A Theory of *Ex Post* Rationalization

Erik Eyster, Shengwu Li, and Sarah Ridout*

December 2, 2021

Abstract

People rationalize their past choices, even those that were mistakes in hindsight. We propose a formal theory of this behavior. The theory predicts sunk-cost effects, as well as 'unsunk-benefit' effects. Its model primitives are identified by choice behavior and it yields tractable comparative statics.

*Eyster: UC Santa Barbara, erikeyster@ucsb.edu; Li: Harvard University, shengwu_li@fas.harvard.edu; Ridout: Vanderbilt University, sarah.e.ridout@vanderbilt.edu. We thank Mohammad Akbarpour, Nicholas Barberis, Douglas Bernheim, Ben Bushong, Sam Gershman, Ben Golub, Muriel Niederle, Matthew Rabin, Gautam Rao, Philipp Strack, Bruno Strulovici, Tomasz Strzalecki, and Richard Thaler for valuable comments. All errors remain our own.

1 Introduction

People rationalize their past choices. We look back on our lives and try to make sense of what we have done. Upon realizing that, in hindsight, we have made a mistake, we can adapt our goals, attitudes, or beliefs to justify the decision.

Classical economic theory rules out rationalization. It assumes that people make forward-looking choices according to time-stable preferences, rather than adapting their preferences to rationalize past decisions. Introspection, common sense, and psychological research all suggest that the classical approach omits a key aspect of human decision-making.[1]

How can economic models capture *ex post* rationalization? We develop a theory that accommodates this behavior. To fix ideas, consider the following example from Thaler (1980).[2]

> **Example 1**: Bob pays \$100 for a ticket to a basketball game to be played 60 miles from his home. On the day of the game there is a snowstorm. He decides to go anyway. If the ticket had been free-of-charge, he would have stayed home.

The \$100 that Bob paid is a sunk cost. It is not worth going to the basketball game during a snowstorm. In hindsight, it was a mistake to have bought the ticket. But if Bob goes to the game, then he can avoid acknowledging the mistake, by exaggerating his enthusiasm for basketball or by downplaying the hazards of driving through a snowstorm. If he stays home, then he is inarguably worse off than if he had not bought a ticket in the first place.

There are two key ingredients for this behavior. First, Bob must have made a choice that was an *ex post* mistake. Hence, this modified example is far-fetched:

> **Example 2**: Bob receives a free ticket to a basketball game and loses \$100 due to an unusually high utility bill. On the day of the game there is a snowstorm. He decides to go anyway. If he had not lost the \$100, he would have stayed home.

Second, there must be plausible preferences that, if adopted, would justify Bob's earlier decision. To illustrate this, we replace the physical consequences in Example 1 with monetary gains and losses.

> **Example 3**: Bob pays \$100 for a financial option that can only be exercised on the day of the basketball game. It yields \$180 if exercised in good weather and *loses* \$20 if exercised in a snowstorm. On the day of the

[1] We review psychological research on rationalization in Section 2.
[2] The ticket cost \$40 in Thaler's example; we have raised the price due to inflation.

game there is a snowstorm. He decides to exercise the option anyway, for a net loss of $\$100 + \$20$.

Example 3 is unnatural because there is no way for Bob to rationalize his initial purchase. Letting the option expire results in a net loss of \$100, whereas exercising the option results in a net loss of \$120. More money is better, so Bob has to acknowledge the mistake and cut his losses.

Even at high stakes, decision-makers sometimes rationalize sunk costs rather than acknowledge error. For instance, a senior Irish Republican Army leader was asked in 1978 whether the years of violent resistance had been worth it. He replied, "Virtually nothing has been achieved. We can't give up now and admit that men and women who were sent to their graves died for nothing." (Smith, 1997, p. 225)[3]

Motivated by these examples, we propose a theory about agents who seek to rationalize their past choices by adapting their preferences. We model an agent facing a decision problem with this structure:

1. The agent chooses action a_1 from menu A_1.
2. The agent learns the state of the world $s \in S$.
3. The agent chooses action a_2 from menu $A_2(a_1)$, which can depend on his first action.

A utility function takes as arguments a_1, a_2, and s. The agent's *material utility function* is denoted u. The agent may adopt any utility function in the set $\mathcal{V}$, which we call *rationales*. u and $\mathcal{V}$ are primitives of the model. We assume that $u \in \mathcal{V}$.

We start by describing the agent's choice from menu $A_2(a_1)$, after having chosen a_1 from menu A_1 and learned that the state is s. The agent maximizes a weighted sum of his material utility and a regret term that is assessed according to the agent's chosen rationale. Formally, the agent chooses action a_2 from menu $A_2(a_1)$ and rationale v from $\mathcal{V}$ to maximize *total utility*, that is

$$(1-\gamma)\underbrace{u(a_1, a_2, s)}_{\substack{\text{material}\\ \text{utility}}} + \gamma \underbrace{\left[v(a_1, a_2, s) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} v(\hat{a}_1, \hat{a}_2, s) \right]}_{\textit{ex post} \text{ optimality under chosen rationale}}$$

where $\gamma \in [0, 1)$ is the weight on the regret term.

Observe that if a_1 is *ex post* optimal, that is

$$\max_{\hat{a}_2 \in A_2(a_1)} u(a_1, \hat{a}_2, s) = \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} u(\hat{a}_1, \hat{a}_2, s),$$

[3]For further reading on how rationalizations by combatants prolonged the Troubles, see Chapter 3 of Alonso (2007).

then the theory predicts that the agent chooses a_2 to maximize material utility. Moreover, if A_1 is a singleton then a_1 is trivially *ex post* optimal. Hence, the theory departs from the classical prediction only when the agent has made a choice and that choice was an *ex post* mistake.

When the agent has made an *ex post* mistake, he may be able to reduce regret by choosing rationale $v \neq u$. By construction, a_2 maximizes a weighted sum of his material utility $u(a_1, a_2, s)$ and his chosen rationale $v(a_1, a_2, s)$, which distorts his choices compared to the classical benchmark.

We now apply the model to the earlier examples. The rationales $\mathcal{V}$ are parameterized by $\theta \in [0, 400]$. Utility function v_θ specifies that the agent gets θ utils for attending the game, -200 utils for driving through a snowstorm, and $-p$ utils for paying p dollars. Material utility is $u = v_{180}$, so a classical agent ($\gamma = 0$) is willing to pay \$180 to attend the game in good weather, but will stay home in a snowstorm.

In Example 1, the menu A_1 has two alternatives; the agent can get a ticket and lose \$100 or he can decline. After buying the ticket for \$100, the agent learns that there is a snowstorm. If he stays home, then his material utility is -100, and his regret term is $-100 - \max\{\theta - 200 - 100, 0\}$. It is optimal to choose $\theta \leq 300$, yielding total utility $(1-\gamma)(-100) + \gamma(-100) = -100$. By contrast, if the agent attends the game, then his material utility is $180 - 200 - 100 = -120$ and his regret term is $\theta - 200 - 100 - \max\{\theta - 200 - 100, 0\}$. Now it is optimal to choose $\theta \geq 300$, for a regret term of 0 and a total utility of $(1-\gamma)(-120)$. By attending the game and exaggerating his enthusiasm, the agent is able to reduce regret at the cost of material utility. For $\gamma > \frac{1}{6}$, his total utility is strictly higher when he attends the game.

Suppose instead that the ticket was free-of-charge. Then staying home in a snowstorm leads to no regret under the agent's material utility function u. Hence, the agent maximizes total utility by adopting rationale $v = u$ and staying home. The agent's behavior exhibits sunk-cost effects; his choice on the day of the basketball game depends on upfront costs that he cannot recover.

In Example 2, the agent has no choice initially, so the menu A_1 contains only one alternative: the agent gets a ticket and loses \$100. This is trivially *ex post* optimal, so the agent maximizes total utility by staying home in the snowstorm. Hence, removing unchosen alternatives from the menu A_1 can alter the agent's later choice from A_2.

In Example 3, we have taken the agent's material utility for the outcomes in Example 1 and converted utils to dollars. The agent bought the financial option for \$100. Exercising the option in good weather yields \$180, and exercising it in a snowstorm loses \$20. But every rationale agrees about money, so there is no room to reduce regret and the agent does not exercise the option in a snowstorm.

A plausible objection is that Bob's behavior is not due to rationalization but due to 'waste aversion'. If Bob buys the ticket and stays home, then he has *wasted* \$100, whereas if he goes to the game in the snowstorm then it was not a waste. But the same logic applies to Example 3—if Bob buys the financial option and does not use it, then he has wasted \$100. Why cannot he avoid waste by exercising the option (and losing a further \$20)?

Ex post rationalization extends and formalizes these intuitions about waste aversion. In some situations, people can 'avoid waste' by doubling down on bad decisions, and in other situations they have to recognize losses. This distinction is determined, in the theory, by the set of rationales. In Example 1, Bob can rationalize that attending the game means that the \$100 was not wasted, whereas in Example 3 no such rationale is available.[4]

The theory predicts not only that the agent counts sunk costs as reasons to act, but also that he counts *unsunk benefits* as reasons to refrain, as the next example illustrates.

> **Example 4**: Bob has an opportunity to buy a discounted ticket to the basketball game for \$20. He declines, believing that there will be a snowstorm. On the day of the game the weather is warm and sunny. An acquaintance offers to sell him a ticket for \$100. Bob decides to stay home. If the tickets had never been on discount, he would have gone to the game.

Under our utility specification, accepting the acquaintance's offer and attending the game yields a material utility of $180 - 100 = 80$ but a regret term of -80. Staying home and adopting the rationale v_0 yields a material utility of 0 and a regret term of 0. For $\gamma > \frac{1}{2}$, the agent declines the offer, even though he would have accepted had the tickets never been on discount.[5]

To complete the model, we specify the agent's behavior when choosing from the first menu A_1. At this point, the agent has no earlier choices to rationalize, so we assume that the agent evaluates choices from A_1 according to his expected material utility under some prior on the states S. This depends on the agent's beliefs about his future choice from A_2. We study two benchmarks: a *naïve* agent believes he will maximize material utility when choosing from A_2; a *sophisticated* agent correctly foresees his choices from A_2, but evaluates them according to material utility.

Given material utility u and the rationales $\mathcal{V}$, the theory predicts the agent's choices. But how is the modeler to specify the primitives? In some situations, we can use standard restrictions on the preferences agents may plausibly hold. For instance, for an

[4] Of course, this is only a partial account of the psychology at work. Imas (2016) finds that lab subjects take on more risk after paper losses and less risk after realized losses. The present theory neglects this dependence on framing.

[5] Tykocinski et al. (1995) find evidence of such behavior in a vignette study.

agent choosing between money lotteries, we could assume that $\mathcal{V}$ is a class of preferences with constant relative risk aversion. Similarly, for an agent bidding in an auction, we could assume that the rationales $\mathcal{V}$ have different valuations for the object, but are all quasi-linear in money.

If we do not make *a priori* restrictions on $\mathcal{V}$, can we nonetheless deduce the rationales? We prove that the model primitives u, $\mathcal{V}$, and γ are fully identified by choice behavior. That is, suppose we start with finitely many outcomes, and each utility function depends on the outcome and the state. We then construct objective lotteries over outcomes, and extend utility functions by taking expectations. The agent faces decision problems of this form:

1. The agent selects a menu M from a collection of menus of lotteries.
2. The agent learns the state of the world $s \in S$.
3. The agent chooses a lottery from M.

Under a regularity condition, we prove that whenever the agent's choice correspondence is consistent with material utility u and rationales $\mathcal{V}$, those primitives are unique up to a state-specific affine transformation.[6] Moreover, the weight γ on the regret term is identified. Hence, statements about the agent's rationales can be reduced to statements about the agent's choice behavior.

Next, we compare the behavior of rationalizers and classical agents. We impose more structure to yield comparative statics, by assuming that first actions, second actions, and rationales are complements. Let the first actions, the second actions, and parameter set Θ be totally ordered sets.[7] We assume that the rationales $\mathcal{V}$ have the form $\{w(a_1, a_2, \theta, s) : \theta \in \Theta\}$, for some function w that is supermodular in (a_1, a_2, θ). For instance, this includes the rationales we posited for Example 1, if we impose that buy a ticket is a higher action than don't buy a ticket, and that go to the game is a higher action than stay home. It also includes time-separable utility functions, of the form $w(a_1, a_2, \theta, s) = w_1(a_1, \theta, s) + w_2(a_2, \theta, s)$, with w_t supermodular in (a_t, θ). We assume that the menu $A_2(a_1)$ is monotone non-decreasing in a_1.

We prove that if the rationalizer's first action was *ex post* too high, then his second action is distorted upwards compared to the classical benchmark. Symmetrically, if the rationalizer's first action was *ex post* too low, then his second action is distorted downwards compared to the classical benchmark, as happens in Example 4.

This result yields comparative statics for a variety of settings. It predicts sunk-cost effects in multi-part projects — when first-period effort and second-period effort are complements, the agent responds to first-period cost shocks by exaggerating the

[6]Even classical models of state-dependent utility are only unique up to a state-specific affine transformation, because beliefs and utilities are not separately identified.

[7]Our results also hold under weaker order-theoretic assumptions, that we state in Section 5.

value of the project and raising second-period effort. It predicts that agents repeatedly facing identical decisions will have 'sticky' choice behavior, responding too little to new information. In particular, lab subjects who make incentivized reports of priors and posteriors will report posteriors biased towards their priors, and will underweight informative signals compared to subjects who report only posteriors.

As another application, suppose a consumer can pay an upfront fee for the option to buy goods at a per-unit price, then has a taste shock, then chooses how much to consume. The consumer's marginal utility of consumption is increasing in θ. Conditional on paying the fee, a classical agent's consumption does not depend on whether the fee is high or low, because the fee is a sunk cost. By contrast, our comparative statics result implies that when the upfront fee is high enough, the rationalizer's consumption is distorted upwards compared to the material optimum.

Next, we study the effect of *unchosen* time-1 alternatives on time-2 choice. In the classical model, such alternatives are irrelevant for time-2 behavior. By contrast, for a rationalizer facing a supermodular decision problem, raising the unchosen time-1 alternatives lowers the agent's time-2 choices. That is, suppose we raise the first menu A_1 in the strong set order, while holding the chosen action a_1 constant. We prove that this change weakly lowers the agent's second action in every state. This prediction compares cleanly to the zero effect predicted by the classical model.

So far we have interpreted the theory as capturing individual psychological motives, but it has another interpretation in the context of organizations. On this interpretation, the theory describes a rational agent who is rewarded for past performance, but can influence the principal's criteria for performance evaluation. The agent's need to defend past decisions generates sunk-cost effects. This formalizes an observation by Staw (1980) about the perverse incentives of retrospective performance evaluation.

2 Literature review

Many experimental paradigms in psychology have found evidence of rationalization. Cushman (2020) writes, "Among psychologists, [rationalization] is one of the most exhaustively documented and relentlessly maligned acts in the human repertoire." For brevity, we focus on two topics that pertain closely to the present theory: cognitive dissonance and confabulation.

Cognitive dissonance theory posits that people adapt their cognitions so as to achieve internal consistency (Festinger, 1957). It predicts that the more a person suffers to obtain achieve some result, the more they will value that result. For instance, some experiments find that unexpectedly severe initiation rituals cause new members to evaluate the group more positively, a result that is also consistent with *ex post* rationalization (Aronson and Mills, 1959; Gerard and Mathewson, 1966). Harmon-Jones and Harmon-Jones (2007) review experimental paradigms testing the theory of cognitive dissonance. However, one of these—the free choice paradigm of Brehm (1956)—is confounded by self-selection bias and requires caution in interpreting the results (Chen and Risen, 2010; Risen and Chen, 2010).

The literature on confabulation finds that introspection does not reliably detect the reasons for one's past choices. Instead of accurately recalling the reasons for their choices, people sometimes generate *post hoc* explanations and sincerely believe them. Such confabulation has been documented in split-brain patients (Gazzaniga, 1967, 2005) and in ordinary people manipulated to misremember their own past choices (Johansson et al., 2005, 2006). Nisbett and Wilson (1977) find that lab subjects are often unaware of the effects of experimental stimuli on their own behavior, and offer spurious explanations when queried by experimenters.

Various economic theories allow that agents choose their beliefs or preferences.[8] In particular, Yariv (2005), Acharya et al. (2018), and a preprint of Bernheim et al. (2021) model agents who change their beliefs or preferences to align with past actions. Unlike these theories, *ex post* rationalization is about choosing preferences to reduce regret. Regret depends not only on what the agent chose, but also on what else she could have chosen. Suzuki (2019) studies an agent who chooses a project, observes a signal, and then distorts her beliefs and effort to justify her choice of project. Our approach is similar, but we study general decision problems and identify the model from choice data.

An *ex post* rationalizer distorts today's choices so as to justify yesterday's choices. This contrasts with the canonical approach to regret theory, in which the agent makes choices today so as to reduce regret tomorrow (Savage, 1951; Loomes and Sugden, 1982; Bell, 1982; Wong and Kwong, 2007; Sarver, 2008). Also, regret theory does not permit the agent to adapt her preferences to reduce regret, which is crucial for rationalization.

The present theory posits that the agent seeks to justify her choice *ex post*. She does not adopt the perspective that her choice, while mistaken in hindsight, was sensible given what she knew at the time. After an event has occurred, people are overconfident that they could have predicted it in advance. This phenomenon, known as hindsight bias, is the subject of a vast literature (Fischhoff, 1975; Blank et al., 2007). People even misremember their own *ex ante* predictions, falsely believing that they correctly predicted what came to pass (Fischhoff and Beyth, 1975; Fischhoff, 1977).

Sunk-cost effects have been studied extensively in organizational behavior (Staw, 1976) and in economics (Thaler, 1980). Sunk-cost effects have been documented in

[8]For examples, see Rabin (1994), Rotemberg (1994), Becker and Mulligan (1997), Akerlof and Kranton (2000), Brunnermeier and Parker (2005), and Bernheim et al. (2021).

many settings, such as business decisions[9], usage of durable goods[10], professional sports[11], and auctions[12]. Some studies do not find evidence of sunk-cost effects[13]. For a meta-analysis, see Roth et al. (2015).

There is no broad consensus as to why human behavior exhibits sunk-cost effects. Some studies have offered explanations based on prospect theory (Thaler, 1980; Arkes and Blumer, 1985). These explanations depend on losses relative to a reference point, whereas the theory of *ex post* rationalization predicts that sunk-cost effects depend on whether the agent made a choice to incur those costs. Other studies have suggested that sunk-cost effects may be rational due to reputation concerns (Prendergast and Stole, 1996; McAfee et al., 2010), limited memory (Baliga and Ely, 2011), and self-signaling motives (Hong et al., 2019). Closest to our explanation, Staw (1980) argues that sunk-cost effects arise because people "re-evaluate alternatives and outcomes to make it *appear* that that they have acted in a competent or intelligent manner".

Recent work finds that sunk-cost effects depend on whether the decision-maker was responsible for the initial decision, as predicted by *ex post* rationalization. Martens and Orzen (2021) study the effect of sunk costs on follow-up investment decisions, in a laboratory experiment. They find that sunk-cost effects increase when subjects are responsible for the initial investment decision. Similarly, Guenzel (2021) finds that for high-stakes corporate acquisitions: exogenous cost shocks (occurring *after* the acquisition decision) decrease the acquiring company's willingness to divest, but this effect is substantially reduced if the CEO who led the acquisition steps down.

Most directly, the present study builds on ideas from Eyster (2002) and Ridout (2020).

Eyster (2002) studies a two-period model in which an agent wishes to maximize a weighted average of material utility and *ex post* regret according to a fixed utility function, but limits attention to alternative first actions that are consistent with the chosen second action. To illustrate, the theory of Eyster (2002) would explain Example 1 by positing that if Bob attends the game, then only $\underline{\text{buy a ticket}}$ is consistent, so he feels no regret. On the other hand, if Bob stays home, then both $\underline{\text{buy a ticket}}$ and $\underline{\text{don't buy a ticket}}$ are consistent, so he feels regret for having bought the ticket. One limitation of this approach is that the modeler's intuitions about consistency may vary with how the actions are framed – $\underline{\text{stay home}}$ seems consistent with $\underline{\text{don't buy a ticket}}$, but $\underline{\text{stay home with a ticket in hand}}$ does not seem consistent with $\underline{\text{don't buy a ticket}}$, so Bob can also avoid regret by staying home with a ticket in hand. One advantage of

[9] McCarthy et al. (1993); Schoorman (1988); Staw et al. (1997).
[10] Ho et al. (2018).
[11] Staw and Hoang (1995); Camerer and Weber (1999); Keefer (2017).
[12] Herrmann et al. (2015); Augenblick (2016).
[13] Ashraf et al. (2010); Friedman et al. (2007); Ketel et al. (2016); Negrini et al. (2020).

our present approach is that it overcomes this framing objection. Instead of a frame-dependent consistency relation, our main primitive is a set of rationales, *i.e. post hoc* reasons for the agent's choice, and the rationales are identified from choice data.

Ridout (2020) studies a model of one-shot choice, with an agent who has a set of 'justifiable' preferences and a material preference, and is constrained to choose only alternatives that maximize some justifiable preference. As in the present study, the agent sometimes fails to maximize his material preference, but the mechanism is different. Ridout's agent may forego an alternative he prefers because he does not consider his material preference justifiable. By contrast, our agent's material preference belongs to the set of rationales, so there is no distortion in the absence of past mistakes. The agent foregoes a materially preferred alternative only if acting on his material preference would lead him to regret a past decision.

3 Statement of theory

In our model, an agent chooses an action from a menu, then learns the state of the world, and then finally chooses an action from a second menu, which can depend on the first action.

We now define the model primitives. $\mathcal{A}_1$ denotes the *first actions*; $\mathcal{A}_2$ denotes the *second actions*; and S denotes the *states of the world*, with representative elements a_1, a_2, and s, respectively.

A *decision problem* $D \equiv (A_1, A_2, F)$ consists of

1. a first-period menu $A_1 \subseteq \mathcal{A}_1$,
2. and a second-period menu correspondence $A_2 : A_1 \rightrightarrows \mathcal{A}_2$.
3. a prior over states $F \in \Delta S$,

We require that A_1 and A_2 be non-empty.

A *utility function* is a function $v : \mathcal{A}_1 \times \mathcal{A}_2 \times S \to \mathbb{R}$. The *rationales* are denoted $\mathcal{V}$; these are a set of utility functions that the agent may adopt to justify her actions. The agent's *material utility function* is denoted u, and we assume that $u \in \mathcal{V}$.

The set of rationales $\mathcal{V}$ captures the preferences that the agent regards as reasonable. For instance, the rationales could specify the agent's utility from consuming a good or service. Alternatively, rationales could specify the agent's subjective beliefs about some payoff-relevant event, with the observed state s being a noisy signal about that event.

We start by describing choice in the second period. The agent facing decision problem D has chosen a_1 from menu A_1 and learned that the state is s. She chooses

$a_2 \in A_2(a_1)$ and $v \in \mathcal{V}$ to maximize

$$U_D(a_2, v \mid a_1, s) \equiv (1-\gamma)\underbrace{u(a_1, a_2, s)}_{\substack{\text{material}\\ \text{utility}}} + \gamma \underbrace{\left[v(a_1, a_2, s) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} v(\hat{a}_1, \hat{a}_2, s) \right]}_{\textit{ex post} \text{ optimality under chosen rationale}} \quad (1)$$

for parameter $\gamma \in [0, 1)$. Equation (1) states that the agent places weight $(1-\gamma)$ on maximizing material utility u, and γ on rationalizing her choice *ex post*. The second term in (1) measures how close her course of action is to the *ex post optimum* under her chosen rationale v. When a_1 was *ex post* sub-optimal according to u, the second term might be increased by adopting rationale $v \neq u$. This distorts the agent's choice of a_2, which maximizes $(1-\gamma)u(a_1, a_2, s) + \gamma v(a_1, a_2, s)$.

We restrict attention to decision problems for which the relevant maxima are well-defined. This is implied, for instance, if every $v \in \mathcal{V}$ is continuous in the actions, and the sets $A_2(a_1)$, $\{(a'_1, a'_2) : a'_1 \in A_1 \text{ and } a'_2 \in A_2(a'_1)\}$, and $\mathcal{V}$ are compact.

We discuss some natural benchmarks for first-period behavior. A *naïf* chooses a_1 to maximize $\mathbb{E}_F[u(a_1, a_2^*(a_1, s), s)]$ where $a_2^*(a_1, s)$ is a selection from

$$\operatorname*{argmax}_{a_2 \in A_2(a_1)} u(a_1, a_2, s).$$

A *sophisticate* chooses a_1 to maximize $\mathbb{E}_F[u(a_1, \tilde{a}_2(a_1, s), s)]$ where $\tilde{a}_2(a_1, s)$ is a selection from

$$\operatorname*{argmax}_{a_2 \in A_2(a_1)} \max_{v \in \mathcal{V}} U_D(a_2, v \mid a_1, s).$$

Naïfs and sophisticates both maximize expected material utility *ex ante*, albeit with different beliefs about *ex post* behavior.[14]

Another natural benchmark is the *empathetic sophisticate*, whose first action maximizes expected total utility, *i.e.*

$$\mathbb{E}_F \left[\max_{a_2 \in A_2(a_1)} \max_{v \in \mathcal{V}} U_D(a_2, v \mid a_1, s) \right].$$

This agent both correctly foresees his *ex post* behavior and desires to reduce regret when choosing *ex ante*.

[14]At the start of the Vietnam War, undersecretary-of-state George Ball gave this warning to President Lyndon Johnson: "Once we suffer large casualties, we will have started a well-nigh irreversible process. Our involvement will be so great that we cannot—without national humiliation—stop short of achieving our complete objectives." (Sheehan and Kenworthy, 1971) Ball was a sophisticate about rationalizing behavior.

3.1 Discussion of modeling choices

Plausibly, the agent's rationalization motive depends on the kind of *ex ante* uncertainty she faced. Choosing a risky investment is not like choosing a bet in roulette. It is easier to remember the *ex ante* perspective when evaluating choices with objective risks. By contrast, people are more likely to say, "I should have known it!" for decisions that involved Knightian uncertainty or required deliberation to weigh competing considerations. Our model abstracts from this nuance, representing uncertainty using only a distribution over states. However, we interpret the scope of the theory to be confined to those kinds of uncertainty which seem predictable in hindsight.[15]

For the theory to depart from the classical prediction, the available rationales $\mathcal{V}$ must be limited. For instance, if $\mathcal{V}$ includes a 'stoic' rationale that is indifferent between all action sequences, then the second term in (1) can always be set to zero, and the theory predicts material utility maximization. Thus, the theory's novel predictions depend on plausible restrictions on the rationales that the agent can adopt.

For specific applications, we suggest that $\mathcal{V}$ should be some standard class of preferences for the setting under consideration, if such standards exist. This serves to prevent *ad hoc* explanations and to make the theory a portable extension of existing models, in the sense of Rabin (2013). Other economic exercises also take restrictions on preferences as given. For instance, in mechanism design, positive results often depend on restricting the agent's preferences to lie within certain *a priori* limits (Hurwicz, 1972). As another example, structural estimation methods often require functional-form restrictions on preferences.

Nonetheless, because the theory's predictions depend on $\mathcal{V}$, its general applicability depends on whether $\mathcal{V}$ can be identified from choice behavior. We take up this challenge in the next section.

4 Identification of preference parameters

In this section, we find that the theory's preference parameters are identified from choice behavior, by extending the environment to include objective lotteries over outcomes.

[15]We suggest that experimental tests of the theory use forms of uncertainty that require the subject to exercise judgment, rather than objective risks such as coin flips or dice rolls. Experimenters might also consider designs that give subjects the illusion of control (Langer, 1975; Presson and Benassi, 1996).

4.1 Extending the environment

We start with a finite set of outcomes Z. For instance, in Example 1, the outcomes are $\underline{\text{attend the game \& pay \$100}}$, $\underline{\text{stay home \& pay \$100}}$, and $\underline{\text{stay home \& pay nothing}}$. The lotteries over outcomes are denoted $\Delta(Z)$.

For any set B, let $\mathcal{K}(B)$ denote the collection of nonempty subsets of B. Let $\mathcal{K}_f(B)$ denote the collection of finite nonempty subsets of B.

The agent faces decision problems of this form:

1. At $t = 1$, the agent selects a menu M from a finite collection of menus $\mathcal{M} \subset \mathcal{K}_f(\Delta(Z))$.
2. The agent learns the state s.
3. At $t = 2$, the agent chooses a lottery q from the selected menu M.

We take as data the agent's time-2 choice correspondence in each state. Having selected menu M from collection $\mathcal{M}$ and learned that the state is is s, the agent's choices from M are denoted $c_2^s(M \mid \mathcal{M})$, which satisfies $c_2^s(M \mid \mathcal{M}) \subseteq M$. Observe that the agent can, by choosing differently at time-1, achieve any lottery in $\bigcup \mathcal{M}$. We use $\mathcal{U}$ to denote the set of all functions from Z to $\mathbb{R}$, and extend these to $\Delta(Z)$ by taking expectations.

Definition 4.1 (Representation)**.** *$(\gamma, u^s, \mathcal{V}^s) \in [0,1] \times \mathcal{U} \times \mathcal{K}(\mathcal{U})$ **represents** choice correspondence c_2^s if*

$$c_2^s(M \mid \mathcal{M}) = \operatorname*{argmax}_{q \in M} \left\{ (1-\gamma) u^s(q) + \gamma \max_{v^s \in \mathcal{V}^s} \left\{ v^s(q) - \max_{\hat{q} \in \bigcup \mathcal{M}} v^s(\hat{q}) \right\} \right\}$$

for all $\mathcal{M} \subset \mathcal{K}_f(\Delta(Z))$ and all $M \in \mathcal{M}$.

Our identification exercise relies only on time-2 choice behavior, so the resulting theorem applies equally to naïfs, sophisticates, and empathetic sophisticates.

By taking the choice correspondence c_2^s as data, we are asking whether the preference parameters are identified from a *complete* description of time-2 choice behavior. That is, even setting aside practical questions of elicitation, are the parameters pinned down by knowing how the agent would choose, for each menu M and each collection $\mathcal{M}$?

An important caveat is that we assume that the analyst observes the agent's time-2 behavior contingent on each time-1 choice, including for time-1 choices that do not maximize expected material utility. Trembles by the agent might justify access to this contingent choice data. Alternatively, the analyst may be able to induce different choices by manipulating the agent's prior, provided that each outcome is optimal in some state.

As with classical models of state-dependent utility, the agent's behavior in both periods remains the same if we scale up the utility functions in state s and scale down the prior probability on s to compensate. Hence, in general, utility can at best be identified up to a *state-specific* affine transformation. We will shortly state conditions under which this is possible.

4.2 Identification result

For the identification result, we restrict attention to choice correspondences with regular representations. In the definition that follows, we treat each element of $\mathcal{U}$ as a point in $\mathbb{R}^Z$.

Definition 4.2 (Regularity)**.** $(\gamma, u^s, \mathcal{V}^s) \in [0,1] \times \mathcal{U} \times \mathcal{K}(\mathcal{U})$ *is* ***regular*** *if it satisfies the following conditions:*

1. $0 < \gamma < 1$.
2. $\mathcal{V}^s$ *is compact, convex, and non-singleton.*
3. u^s *is in the relative interior of* $\mathcal{V}^s$*.*
4. *(Non-redundancy) There do not exist distinct* $v^s, \tilde{v}^s \in \mathcal{V}^s$ *such that* v^s *is a positive affine transformation of* $\tilde{v}^s$*.*[16]

Clause 3 of Definition 4.2 means that if the agent can distort his rationale in one direction, then he can distort his rationale (at least slightly) in the opposite direction.

Non-redundancy requires that $\mathcal{V}^s$ does not contain extra rationales that are irrelevant for behavior. In particular, consider an agent with rationales v^s and $\alpha v^s + \beta$, for $\alpha \geq 1$ and $\beta \in \mathbb{R}$. By inspection of (1), switching from $\alpha v^s + \beta$ to v^s weakly increases total utility, so deleting $\alpha v^s + \beta$ has no effect on choice behavior. Non-redundancy requires that V^s includes at most one rationale in each positive affine equivalence class.

Regularity rules out that the agent always maximizes material utility u^s in state s.[17] In that case, it is not possible to fully identify the rationales.

Suppose that the agent's second-period choice correspondence in state s has a regular representation. The next theorem states that this representation is unique up to a positive affine transformation.

Theorem 4.3. *For any state s, if choice correspondence c_2^s has regular representations $(\gamma, u^s, \mathcal{V}^s)$ and $(\hat{\gamma}, \hat{u}^s, \hat{\mathcal{V}}^s)$, then $\hat{\gamma} = \gamma$ and there exist scale factor $\alpha > 0$ and additive*

[16] $\mathcal{V}^s$ satisfies non-redundancy if and only if no two distinct utilities in $\mathcal{V}^s$ represent the same preference relation on $\Delta(Z)$.

[17] For instance, regularity rules out the 'stoic' rationale $\bar{v}^s$ that is indifferent between all outcomes. Suppose $\bar{v}^s \in \mathcal{V}^s$. By $\mathcal{V}^s$ non-singleton, $\mathcal{V}^s$ contains some rationale $v^s \neq \bar{v}^s$. By convexity, we have $.5v^s + .5\bar{v}^s \in \mathcal{V}^s$. But $.5v^s + .5\bar{v}^s$ is a positive affine transformation of v^s, contradicting non-redundancy.

constants $\beta : \mathcal{V}^s \to \mathbb{R}$ *such that*

$$\hat{u}^s = \alpha u^s + \beta(u^s),$$
$$\hat{\mathcal{V}}^s = \{\alpha v^s + \beta(v^s) : v^s \in \mathcal{V}^s\}.$$

The proof is in Appendix A.

4.2.1 Proof sketch for Theorem 4.3

The key step in the proof is to recover $\mathcal{V}^s$ from choice data; we sketch the method here.

An example illustrates the main idea. The agent is contemplating a two-part project. There are three outcomes: <u>decline</u> yields 0 utils, <u>start & quit</u> yields $-s$ utils, and <u>start & finish</u> yields $\theta - 1 - s$ utils. The available rationales have $\theta \in [0, 3]$.

Let us suppose that $s = 1$, and fix some fully mixed lottery q (over outcomes), which is in the unit 3-simplex. Figure 1 depicts the indifference curves that pass through this point, for rationales with $\theta = 0$, $\theta = 2$, and $\theta = 3$. The shaded area consists of lotteries that *every* rationale regards as weakly worse than q. This set is a convex cone with vertex at q, and its supporting hyperplanes at q are the rationales' indifference curves through q. Thus, we can recover the preferences represented by the rationales by recovering the shaded area.

Returning to generality, let us fix some fully mixed lottery q and some state s. Let us consider the lotteries that every rationale regards as weakly worse than q, that is

$$\bar{W}_{\text{inner}}(q, s) \equiv \bigcap_{v^s \in \mathcal{V}^s} \{r \in \Delta(Z) : v^s(q) \geq v^s(r)\} \tag{2}$$

which corresponds to the shaded area in Figure 1. Since each rationale evaluates lotteries according to expected utility, (2) is a convex cone with vertex at q. It will turn out that the cone's supporting hyperplanes at q pin down $\mathcal{V}^s$.

Recall from Example 2 in Section 1 that a rationalizer's second-period choice can depend on the availability of *unchosen* first-period alternatives. We use this phenomenon to fully identify $\bar{W}_{\text{inner}}(q, s)$. Let x be a lottery. We say that x **matters for**[18] q (in state s) if there exists a menu M containing q and a collection of menus $\mathcal{M}$ containing M such that

1. after choosing M and foregoing the other members of $\mathcal{M}$, and learning the state is s, the agent chooses q,

2. but after choosing M and foregoing the other members of $\mathcal{M}$ and $\{x\}$, and learning the state is s, the agent *does not* choose q.

[18] In the proof, we use a definition of "matters for" that puts more structure on M and $\mathcal{M}$, but the definitions are equivalent.

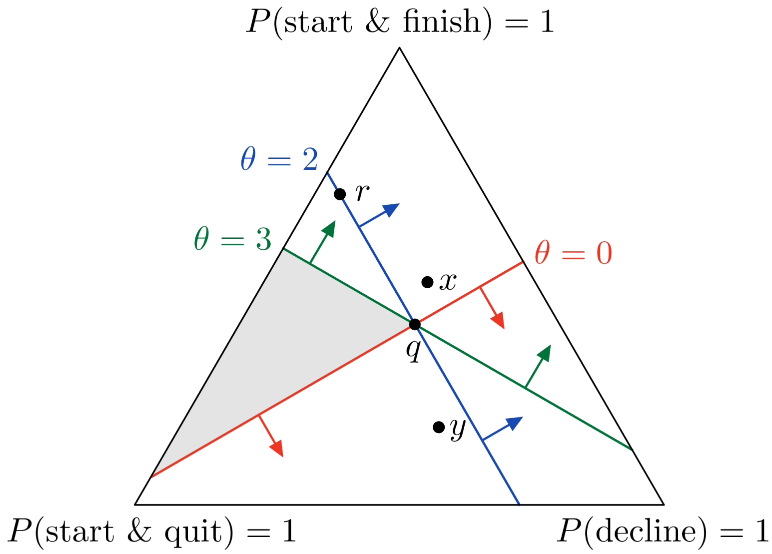

Figure 1: Indifference curves for lotteries over outcomes when $s = 1$. Arrows indicate direction of increasing utility for each rationale.

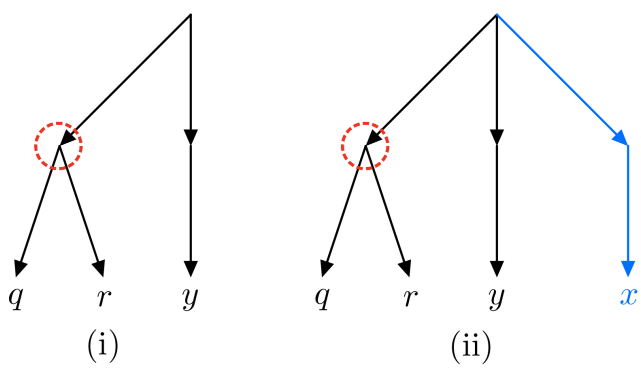

Figure 2: x matters for q (in state s) if q is chosen at the circled menu in decision problem (i), but q is not chosen at the circled menu in decision problem (ii).

Intuitively, the availability of x in the first period increases the regret cost of choosing q from M in the second period, which changes the agent's choice. Figure 2 provides an illustration with particular $M = \{q, r\}$ and $\mathcal{M} = \{\{q, r\}, \{y\}\}$.

Notice that if $x \in \bar{W}_{\text{inner}}(q, s)$, then *every* rationale weakly prefers q to x in state s, so the availability of x does not ever alter the regret cost of choosing q. It follows that x does not matter for q in state s. Thus, $x \notin \bar{W}_{\text{inner}}(q, s)$ if x matters for q in state s.

The converse also holds. Suppose $x \notin \bar{W}_{\text{inner}}(q, s)$. We want to show that x matters for q in s, that is, to find a menu $M \ni q$ and a collection of menus $\mathcal{M} \ni M$ such that offering $\{x\}$ in addition to $\mathcal{M}$ in the first period causes the agent to switch away from

q in the second period. Figure 1 illustrates the construction. Assume that material utility has $\theta = 2$. Material utility weakly prefers q to both r and y, so the agent will choose q from $\{q, r\}$ after foregoing $\{y\}$.[19] The rationale with $\theta = 3$ prefers r to all three alternatives, so the agent can always choose r without regret. Since every rationale strictly prefers at least one of y or x to q, choosing q results in positive regret when y and x were available. Since q and r have equal material utility, the agent will not choose q from $\{q, r\}$ after foregoing $\{y\}$ and $\{x\}$.

We find that similar constructions work for all regular representations and for all $x \notin \bar{W}_{\text{inner}}(q, s)$. This yields the lemma that $x \notin \bar{W}_{\text{inner}}(q, s)$ if and only if x matters for q in state s (Lemma A.3). This lemma establishes that $\bar{W}_{\text{inner}}(q, s)$ is identified from choice behavior. With this in hand, we then prove that the supporting hyperplanes of $\bar{W}_{\text{inner}}(q, s)$ at q are the indifference curves of the utility functions in $\mathcal{V}^s$. Thus, the preferences represented by the rationales can be deduced from choice data.

The preference represented by material utility function u^s is easy to deduce, as the agent will always choose the materially best alternative from a binary menu $\{q, r\}$ when he was not offered any other menu in the first period. We leave further details (including the scaling of rationales relative to material utility) to Appendix A.

4.3 Empirical content

The identification procedure discussed above can also be used to study the general empirical content of the theory—that is, the predictions that follow directly from the existence of a regular representation, without any further assumptions about material utility or rationales.

One simple prediction regards the agent's behavior when he has not made an *ex post* mistake. As noted in the last section, we can use binary menus to fully recover the agent's material preferences. This allows us to check whether a given first-period choice is an *ex post* mistake. We can then check whether the agent maximizes his material preference after any *ex post* optimal first-period choice, as the theory predicts.

A more interesting prediction uses Lemma A.3, discussed in the previous section. Suppose that the agent chooses r over q in state s when no other alternatives were available in the first period. We have seen that this implies $u^s(r) > u^s(q)$. Since material utility u is one of the rationales, it cannot be that every rationale weakly prefers q to r. By Lemma A.3, r matters for q. In summary: if the agent would choose r over q in the second period when no other alternatives were available in the first period, then the sunk cost of foregoing r in the first period can, under some circumstances, induce the decision maker to forego q in the second period.

[19] q and r yield identical total utility in this case. We could break the tie by perturbing r, at the cost of a longer argument.

The converse does not hold, as r may matter for q in state s even if $u^s(q) > u^s(r)$. Intuitively, r alone cannot make it costly to choose q, but r in combination with other, better alternatives can. This is not to say that any alternative can matter for q. In the proof of Theorem 4.3, we show that the set $\bar{W}_{\text{inner}}(q, s)$ must be nonempty. Applying Lemma A.3 delivers another prediction: there exist alternatives that do not matter for q.[20]

5 Comparative statics for complements

In this section, we impose additional structure on the decision problem to yield comparative statics results. With this structure, the theory predicts systematic deviations from the classical model.

We now assume that the rationales $\mathcal{V}$ have the form $\{w(a_1, a_2, \theta, s) : \theta \in \Theta\}$, for some set Θ and some function w. We use θ^* to denote the parameter value that corresponds to material utility, so $u(a_1, a_2, s) = w(a_1, a_2, \theta^*, s)$. Hence, the agent facing some decision problem D, having chosen a_1 from menu A_1 and observed state s, chooses $a_2 \in A_2(a_1)$ and $\theta \in \Theta$ to maximize

$$U_D(a_2, \theta \mid a_1, s) \equiv (1-\gamma)w(a_1, a_2, \theta^*, s) + \gamma \left[w(a_1, a_2, \theta, s) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta, s) \right].$$

We assume that w and U_D have non-empty maxima with respect to (a_1, a_2, θ), and similarly for subsets of these arguments.

We will assume that the choice variables are complements—the marginal return of raising one variable is non-decreasing in the other variables. All our results cover the case of $\mathcal{A}_1$, $\mathcal{A}_2$, and Θ totally ordered, and w supermodular in (a_1, a_2, θ). If additionally w is differentiable, then this reduces to the requirement that the cross partial derivatives are all non-negative. However, we state our results under weaker order-theoretic assumptions to expand their scope.

We now state some standard definitions; for more detail see Milgrom and Shannon (1994). Suppose X and Y are partially ordered sets. Function $f : X \times Y \times S \to \Re$ has *increasing differences* between x and y if for all $\tilde{x} \le \tilde{x}'$, all $\tilde{y} \le \tilde{y}'$, and all s, we have

$$f(\tilde{x}, \tilde{y}', s) - f(\tilde{x}, \tilde{y}, s) \le f(\tilde{x}', \tilde{y}', s) - f(\tilde{x}', \tilde{y}, s).$$

Suppose X is a lattice and Y is an arbitrary set. Function $f : X \times Y \to \Re$ is *super-*

[20] This prediction is easy to test if we assume structure on that rationales, such as that they are strictly increasing in money: r should not matter for q if every rationale regards r as worse than q.

modular in x if for all $\tilde{x}$, all $\tilde{x}'$, and all y, we have

$$f(\tilde{x}, y) + f(\tilde{x}', y) \leq f(\tilde{x} \wedge \tilde{x}', y) + f(\tilde{x} \vee \tilde{x}', y).$$

Given two lattices X and Y, we order $X \times Y$ according to the component-wise order. Given any lattice, we order subsets X and Y with the *strong set order*, writing $X \ll Y$ if for any $x \in X$ and $y \in Y$, we have $x \wedge y \in X$ and $x \vee y \in Y$. Given a partially ordered set X and a lattice Y, we say that a correspondence $J : X \rightrightarrows Y$ is *monotone non-decreasing* if $x \leq x'$ implies that $J(x) \ll J(x')$.[21]

In decision problems with complements, Topkis's theorem implies that raising the first action, *ceteris paribus*, raises the classical agent's chosen second action. For rationalizers, we find that raising the first action raises both the chosen second action and the chosen rationale.

Proposition 5.1. *Let $\mathcal{A}_1$ be a partially ordered set, let $\mathcal{A}_2$ be a lattice and let Θ be totally ordered. Suppose that w has increasing differences between a_1 and (a_2, θ) and is supermodular in (a_2, θ). Suppose that $A_2(a_1)$ is monotone non-decreasing. For any decision problem D and any state s, the correspondence*

$$\underset{\substack{a_2 \in A_2(a_1) \\ \theta \in \Theta}}{\operatorname{argmax}} \; U_D(a_2, \theta \mid a_1, s),$$

is monotone non-decreasing in a_1.

The proof is in Appendix B.1. The conclusion of Proposition 5.1 implies that the agent's chosen actions,

$$\underset{a_2 \in A_2(a_1)}{\operatorname{argmax}} \left\{ \max_{\theta \in \Theta} U_D(a_2, \theta \mid a_1, s) \right\},$$

are monotone non-decreasing in a_1.

The next theorem shows that under appropriate complementarities, the theory yields systematic deviations from the classical benchmark. In the theory, sunk-cost effects are one instance of a more general phenomenon, stated as follows.

Theorem 5.2. *Let $\mathcal{A}_1$ be a partially ordered set, let $\mathcal{A}_2$ be a lattice and let Θ be totally ordered. Suppose that w has increasing differences between a_1 and (a_2, θ) and is supermodular in (a_2, θ). Suppose that $A_2(a_1)$ is monotone non-decreasing. If the agent's time-*1 *choice was ex post weakly higher than optimal, i.e. $\bar{a}_1 \geq a_1^*$ for*

$$a_1^* \in \underset{a_1 \in A_1}{\operatorname{argmax}} \left\{ \max_{a_2 \in A_2(a_1)} w(a_1, a_2, \theta^*, s) \right\},$$

[21] The relation $\leq$ is reflexive, so this implies that for all x, $J(x)$ is a sublattice of Y.

then the agent's time-2 choice is weakly higher than materially optimal, i.e.

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \left\{ \max_{\theta \in \Theta} U_D(a_2, \theta \mid \bar{a}_1, s) \right\} \gg \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*, s), \tag{3}$$

and for any selection $\bar{a}_2$ from the left-hand side of (3), there exists $\bar{\theta} \geq \theta^*$ such that

$$\bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} U_D(a_2, \bar{\theta} \mid \bar{a}_1, s).$$

Symmetrically, if the agent's time-1 choice was ex post weakly lower than optimal, then the agent's time-2 choice is weakly lower than materially optimal, and there exists $\bar{\theta} \leq \theta^*$ such that $\bar{a}_2 \in \operatorname{argmax}_{a_2 \in A_2(\bar{a}_1)} U_D(a_2, \bar{\theta} \mid \bar{a}_1, s)$.

The proof is in Appendix B.2.

Whether first actions, second actions, and rationales are complements depends not only on the agent's preferences, but also on the structure of the decision problem.[22] As is usual for comparative statics results, applying Theorem 5.2 requires good judgement in defining orders.

Now we consider changing the unchosen alternatives at $t = 1$. Observe that for a classical agent, unchosen alternatives from the first menu have no effect on second-period choice. Under complements, the next theorem predicts that increasing the first menu (in the strong set order), while leaving the first choice unchanged, *decreases* the rationalizer's second-period choice as well as her rationale.

Theorem 5.3. *Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be lattices and let Θ be totally ordered. Suppose that w is supermodular in (a_1, a_2, θ). Suppose that the correspondence A_2 satisfies*

$$a_2 \in A_2(a_1) \text{ and } a_2' \in A_2(a_1') \implies a_2 \wedge a_2' \in A_2(a_1 \wedge a_1') \text{ and } a_2 \vee a_2' \in A_2(a_1 \vee a_1'). \tag{4}$$

Take any $A_1, A_1' \subseteq \mathcal{A}_1$ such that $A_1 \ll A_1'$. Let D and D' denote the decision problems with A_1 and A_1' respectively. For any $\bar{a}_1 \in A_1 \cap A_1'$ and any s we have

$$\operatorname*{argmax}_{\substack{a_2 \in A_2(\bar{a}_1) \\ \theta \in \Theta}} U_D(a_2, \theta \mid \bar{a}_1, s) \gg \operatorname*{argmax}_{\substack{a_2 \in A_2(\bar{a}_1) \\ \theta \in \Theta}} U_{D'}(a_2, \theta \mid \bar{a}_1, s). \tag{5}$$

The proof is in Appendix B.3.

Condition (4) is stronger that A_2 monotone non-decreasing. It is implied by $\mathcal{A}_1$ totally ordered and A_2 monotone non-decreasing. Alternatively, it is implied by constant A_2.

[22]Theorem 4.3 took as data the class of all decision problems in which the first action chooses a menu of lotteries, and the second action chooses a lottery. Theorem 5.2 delivers comparative statics for decision problems in which first actions, second actions, and rationales are complements. Some decision problems taken as data in Theorem 4.3 may not have the relevant complementarities for the application of Theorem 5.2.

Note that (5) implies that the rationalizer's chosen actions decrease *i.e.*

$$\underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}} \left\{ \max_{\theta \in \Theta} U_D(a_2, \theta \mid \bar{a}_1, s) \right\} \gg \underset{a_2 \in A_2(\bar{a}_1)}{\operatorname{argmax}} \left\{ \max_{\theta \in \Theta} U_{D'}(a_2, \theta \mid \bar{a}_1, s) \right\}.$$

This provides a testable prediction of the theory that does not require us to separately identify the adopted rationales or the materially optimal benchmark.

5.1 Applications of results

We examine some natural decision problems that satisfy the assumptions of Proposition 5.1, Theorem 5.2, and Theorem 5.3.

5.1.1 The sunk-cost effect for two-part projects

The agent chooses effort levels $a_1 \in [0, 1]$ and $a_2 \in [0, 1]$. The project succeeds with probability $a_1 a_2$, he receives a reward valued at $\theta \geq 0$ if it succeeds, and he pays effort cost $sc_1(a_1) + c_2(a_2)$, for continuous non-decreasing cost functions c_1 and c_2, where s is a cost shock for first-period effort. Hence

$$w(a_1, a_2, \theta) = \theta a_1 a_2 - sc_1(a_1) - c_2(a_2)$$

The agent chooses a_1 before learning s, so the materially optimal choice of a_2 does not depend on the realized s. Theorem 5.2 implies that when s has a high enough realization, so that a_1 was *ex post* too high, the rationalizer's choice of a_2 is distorted upwards compared to the classical benchmark. Hence, the agent persists more than is materially optimal for projects that turned out to be unexpectedly costly.

Theorem 5.3 implies that if we mandate a floor for first-period effort, and this floor does not bind, then the agent's second-period effort weakly decreases. Similarly, if we mandate a ceiling for first-period effort, and this ceiling does not bind, then the agent's second-period effort weakly increases.

5.1.2 Encountering the same problem twice

The agent faces a decision problem, chooses an action, then learns the state, and then faces the same problem again. That is, $A_1 = A_2 = A$ and

$$w(a_1, a_2, \theta, s) = \phi(a_1, \theta, s) + \phi(a_2, \theta, s),$$

for some function $\phi : A \times \Theta \times S \to \mathbb{R}$ that is supermodular in (a, θ). Let $a^*(s)$ be an *ex post* optimal choice in state s, *i.e.* $a^*(s) \in \operatorname{argmax}_a \phi(a, \theta^*, s)$.

Material utility is time-separable, so upon learning the state, a classical agent's time-2 choice is $\text{argmax}_a\, \phi(a, \theta^*, s)$; his second-period choice does not depend on his first-period choice. By contrast, Theorem 5.2 implies that a rationalizer chooses a_2 to maximize $(1-\gamma)\phi(a_2, \theta^*, s) + \gamma\phi(a_2, \bar{\theta}, s)$, with $\bar{\theta} \geq \theta^*$ when $a_1 \geq a^*(s)$ and $\bar{\theta} \leq \theta^*$ when $a_1 \leq a^*(s)$. Attempting to rationalize the earlier decision creates a link between otherwise-separate decisions, pulling the rationalizer's second-period choice away from $a^*(s)$ in the direction of his initial choice a_1. Thus, the rationalizer's choice is 'stickier' than a classical agent's choice, responding less to the new information about the state.

5.1.3 Belief elicitation

In many laboratory experiments, subjects provide point estimates of some quantity, then learn some information, and finally report updated estimates. They are paid for one decision drawn at random, so they encounter the same problem twice, in the sense of Section 5.1.2.[23]

The outcome Y is a real-valued random variable with countable support.[24] The subject makes an incentivized report of $\mathbb{E}[Y]$, then observes a signal X with known conditional distribution $g(x \mid y)$, then makes an incentivized report of $\mathbb{E}[Y \mid X = x]$. The available rationales are priors on Y; these are a set of probability mass functions indexed by θ, denoted $(h_\theta)_{\theta\in\Theta}$. We assume that this set is totally ordered by the monotone likelihood ratio property (MLRP) (Milgrom, 1981), that is, for any $\theta > \theta'$ and any $y > y'$

$$h_\theta(y)h_{\theta'}(y') > h_{\theta'}(y)h_\theta(y').$$

This restriction is without loss of generality if Y is a Bernoulli random variable, *i.e.* when the agent is being asked to report the probability of some event. We assume that each h_θ has full support, so that no rationale is ruled out by some signal realization.

Given prior h_θ and signal realization x, we denote the posterior probability mass function $h_\theta(y \mid X = x)$. The agent reports a_1, then observes the signal realization, then reports a_2. For each report a_t, the agent faces quadratic loss (conditional on the signal realization), resulting in the payoff

$$\phi(a_t, \theta, x) = -\sum_y (a_t - y)^2 h_\theta(y \mid X = x).$$

This captures the interim expected utility of a risk-neutral agent facing a quadratic scoring rule. It also captures the interim expected utility of an agent with general risk preferences facing an appropriate binarized scoring rule (Hossain and Okui, 2013).

[23] Azrieli et al. (2018) study the merits of paying one decision drawn at random.

[24] A parallel construction works if Y has support in some interval and each available rationale is an atomless distribution with strictly positive density.

Given the same signal realization, MLRP-ordered priors induce posteriors that are ordered by first-order stochastic dominance (Milgrom, 1981; Klemens, 2007). Thus, if $\theta > \theta'$ then $h_\theta(y \mid X = x)$ first-order stochastically dominates $h_{\theta'}(y \mid X = x)$. It follows that ϕ is supermodular in (a_t, θ).

Our analysis in Section 5.1.2 implies that when $a_1 \geq \mathbb{E}[Y \mid X = x]$, then the agent's reported posterior beliefs are distorted upwards, $a_2 \geq \mathbb{E}[Y \mid X = x]$.[25] Such preference for consistency in belief elicitation is folk wisdom amongst experimenters.[26]

5.1.4 Consumption under two-part tariffs

Consider a consumer facing two-part tariffs, each consisting of a lump-sum payment L and a per-unit price p (both non-negative), as in Thaler (1980). The consumer faces a finite list of such tariffs, denoted $(L_k, p_k)_{k \in K}$. Without loss of generality, we assume that the list contains no dominated tariffs and no duplicates. We order the tariffs so that $L_1 < L_2 < \cdots$ and $p_1 > p_2 > \cdots$.

The timing is as follows:

1. The consumer chooses a tariff (L_k, p_k) from the list or declines.
2. The consumer learns the taste shock $s \in [0, 1]$.
3. If the consumer chose a tariff, the consumer chooses quantity $q \in [0, 1]$.

The set of rationales is indexed by $\Theta = [0, 1]$. The consumer's utility from tariff (L, p) under rationale θ is

$$u(q, s, \theta, L, p) = s\psi(q, \theta) - pq - L$$

where ψ is continuous and supermodular in both arguments, and $\psi(0, \theta) = 0$ for all θ. The utility from declining is 0.

Consider a list comprised of tariffs (L_1, p_1) and (L_2, p_2), with $L_1 < L_2$ and $p_1 > p_2$. By Proposition 5.1, changing the chosen tariff from (L_1, p_1) to (L_2, p_2) weakly raises the quantity consumed, for every realization of the taste shock.

Observe that for a classical consumer, once a tariff has been chosen, the lump-sum L is sunk and has no effect on the quantity demanded. By contrast, (Thaler, 1980) proposes that when a consumer responds to sunk costs, raising the lump-sum payment can increase the quantity demanded. We formalize this observation in the context of our model.

[25] Also in that case, when $a_1 \in \operatorname{argmax}_{a \in A_1} \{\phi(a, \theta', x)\}$ for some θ', then $a_2 \leq a_1$. (If $a_2 > a_1$, then $\phi(a_1, \theta^*, x) > \phi(a_2, \theta^*, x)$, in which case by reducing a_2 to a_1 and adopting the rationale θ', the agent can achieve higher material utility and zero regret, which contradicts the optimality of a_2.) This assumption is satisfied whenever the rationales include all full-support priors on Y.

[26] See Falk and Zimmermann (2018) for experiments finding that laboratory subjects report beliefs that are distorted towards their prior reports.

Take any per-unit price p and taste shock s. Take any $L < L'$ such that

$$\max_{q\in[0,1]} \{s\psi(q,\theta^*) - pq - L\} \geq 0 > \max_{q\in[0,1]} \left\{s\psi(q,\theta^*) - pq - L'\right\}.$$

If the consumer was offered only tariff (L, p) and accepted, then in state s this was not an *ex post* mistake, so he demands the materially optimal quantity. By contrast, if he was offered only tariff (L', p) and accepted, then his first choice was *ex post* too high. Theorem 5.2 implies that his demand is weakly higher after (L', p) than after (L, p). Moreover, it is strictly higher for various simple functional forms, such as $\psi(q,\theta) = \theta\sqrt{q}$ with interior θ^*. Thus, high enough lump-sum payments can raise demand compared to the material optimum, provided that they do not cause the consumer to decline the tariff.

Finally, let us take compare two lists, one of which is produced by truncating the other from above. That is, we have $(L_k, p_k)_{k=1}^{K}$ and $(L_k, p_k)_{k=1}^{K'}$, for $K < K'$. Suppose we fix the agent's chosen tariff at (L_j, p_j) for $j \leq K$ and switch from the full list $(L_k, p_k)_{k=1}^{K'}$ to the truncated list $(L_k, p_k)_{k=1}^{K}$. By Theorem 5.3, this change weakly increases demand in every state. This suggests that a firm selling to rationalizing consumers may find it beneficial to withdraw options with high lump-sums and low marginal prices, especially if those options are seldom chosen.

6 Extensions

In our model, all uncertainty is resolved before the second action. If instead the agent observes a random variable X correlated with the state, then it is natural to stipulate that she assesses regret conditional on the signal realization. This can be accommodated by transforming the state space and utility functions, so our previous results also apply to decisions with noisy signals. Namely, given each signal realization x and each rationale v, we define

$$\bar{v}(a_1, a_2, x) \equiv \mathbb{E}_s[v(a_1, a_2, s) \mid X = x],$$

and transform the objective function (1) by substituting $\bar{v}$ for v and x for s. Moreover, each rationale can represent a different prior about s, so that the agent can rationalize her actions by adjusting her beliefs, as in Section 5.1.3.

More subtly, the agent's first action could determine not only her payoffs, but also what signal she sees. In this case, sophisticated rationalizers have a novel motive for information avoidance: If the sophisticate avoids new information, then *ex ante* optimal actions are also *ex post* optimal, so there is no loss of material utility from rationalizing behavior.

The theory can be extended beyond two periods. To do so, we stipulate that at each time t, the agent chooses some rationale v, and compares the expected utility of her strategy under v to the expected utility of the interim-optimal strategy for v, with both expectations conditional on the information available at t. At each t, she chooses a continuation strategy that maximizes a weighted sum of expected material utility and this generalized regret term. For sophisticates, we restrict the continuation strategies to be consistent with future rationalizing behavior.

Should we additionally require that the chosen rationales are stable over time? That is, if the agent acts on Monday, Tuesday, and Wednesday, can she adopt one rationale on Tuesday and then a different rationale on Wednesday? The theory does not forbid such inconsistent rationalizations, but it does weigh against them, because Tuesday's rationale affects Tuesday's action, which then affects Wednesday's rationale. If we directly required Tuesday's rationale to be equal to Wednesday's rationale, then the theory's predictions would vary with the addition of 'dummy periods' with singleton action menus. In the interest of invariance, we do not impose this requirement.

7 Conclusion

Standard economic theory holds that people make decisions to satisfy their preferences. By contrast, the present theory allows that people adopt preferences to rationalize their past decisions. Reassuringly, this is also tractable for formal economic analysis.

Having come this far, it is for the reader to decide whether the exercise was worthwhile.

References

Acharya, A., M. Blackwell, and M. Sen (2018): "Explaining preferences from behavior: A cognitive dissonance approach," *The Journal of Politics*, 80, 400–411.

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

Alonso, R. (2007): *The IRA and Armed Struggle*, Political Violence, Taylor & Francis.

Arkes, H. R. and C. Blumer (1985): "The psychology of sunk cost," *Organizational Behavior and Human Decision Processes*, 35, 124–140.

Aronson, E. and J. Mills (1959): "The effect of severity of initiation on liking for a group." *The Journal of Abnormal and Social Psychology*, 59, 177.

Ashraf, N., J. Berry, and J. M. Shapiro (2010): "Can higher prices stimulate product use? Evidence from a field experiment in Zambia," *American Economic Review*, 100, 2383–2413.

Augenblick, N. (2016): "The sunk-cost fallacy in penny auctions," *The Review of Economic Studies*, 83, 58–86.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018): "Incentives in experiments: A theoretical analysis," *Journal of Political Economy*, 126, 1472–1503.

Baliga, S. and J. C. Ely (2011): "Mnemonomics: the sunk cost fallacy as a memory kludge," *American Economic Journal: Microeconomics*, 3, 35–67.

Becker, G. S. and C. B. Mulligan (1997): "The endogenous determination of time preference," *The Quarterly Journal of Economics*, 112, 729–758.

Bell, D. E. (1982): "Regret in decision making under uncertainty," *Operations research*, 30, 961–981.

Bernheim, B. D., L. Braghieri, A. Martínez-Marquina, and D. Zuckerman (2021): "A theory of chosen preferences," *American Economic Review*, 111, 720–54.

Blank, H., J. Musch, and R. F. Pohl (2007): "Hindsight bias: On being wise after the event," *Social Cognition*, 25, 1–9.

Brehm, J. W. (1956): "Postdecision changes in the desirability of alternatives." *The Journal of Abnormal and Social Psychology*, 52, 384.

Brunnermeier, M. K. and J. A. Parker (2005): "Optimal expectations," *American Economic Review*, 95, 1092–1118.

Camerer, C. F. and R. A. Weber (1999): "The econometrics and behavioral economics of escalation of commitment: A re-examination of Staw and Hoang's NBA data," *Journal of Economic Behavior & Organization*, 39, 59–82.

Chen, M. K. and J. L. Risen (2010): "How choice affects and reflects preferences: revisiting the free-choice paradigm." *Journal of personality and social psychology*, 99, 573.

Cushman, F. (2020): "Rationalization is rational," *Behavioral and Brain Sciences*, 43.

Eyster, E. (2002): "Rationalizing the past: A taste for consistency," *Nuffield College Mimeograph.*

Falk, A. and F. Zimmermann (2018): "Information Processing and Commitment," *The Economic Journal*, 613, 1983–2002.

Festinger, L. (1957): *A Theory of Cognitive Dissonance.*, California: Stanford University Press.

Fischhoff, B. (1975): "Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty." *Journal of Experimental Psychology: Human perception and performance*, 1, 288.

——— (1977): "Perceived informativeness of facts." *Journal of Experimental Psychology: Human Perception and Performance*, 3, 349.

Fischhoff, B. and R. Beyth (1975): "I knew it would happen: Remembered probabilities of once—future things," *Organizational Behavior and Human Performance*, 13, 1–16.

Friedman, D., K. Pommerenke, R. Lukose, G. Milam, and B. A. Huberman (2007): "Searching for the sunk cost fallacy," *Experimental Economics*, 10, 79–104.

Gazzaniga, M. S. (1967): "The split brain in man," *Scientific American*, 217, 24–29.

——— (2005): "Forty-five years of split-brain research and still going strong," *Nature Reviews Neuroscience*, 6, 653–659.

Gerard, H. B. and G. C. Mathewson (1966): "The effects of severity of initiation on liking for a group: A replication," *Journal of Experimental Social Psychology*, 2, 278–287.

Guenzel, M. (2021): "In too deep: The effect of sunk costs on corporate investment," Tech. rep., University of Pennsylvania Working Paper.

Harmon-Jones, E. and C. Harmon-Jones (2007): "Cognitive dissonance theory after 50 years of development," *Zeitschrift für Sozialpsychologie*, 38, 7–16.

Herrmann, P. N., D. O. Kundisch, and M. S. Rahman (2015): "Beating irrationality: does delegating to IT alleviate the sunk cost effect?" *Management Science*, 61, 831–850.

Ho, T.-H., I. P. Png, and S. Reza (2018): "Sunk cost fallacy in driving the world's costliest cars," *Management Science*, 64, 1761–1778.

Hong, F., W. Huang, and X. Zhao (2019): "Sunk cost as a self-management device," *Management Science*, 65, 2216–2230.

Hossain, T. and R. Okui (2013): “The binarized scoring rule,” *Review of Economic Studies*, 80, 984–1001.

Hurwicz, L. (1972): “On informationally decentralized systems,” in *Decision and Organization*, ed. by C. B. McGuire and R. Radner, Amsterdam: North-Holland, chap. 14, 297–336.

Imas, A. (2016): “The realization effect: Risk-taking after realized versus paper losses,” *American Economic Review*, 106, 2086–2109.

Johansson, P., L. Hall, S. Sikström, and A. Olsson (2005): “Failure to detect mismatches between intention and outcome in a simple decision task,” *Science*, 310, 116–119.

Johansson, P., L. Hall, S. Sikström, B. Tärning, and A. Lind (2006): “How something can be said about telling more than we can know: On choice blindness and introspection,” *Consciousness and cognition*, 15, 673–692.

Keefer, Q. A. (2017): “The sunk-cost fallacy in the National Football League: Salary cap value and playing time,” *Journal of Sports Economics*, 18, 282–297.

Ketel, N., J. Linde, H. Oosterbeek, and B. van der Klaauw (2016): “Tuition fees and sunk-cost effects,” *The Economic Journal*, 126, 2342–2362.

Klemens, B. (2007): “When Do Ordered Prior Distributions Induce Ordered Posterior Distributions?” *Available at SSRN 964720.*

Langer, E. J. (1975): “The illusion of control.” *Journal of personality and social psychology*, 32, 311.

Loomes, G. and R. Sugden (1982): “Regret theory: An alternative theory of rational choice under uncertainty,” *The economic journal*, 92, 805–824.

Martens, N. and H. Orzen (2021): “Escalating commitment to a failing course of action—A re-examination,” *European Economic Review*, 103811.

McAfee, R. P., H. M. Mialon, and S. H. Mialon (2010): “Do sunk costs matter?” *Economic Inquiry*, 48, 323–336.

McCarthy, A. M., F. D. Schoorman, and A. C. Cooper (1993): “Reinvestment decisions by entrepreneurs: rational decision-making or escalation of commitment?” *Journal of business venturing*, 8, 9–24.

MILGROM, P. AND C. SHANNON (1994): "Monotone comparative statics," *Econometrica*, 157–180.

MILGROM, P. R. (1981): "Good news and bad news: Representation theorems and applications," *The Bell Journal of Economics*, 380–391.

NEGRINI, M., A. RIEDL, AND M. WIBRAL (2020): "Still in search of the sunk cost bias," Tech. rep., CESifo Working Paper.

NISBETT, R. E. AND T. D. WILSON (1977): "Telling more than we can know: Verbal reports on mental processes." *Psychological review*, 84, 231.

PRENDERGAST, C. AND L. STOLE (1996): "Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning," *Journal of political Economy*, 104, 1105–1134.

PRESSON, P. K. AND V. A. BENASSI (1996): "Illusion of control: A meta-analytic review," *Journal of social behavior and personality*, 11, 493.

RABIN, M. (1994): "Cognitive dissonance and social change," *Journal of Economic Behavior & Organization*, 23, 177–194.

——— (2013): "An approach to incorporating psychology into economics," *American Economic Review*, 103, 617–22.

RIDOUT, S. (2020): "A Model of Justification," *arXiv preprint arXiv:2003.06844*.

RISEN, J. L. AND M. K. CHEN (2010): "How to study choice-induced attitude change: Strategies for fixing the free-choice paradigm," *Social and Personality Psychology Compass*, 4, 1151–1164.

ROTEMBERG, J. J. (1994): "Human relations in the workplace," *Journal of Political Economy*, 102, 684–717.

ROTH, S., T. ROBBERT, AND L. STRAUS (2015): "On the sunk-cost effect in economic decision-making: a meta-analytic review," *Business research*, 8, 99–138.

SARVER, T. (2008): "Anticipating regret: Why fewer options may be better," *Econometrica*, 76, 263–305.

SAVAGE, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

Schoorman, F. D. (1988): "Escalation bias in performance appraisals: An unintended consequence of supervisor participation in hiring decisions." *Journal of Applied Psychology*, 73, 58.

Sheehan, N. and E. Kenworthy (1971): *The Pentagon Papers*, Times Books.

Smith, M. L. R. (1997): *Fighting for Ireland?: the military strategy of the Irish Republican movement*, London: Routledge.

Staw, B. M. (1976): "Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action," *Organizational behavior and human performance*, 16, 27–44.

——— (1980): "Rationality and justification in organizational life," *Research in organizational behavior*, 2, 45–80.

Staw, B. M., S. G. Barsade, and K. W. Koput (1997): "Escalation at the credit window: A longitudinal study of bank executives' recognition and write-off of problem loans." *Journal of Applied Psychology*, 82, 130.

Staw, B. M. and H. Hoang (1995): "Sunk costs in the NBA: Why draft order affects playing time and survival in professional basketball," *Administrative Science Quarterly*, 474–494.

Suzuki, T. (2019): "Choice set dependent performance and post-decision dissonance," *Journal of Economic Behavior & Organization*, 163, 24–42.

Thaler, R. (1980): "Toward a positive theory of consumer choice," *Journal of Economic Behavior & Organization*, 1, 39–60.

Tykocinski, O. E., T. S. Pittman, and E. E. Tuttle (1995): "Inaction inertia: Foregoing future benefits as a result of an initial failure to act." *Journal of personality and social psychology*, 68, 793.

Wong, K. F. E. and J. Y. Kwong (2007): "The role of anticipated regret in escalation of commitment." *Journal of Applied Psychology*, 92, 545.

Yariv, L. (2005): "I'll See it When I Believe it - A Simple Model of Cognitive Consistency," *working paper.*

A Proof of Theorem 4.3

We establish some notation.

- Since we will only compare choices within a given state, we drop the superscript s throughout.
- For any $G \subseteq \mathcal{U}$, let G_{pref} denote the set of preferences with representations in G.
- For any set X, let $\text{co}(X)$ denote the convex hull, and let $\bar{X}$ denote the closure.[27]
- For any $q \in \Delta(Z)$ and any $M \in \mathcal{K}_f(\Delta(Z))$ such that $q \in M$, let

$$U\left(q \mid M\right) \equiv (1-\gamma)u(q) + \gamma \max_{v\in\mathcal{V}} \left(v(q) - \max_{\hat{q}\in M} v(\hat{q}) \right).$$

- For any $q \in \Delta(Z)$, let

$$W_{\text{inner}}(q) \equiv \bigcap_{v\in\mathcal{V}} \{r \in \Delta(Z) : v(q) > v(r)\}$$
$$W_{\text{outer}}(q) \equiv \bigcap_{v\in\mathcal{V}} \{r \in \Delta(Z) : ((1-\gamma)u + \gamma v)(q) > ((1-\gamma)u + \gamma v)(r)\}.$$

The next lemma establishes some properties of $W_{\text{inner}}(q)$ and $W_{\text{outer}}(q)$.

Lemma A.1. *For any regular representation $(\gamma, u, \mathcal{V})$ and any $q \in int(\Delta(Z))$, $W_{inner}(q)$ and $W_{outer}(q)$ are open and nonempty, and satisfy*

$$\bar{W}_{inner}(q) = \bigcap_{v\in\mathcal{V}} \{r \in \Delta(Z) : v(q) \geq v(r)\} \tag{6}$$
$$\bar{W}_{outer}(q) = \bigcap_{v\in\mathcal{V}} \{r \in \Delta(Z) : ((1-\gamma)u + \gamma v)(q) \geq ((1-\gamma)u + \gamma v)(r)\}. \tag{7}$$

Proof. For openness of $W_{\text{inner}}(q)$: Consider a sequence $r_i \to r$ such that $r_i \notin W_{\text{inner}}(q)$ for all i. By definition of $W_{\text{inner}}(q)$, for each r_i, there exists some $v_i \in \mathcal{V}$ such that $v_i(r_i) \geq v_i(q)$. Since $\mathcal{V}$ is compact, we can pass to a convergent subsequence $\bar{v} \in \mathcal{V}$. We must have $\bar{v}(r) \geq \bar{v}(q)$, so $r \notin W_{\text{inner}}(q)$. The argument for openness of $W_{\text{outer}}(q)$ is the same since $\{(1-\gamma)u + \gamma v : v \in \mathcal{V}\}$ inherits compactness from $\mathcal{V}$.

For nonemptiness of $W_{\text{inner}}(q)$: Fix $q \in \Delta(Z)$ and $R \in \mathcal{K}_f(\Delta(Z))$. Suppose there does not exist $v \in \mathcal{V}$ such that $v(q) \geq \max_{r\in R} v(r)$. We show that there exists $\bar{r} \in \text{co}(R)$ such that $v(q) < v(\bar{r})$ for all $v \in \mathcal{V}$. For each $v \in \mathcal{V}$, we have some $r \in R$ such that $v(r) < v(q)$. It without loss to assume that $v(q) = 0$ for all $v \in \mathcal{V}$. We want to find a

[27] When X is a set of preferences, a preference $\succsim \in \text{co}(X)$ if some representation of $\succsim$ is a convex combination of representations of preferences in S. A non-constant preference $\succsim \in \bar{X}$ if some representation of $\succsim$ is the limit of representations of preferences in X.

set of weights α such that

$$\sum_{r \in R} \alpha(r) v(r) > 0 \tag{8}$$

for all $v \in \mathcal{V}$. Enumerate the elements of R: $(r_1, \ldots, r_{|R|})$. For each $v \in \mathcal{V}$, let v_R denote the vector with $v(r_i)$ as its i-th entry. Let $\mathcal{V}_R$ denote $\{v_R : v \in \mathcal{V}\}$. As with $\mathcal{V}$, $\mathcal{V}_R$ is nonempty, compact and convex. Let $N \equiv \mathbb{R}^{|R|}_{\leq 0}$, which is nonempty, closed and convex. Notice that no element of $\mathcal{V}_R$ can be less than or equal to 0. (Otherwise, some $v \in \mathcal{V}$ would rank q weakly higher than each member of R.) Thus, $\mathcal{V}_R$ and N are disjoint, and we can apply the separating hyperplane theorem. This delivers a nonzero $\alpha \in \mathbb{R}^{|R|}$ and $\psi \in \mathbb{R}$ such that $\alpha' n < \psi < \alpha' v_R$ for all $n \in N$, $v_R \in \mathcal{V}_R$. Since the zero vector belongs to N, we must have $\psi > 0$. Suppose the ith element of α is strictly negative. By choosing n with a sufficiently negative number in ith position and zeros elsewhere, we get $\alpha' n > \psi$, a contradiction. Thus, each element of α is weakly positive. If we rescale α to a unit sum, we still have $\alpha' v_R > 0$ for all $v_R \in \mathcal{V}_R$. This inequality is equivalent to (8).

Now fix some $q \in \text{int}(\Delta(Z))$, and suppose that $W_{\text{inner}}(q) = \emptyset$. Since there is no $r \in \Delta(Z)$ such that $v(q) > v(r)$ for all $v \in \mathcal{V}$, there cannot be any $\bar{r} \in \Delta(Z)$ such that $v(\bar{r}) > v(q)$ for all $v \in \mathcal{V}$. Fix any $R \in \mathcal{K}_f(\Delta(Z))$ such that $q \in \text{int}(\text{co}(R))$. By the previous argument, there must be some $v \in \mathcal{V}$ such that $v(q) \geq \max_{r \in R} v(r)$. Since $q \in \text{int}(\text{co}(R))$, v must be a constant. By non-redundancy and $|\mathcal{V}| > 1$, $\mathcal{V}$ must also contain some non-constant utility v'. Since $\mathcal{V}$ is convex, it contains all convex combinations of v' and v, which violates non-redundancy. Conclude that $\mathcal{V}$ does not contain a constant preference, so $W_{\text{inner}}(q) \neq \emptyset$. Moreover, $W_{\text{inner}}(q) \subseteq W_{\text{outer}}(q)$, so $W_{\text{outer}}(q) \neq \emptyset$.

For (6): Fix some $q \in \text{int}(\Delta(Z))$ and some $r \in \Delta(Z)$ such that $v(q) \geq v(r)$ for all $v \in \mathcal{V}$. We want to show that r is a limit point of $W_{\text{inner}}(q)$. Let $\epsilon_k \equiv \frac{1}{2^k}$ for $k \in \mathbb{N}$. By q interior, $W_{\text{inner}}(q)$ is non-empty. Let x be some element of $W_{\text{inner}}(q)$, and define $q_k \equiv \epsilon_k x + (1 - \epsilon_k) q$. Observe that $q_k \to q$ and for all k, $q_k \in W_{\text{inner}}(q)$. For all k and all $v \in \mathcal{V}$, we have

$$v(q) > \epsilon_k v(q_k) + (1 - \epsilon_k) v(r) = v(\epsilon_k q_k + (1 - \epsilon_k) r)$$

and thus $\epsilon_k q_k + (1 - \epsilon_k) r \in W_{\text{inner}}(q)$. Moreover, $\epsilon_k q_k + (1 - \epsilon_k) r \to r$, so r is a limit point of $W_{\text{inner}}(q)$.

Now fix some $r \in \bar{W}_{\text{inner}}(q)$. We have a sequence $r_k \to r$ such that, for each k, $v(q) > v(r_k)$ for all $v \in \mathcal{V}$. By continuity of v, $v(q) \geq v(r)$ for all $v \in \mathcal{V}$. The same arguments work for (7). $\square$

Next, we show that $W_{\text{inner}}(q)$ can be fully identified from choice data. Since the agent's $t = 2$ choices depend on the collection $\mathcal{M}$ only via $\bigcup \mathcal{M}$, we abuse notation and write $c_2(M|\mathcal{M})$ as $c_2(M|\bigcup \mathcal{M})$.

Definition A.2. *Fix $q, x \in \Delta(Z)$. Say that x* ***matters for*** *q if there exist $y, r \in \Delta(Z)$ such that*

$$\begin{aligned} q &\in c_2\left(q, r \mid \; q, r, y\right) \\ q &\notin c_2\left(q, r \mid q, r, y, x\right). \end{aligned}$$

Lemma A.3. *For any regular representation $(\gamma, u, \mathcal{V})$, any $q \in int(\Delta(Z))$, and any $x \in \Delta(Z)$, we have that $x \notin \bar{W}_{inner}(q)$ if and only if x matters for q.*

Proof. Suppose that $x \notin \bar{W}_{\text{inner}}(q)$. We show that x matters for q.

First, suppose that $q \in \bar{W}_{\text{inner}}(x)$. Let v^* be any member of $\mathcal{V}$ that maximizes $v^*(q) - v^*(x)$, which exists by compactness of $\mathcal{V}$. We have

$$U\left(q \mid q, x\right) = (1-\gamma)u(q) + \gamma\left(v^*(q) - v^*(x)\right).$$

Since u is in the relative interior of $\mathcal{V}$, it is without loss to assume that $v^* \neq u$. Since q is interior and $v^* \neq u$, by non-redundancy there exists $r \in \Delta(Z)$ such that $u(q) = u(r)$ and $v^*(r) > v^*(q)$. We can always choose r close enough to q that $v^*(x) > v^*(r)$. Then,

$$\begin{aligned} U\left(r \mid q, r, x\right) &\geq (1-\gamma)u(q) + \gamma\left(v^*(r) - v^*(x)\right) \\ &> (1-\gamma)u(q) + \gamma\left(v^*(q) - v^*(x)\right) \\ &= U\left(q \mid q, x\right) \\ &\geq U\left(q \mid q, r, x\right) \\ \implies \quad & q \notin c_2\left(q, r \mid q, r, x\right). \end{aligned}$$

Since $u(q) = u(r)$,

$$\begin{aligned} U\left(q \mid q, r\right) &= (1-\gamma)u(q) = U\left(r \mid q, r\right) \\ \implies \quad & q \in c_2(q, r|q, r). \end{aligned}$$

Conclude that x matters for q. (To fit the definition, set y equal to q or r.)

Now we consider the case $q \notin \bar{W}_{\text{inner}}(x)$. Since $x \notin \bar{W}_{\text{inner}}(q)$, there exists $v^* \in \mathcal{V}$ such that $v^*(x) > v^*(q)$. It is without loss to assume that v^* is extreme. Fix $\hat{y} \in W_{\text{inner}}(q)$, which is nonempty by Lemma A.1. Since q is interior, $q + \epsilon(q - \hat{y}) \in \Delta(Z)$ for $\epsilon > 0$ sufficiently small. For all such ϵ, let $\bar{y}_\epsilon := q + \epsilon(q - \hat{y})$. Since $\hat{y} \in W_{\text{inner}}(q)$, we have $q \in W_{\text{inner}}(\bar{y}_\epsilon)$ for all ϵ such that $\bar{y}_\epsilon$ is defined. For all such ϵ, let

$$\lambda_\epsilon := \frac{v^*(\bar{y}_\epsilon) - v^*(q)}{v^*(x) - v^*(\bar{y}_\epsilon)}.$$

Since $\bar{y}_\epsilon \to q$ and $v^*(x) > v^*(q)$, we have $\lim_{\epsilon \to 0} \lambda_\epsilon = 0$. Since q is interior, $\bar{y}_\epsilon + \lambda_\epsilon(\bar{y}_\epsilon - x) \in \Delta(Z)$ for $\epsilon > 0$ sufficiently small. Let $y_\epsilon := \bar{y}_\epsilon + \lambda_\epsilon(\bar{y}_\epsilon - x)$ for all such ϵ. By construction, $v^*(y_\epsilon) = v^*(q)$ for all ϵ such that y_ϵ is defined.

For every ϵ for which y_ϵ is defined, $\bar{y}_\epsilon$ is a convex combination of y_ϵ and x. Since every $v \in \mathcal{V}$ has $v(\bar{y}_\epsilon) > v(a)$ for all ϵ such that y_ϵ is defined, we have $\max_{y_\epsilon, x} v > v(q)$ for all $v \in \mathcal{V}$ and all such ϵ. This implies

$$U\left(q \mid q, y_\epsilon, x\right) < (1-\gamma)u(q).$$

for all $\epsilon > 0$ for which y_ϵ is defined.

Since $q \notin \bar{W}_{\text{inner}}(x)$, there exists $\bar{v} \in \mathcal{V}$ such that $\bar{v}(q) > \bar{v}(x)$. It is without loss to assume that $\bar{v}$ is extreme. We claim that there exists $r \in \Delta(Z)$ such that

$$\begin{aligned} u(r) &= u(q) \\ v^*(r) &\leq v^*(q) \\ \bar{v}(r) &> \bar{v}(q). \end{aligned}$$

Suppose not. Then, v^* and $\bar{v}$ must agree when restricted to $\{r \in \Delta(Z) : u(r) = u(q)\}$. But this implies that $v^* \in \text{co}(\{\bar{v}, u\})$ or that $\bar{v} \in \text{co}(\{v^*, u\})$. The first case contradicts extremeness of v^*, and the second contradicts extremeness of $\bar{v}$.

Since $\bar{v}(r) > \bar{v}(q) > \bar{v}(x)$, and since $\lim_{\epsilon \to 0} y_\epsilon = q$,

$$\bar{v}(r) = \max_{q, r, y_\epsilon, x} \bar{v}$$

for all ϵ sufficiently small. This implies

$$\begin{aligned} U\left(r \mid q, r, y_\epsilon, x\right) &= (1-\gamma)u(r) \\ U\left(r \mid q, r, y_\epsilon\right) &= (1-\gamma)u(r) \end{aligned}$$

for all ϵ sufficiently small. Since $U(q \mid Q) \leq U(q \mid R)$ whenever $q \in R \subset Q$,

$$\begin{aligned} U(q \mid q, r, y_\epsilon, x) &\leq U(q \mid q, y_\epsilon, x) \\ &< (1-\gamma)u(q) \\ &= (1-\gamma)u(r) \\ &= U(r \mid q, r, y_\epsilon, x) \\ \implies \quad & q \notin c_2(q, r \mid q, r, y_\epsilon, x) \end{aligned}$$

for all ϵ sufficiently small. Since $v^*(q) = v^*(y_\epsilon) \geq v^*(r)$ for all ϵ such that y_ϵ is defined,

$$\begin{aligned} U(q \mid q, r, y_\epsilon) &= (1-\gamma)u(q) \\ &= (1-\gamma)u(r) \\ &= U(r \mid q, r, y_\epsilon) \\ \implies \quad & q \in c_2(q, r \mid q, r, y_\epsilon) \end{aligned}$$

for all ϵ sufficiently small. Conclude that x matters for q. □

The next lemma shows that we can use $W_{\text{inner}}(q)$ to fully recover $\mathcal{V}_{\text{pref}}$.

Lemma A.4. *Fix any regular representation $(\gamma, u, \mathcal{V})$ and any $q \in int(\Delta(Z))$.*

1. *For any $G_{pref} \subset \mathcal{U}_{pref}$ such that $W_{inner}(q) = \bigcap_{\succsim \in G_{pref}} \{r \in \Delta(Z) : q \succ r\}$,*

$$\bar{co}(G_{pref}) = \{\succsim \in \mathcal{U}_{pref} : q \succ r \text{ for all } r \in W_{inner}(q)\}.$$

2. *$\mathcal{V}_{pref} = \{\succsim \in \mathcal{U}_{pref} : q \succ r$ for all $r \in W_{inner}(q)\}$.*

Proof. First part: We show that $\succsim^* \in \text{co}(G_{\text{pref}})$ if $q \succ^* r$ for all $r \in \bar{W}_{\text{inner}}(q) \setminus \{q\}$. By definition of G_{pref}, we have

$$\{r \in \Delta(Z) : r \succsim^* q\} \setminus \{q\} \subset \bigcup_{\succsim \in \mathcal{V}_{\text{pref}}} \{r \in \Delta(Z) : r \succ q\}.$$

Fix some $\epsilon > 0$ such that the ϵ-ball $B_\epsilon(q) \subset \text{int}(\Delta(Z))$. We have

$$\{r \in \Delta(Z) : r \succsim^* q\} \setminus B_\epsilon(q) \subset \bigcup_{\succsim \in \mathcal{V}_{\text{pref}}} \{r \in \Delta(Z) : r \succ q\}.$$

By the Heine-Borel theorem, there exists some finite $\hat{G}_{\text{pref}} \subset G_{\text{pref}}$ such that

$$\{r \in \Delta(Z) : r \succsim^* q\} \setminus B_\epsilon(q) \subset \bigcup_{\succsim \in \hat{G}_{\text{pref}}} \{r \in \Delta(Z) : r \succ q\}. \tag{9}$$

This implies

$$\{r \in \Delta(Z) : r \succsim^* q\} \setminus \{q\} \subset \bigcup_{\succsim \in \hat{G}_{\text{pref}}} \{r \in \Delta(Z) : r \succ q\}. \tag{10}$$

Suppose not. Then, there must be some $r \in B_\epsilon(q) \setminus \{q\}$ such that $r \succsim^* q$, but $r \precsim q$ for all $\succsim \in \hat{G}_{\text{pref}}$. For some $\lambda > 0$, we will have $r + \lambda(r - q) \in \Delta(Z) \setminus B_\epsilon(q)$. By linearity, $r + \lambda(r - q) \succsim^* q$, but $r + \lambda(r - q) \precsim q$ for all $\succsim \in \hat{G}_{\text{pref}}$. Since $r + \lambda(r - q) \notin B_\epsilon(q)$, this contradicts (9).

Assign $\succsim^*$ a representation v^* such that $v^*(q) = 0$. For each $\succsim \in \hat{G}_{\text{pref}}$, assign a representation v such that $v(q) = 0$, and let $\hat{G}$ denote the resulting set of utilities. By Farkas' lemma, we can find $\alpha > 0$ such that $\alpha v^* \in \text{co}(\hat{G})$ provided there does not exist $\rho \in \mathbb{R}^{|Z|}$ such that $\rho' v^* < 0$ and $\rho' v \geq 0$ for all $v \in S$. Suppose there exists some such ρ. Consider the case $\sum_{k:\rho_k > 0} \rho_k \geq \sum_{k:\rho_k < 0}(-\rho_k)$. Normalize ρ by dividing each ρ_k by $\sum_{k:\rho_k > 0} \rho_k$. (Since $\rho \neq 0$, this term must be strictly positive.) We have

$$\begin{aligned} \sum_{k:\rho_k > 0} \rho_k v^*(q_k) &< \sum_{k:\rho_k < 0} (-\rho_k) v^*(q_k) \\ \sum_{k:\rho_k > 0} \rho_k v(q_k) &\geq \sum_{k:\rho_k < 0} (-\rho_k) v(q_k) \text{ for all } v \in S. \end{aligned}$$

Since the normalization ensures that $\sum_{k:\rho_k > 0} \rho_k = 1$, the left-hand side is the valuation of a lottery, which we label ρ^+. Since $v^*(q) = v(q) = 0$,

$$\begin{aligned} v^*(\rho^+) &< \sum_{k:\rho_k < 0} (-\rho_k) v^*(q_k) + \left(1 - \sum_{k:\rho_k < 0} (-\rho_k)\right) v^*(q) \\ v(\rho^+) &\geq \sum_{k:\rho_k < 0} (-\rho_k) v(q_k) + \left(1 - \sum_{k:\rho_k < 0} (-\rho_k)\right) v(q) \text{ for all } v \in \hat{G}. \end{aligned}$$

Now the right-hand side is also the valuation of a lottery, which we label ρ^-. We have

$$\begin{aligned} v^*(\rho^+) &< v^*(\rho^-) \\ v(\rho^+) &\geq v(\rho^-) \text{ for all } v \in \hat{G}. \end{aligned}$$

For $\epsilon > 0$ sufficiently small, we have

$$\begin{aligned} v^*(q) &< v^*(q + \epsilon(\rho^- - \rho^+)) \\ v(q) &\geq v(q + \epsilon(\rho^- - \rho^+)) \text{ for all } v \in \hat{G}. \end{aligned}$$

This contradicts (10). An exactly similar argument covers the case $\sum_{k:\rho_k > 0} \rho_k \geq$

$\sum_{k:\rho_k<0}(-\rho_k)$. Conclude that $\alpha v^* \in \text{co}(\hat{G})$ for some $\alpha > 0$, so $\succsim^* \in \text{co}(\hat{G}_{\text{pref}})$. Since $\hat{G}_{\text{pref}} \subset G_{\text{pref}}$, $\succsim^* \in \text{co}(G_{\text{pref}})$.

Now take any $\bar{\succsim} \in \mathcal{U}_{\text{pref}}$ such that $q \mathrel{\bar{\succ}} r$ for all $r \in W_{\text{inner}}(q)$. Suppose that there is some $\succsim^* \in \mathcal{U}_{\text{pref}}$ such that $q \succ^* r$ for all $r \in \bar{W}_{\text{inner}}(q) \setminus \{q\}$. We already showed that $\succsim^* \in \text{co}(G_{\text{pref}})$. Assign $\bar{\succsim}$ a representation $\bar{v}$, and assign $\succsim^*$ a representation v^*. Fix any sequence $\epsilon_i \to 0$ such that $\epsilon_i \in (0,1)$ for each i. Each $\epsilon_i v^* + (1-\epsilon_i)\bar{v}$ represents some $\succsim_i \in \mathcal{U}_{\text{pref}}$. For each i, $q \succ_i r$ for all $r \in \bar{W}_{\text{inner}}(q) \setminus \{q\}$. Thus, $\succsim_i \in \mathcal{V}_{\text{pref}}$ for all i. Since the utilities that represent $\succsim_i$ converge to $\bar{v}$, which represents $\bar{\succsim}$, we have $\bar{\succsim} \in \bar{\text{co}}(G_{\text{pref}})$.

Finally, suppose there is no $\succsim^* \in \mathcal{U}_{\text{pref}}$ such that $q \succ^* r$ for all $r \in \bar{W}_{\text{inner}}(q) \setminus \{q\}$. For any $J \subset Z$, we will use the superscript J to denote the restriction of a set or preference to $\Delta(J)$. We will find $J^* \subset Z$ such that $|J^*| \geq 3$ for which there exists an EU preference $\succsim^*$ on $\Delta(J^*)$ with the following property: for any $q^* \in \Delta(J^*)$,

$$r^* \in \bar{W}^{J^*}_{\text{inner}}(q^*) \setminus \{q^*\} \quad \Longrightarrow \quad q^* \succ^* r^*. \tag{11}$$

Since there is no preference with this property when $J^* = Z$, there must be $q^0, r^0 \in \Delta(Z)$ such that

$$\begin{aligned} r^0 &\in \bar{W}_{\text{inner}}(q^0) \setminus \{q^0\} \\ q^0 + (q^0 - r^0) &\in \bar{W}_{\text{inner}}(q^0) \setminus \{q^0\}. \end{aligned}$$

This implies $r^0 \sim q^0$ for all $\succsim \in G_{\text{pref}}$. Since $r^0 \neq q^0$, we can always relabel the elements of Z so that $r^0(z_1) \neq q^0(z_1)$. Let $J^1 \equiv Z \setminus \{z_1\}$, and check whether there is any preference satisfying (11) with J^1 in place of J^*. If not, then we can repeat the argument, deriving a new indifference condition $q^1 \sim r^1$ and setting $J^2 \equiv Z \setminus \{z_1, z_2\}$. We repeat the process until we arrive at the desired J^*. To see why some such J^* must exist, suppose that we have iterated the process to $J^{n-3} = \{z_{n-2}, z_{n-1}, z_n\}$. Suppose there is no preference satisfying (11) with J^{n-3} in place of J^*. This can only happen if $\bar{W}^{J_{n-3}}_{\text{inner}}(q^{n-3})$ is a half-plane, which in turn only happens if $|\mathcal{V}^{J_{n-3}}_{\text{pref}}| = 1$. Since each preference in $\mathcal{V}_{\text{pref}}$ is pinned down by its restriction to $\mathcal{V}^{J_{n-3}}_{\text{pref}}$ (and the indifference conditions), we have $|\mathcal{V}_{\text{pref}}| = 1$. This contradicts non-redundancy and $|\mathcal{V}| > 1$. Conclude that J^* exists.

Fix any $q^* \in \text{int}(\Delta(J^*))$. The first part of the proof implies, for any EU preference $\succsim^{J^*}$ on $\Delta(J^*)$,

$$\succsim^{J^*} \in G^{J^*}_{\text{pref}} \quad \Longleftrightarrow \quad q^* \succ^{J^*} r^* \text{ for all } r^* \in W^{J^*}_{\text{inner}}(q^*). \tag{12}$$

Take any $\succsim \in \mathcal{U}_{\text{pref}}$ such that, for all $q \in Z$, $q \succ r$ for all $r \in W_{\text{inner}}(q)$. By (12), the restriction of $\succsim$ to J^* belongs to $G_{\text{pref}}^{J^*}$. Thus, there is some preference in G_{pref} that agrees with $\succsim$ in its restriction to J^*. This preference is pinned down by the indifference conditions derived at the last step: it must be indifferent between q^0 and r^0, q^1 and r^1, and so on. Notice that $\succsim$ must satisfy all these indifference conditions. (For instance, $q^0 \sim r^0$ because $r^0 \in \bar{W}_{\text{inner}}(q^0)$ implies $q^0 \succsim r^0$, while $q^0 + (q^0 - r^0) \in \bar{W}_{\text{inner}}(q^0)$ implies $r^0 \succsim q^0$. An exactly similar argument covers any remaining indifference conditions.) Conclude that $\succsim \in G_{\text{pref}}$.

Second part: By definition,

$$W_{\text{inner}}(q) = \bigcap_{\succsim \in \mathcal{V}_{\text{pref}}} \{r \in \Delta(Z) : q \succ r\}.$$

By the first part, $\succsim \in \bar{\text{co}}(\mathcal{V}_{\text{pref}})$ if and only if $q \succ r$ for all $r \in W_{\text{inner}}(q)$. Since $\mathcal{V}_{\text{pref}}$ is closed and convex, $\succsim \in \mathcal{V}_{\text{pref}}$ if and only if $q \succ r$ for all $r \in W_{\text{inner}}(q)$. $\square$

We now recover $\succsim_u$, the preference relation represented by the material utility function. For any $M \in \mathcal{K}_f(\Delta(Z))$, we have $c_2(M \mid M) = \text{argmax}(M, \succsim_u)$. For any $q, r \in \Delta(Z)$, we have

$$q \succsim_u r \quad \Longleftrightarrow \quad q \in c_2(q, r \mid q, r).$$

This allows us to recover $\succsim_u$. As usual, the representation of $\succsim_u$ is pinned down up to a positive affine transformation.

We now show that, for any two representations with the same u and γ, for any $\succsim \in \mathcal{V}_{\text{pref}}$, the utilities that represent $\succsim$ differ (at most) by an additive constant. For this step, we need to know $W_{\text{outer}}(q)$. The following lemma shows that $W_{\text{outer}}(q)$ can be fully recovered from choice data.

Definition A.5. *For $q, x \in \Delta(Z)$, lottery x is **never chosen over** lottery q if there is no $y \in \Delta(Z)$ such that*

$$\{x\} = c_2\left(q, x \mid q, x, y\right).$$

Lemma A.6. *For any regular representation $(\gamma, u, \mathcal{V})$ and any $q, x \in int(\Delta(Z))$: $x \in \bar{W}_{outer}(q)$ if and only if, for all $\epsilon \in (0, 1]$, $\epsilon x + (1 - \epsilon)q$ is never chosen over q.*

Proof. First, we show that if $x \in \bar{W}_{\text{outer}}(q)$, then x_ϵ is never chosen over q for any $\epsilon \in (0, 1]$. If $x \in \bar{W}_{\text{outer}}(q)$, then $x_\epsilon \in \bar{W}_{\text{outer}}(q)$ as well. Thus, it suffices to show that if $x \in \bar{W}_{\text{outer}}(q)$, then x is never chosen over q. Let v' denote an element of $\text{argmax}_{v \in \mathcal{V}} \{v(x) - \max_{q,x,y} v\}$, which is nonempty by continuity and compactness. We have

$$U\left(x \mid q,x,y\right) = (1-\gamma)u(x) + \gamma\left(v'(x) - \max_{q,x,y} v'\right)$$
$$\leq (1-\gamma)u(q) + \gamma\left(v'(q) - \max_{q,x,y} v'\right)$$
$$\leq U(q \mid q,x,y),$$

which is the desired result.

Now we show: if $x \notin \bar{W}_{\text{outer}}(q)$, then there is some $\epsilon \in (0,1]$ for which x_ϵ is sometimes chosen over q. This is clearly the case if $u(x) > u(q)$. Suppose $u(x) = u(q)$. By assumption, there is some $v \in \mathcal{V}$ such that $((1-\gamma)u + \gamma v)(x) > ((1-\gamma)u + \gamma v)(q)$, so $v(x) > v(q)$. Choose any $\hat{y}$ such that $q \in W_{\text{inner}}(\hat{y})$. Since $v(x) > v(q)$, we have

$$v(x) > v(\epsilon\hat{y} + (1-\epsilon)q)$$

for ϵ close enough to 0. Let $y \equiv \epsilon\hat{y} + (1-\epsilon)q$ for some such ϵ. Since $v(x) > \max_{y,q} v$,

$$U(x \mid q,x,y) = (1-\gamma)u(x) = (1-\gamma)u(q).$$

Since $q \in W_{\text{inner}}(\hat{y})$, we have $q \in W_{\text{inner}}(y)$, so

$$U(q \mid q,x,y) < (1-\gamma)u(q) = U(x \mid q,x,y).$$

Conclude that x is sometimes chosen over q.

Now we cover the final case: $u(x) < u(q)$. Let

$$G \equiv \{v \in \mathcal{V} : ((1-\gamma)u + \gamma v)(q) \geq ((1-\gamma)u + \gamma v)(x)\}.$$

We want to find $\hat{y}$ on the boundary of $W_{\text{inner}}(q)$ such that $v(q) > v(\hat{y})$ for all $v \in G$. Suppose there is no such $\hat{y}$. Then,

$$W_{\text{inner}}(q) = \bigcup_{\succsim \in G_{\text{pref}}} \{r \in \Delta(Z) : q \succ r\}.$$

By Lemma A.4, $G_{\text{pref}} = \bar{\text{co}}(G_{\text{pref}}) = \mathcal{V}_{\text{pref}}$. That is, every $v \in \mathcal{V}$ represents the same preference as some $g \in G \subseteq \mathcal{V}$. By non-redundancy, we must have $v = g$, so $\mathcal{V} = G$. But since $((1-\gamma)u + \gamma v)(x) > ((1-\gamma)u + \gamma v)(q)$ for some $v \in \mathcal{V}$, this cannot be the case. Conclude that there exists $\hat{y}$ on the boundary of $W_{\text{inner}}(q)$ such that $v(q) > v(\hat{y})$ for all $v \in G$. Since $\hat{y} \in \bar{W}_{\text{inner}}(q)$, we must have $v(q) = v(\hat{y})$ for some $v \in \mathcal{V} \setminus G$. For some $\hat{\delta} > 0$, we have $q + \hat{\delta}(q - \hat{y}) \in \Delta(Z)$. Set $y \equiv q + \hat{\delta}(q - \hat{y})$ for some such $\hat{\delta}$. Notice that

$$
\begin{aligned}
&v(y) > v(q) \text{ for all } v \in G \\
&v(q) = v(y) \text{ for some } v \in \mathcal{V} \setminus G.
\end{aligned}
$$

Since no $v \in \mathcal{V}$ prefers q to both x and y, we have

$$
U(q \mid q, x_\epsilon, y) = (1-\gamma)u(q) + \gamma \left(v_\epsilon(q) - \max_{x_\epsilon, y} v_\epsilon \right)
$$

for all $\epsilon > 0$, where $v_\epsilon \in \operatorname{argmax}_{v \in \mathcal{V}} \{v(q) - \max_{x_\epsilon, y} v\}$. Take a sequence $\{\epsilon_k\} \to 0$. Since $\mathcal{V}$ is compact, we can pass to a convergent subsequence of v_{ϵ_k}. Denote the limit v_0. Fix some $v^* \in \mathcal{V}$ such that $v^*(q) = v^*(y)$. We have

$$
\begin{aligned}
v_{\epsilon_k}(q) - v_{\epsilon_k}(y) &\geq v_{\epsilon_k}(q) - \max_{x_{\epsilon_k}, y} v_{\epsilon_k} \\
&= \max_{v \in \mathcal{V}} \left\{ v(q) - \max_{x_{\epsilon_k}, y} v \right\} \\
&\geq v^*(q) - \max_{x_{\epsilon_k}, y} v^* \\
&= \min \{ v^*(q) - v^*(x_{\epsilon_k}), v^*(q) - v^*(y) \} \\
&= \min \{ \epsilon_k (v^*(q) - v^*(x)), 0 \}.
\end{aligned}
$$

Moreover, the limit of the left-hand side is

$$
\lim_{k \to \infty} \left(v_{\epsilon_k}(q) - v_{\epsilon_k}(y) \right) = v_0(q) - v_0(y)
$$

and the limit of the right-hand side is

$$
\lim_{k \to \infty} \min \{ \epsilon_k (v^*(q) - v^*(x)), 0 \} = 0
$$

so $v_0(q) \geq v_0(y)$, which implies $v_0 \notin G$, and hence $((1-\gamma)u + \gamma v_0)(x) > ((1-\gamma)u + \gamma v_0)(q)$. For k sufficiently large, the same will be true with v_{ϵ_k} in place of v_0. Using this fact,

$$
\begin{aligned}
U(q \mid q, x_{\epsilon_k}, y) &= (1-\gamma)u(q) + \gamma \left(v_{\epsilon_k}(q) - \max_{x_{\epsilon_k}, y} v_{\epsilon_k} \right) \\
&\leq (1-\gamma)u(q) + \gamma \left(v_{\epsilon_k}(q) - v_{\epsilon_k}(x_{\epsilon_k}) \right) \\
&= (1-\gamma)(1-\epsilon_k)u(q) + \epsilon_k \left[(1-\gamma)u(q) + \gamma \left(v_{\epsilon_k}(q) - v_{\epsilon_k}(x) \right) \right] \\
&< (1-\gamma) \left[(1-\epsilon_k)u(q) + \epsilon_k u(x) \right] \\
&= (1-\gamma)u(x_{\epsilon_k})
\end{aligned}
$$

for k sufficiently large. Since $v_0(x) > v_0(q) = v_0(y)$, we have

$$U(x_{\epsilon_k} \mid q, x_{\epsilon_k}, y) = (1-\gamma)u(x_{\epsilon_k}).$$

Conclude that, for some $\epsilon \in (0,1]$, x_ϵ is sometimes chosen over q. $\square$

Fix some $q \in \text{int}(\Delta(Z))$, and let

$$H_{\text{pref}} \equiv \{\succsim \in \mathcal{V}_{\text{pref}} : q \sim r \text{ for some } r \in \bar{W}_{\text{inner}}(q) \text{ s.t. } u(q) > u(r)\}.$$

We will show that, for any two representations with the same u, for any $\succsim \in H_{\text{pref}}$, the utilities that represent $\succsim$ differ (at most) by an additive constant.

Lemma A.7. *Suppose $(\gamma, u, \mathcal{V})$ is regular. Fix some $\succsim \in H_{pref}$ and some $q \in int(\Delta(Z))$. A representation v of $\succsim$ belongs to $\mathcal{V}$ only if*

$$((1-\gamma)u + \gamma v)(q) > ((1-\gamma)u + \gamma v)(r) \text{ for all } r \in W_{outer}(q) \tag{13}$$

$$((1-\gamma)u + \gamma v)(q) = ((1-\gamma)u + \gamma v)(r) \text{ for some } r \in \bar{W}_{outer}(q) \text{ s.t. } u(q) > u(r). \tag{14}$$

Proof. Each $v \in \mathcal{V}$ must satisfy (13) by definition of $W_{\text{outer}}(q)$. For (14), suppose that $\hat{v} \in \mathcal{V}$ represents $\succsim \in H_{\text{pref}}$ but $((1-\gamma)u + \gamma\hat{v})(q) > ((1-\gamma)u + \gamma\hat{v})(r)$ for all $r \in \bar{W}_{\text{outer}}(q)$ such that $u(q) > u(r)$. Then, there exists[28] some $\alpha > 1$ such that

$$((1-\gamma)u + \alpha\gamma\hat{v})(q) > ((1-\gamma)u + \alpha\gamma\hat{v})(r) \text{ for all } r \in W_{\text{outer}}(q). \tag{15}$$

We can use exactly the same arguments in the proof of Lemma A.4 to show the following: any preference on $\Delta(Z)$ that strictly prefers q to everything in $W_{\text{outer}}(q)$ must have a representation in $\{(1-\gamma)u + \gamma v : v \in \mathcal{V}\}$. By (15), the preference represented by $(1-\gamma)u + \alpha\gamma\hat{v}$ satisfies this condition. Thus, there exist $\beta_1 > 0$ and $\beta_2 \in \mathbb{R}$ such that

$$\beta_1((1-\gamma)u + \alpha\gamma\hat{v}) + \beta_2 \in \{(1-\gamma)u + \gamma v : v \in \mathcal{V}\}.$$

This implies

$$(\beta_1 - 1)\left(\frac{1-\gamma}{\gamma}\right)u + \alpha\beta_1\hat{v} + \frac{\beta_2}{\gamma} \in \mathcal{V}, \tag{16}$$

[28]This may not be obvious. For a proof, see Appendix A.1.

which is conveniently rewritten

$$\left(\beta_1\left(\frac{1}{\gamma}-1+\alpha\right)-\left(\frac{1}{\gamma}-1\right)\right)\times \\ \left(\frac{(\beta_1-1)\left(\frac{1}{\gamma}-1\right)}{\beta_1(\frac{1}{\gamma}-1+\alpha)-\left(\frac{1}{\gamma}-1\right)}u+\frac{\alpha\beta_1}{\beta_1(\frac{1}{\gamma}-1+\alpha)-\left(\frac{1}{\gamma}-1\right)}\hat{v}\right)+\frac{\beta_2}{\gamma}\in\mathcal{V}. \quad (17)$$

Suppose $\beta_1\geq 1$. Since $\mathcal{V}$ is convex and both u and $\hat{v}$ belong to it,

$$\frac{(\beta_1-1)\left(\frac{1}{\gamma}-1\right)}{\beta_1(\frac{1}{\gamma}-1+\alpha)-\left(\frac{1}{\gamma}-1\right)}u+\frac{\alpha\beta_1}{\beta_1(\frac{1}{\gamma}-1+\alpha)-\left(\frac{1}{\gamma}-1\right)}\hat{v}\in\mathcal{V}. \quad (18)$$

Together with non-redundancy, (17) and (18) imply that

$$\beta_1\left(\frac{1}{\gamma}-1+\alpha\right)-\left(\frac{1}{\gamma}-1\right)=1.$$

Rearranging gives

$$\frac{1}{\gamma}(\beta_1-1)+\beta_1(\alpha-1)=0.$$

But since $\alpha>1$ and $\beta_1\geq 1$, this cannot hold.

Now suppose $\beta_1<1$. By (16), we know there exists $\bar{v}\in\mathcal{V}$ such that

$$\bar{v}=(\beta_1-1)\left(\frac{1}{\gamma}-1\right)u+\alpha\beta_1\hat{v}+\beta_2$$
$$\implies\quad \frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v}+\frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u=\frac{\alpha\beta_1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\hat{v}+\frac{\beta_2}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}.$$

Since u and $\bar{v}$ both belong to $\mathcal{V}$, and since $\mathcal{V}$ is convex,

$$\frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v}+\frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u\in\mathcal{V}.$$

Since this utility represents the same preference on $\Delta(Z)$ that $\hat{v}$ does, and since $\hat{v}\in\mathcal{V}$, non-redundancy implies

$$\frac{1}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}\bar{v}+\frac{(1-\beta_1)\left(\frac{1}{\gamma}-1\right)}{\frac{1}{\gamma}(1-\beta_1)+\beta_1}u=\hat{v}. \quad (19)$$

Since $\hat{v}$ represents a preference in H_{pref}, by definition of H_{pref} there exists $r\in\bar{W}_{\text{inner}}(q)$

such that $u(q) > u(r)$ but $\hat{v}(q) = \hat{v}(r)$. Plugging this into (19) and rearranging gives

$$\bar{v}(q) - \bar{v}(r) = (\beta_1 - 1)\left(\frac{1}{\gamma} - 1\right)(u(q) - u(r)) < 0.$$

We have $\bar{v}(q) < \bar{v}(r)$ for some $r \in \bar{W}_{\text{inner}}(q)$, which contradicts $\bar{v} \in \mathcal{V}$. Conclude that (14) cannot be violated. $\square$

Fix some $\succsim \in H_{\text{pref}}$ and an arbitrary representation $\hat{v}$ of $\succsim$. Suppose that $\alpha\hat{v} + \beta$ and $\alpha'\hat{v} + \beta'$ represent $\succsim$ in $(\gamma, u, \mathcal{V})$ and $(\gamma, u, \mathcal{V}')$ respectively. By Lemma A.7,

$$((1-\gamma)u + \alpha\gamma\hat{v})(q) = ((1-\gamma)u + \alpha\gamma\hat{v})(r) \text{ for some } r \in \bar{W}_{\text{outer}}(q) \text{ s.t. } u(q) > u(r)$$
$$((1-\gamma)u + \alpha'\gamma\hat{v})(q) = ((1-\gamma)u + \alpha'\gamma\hat{v})(r') \text{ for some } r' \in \bar{W}_{\text{outer}}(q) \text{ s.t. } u(q) > u(r').$$

If $\alpha' > \alpha$, we have

$$\begin{aligned} & (1-\gamma)(u(q) - u(r)) = \alpha\gamma(\hat{v}(r) - \hat{v}(q)) < \alpha'\gamma(\hat{v}(r) - \hat{v}(q)) \\ \implies & ((1-\gamma)u + \alpha'\gamma\hat{v})(q) < ((1-\gamma)u + \alpha'\gamma\hat{v})(r). \end{aligned}$$

Since $r \in \bar{W}_{\text{outer}}(q)$, this contradicts (13). The case $\alpha > \alpha'$ is exactly the same, so we have $\alpha = \alpha'$. Conclude that, conditional on u and γ, two representations of $\succsim$ differ at most by an additive constant.

Now we extend the argument from H_{pref} to the remainder of $\mathcal{V}_{\text{pref}}$. Let

$$H \equiv \{v \in \mathcal{V} : v \text{ represents } \succsim \in H_{\text{pref}}\}.$$

Lemma A.8. *For any regular representation $(\gamma, u, \mathcal{V})$,*

1. $\bar{co}(H_{pref}) = \mathcal{V}_{pref}$.
2. $\bar{co}(H) = \mathcal{V}$.

Proof. We first show that $\bar{\text{co}}(H_{\text{pref}} \cup \{\succsim_u\}) = \mathcal{V}_{\text{pref}}$. By Lemma A.4, it suffices to show

$$W_{\text{inner}}(q) = \bigcap_{\succsim \in H_{\text{pref}} \cup \{\succsim_u\}} \{r \in \Delta(Z) : q \succ r\}. \tag{20}$$

Since $H_{\text{pref}} \cup \{\succsim_u\} \subseteq \mathcal{V}_{\text{pref}}$, the left-hand side of (20) is a subset of the right-hand side. Suppose the set inclusion is strict. Since the right-hand side is convex, it includes a boundary point y of $W_{\text{inner}}(q)$. Notice that $u(q) > u(y)$. Since $W_{\text{inner}}(q)$ is an open convex cone, there exists some utility $\hat{v}$ such that $\hat{v}(q) = \hat{v}(y) > \hat{v}(r)$ for all $r \in W_{\text{inner}}(q)$. By Lemma A.4, $\hat{v}$ represents a preference that belongs to $\mathcal{V}_{\text{pref}}$. Call this preference $\hat{\succsim}$. Since $q \hat{\sim} y$ for some $y \in \bar{W}_{\text{inner}}(q)$ such that $u(q) > u(y)$, $\hat{\succsim} \in H_{\text{pref}}$. But this contradicts the assumption that y belongs to the right-hand side of (20).

Next, we show that $\bar{\text{co}}(H \cup \{u\}) = \mathcal{V}$. Since $\mathcal{V}$ is closed and convex and $H \cup \{u\} \subset \mathcal{V}$, we have $\bar{\text{co}}(H \cup \{u\}) \subseteq \mathcal{V}$. Now take any $v \in \mathcal{V}$. By the first part, v represents a preference in $\bar{\text{co}}(H_{\text{pref}} \cup \{\succsim_u\})$. Every such preference has a representation in $\bar{\text{co}}(H \cup \{u\})$. By non-redundancy, that representation can only be v. Thus, $\mathcal{V} \subseteq \bar{\text{co}}(H \cup \{u\})$.

Fix any $v^* \in \text{ext}(\mathcal{V})$. Since $\mathcal{V} = \bar{\text{co}}(H \cup \{u\})$, v^* can be written

$$v^* = \lim_{i \to \infty} v^i,$$

where each $v^i \in \text{co}(H \cup \{u\})$. In turn, each v^i can be written

$$v^i = \left(1 - \sum_{j=1}^{J^i} \lambda_j^i\right) u + \sum_{j=1}^{J^i} \lambda_j^i h_j^i$$

where the λ_j^i are positive and sum to no more than unity, and each $h_j^i \in H$. Pass to a subsequence such that $\lim_i \sum_{j=1}^{J^i} \lambda_j^i$ converges, and let $\bar{\lambda}$ denote the limit. If $\bar{\lambda} = 0$, then $v^* = u$. This contradicts $u \in \text{relint}(\mathcal{V})$, so $\bar{\lambda} > 0$. Substituting for each v^i gives

$$\begin{aligned}
v^* &= (1 - \bar{\lambda})u + \lim_{i \to \infty} \sum_{j=1}^{J^i} \lambda_j^i h_j^i \\
&= (1 - \bar{\lambda})u + \lim_{i \to \infty} \left(\sum_{j=1}^{J^i} \lambda_j^i \times \sum_{j=1}^{J^i} \left(\frac{\lambda_j^i}{\sum_{k=1}^{J^i} \lambda_k^i} \right) h_j^i \right) \\
&= (1 - \bar{\lambda})u + \bar{\lambda} \lim_{i \to \infty} \sum_{j=1}^{J^i} \left(\frac{\lambda_j^i}{\sum_{k=1}^{J^i} \lambda_k^i} \right) h_j^i \\
&= (1 - \bar{\lambda})u + \bar{\lambda}\bar{h} \text{ for some } \bar{h} \in \bar{\text{co}}(H).
\end{aligned}$$

Since $u, \bar{g} \in \mathcal{V}$, both can be written as convex combinations of extremes of $\mathcal{V}$. Since u is non-extreme, it must place weight on at least two distinct extremes. If $\bar{\lambda} < 1$, then the same is true of v^*. Since v^* is extreme, it must be that $\bar{\lambda} = 1$, so $v^* \in \bar{\text{co}}(H)$. Since v^* was an arbitrary member of $\text{ext}(\mathcal{V})$, we have $\mathcal{V} = \bar{\text{co}}(H)$. This implies $\mathcal{V}_{\text{pref}} = \bar{\text{co}}(H_{\text{pref}})$. $\square$

Fix any $\succsim \in \mathcal{V}_{\text{pref}}$. Suppose that v and v' represent $\succsim$ in $(\gamma, u, \mathcal{V})$ and $(\gamma, u, \mathcal{V}')$ respectively. Since $\mathcal{V} = \bar{\text{co}}(H)$, each $v \in \mathcal{V}$ can be written

$$v = \lim_{i \to \infty} v^i$$

where each $v^i \in \text{co}(H)$. In turn, each v^i can be written

$$v^i = \sum_{j=1}^{J^i} \lambda_j^i h_j^i$$

where λ^i is a vector of weights and h^i is a vector of utilities in H. Substituting for each v_i, we have

$$v = \lim_{i\to\infty} \sum_{j=1}^{J^i} \lambda_j^i h_j^i$$

where each $h_j^i \in H$. We already showed that, for each $h_j^i \in H$, there exists a constant β_j^i such that $h_j^i + \beta_j^i \in H'$. Since $\mathcal{V}'$ is convex, it must contain each $\sum_{j=1}^{J^i} \lambda_j^i (h_j^i + \beta_j^i) \in \mathcal{V}'$. Since $\mathcal{V}'$ is compact, we can pass to a convergent subsequence of $\sum_{j=1}^{J^i} \lambda_j^i (h_j^i + \beta_j^i)$. Let v^∞ denote the limit. Since $\mathcal{V}'$ is closed, we have $v^\infty \in \mathcal{V}'$. Moreover,

$$v^\infty - v = \lim_{i\to\infty} \sum_{j=1}^{J_i} \lambda_j^i \beta_j^i \in \mathbb{R}$$

so v^∞ represents $\succ$. Since v^∞ and v' both represent $\succ$ and belong to $\mathcal{V}'$, non-redundancy implies that they are identical. Substituting v' for v^∞ gives

$$v' - v = \lim_{i\to\infty} \sum_{j=1}^{J_i} \lambda_j^i \beta_j^i \in \mathbb{R},$$

so, conditional on u and γ, v' and v differ by an additive constant at most.

Finally, we show that γ is identified. Suppose that $(\gamma, u, \mathcal{V})$ and $(\gamma', u', \mathcal{V}')$ are both regular representations. Since we can always construct a new regular representation by dividing u' and $\mathcal{V}'$ by a positive constant and adding another constant to u', it is without loss to assume that $u = u'$. Since $\mathcal{V}_{\text{pref}}$ and W_{inner} do not differ across representations, we must also have $H_{\text{pref}} = H'_{\text{pref}}$. Fix some $\succsim \in H_{\text{pref}}$. Suppose that h and $h' \equiv \alpha h + \beta$ represent $\succsim$ in H and H' respectively. By Lemma A.7, we must have

$$((1-\gamma')u + \gamma' h')(q) = ((1-\gamma')u + \gamma' h')(r') \text{ for some } r' \in \bar{W}_{\text{outer}}(q) \text{ s.t. } u(q) > u(r').$$

This implies

$$((1-\gamma)u + \gamma k h)(q) = ((1-\gamma)u + \gamma k h)(r') \tag{21}$$

where

$$k \equiv \alpha \frac{\gamma'}{\gamma} \frac{1-\gamma}{1-\gamma'}.$$

Again by Lemma A.7, we must have

$$((1-\gamma)u+\gamma h)(q)=((1-\gamma)u+\gamma h)(r) \text{ for some } r\in \bar{W}_{\text{outer}}(q) \text{ s.t. } u(q)>u(r). \quad (22)$$

We showed below Lemma A.7 that (21) and (22) together imply $k=1$. Solving for α gives

$$h'=\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}h+\beta.$$

Since $\mathcal{V}=\bar{\text{co}}(H)$ and $u\in\mathcal{V}$, we have

$$u=\sum_i \lambda_i h_i$$

for some weight vector λ and some vector h of utilities in H. By the previous argument, $h_i\in H$ implies

$$\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}h_i+\beta_i\in H'$$

for some constant β_i. Since $\mathcal{V}'$ is convex,

$$\sum_i \lambda_i\left(\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}h_i+\beta_i\right)=\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}u+\sum_i\lambda_i\beta_i\in H'.$$

Since $u\in\mathcal{V}'$, non-redundancy implies

$$\frac{1-\gamma'}{1-\gamma}\frac{\gamma}{\gamma'}=1,$$

which in turn implies $\gamma=\gamma'$.

A.1 Proof of claim in Lemma A.7

Lemma A.9. *Fix a regular representation $(\gamma,u,\mathcal{V})$ and $q\in int(\Delta(Z))$. Suppose that $\hat{v}\in\mathcal{V}$ represents $\succsim\in H_{pref}$, but $((1-\gamma)u+\gamma\hat{v})(q)>((1-\gamma)u+\gamma\hat{v})(r)$ for all $r\in\bar{W}_{outer}(q)$ such that $u(q)>u(r)$. Then, there exists some $\alpha>1$ such that (15) holds.*

Proof. The result is clearly true if $u=\hat{v}$, so suppose $u\neq\hat{v}$. Let

$$X\equiv\{r\in\Delta(Z): u(q)=u(r) \text{ and } \hat{v}(q)=\hat{v}(r)\}.$$

Let

$$Y\equiv \text{co}(X\cup\bar{W}_{\text{outer}}(q)).$$

Using the fact that X and $\bar{W}_{\text{outer}}(q)$ are compact and convex, it is straightforward to show that Y is closed.

Since $u \neq \hat{v}$ and $q \in \text{int}(\Delta(Z))$, by non-redundancy there must exist $r^* \in \text{int}(\Delta(Z))$ such that

$$u(q) > u(r^*)$$
$$((1-\gamma)u + \gamma\hat{v})(q) = ((1-\gamma)u + \gamma\hat{v})(r^*).$$

Suppose $r^* \in Y$. Then, by definition of X and $\bar{W}_{\text{outer}}(q)$,

$$r^* \in \text{co}(\{r \in Y : ((1-\gamma)u + \gamma\hat{v})(q) = ((1-\gamma)u + \gamma\hat{v})(r)\}).$$

Moreover, r^* must place positive weight on some $\bar{r} \in \bar{W}_{\text{outer}}(q)$ with $u(q) > u(\bar{r})$ as well as $((1-\gamma)u+\gamma\hat{v})(q) = ((1-\gamma)u+\gamma\hat{v})(\bar{r})$. This contradicts the assumption about $\hat{v}$. Conclude that $r^* \notin Y$.

We extend Y beyond the simplex as follows:

$$Y^{\text{ext}} \equiv \left\{ y \in \mathbb{R}^Z : \sum_Z y(z) = 1 \text{ and } \epsilon y + (1-\epsilon)a \in Y \text{ for some } \epsilon > 0 \right\}.$$

It is straightforward to show that Y^{ext} inherits closedness and convexity from Y. Moreover, since $r^* \in \text{int}(\Delta(Z))$ and Y and Y^{ext} agree on $\Delta(Z)$, $r^* \notin Y^{\text{ext}}$. By the separating hyperplane theorem, there exists some utility b and some constant $\bar{b}$ such that

$$b(r^*) > \bar{b} \geq b(y) \text{ for all } y \in Y^{\text{ext}}. \tag{23}$$

Suppose that $b(x) > b(q)$ for some $x \in X$. We have $x + \lambda(x-q) \in Y^{\text{ext}}$ for all $\lambda > 0$. But for λ sufficiently large, $b(x+\lambda(x-q)) > \bar{b}$, which contradicts (23). Now suppose $b(x) < b(q)$ for some $x \in X$. By definition of x, we must have $q + \epsilon(q-x) \in X$ for $\epsilon > 0$ sufficiently small. Thus, $q + \lambda(q-x) \in Y^{\text{ext}}$ for all $\lambda > 0$. But for λ sufficiently large, $b(q+\lambda(q-x)) > \bar{b}$, which contradicts (23). Conclude that $b(q) = b(x)$ for all $x \in X$. This implies that b can be written as a linear combination of $(1-\gamma)u+\gamma\hat{v}$ and u: for some $\alpha_1, \alpha_2 \in \mathbb{R}$,

$$b = \alpha_1((1-\gamma)u + \gamma\hat{v}) + \alpha_2(u).$$

Since $b(r^*) > b(q)$ but $((1-\gamma)u + \gamma\hat{v})(r^*) = ((1-\gamma)u + \gamma\hat{v})(q)$ and $u(q) > u(r^*)$, we must have $\alpha_2 < 0$.

Let

$$b^* \equiv (1-\epsilon)((1-\gamma)u + \gamma\hat{v}) + \epsilon b$$

for $\epsilon > 0$ small enough that $\epsilon(1 - \alpha_1 - \alpha_2) < 1$. We can rewrite b^* as follows:

$$\begin{aligned} b^* &= (1-\gamma)(1-\epsilon(1-\alpha_1-\alpha_2))u + \gamma(1-\epsilon(1-\alpha_1))\hat{v} \\ &\propto (1-\gamma)u + \gamma\frac{1-\epsilon(1-\alpha_1)}{1-\epsilon(1-\alpha_1-\alpha_2)}\hat{v} \\ &= (1-\gamma)u + \alpha\gamma\hat{v} \text{ for some } \alpha > 1 \end{aligned}$$

where the final equality uses $\alpha_2 < 0$. Since b^* is a convex combination of $(1-\gamma)u+\gamma v$ and b, and since both of these utilities weakly prefer q to everything in $\bar{W}_{\text{outer}}(q)$, (15) holds. □

B Proofs of comparative statics results

B.1 Proof of Proposition 5.1

In this proof, we fix the state s and the decision problem D with $A_2(a_1)$ monotone non-decreasing, and we suppress the dependence of w on s to reduce notation. We define $\tilde{U}(a_1, a_2, \theta) \equiv U_D(a_2, \theta \mid a_1, s)$.

Lemma B.1. *If w has increasing differences between a_1 and (a_2, θ), then $\tilde{U}$ has increasing differences between a_1 and (a_2, θ).*

Proof. Take any $a_1' \geq a_1$ and any $(a_2', \theta') \geq (a_2, \theta)$. Canceling terms, we have

$$\begin{aligned} &\tilde{U}(a_1', a_2', \theta') - \tilde{U}(a_1', a_2, \theta) - \left[\tilde{U}(a_1, a_2', \theta') - \tilde{U}(a_1, a_2, \theta)\right] \\ =&(1-\gamma)\left[w(a_1', a_2', \theta^*) - w(a_1', a_2, \theta^*) - \left[w(a_1, a_2', \theta^*) - w(a_1, a_2, \theta^*)\right]\right] \\ &+\gamma\left[w(a_1', a_2', \theta') - w(a_1', a_2, \theta) - \left[w(a_1, a_2', \theta') - w(a_1, a_2, \theta)\right]\right]. \end{aligned} \tag{24}$$

By increasing differences for w, the term multiplied by $(1-\gamma)$ and the term multiplied by γ are both non-negative, so the right-hand side of (24) is non-negative. □

Lemma B.2. *If w is supermodular in (a_2, θ), then $\tilde{U}$ is supermodular in (a_2, θ).*

Proof. Take any (a_2, θ) and (a_2', θ'). By substitution, we have

$$\begin{aligned} &\tilde{U}(a_1, (a_2, \theta) \wedge (a_2', \theta')) + \tilde{U}(a_1, (a_2, \theta) \vee (a_2', \theta')) \\ &\quad - \tilde{U}(a_1, a_2, \theta) - \tilde{U}(a_1, a_2', \theta') \\ =&(1-\gamma)[w(a_1, a_2 \wedge a_2', \theta^*) + w(a_1, a_2 \vee a_2', \theta^*) \\ &\quad - w(a_1, a_2, \theta^*) - w(a_1, a_2', \theta^*)] \\ &+\gamma[w(a_1, (a_2, \theta) \wedge (a_2', \theta')) + w(a_1, (a_2, \theta) \vee (a_2', \theta')) \\ &\quad - w(a_1, a_2, \theta) - w(a_1, a_2', \theta')] - \gamma\Upsilon, \end{aligned} \tag{25}$$

for

$$\Upsilon \equiv \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta \wedge \theta') + \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta \vee \theta')$$
$$- \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta) - \max_{\substack{\hat{a}_1 \in A_1 \\ \hat{a}_2 \in A_2(\hat{a}_1)}} w(\hat{a}_1, \hat{a}_2, \theta').$$

By Θ totally ordered, $\Upsilon = 0$. Thus, supermodularity of w in (a_2, θ) implies that the right-hand side of (25) is non-negative. □

Lemma B.1, Lemma B.2, and Topkis's theorem yield Proposition 5.1. □

B.2 Proof of Theorem 5.2

Suppose $\bar{a}_1$ was *ex post* too high. We prove that if some action $\bar{a}_2$ and some rationale $\bar{\theta}$ maximizes total utility, then there exists $\theta \geq \theta^*$ such that $(\bar{a}_2, \theta)$ maximizes total utility. Thus, when we are only concerned with actions a_2 that maximize total utility, it is without loss of generality to restrict the rationales to be at least θ^*. An argument using Topkis's theorem then yields Theorem 5.2.

In this proof, we fix the state s and the decision problem D with $A_2(a_1)$ monotone non-decreasing, and we suppress the dependence of w on s to reduce notation. As before, we define $\tilde{U}(a_1, a_2, \theta) \equiv U_D(a_2, \theta \mid a_1, s)$.

Let a_1^* be as defined in Theorem 5.2 and let $\bar{a}_1 \geq a_1^*$. We now define

$$\bar{\Theta} \equiv \operatorname*{argmax}_{\theta \in \Theta} \max_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \theta),$$
$$\bar{\Theta}_{\geq} \equiv \left\{\theta \in \bar{\Theta} : \theta \geq \theta^*\right\}.$$

Observe that

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \max_{\theta \in \Theta} \tilde{U}(\bar{a}_1, a_2, \theta) = \bigcup_{\bar{\theta} \in \bar{\Theta}} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}). \tag{26}$$

Lemma B.3. *Under the assumptions of Theorem 5.2, we have*

$$\bigcup_{\bar{\theta} \in \bar{\Theta}} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) = \bigcup_{\bar{\theta} \in \bar{\Theta}_{\geq}} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \tag{27}$$

Proof. Take any

$$(\bar{a}_2, \bar{\theta}) \in \operatorname*{argmax}_{\substack{a_2 \in A_2(\bar{a}_1) \\ \theta \in \Theta}} \tilde{U}(\bar{a}_1, a_2, \theta). \tag{28}$$

We will show that

$$(\bar{a}_2, \bar{\theta} \vee \theta^*) \in \operatorname*{argmax}_{\substack{a_2 \in A_2(\bar{a}_1) \\ \theta \in \Theta}} \tilde{U}(\bar{a}_1, a_2, \theta).$$

Let us take any $a_2^* \in \operatorname{argmax}_{a_2 \in A_2(a_1^*)} w(a_1^*, a_2, \theta^*)$. By a_1^* *ex post* optimal, we have

$$\tilde{U}(a_1^*, a_2^*, \theta^*) = (1-\gamma) \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*) = \max_{\substack{a_2 \in A_2(a_1^*) \\ \theta \in \Theta}} \tilde{U}(a_1^*, a_2, \theta). \tag{29}$$

By $a_1^* \leq \bar{a}_1$, (28), (29), and Proposition 5.1, we have

$$(a_2^* \wedge \bar{a}_2, \theta^* \wedge \bar{\theta}) \in \operatorname*{argmax}_{\substack{a_2 \in A_2(a_1^*) \\ \theta \in \Theta}} \tilde{U}(a_1^*, a_2, \theta). \tag{30}$$

By (30) and then (29), we have

$$\begin{aligned} (1-\gamma) w(a_1^*, a_2^* \wedge \bar{a}_2, \theta^*) &\geq \tilde{U}(a_1^*, a_2^* \wedge \bar{a}_2, \theta^* \wedge \bar{\theta}) \\ &\geq \tilde{U}(a_1^*, a_2^*, \theta^*) = (1-\gamma) \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \theta^*). \end{aligned}$$

This implies that the action sequence $(a_1^*, a_2^* \wedge \bar{a}_2)$ yields no regret under rationale θ^*, and thus

$$(a_2^* \wedge \bar{a}_2, \theta^*) \in \operatorname*{argmax}_{\substack{a_2 \in A_2(a_1^*) \\ \theta \in \Theta}} \tilde{U}(a_1^*, a_2, \theta). \tag{31}$$

By $a_1^* \leq \bar{a}_1$, (28), (31), and Proposition 5.1, we have

$$(\bar{a}_2, \bar{\theta} \vee \theta^*) = (\bar{a}_2, \bar{\theta}) \vee (a_2^* \wedge \bar{a}_2, \theta^*) \in \operatorname*{argmax}_{\substack{a_2 \in A_2(\bar{a}_1) \\ \theta \in \Theta}} \tilde{U}(\bar{a}_1, a_2, \theta).$$

Our argument holds for any $(\bar{a}_2, \bar{\theta})$ satisfying (28), which yields (27). □

Lemma B.4. *Under the assumptions of Theorem 5.2, we have*

$$\bigcup_{\bar{\theta} \in \bar{\Theta}_{\geq}} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) \gg \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*). \tag{32}$$

Proof. Take any $\bar{\theta} \in \bar{\Theta}_{\geq}$. We define

$$g_{\bar{\theta}}(a_2, \gamma) \equiv (1-\gamma) w(\bar{a}_1, a_2, \theta^*) + \gamma w(\bar{a}_1, a_2, \bar{\theta}).$$

By $\bar{\theta} \geq \theta^*$ and w supermodular, $g_{\bar{\theta}}(a_2, \gamma)$ is supermodular. Thus, by Topkis's theorem,

$\operatorname{argmax}_{a_2 \in A_2(\bar{a}_1)} g_{\bar{\theta}}(a_2, \gamma)$ is monotone non-decreasing in γ. Thus we have,

$$\begin{aligned} \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}) = \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} g_{\bar{\theta}}(a_2, \gamma) \gg \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} g_{\bar{\theta}}(a_2, 0) \\ = \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*). \end{aligned} \tag{33}$$

(33) holds for all $\bar{\theta} \in \bar{\Theta}_{\geq}$, which implies (32). $\square$

By Lemma B.3, Lemma B.4, and (26), we have

$$\operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \max_{\theta \in \Theta} \tilde{U}(\bar{a}_1, a_2, \theta) \gg \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} w(\bar{a}_1, a_2, \theta^*). \tag{34}$$

Moreover, by Lemma B.3, for any a_2 in the left-hand side of (34), there exists $\bar{\theta} \geq \theta^*$ such that

$$\bar{a}_2 \in \operatorname*{argmax}_{a_2 \in A_2(\bar{a}_1)} \tilde{U}(\bar{a}_1, a_2, \bar{\theta}).$$

which completes the proof of Theorem 5.2 for the case $\bar{a}_1 \geq a_1^*$.

Finally, note that if w has increasing differences and is supermodular with respect to some order relations on $\mathcal{A}_1$, $\mathcal{A}_2$, and Θ, then it has increasing differences and is supermodular with respect to the inverse orders on $\mathcal{A}_1$, $\mathcal{A}_2$, and Θ. Similarly, if $A_2(a_1)$ is monotone non-decreasing with respect to some order relations on $\mathcal{A}_1$ and $\mathcal{A}_2$, then it is also monotone non-decreasing with respect to the inverse orders. Thus, our proof covers the case $\bar{a}_1 \leq a_1^*$. $\square$

B.3 Proof of Theorem 5.3

In this proof, we fix the state s and suppress the dependence of U and w on s to reduce notation.

We first establish that switching from D' to D increases the difference from raising a_2 and θ.

Lemma B.5. *Under the assumptions of Theorem 5.3, for any $(\bar{a}_2, \bar{\theta}) \leq (\bar{a}_2', \bar{\theta}')$, we have*

$$U_{D'}(\bar{a}_2', \bar{\theta}' \mid \bar{a}_1) - U_{D'}(\bar{a}_2, \bar{\theta} \mid \bar{a}_1) \leq U_D(\bar{a}_2', \bar{\theta}' \mid \bar{a}_1) - U_D(\bar{a}_2, \bar{\theta} \mid \bar{a}_1). \tag{35}$$

Proof. For $\gamma = 0$, we have (35) trivially. Otherwise, by substitution and then some algebra, (35) reduces to

$$\max_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) + \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}') \leq \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) + \max_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}').$$

Let us define

$$(\hat{a}_1, \hat{a}_2) \in \operatorname*{argmax}_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}),$$

$$(\tilde{a}_1, \tilde{a}_2) \in \operatorname*{argmax}_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}'),$$

By $A_1 \ll A_1'$, we have $\hat{a}_1 \wedge \tilde{a}_1 \in A_1$ and $\hat{a}_1 \vee \tilde{a}_1 \in A_1'$. By (4), we have $\hat{a}_2 \wedge \tilde{a}_2 \in A_2(\hat{a}_1 \wedge \tilde{a}_1)$ and $\hat{a}_2 \vee \tilde{a}_2 \in A_2(\hat{a}_1 \vee \tilde{a}_1)$. Thus, we have

$$w(\hat{a}_1 \wedge \tilde{a}_1, \hat{a}_2 \wedge \tilde{a}_2, \bar{\theta}) \leq \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta})$$

$$w(\hat{a}_1 \vee \tilde{a}_1, \hat{a}_2 \vee \tilde{a}_2, \bar{\theta}') \leq \max_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}'),$$

Combining inequalities yields

$$\begin{aligned}
&\max_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) + \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}') \\
&= w(\hat{a}_1, \hat{a}_2, \bar{\theta}) + w(\tilde{a}_1, \tilde{a}_2, \bar{\theta}') \\
&\leq w(\hat{a}_1 \wedge \tilde{a}_1, \hat{a}_2 \wedge \tilde{a}_2, \bar{\theta}) + w(\hat{a}_1 \vee \tilde{a}_1, \hat{a}_2 \vee \tilde{a}_2, \bar{\theta}') && \text{by supermodularity} \\
&\leq \max_{\substack{a_1 \in A_1 \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}) + \max_{\substack{a_1 \in A_1' \\ a_2 \in A_2(a_1)}} w(a_1, a_2, \bar{\theta}'),
\end{aligned}$$

which implies (35). □

By Lemma B.2, U is supermodular in (a_2, θ). Thus, by Lemma B.5 and Topkis's theorem, we have Theorem 5.3. □