \in X$. Define four real-valued functions u, h_0, v of lotteries, and f of sets, by

$$\begin{aligned} u(x) &:= U(x, \{x\}); \\ h_0(x) &:= U(x, X) - U(x_w, X); \\ v(x) &:= h_0(x) - u(x); \\ f_0(A) &:= h_0(x) - U(x, A). \end{aligned}$$

By construction, we have $f_0(A) = h_0(x) - U(x, A) = h(x) - h(x_w) - [h(x) - f(A)] = f(A) - h(x_w) = h_0(y) - U(y, A)$ for any $x, y \in A$. Also, we have

$$f_0(\{x\}) = h_0(x) - U(x, \{x\}) = h_0(x) - u(x) = v(x) \quad \forall x \in X. \tag{5}$$

Finally, we have $U(x, A)$ is equal to $u(x)$ if A is a singleton, and $u(x) + v(x) - f_0(A)$ otherwise. Using these definitions, we have the following result:

Lemma 6. *The functions* u, h_0, v *and* f_0 *are affine.*

Proof of Lemma 6. Axioms 1, 2, 3 imply that U restricted to either the vacuous choices or the choices from X is also affine, and thus u and h_0 are affine. In fact, h_0 represents $\succcurlyeq_1^c$ due to Consistency. Note that $h_0(x) = h(x) - f(X) - [h(x_w) -$

$f(X)] = h(x) - h(x_w)$ which is an affine transformation of h. The function v is also affine because it is the sum of two affine functions h_0 and $-u$. For f_0, note that $f_0(A) = h(x) - h(x_w) - [h(x) - f(A)] = f(A) - h(x_w)$. Since f is an affine function and $h(x_w)$ is a constant, f_0 is also affine. *Q.E.D.*

We also have the following result:

Lemma 7. *For any* A, $\min_{x\in A} f_0(\{x\}) \leq f_0(A) \leq \max_{x\in A} f_0(\{x\})$.

Proof of Lemma 7. By Relativity(a), $U(x, A) = U(x, A \cup \{r(A)\})$ is equivalent to $h_0(x) - f_0(A) = h_0(x) - f_0(A \cup \{r(A)\})$, and thus $f_0(A) = f_0(A \cup \{r(A)\})$. Relativity(b) implies that $U(r(A), \{r(A)\}) = U(r(A), A \cup \{r(A)\})$ which means $u(r(A)) = h_0(r(A)) - f_0(A \cup \{r(A)\})$. Then, it follows that $u(r(A)) = h_0(r(A)) - f_0(A)$. Finally, by the definition of v and (5), $f_0(A) = h_0(r(A)) - u(r(A)) = v(r(A)) = f_0(\{r(A)\})$.

Since $r(A) \in \text{conv}(A)$ and v is affine, we have

$$\min_{x\in A} v(x) \leq v(r(A)) \leq \max_{x\in A} v(x). \tag{6}$$

By (5), (6) is equivalent to our desired result. *Q.E.D.*

The proof of Lemma 7 also shows that the choice function $r(\cdot)$ is well-defined, and

$$U(x, A) = u(x) + v(x) - f_0(A) = u(x) + v(x) - v(r(A)).$$

It is clear that the pair (v, r) is a SPA representation. Then, let $U_{u,v,r} := U$. This completes the proof of the existence result in Theorem 2.

For the uniqueness result, the "if" part is straightforward. For the "only if" part, suppose (u, v, r) and (u', v', r') represent $\succeq$ as in Theorem 2. Due to the standard EU theory applied to vacuous choices, u' is an affine transformation of u. Since both v and v' represent $\succcurlyeq_1^*$ induced by $\succeq$, v' must be an affine transformation of v as well. Lastly, to show that $r(A) \sim_1^* r'(A)$ for all $A \in \mathbb{M}$, note that given any menu A, we have $r(A) \sim_1^* r(\text{conv}(A))$ and $r'(A) \sim_1^* r'(\text{conv}(A))$ since $v(r(\cdot))$ and $v(r'(\cdot))$ are affine functions of sets. Hence, I assume A is already convex so that $r(A), r'(A) \in A$. Note that $(r(A), \{r(A)\}) \sim (r(A), A)$ and $(r'(A), \{r'(A)\}) \sim (r'(A), A)$. If $u(r(A)) = u(r'(A))$, then by transitivity, $(r(A), A) \sim (r'(A), A)$. Then,

$$\tfrac{1}{2}(r(A), A) + \tfrac{1}{2}(r'(A), \{r'(A)\}) \sim \tfrac{1}{2}(r'(A), A) + \tfrac{1}{2}(r(A), \{r(A)\})$$

which is equivalent to

$$\left(\tfrac{1}{2}r(A)+\tfrac{1}{2}r'(A),\tfrac{1}{2}A+\tfrac{1}{2}\{r'(A)\}\right)\sim\left(\tfrac{1}{2}r'(A)+\tfrac{1}{2}r(A),\tfrac{1}{2}A+\tfrac{1}{2}\{r(A)\}\right) \quad (7)$$

and thus, $r(A)\sim_1^* r'(A)$. If $u(r(A))\neq u(r'(A))$, then we can assume $u(r(A))>u(r'(A))$ without loss of generality, which means $(r(A),\{r(A)\})\sim(r(A),A)\succ(r'(A),A)\sim(r'(A),\{r'(A)\})$. In this case, (7) still holds, and thus, $r(A)\sim_1^* r'(A)$. *Q.E.D.*

References

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

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

Arrow, K. J. and Hurwicz, L. (1972). An optimality criterion for decision-making under ignorance. *Uncertainty and expectations in economics : essays in honour of G.L.S. Shackle*.

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

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

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

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

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

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

Dekel, E., 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.

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

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

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

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18 (2): 141–153.

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

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

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

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

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

Hurwicz, L. (1951). Some specification problems and applications to econometric methods. *Econometrica*, 19: 343–344.

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.

Kopylov, I. and Noor, J. (2018). Commitments and weak resolve. *Economic Theory*, 66 (1): 1–19.

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.

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

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

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

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

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

Noor, J. and Takeoka, N. (2015). Menu-dependent self-control. *Journal of Mathematical Economics*, 61: 1–20.

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.

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.

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

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

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

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

Supplemental Appendix

This Supplemental Appendix provides three approaches to observing preferences over the acts of choosing, in a dictator game context. Section SA1 presents a menu-choice approach, analogous to the standard social settings designed to test the models of menu preference. To address the potential non-comparability problem in menu-choice environments, the second approach in Section SA2 uses my model to fit Bernheim et al. (2024)'s econometric estimation of the dictator's welfare. Section SA3 presents a direct revealed-preference approach based on *choices over decision-makers*.

SA1 Menu-choice approach: a dictator's menu choice

Consider the two-period dictator game context adopted by Dillenberger and Sadowski (2012) and Saito (2015). In period 1, the dictator publicly chooses an option from a menu A. The option refers to an allocation $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ of wealth between himself (who gets x_1) and a passive recipient (who gets x_2). When the allocation is chosen, the wealth is distributed, and the game ends. Additionally, there is an ex ante stage that the recipient is unaware of: in period 0, the dictator is allowed to privately choose a menu A from an exogenously given set $\mathcal{A}$ that contains menus of allocations (see Figure 1). Hence, the recipient perceives the dictator's menu A as exogenous.

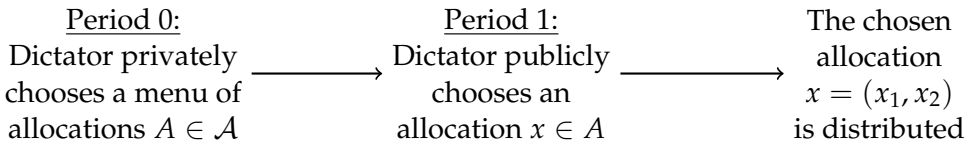

Figure 1: Timeline of the dictator game

Consider three allocations: a fair allocation $x_f = (5, 5)$, a selfish allocation $x_s = (6, 4)$, and let $x_p^w = (6 + w, 6 + w)$ be called a Pareto optimal allocation with an increment $w > 0$. Suppose the set of menus of allocations is

$\mathcal{A} = \{A_0, B^w\}$ where $A_0 = \{x_f, x_s\}$ and $B^w = \{x_f, x_s, x_p^w\}$. A standard dictator—who only has a first-order preference for his own wealth—would obviously prefer the choice $(x_p^w, \{x_f, x_s, x_p^w\})$ to any other possible act of choosing. Thereby, his menu choice would be B^w trivially for any increment $w > 0$.

Alternatively, assume the dictator chooses a menu in period 0 to maximize his utility of the act of choosing in period 1, whose preference has a TPA representation $U_{u,v,\tau}^{Temp}$ such that he prefers "preferring being altruistic to being selfish," but is tempted to be selfish. We can assume that his outcome preference is $u(x) = x_1$; his ideal preference is $v(x) = kx_2$ reflecting altruism with a constant weight $k \geq 0$; his temptation is $\tau(x) = \delta u(x)$ with a constant weight $\delta \geq 0$; and he anticipates that his choice will conform to the ranking $v + \tau$. The utility function can be written as

$$U_{\delta,k}(x, A) := x_1 + kx_2 - \left[\max_{(y_1,y_2)\in A} \left\{\delta y_1 + ky_2\right\} - \max_{(z_1,z_2)\in A} \delta z_1\right].$$

Given his period 1 preference $\succeq$ represented by $U_{\delta,k}$, his induced menu preference, denoted by $\succeq_M$, can be defined as follows: $A \succeq_M B$ if and only if there exists $x \in A$ such that $(x, A) \succeq (y, B)$ for all $y \in B$. That is, his menu choice is $\arg\max_{A\in\mathcal{A}}\{\max_{x\in A} U_{\delta,k}(x, A)\}$. Thus, the dictator's preference over the acts of choosing the three allocations can be summarized as follows:

	Outcome	Ideal	Temptation
Allocations	u	v	τ
$x_f = (5, 5)$	5	$5k$	5δ
$x_s = (6, 4)$	6	$4k$	6δ
$x_p^w = (6 + w, 6 + w)$	$6 + w$	$(6 + w)k$	$(6 + w)\delta$

Suppose $k \geq 1$. Then, it follows that $u(x_f) + v(x_f) \geq u(x_s) + v(x_s)$. That is, the dictator facing the menu A_0 would choose the fair allocation over the selfish one because the act of being altruistic is more valuable than the outcome of x_s. When facing B^w, he would obviously choose the Pareto optimal allocation x_p^w because it is both the most tempting and the most ideal allocation.

However, in period 0, the dictator would choose the menu A_0 over B^w if the increment w is small enough so that $U_{\delta,k}(x_f, A_0) \geq U_{\delta,k}(x_p^w, B^w)$. Note that if $k \geq 1$, then

$$U_{\delta,k}(x_f, A_0) = 5 + 5k - \max\{5(\delta + k), 6\delta + 4k\} + 6\delta = 5 + \min\{k, \delta\};$$
$$U_{\delta,k}(x_p^w, B^w) = (6 + w)(1 + k) - (6 + w)(\delta + k) - (6 + w)\delta = 6 + w.$$

Hence, $k \geq 1$ implies the following:

$$\min\{k, \delta\} > 1 + w \iff A_0 \succ_M B^w. \tag{8}$$

That is, if either conforming to his ideal preference is important enough (a high k), or the temptation of being selfish is strong enough (a high δ), then he deliberately removes the best outcome (x_p^w) from his menu to pursue the best act of choosing. To see why this holds, notice that the utility of the act of choosing x_f from A_0 involves a non-monetary benefit of $\min\{k, \delta\}$: $U_{\delta,k}(x_f, A_0) = u(x_f) + \min\{k, \delta\}$. On the other hand, the act of choosing the Pareto optimal allocation from B^w does not: $U_{\delta,k}(x_p^w, B^w) = u(x_p^w)$ for all $w > 0$. Intuitively, when the menu is A_0, the dictator feels a sense of pride in making the hard choice—sacrificing his own wealth to willingly pursue altruism. In contrast, he feels no pride at all in choosing x_p^w from B^w because he is not willingly giving up anything: he is neither giving up being altruistic nor giving up being selfish.

The result (8) also allows us to capture the monetary value of the sense of pride in choosing x_f over x_s. Let $\overline{w}$ satisfy $1 + \overline{w} = \min\{k, \delta\}$, then the dictator is indifferent between A_0 and $B^{\overline{w}}$, which means he is willing to pay $6 + \overline{w} - 5$ amount of wealth to pursue the sense of pride and give up the monetary benefit of the Pareto optimal allocation. Notice that $U_{\delta,k}$ is decomposed into its first- and second-order preference representations, as follows: $U_{\delta,k}(x, A) = u(x) + V_{\delta,k}(x, A)$. Then, the threshold $\overline{w}$ implies that

$$u(x_f) + V_{\delta,k}(x_f, A_0) = u(x_p^{\overline{w}}) + V_{\delta,k}(x_p^{\overline{w}}, B^{\overline{w}})$$

Since $V_{\delta,k}(x_p^w, B^w) = 0$ for all w, we have $V_{\delta,k}(x_f, A_0) = u(x_p^{\overline{w}}) - u(x_f)$. That is, the value of "preferring being altruistic to being selfish" is equal to the difference in wealth utilities between the Pareto optimal allocation with increment $\overline{w}$ and the fair allocation.[47]

In terms of the attitude toward the act of choosing, the dictator chooses A_0 over B^w because he becomes purely paternalistic only when facing B^w. To see this, let $R_{k,\delta}(A)$ be the reference value of the menu A: i.e., $R_{k,\delta}(A) := u(x) + v(x) - U_{\delta,k}(x, A)$. We can easily verify that $R_{k,\delta}(A_0) = 4k + \max\{k - \delta, 0\}$.[48]

[47] We could consider a more general setting where $x_f = (a - b, a - b)$, $x_s = (a, a - b - c)$, and $x_p^w = (a + w, a + w)$ for some constants $a, b, c > 0$. If $kc \geq b$, then the dictator would choose x_f over x_s—i.e., $U_{\delta,k}(x_f, A_0) \geq U_{\delta,k}(x_s, A_0)$. Then, it follows that $A_0 \sim_M B^{\overline{w}}$ if and only if $\min\{b\delta, ck\} = b + \overline{w}$.

[48] Since $\max_{y \in A_0}\{y_1 + ky_2\} = 5\delta + 5k$ if $k \geq \delta$, and $\max_{y \in A}\{y_1 + ky_2\} = 6\delta + 4k$ otherwise, it follows that $R_{k,\delta}(A_0) = (5\delta + 5k) - 6\delta = 5k - \delta$ if $k \geq \delta$, and $R_{k,\delta}(A_0) = 4k$ otherwise.

Since $\min_{x \in A_0} v(x) = 4k$ and $\max_{x \in A_0} v(x) = 5k$, the dictator is not purely paternalistic as long as $\delta > 0$. In fact, he becomes purely libertarian if $\delta \geq k$. That is, if the material gain tempts him more than how much the ideal preference is important to him, then choosing a fair allocation evokes pride significantly. In contrast, when facing B^w, he is always purely paternalistic given any increment $w > 0$ since $R_{k,\delta}(B^w) = (6+w)(k+\delta) - (6+w)\delta = (6+w)k = \max_{x \in B^w} v(x)$ for all $w > 0$.

The model most closely related to this dictator's problem is Saito (2015)'s menu preference representation. They adopt the same problem, and capture what they refer to as *impure altruism*, which is exhibited when an "intrinsically selfish" dictator behaves altruistically in order to feel pride and to avoid the shame of acting selfishly. However, even in their model, the menu A_0 is never preferred to B^w. That is, the value of pride cannot outweigh the value of better outcomes.[49]

SA2 Welfare-measure approach: the dictator's welfare

The menu-choice approach faces an identification issue if the dictator's menu choice itself reflects his preference over the acts of choosing. Suppose the dictator chooses the menu A_0 over B^w, and chooses the fair allocation x_f afterwards. Then, instead of experiencing a sense of pride, he might feel responsible for leaving himself and the recipient with the Pareto-inferior allocation x_f because he consequentially gave up x_p^w by giving up the menu B^w. In this case, the dictator would rather choose B^w over A_0 even though he prioritizes pride over wealth.[50]

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

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

for some $\alpha > 0$ where $W : \mathbb{R} \to \mathbb{R}$ is a ranking of wealth outcomes. Here, the ex post choice of allocation is the most altruistic allocation—i.e., $\arg\max_{x \in A} W(x_2)$. The sense of pride is captured by the term $\max_{y \in A} W(y_1) - W(x_1)$ which is the dictator's maximal wealth outcome available on the menu A minus his chosen wealth outcome $W(x_1)$. When the menu is $A_0 = \{x_f, x_s\}$, we have $U_S(A) = W(5) + \alpha W(6)$. By choosing x_f over x_s, the dictator gains the sense of pride measured by $\alpha W(6)$. When the menu is B^w, we have $U_S(B^w) = W(6+w) + \alpha W(6+w)$. Assuming that W is an increasing function, we have $U_S(B^w) \geq U_S(A_0)$.

[50] The menu-choice approach can still effectively identify the dictators whose preferences over the acts of choosing are driven by social motives (e.g., wanting *to appear to others* as someone who prefers being altruistic to being selfish): they would still choose A_0 over B^w.

This identification problem is analogous to the non-comparability problem formally illustrated by Bernheim et al. (2024) (see also Kőszegi and Rabin, 2008). The problem arises when the DM cares about the act of choosing, in which case, his choices and welfare are not necessarily aligned.[51] As a result, the DM's welfare cannot be uniquely recovered from standard choice data. To illustrate using the dictator's example above, suppose a social planner (she) wants to figure out which menu among A_0 and B^w would make the dictator (he) better off. She observes that his choices are B^w in period 0 and x_p^w in period 1. Based on this observation alone, she is unsure whether or not he would feel proud of himself for choosing x_f from A_0, and thus be happier than when B^w is given.

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

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

While their experiments were focused on verifying the existence of the non-

[51] Bernheim et al. (2024) showed that even when the DM cares about the act of choosing *menus* (or any other higher-order menus), the non-comparability problem persists. Suppose the DM faces a n-stage decision problem. An outcome is chosen from a first-order menu (i.e., a menu of outcomes) in the final (nth) stage; the first-order menu is chosen from a second-order menu (i.e., a menu of first-order menus) in the $n-1$th stage, and so on. Even from the choice data in this setting, the social planner cannot recover the DM's preference over the acts of choosing in the first stage when he chooses the nth-order menu from an exogenously given $n+1$th-order menu, due to the same reason why she cannot recover the dictator's preference from his choice of B^w over A_0.

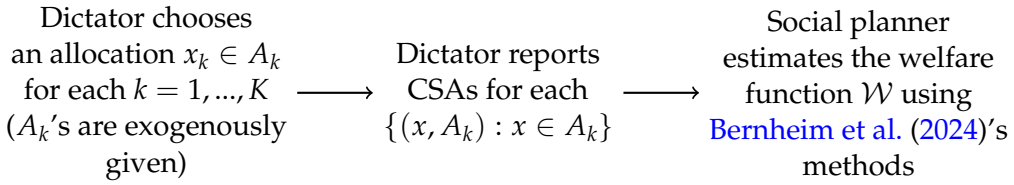

Figure 2: Timeline of experiment using Bernheim et al. (2024)'s design

comparability problem, the focus here is to verify that the best act of choosing can be more valuable than the best outcome—more specifically, verifying that the following inequality is possible:

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

SA3 Choice based on choice data: choosing the right dictator

Second-order preferences are fundamentally difficult to observe directly because it is difficult to observe an individual's choices of *choices*. The two approaches in the previous sections partially address this concern either by observing menu choices or by obtaining the self-reported well-being data, but they still seem to lack observational power in the sense that the axioms in this paper cannot be directly tested.

I address this problem by considering the following problem of *"choice over DMs."* The agent first observes payoff-irrelevant choices made in the past by potential DMs with some consistent preferences, each of whom faced different menus. Next, the agent chooses one DM, who then makes a payoff-relevant choice for the agent. Since these past choices do not affect the agent's own outcome, the agent's concern reduces to choosing the DM whose choices reflect the most preferable preference. Then, the agent's ranking of the candidates, which is conditioned only on their acts of choosing, allows my axioms to be tested without relying on menu choices or econometric models. Plus, since the agent is not responsible for the ultimate outcome, this setting is less susceptible to the non-comparability problem.

Consider the recipient in the dictator game who cares only about his wealth outcomes, and needs to rank preferences for the sake of his outcome. Suppose two dictators (Dictators 1 and 2) played a game: each Dictator $i \in \{1, 2\}$ was exogenously given a menu A_i of allocations, and chose $x_i \in A_i$. (In general, there can be $K \geq 2$ dictators.) The DM who was not involved in the two previous games is about to play a dictator game as a recipient. His decision problem is

as follows: before the game begins, he must choose whether Dictator 1 or Dictator 2 will be the dictator in his own game. Before making this decision, the DM observes the choice data of the past games—the two choices (x_1, A_1) and (x_2, A_2) made by the two dictators. Also, the DM privately knows the menu A^* of allocations that will be given to his dictator. When the game begins, the chosen dictator will choose an allocation from A^*. See Figure 3.

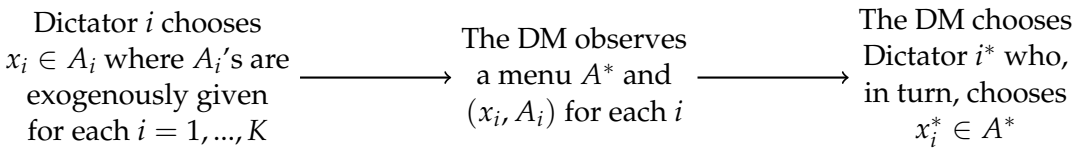

Figure 3: Timeline of the game of choosing the dictator

There are several important assumptions. First, it is common knowledge that all players have no control over the menus. Second, to avoid any strategic motivations, the two dictators are not aware of the possibility of playing another game, or that their choices are being observed by the DM. Third, the dictators do not make mistakes (i.e., there are no cognitive imprecision or trembling hands involved). Also, their preferences over the allocations are stable (i.e., no changing tastes). Alternatively, we can also assume that the dictators are machines that have been programmed to have certain preferences. Fourth, the DM knows that these assumptions hold. Lastly, I allow the dictators to choose more than one allocation if they are indifferent among some options, in which case, the recipient decides the final allocation.[52]

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

Suppose the DM knows that the menu for his chosen dictator will be $A^* = \{x_f, x_s\}$—which means the DM wants an altruistic dictator—and observes five

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

dictators' choices as below:

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

Then, it is reasonable for the DM to choose Dictator 1 to be his dictator. While the choice $(x_f, \{x_f, x_s\})$ suggests that Dictator 1 cares about fairness, the choice $(x_p^w, \{x_f, x_s, x_p^w\})$ does not suggest anything about Dictator 2's preference over $\{x_f, x_s\}$. Similarly, Dictator 5 would be the last person the DM wants because the choice $(x_s, \{x_f, x_s\})$ reveals selfishness. If the DM's preference has a TPA representation $U_{u,v,\tau}^{Temp}$ with $u = 0$, we should observe the following ranking:

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

where the indifference relations are due to IVC.

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

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

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