Intrapersonal Utility Comparisons as Interpersonal Utility Comparisons:

Welfare, Ambiguity, and Robustness in Behavioral Policy Problems*

Canishk Naik
University of California Berkeley

Daniel Reck
University of Maryland

January 14, 2026

Abstract

Inconsistent choice undermines the revealed preference foundations of traditional welfare economics, leading to controversy about policymaking in the presence of behavioral frictions. We model an optimal policy problem wherein a benevolent planner is uncertain which behavioral frame reveals normative preferences. We axiomatize welfarist criteria that are similar to social welfare functions, with intrapersonal frames replacing interpersonal types. Under paternalistic ambiguity aversion or paternalistic risk aversion, the planner values policies that are robust to normative uncertainty. We apply these welfarist criteria and robustness concepts in examples, including default options, manipulation of reference points, present focus, corrective taxation for internalities, and nudging.

JEL Codes: D60, D90, H21

*Contact: dreck@umd.edu and canishk@berkeley.edu. For valuable discussions and comments, we thank Doug Bernheim, Scott Elliott, Emel Filiz-Ozbay, Jacob Goldin, Louis Kaplow, Ben Lockwood, Kristóf Madarász, Yusufcan Masatlioglu, Alex Rees-Jones, Daniel Schaffa, Joel Slemrod, Luminita Stevens, Dmitry Taubinsky, and Damian Vergara. Canishk thankfully acknowledges financial support from the Economic and Social Research Council DTP [Grant No: ES/P000622/1].

There's always a reality in what you are doing
Sometimes it's so hard to see which one is the true one
–Gene Clark

1 Introduction

In many settings, inconsistencies in choice make it difficult to infer preferences. A simple default-effects example illustrates the problem: at a new job, Sam contributes 0% to her employer-sponsored pension when there is no automatic enrollment, but if her employer automatically enrolls employees at a default contribution rate of 5%, Sam is passive and saves 5% of her salary instead of 0%. So, Sam initially revealed a preference for 0% when 5% was available, but later chose 5%. Which choice reflects Sam's normative preference? Is it the choice she makes when she chooses actively, or is there a real welfare cost of opting-out? How does this uncertainty matter for the optimal default?

In this paper, we analyze the problem faced by a benevolent planner who sets policy while facing uncertainty about which choices reveal normative preferences. We develop a framework for analyzing optimal policy problems in such contexts, drawing insights from behavioral welfare economics, decision theory, and the theory of social welfare.

One way of tackling this problem is to formulate the policy-maker's problem in terms of expected-utility maximization. The policy maker seeks to maximize welfare w defined as:

$$w = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta). \tag{1}$$

Each $\theta \in \Theta$ represents a frame: a feature of the decision-making environment that affects choice (x) but not welfare. Meanwhile, $\psi(\theta)$ represents the policy-maker's beliefs about the probability that θ indexes the individual's normative preferences $(u(x, \theta))$, which may disagree with preferences expressed in other frames $(u(x, \theta')$ for $\theta' \neq \theta)$. A behavioral policy problem entails maximizing an objective like (1) under *behavioral incentive compatibility*: the individual maximizes their preferences in some frame, which could be influenced by policy and might not coincide with normative preferences [Rees-Jones and Taubinsky, 2018; Danz et al., 2022].

Under which assumptions can we model welfare using a criterion like Equation (1)? The criterion has an obvious resemblance to utilitarian social welfare. With this parallel in mind, the behavioral welfare criterion of Bernheim and Rangel [2009] – a planner prefers option x to y if the individual always chooses x over y regardless of the menu or the frame – resembles the Pareto criterion. Often, incompleteness makes it impossible to infer optimal policies from such criteria alone [see also Benkert and Netzer, 2018]. We present formal conditions under which we can characterize whether a benevolent planner prefers x to y, even if x is not chosen over y in all frames. We do not advocate a single approach; rather we explore a few welfarist objectives, depending on how we think about normative uncertainty (risk versus ambiguity) and how much structure we can impose on the comparability of welfare across frames.

The core assumptions of our framework are as follows. (i) There are *normative preferences*, which describe what an individual *should* choose in any situation. (ii) Holding the frame

fixed, the individual's revealed preferences are rational, i.e. all inconsistencies are explained by framing effects. (iii) Normative preferences coincide with revealed preferences in some frame, called the *normative frame*.

Most prior work in behavioral public economics assumes normative preferences exist and are known to the social planner [e.g. O'Donoghue and Rabin, 2006; Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019].[1] We focus on settings in which normative preferences exist but are unknown, which we formulate as uncertainty over the normative frame. A limitation of this approach is that the normative frame must be one the planner considers, and they must either observe behavior in all potentially normative frames or extrapolate such behavior from observed choices using non-choice information.

Beginning from these three assumptions, we present conditions under which the objective of a benevolent social planner can be represented by the maximization of an intra-personal welfare function as in Equation (1). First, we show that the welfare criterion of Bernheim and Rangel [2009], which we label BR-dominance, is formally admissible under our core assumptions. Building on the similarity of BR-dominance and Pareto dominance, we adapt an argument from Kaplow and Shavell [2001]. If the planner's normative objective is complete, transitive, continuous, and respects BR-dominance, then the information in revealed preferences in each frame must be sufficient to evaluate the planner's objective. That is, welfare is a functional depending exclusively on the utility functions $\{u(x,\theta)\}_{\theta\in\Theta}$ and it is increasing in each argument.

This first result does not impose structure on how the planner trades off welfare across potentially normative frames, which requires a stronger notion of cardinal comparability of welfare across frames [Debreu, 1959; Wakker, 1984; Sen, 1986]. For example, if policy A is optimal in one frame but policy B is optimal in another, then the planner faces a tradeoff in choosing between A and B. Evaluating this tradeoff requires comparing welfare across frames.

We propose more structured approaches to such tradeoffs, drawing on the theory of normative decision-making under uncertainty. In one approach, we impose that the planner can compare welfare across frames and that their preferences satisfy "tradeoff consistency" [see e.g. Wakker and Zank, 1999]. This assumption implies the planner's objective takes a subjective expected utility(/utilitarian) form like Equation (1). In the context of the introductory example, this means that the planner maximizes Sam's expected utility across the uncertainty about whether the "as-if" cost of opting-out of the default reflects a real cost or a mistake.

With the subjective expected utility approach, we require the planner to have probabilistic beliefs about which frame is normative, which might be objectionable. To relax this assumption, we generalize the approach to the case of *ambiguity aversion*, where the planner does not have a unique prior about which frame is normative but there is a set of priors they find plausible [Knight, 1921; Ellsberg, 1961]. We continue to assume that the planner can compare welfare across frames. If the planner maximizes subjective expected utility over known risks but

[1] This is not a universal property of prior work. Our work in relaxing this assumption builds on Goldin and Reck [2022] and Reck and Seibold [2023]. A few other exceptions, like Handel [2013] and Rafkin and Soltas [2024], analyze welfare under uncertainty about whether a behavioral aspect of revealed preferences reflects a normative concern. In these two studies, behavioral frictions are of sufficiently minor importance that normative ambiguity around them does not matter for the main conclusions. See also Handel and Schwartzstein [2018].

prefers to hedge in the presence of ambiguity, they adopt a max-min expected welfare criterion over plausible priors [Gilboa and Schmeidler, 1989].

How should the modeler specify comparable utility functions, assumed to be known by the planner, in applied settings? We do not claim to solve the comparability problem, but argue that fundamentally, deciding how to compare welfare across frames is very similar to deciding how to compare welfare across interpersonal types [as discussed in Harsanyi, 1955; Sen, 1976; Weymark, 1991; Fleurbaey and Maniquet, 2011, and many others]. Concretely, we examine the conditions under which we can use money-metric equivalent variation for intrapersonal welfare comparisons, under varying degrees of agnosticism about the difference in the value of money across frames. A key assumption underlying comparisons of utility across frames is that we can find an option that delivers the same utility in every frame.

We apply our framework to a few behavioral policy problems, including defaults, the manipulation of reference points, corrective taxation with unknown internalities, and nudges. Prior work focuses on whether an observed behavioral friction reflects a bias or a normative preference [Goldin and Reck, 2022; Reck and Seibold, 2023; Lockwood et al., 2024]. For example, if the planner is unsure whether default adherence reflects a normative preference, then setting the default to minimize opt-outs tends to be a *robust optimum*. In contrast, extreme defaults that force active choice are only optimal if the planner is sure opt-out costs reflect a bias. We find a very similar characterization of optimal manipulation of the reference points when the planner is uncertain about the normative relevance of loss aversion.

Turning to corrective taxes [Mullainathan et al., 2012; Allcott and Taubinsky, 2015], normative uncertainty tilts the robust optimal tax toward the worst-case marginal internality. For an ambiguous negative internality, valuing robustness leads to a higher, more paternalistic corrective tax. On the other hand, when the planner is uncertain whether a nudge operates by imposing a normatively relevant cost, the robust nudge is shaded towards zero (less paternalism) as long as demand is not too price-elastic. All these results rely on assumptions about the comparability of welfare across frames, which we highlight in the formal analysis.

Finally, we review prior work that attempts to resolve the normative uncertainty that is primitive in our model, such as the "Counterfactual Normative Consumer" approach [Goldin and Reck, 2020; Allcott et al., 2019; Lockwood et al., 2024] or meta-choice experiments [e.g. Allcott and Kessler, 2019]. Viewed through the lens of our framework, these are methods for identifying normative weights, and we draw a parallel between many of these approaches and prior work on empirically identifying interpersonal welfare weights. All such strategies require untestable normative assumptions, so we argue that our approach to welfare analysis with unresolved normative uncertainty is complementary to these methods.

Our paper is related to prior work in social choice theory, normative decision theory, and behavioral economics, especially behavioral welfare economics. We discuss this relationship throughout the exposition below. We seek to develop no new decision theory in this paper; rather, our objective is to understand how existing ideas from normative decision theory can justify normative objectives for behavioral policy problems. Recent work on normative uncertainty analyzes uncertainty about the correct social welfare function or other normative

criterion [e.g. Dietrich and Jabarian, 2022; MacAskill et al., 2020]. We take a similar approach to behavioral policy problems. Relative to prior work in behavioral welfare economics, the novelty of our approach primarily lies in the development of criteria for optimizing policy in the presence of uncertainty about normative preferences, and our notion of robustness. This allows us to explore the middle ground between the BR-dominance criterion, which is often of limited practical usefulness due to incompleteness, and most other work in behavioral welfare economics, which relies on much stronger restrictions on normative preferences.

Outline. In Section 2 we set up the primitives of the model. Section 3 builds on the core assumptions to characterize the planner's objective and optimal policy. Section 4 covers approaches to comparing welfare across frames, including money-metric welfare. Section 5 introduces some examples from prior literature and maps them into our framework. We develop characterizations of optimal policy that apply our robustness concepts in Section 6. Section 7 discusses approaches to resolving normative uncertainty. Proofs are in the Appendix, along with some extensions and supplementary analysis.

2 Setup

We begin by introducing the core assumptions of our framework and discussing their relationship to prior work.

2.1 Core Assumptions on Individual Choices and Welfare

Primitives. A choice situation is defined by (X, θ), where X is a set of options within a convex set $\mathcal{X} \subseteq \mathbb{R}^N$, and θ is a frame drawn from a finite set Θ. The individual's choice function is $x : 2^{\mathcal{X}} \times \Theta \to \mathcal{X}$, mapping a choice situation to a selection $x(X, \theta) \in X$. The components of an option are denoted $x = (x_1, \ldots, x_N)$. We use standard notation for preference relations.

Assumption 1. ***Core Assumptions.***

Assumption 1.1. ***Normative Preferences Exist.*** *There is a complete and transitive preference relation $\succeq_*$ defined on $\mathcal{X}$, such that the individual's normative choices are those that maximize $\succeq_*$.*

Assumption 1.2. ***Frame-Dependent Rational Preferences.*** *For each $\theta \in \Theta$, there is a complete and transitive preference relation $\succeq_\theta$ on $\mathcal{X}$. Individual choices maximize these preferences: for any $x \in X$, $x(X, \theta) \succeq_\theta x$.*

Assumption 1.3. ***Revealed Preference Coincidence.*** *There is a $\theta^* \in \Theta$ such that for any $x, x' \in \mathcal{X}$,*

$$x \succeq_{\theta^*} x' \iff x \succeq_* x'.$$

Assumption 1.1 ensures the existence of our normative objective. Assumption 1.2 requires that all choice inconsistencies are captured by framing effects: holding the frame fixed, choices are rational. Assumption 1.3 enables normative revealed preference analysis by requiring that in some frame θ^*, choices reveal normative preferences. We label such a θ^* the *normative frame.*

These core assumptions formalize the definition of a frame: frames affect choices but not normative preferences according to Assumption 1.[2] This definition corresponds to the one stated informally in most prior work [e.g. Bernheim and Rangel, 2009; Bernheim and Taubinsky, 2018].[3] We use the term *frame invariance* to refer to the condition that normative preferences do not depend on the frame.

Our core assumptions also admit the welfare criterion of Bernheim and Rangel [2009]. Specifically, under our core assumptions, if x is revealed preferred to x' in every situation where both are available – which is equivalent to preferring x to x' in every frame under Assumption 1.2 – then the normative preference also favors x over x'.

Observation 1. ***BR-Dominance.*** *Under Assumption 1, for any* $x, x' \in \mathcal{X}$,

$$\forall \theta \in \Theta,\ x \succeq_\theta x' \implies x \succeq_* x'. \tag{2}$$

Core Assumptions in a Running Example. How do we apply this setup to the defaults example from the introduction? We can *rationalize* Sam's choices by supposing she maximizes some utility function over the savings rate she chooses (s) together with a fixed cost ($\gamma > 0$) of opting out of the default (d). Supposing Sam's observed choices come from a frame θ^D, we suppose her revealed preferences in this frame, $\succeq_{\theta^D}$, are represented by a utility function of the following form:

$$u(s, d, \theta^D) = v(s) - \gamma \mathbb{1}\{s \neq d\}.$$

Empirical evidence on default effects [reviewed in Goldin and Reck, 2022] suggests that default effects are often well-explained by this fixed cost structure.[4] In the introduction, we pondered whether Sam's normative preferences are revealed by the choices she makes when she chooses actively. We capture this possibility by introducing an alternative frame, θ^A, in which Sam's preferences are represented by

$$u(s, d, \theta^A) = v(s).$$

With this setup, Assumption 1.1 states Sam has normative preferences that govern what savings rate she *should* choose given the default. Assumption 1.2 says Sam's revealed preferences are rational in the naturally occurring frame and the active choice frame (as we impose when representing these choices with utility functions), and Assumption 1.3 implies that one of these

[2] It may be unclear in some settings whether a particular feature of the choice environment should be regarded as a frame; this possibility can be nested within the model (see Example 3 below). The model has a less behavioral interpretation in which the individual has different information in each frame, normative preferences correspond to the choices the individual would make under full information, and the planner might not know which frame is the full-information frame.

[3] Bernheim and Rangel [2009] do not explicitly assume normative preferences exist. One can use a similar approach to formally defining frames that seems compatible with their discussions of this issue. We would need to relax Assumption 1 to allow that 1) normative preferences may be incomplete where BR-dominance does not apply, and 2) RP-coincidence imposes the existence of a welfare-relevant subset $\Theta^* \subseteq \Theta$ such that $(x, \theta) \succeq_* (x', \theta) \iff \forall \theta \in \Theta^*, x \succeq_\theta x'$.

[4] There are some observed default effects that cannot be explained by a fixed opt-out cost alone. The complexity and opacity of the Medicare Part D plans studied in Brot-Goldberg et al. [2023] suggests that another important factor might be individuals' understanding of the options they could get upon opting out. Our model of welfare can be applied to models that allow the individual to be uncertain or to have mistaken beliefs about the benefits of opting out, but this is not nested by our running example.

two frames reflects Sam's normative choice.

Technical Assumptions. We make three more assumptions that simplify our analysis. We introduce a standard continuity assumption on frame-dependent preferences:

Assumption 2. ***Continuity.*** *For any $x_0 \in \mathcal{X}$ and any $\theta \in \Theta$, the sets $\{x \in \mathcal{X} : x \succeq_\theta x_0\}$ and $\{x \in \mathcal{X} : x \preceq_\theta x_0\}$ are closed.*

This assumption is sufficient to ensure we can write down an ordinal utility function denoted $u(x, \theta)$ that represents frame-dependent preferences $\succeq_\theta$ for given θ. As the running example illustrates, Assumption 2 is not necessary for the existence of such a utility function. When continuity is not met but we can represent preferences with a utility function anyway, our overall approach to welfare analysis still applies.

We also assume there is one good that the individual prefers to consume in strictly increasing amounts regardless of the frame. This allows us to adapt an argument from Kaplow and Shavell [2001] below.

Assumption 3. ***One Dimension of Strict Monotonicity.*** *There is some good x_n such that for every θ, $\succeq_\theta$ is strictly monotonic in x_n.*

For instance, Assumption 3 holds in the running example if we think of the variable s as a vector – e.g. present consumption and future consumption – and the subutility function $v(s)$ is strictly increasing in at least one component of this vector.

Finally, we assume that the grand set $\mathcal{X}$ is sufficiently rich that we can induce arbitrary variation in utility in each frame θ by varying options.

Assumption 4. ***Rich Options.*** *For any two options $x, y \in \mathcal{X}$ and any frame $\theta_0 \in \Theta$, there is another option $z \in \mathcal{X}$ such that $x \sim_{\theta_0} z$ and for any frame $\theta \neq \theta_0$, $y \sim_\theta z$.*

Assumption 4 is not satisfied in the running example as we set it up, but it can be satisfied if we also think of the opt-out cost (γ) as a variable component of the options/outcomes.[5]

2.2 Discussion of Setup and Relationship to Prior Literature

Observed Choices versus Known Preferences We suppose a social planner knows the frame-dependent preferences, and believes our core assumptions. We do not necessarily require that the planner directly observes behavior in every frame. In many practical settings, the planner might believe that the normative frame could be different from all the observed frames, so assuming that all frames are observed would threaten Assumption 1.3. In the running example, for instance, choices under the active-choice frame θ^A may not be directly observed in practice.[6] When some choices are unobserved, we implicitly require that the planner uses

[5] Specifically, it is straightforward to show that regarding γ as a component of the options and allowing any $\gamma \in \mathbb{R}$, including negative values, is sufficient to obtain Assumption 4 in the running example. If this modification of the defaults model is unpalatable, one could pursue an alternative rationale for the normative criteria we propose that does not rely on Assumption 4, as we do in Proposition 8 in the Appendix.

[6] For a setting in which choices *are* observed in an active choice frame, see Carroll et al. [2009].

non-choice information to *infer* unobserved potentially normative preferences from observed ones. Such inferences might rely on knowledge of the structure of preferences and/or the preferences of other individuals, as in the counterfactual normative consumer approach to welfare analysis (see Section 7 and Appendix E).

The Set of Frames Assumption 1 requires that the set of frames includes all policy-relevant decision-making frames and all the frames that might be normative. While the planner may not know the normative frame, we rule out the possibility that the normative frame is something they never considered (a "black swan").

How do we specify a set of frames when applying our model in practice? A limitation of our approach is that the modeler must make normative assumptions about what are the potentially normative frames and hence what type of uncertainty to entertain. That being said, there is a convention in prior work to construct Θ: we begin with traditional preference forms, e.g. satisfying time consistency, axioms of expected utility theory, irrelevance of defaults etc, and then introduce additional frames that rationalize observed deviations from traditional forms. With this approach, Assumption 1 requires that the individual's normative preferences correspond either to the traditional form, or to the rationalizing form(s). The two frames we posited in the running example follows this convention, and a similar approach is applicable to any behavioral policy problem in which we can rationalize choices in this fashion.

Nevertheless, one can think of other potentially normative frames in the running example. If we disregard normatively relevant opt-out costs (supposing θ^D is not normative), the default itself could be regarded as a frame. So a planner who is uncertain whether opt-out costs are normative might also be uncertain whether the default is a frame. We flesh this out further in Example 3 below, but we note that the difficulty we wrestle with here reflects the limits of our framework. One can add frames to incorporate more normative ambiguity into the model, but the modeler must restrict normative ambiguity somehow.

Rectangularity Assuming choices are defined on the Cartesian product $2^{\mathcal{X}} \times \Theta$ imposes a condition Bernheim and Rangel [2009] call *rectangularity*: i.e., the set of possible choice menus does not depend on the frame. We extend the core elements of our framework to allow non-rectangular choice environments in Appendix F and apply this to intertemporal choice. Allowing for non-rectangularity introduces complications involving how we articulate the RP-coincidence assumption or model beliefs about which frame is normative; we view these as mainly technical hurdles. The Appendix also discusses a more substantive difficulty involving how one models welfare when individuals do not compare all options via choices [see also Kőszegi and Rabin, 2008].

Limited Attention and Within-Frame Rationality Assumption 1.2 restricts the framework of Bernheim and Rangel [2009] by requiring that all inconsistencies in choice arise solely from framing effects. Therefore, our BR-dominance criterion (Observation 1) differs from theirs. In particular, Assumptions 1.2 and 1.3 exclude models of limited attention: failure to consider all available options can generate choice inconsistencies over menus within frames. Masatlioglu et al. [2012] show that in models with limited attention, BR-dominance is not generally admissible, because individuals could consistently select the best option they attend to, rather than

the option they truly prefer.

We could extend our approach to cover some forms of limited attention. One simple extension would be to require that Assumption 1.2 holds for a subset of frames in which the individual attends to all their options; then Assumption 1.3 would require that one such attentive frame is the normative frame.[7] However, this approach would not capture the possibility that variation in attention has normative importance because attention is a scarce resource, in which case a key question would seem to be whether attention is allocated optimally by the individual [e.g. Bronchetti et al., 2023]. This echoes the type of question we take up in Example 1 below [see discussion of costly attention and defaults in Goldin and Reck, 2022], but we defer analysis of normative uncertainty and welfare under limited attention to future work.

Role of RP-Coincidence in Prior Literature The"RP-Coincidence" assumption (that a normative frame exists) is maintained implicitly or explicitly in virtually all prior work in behavioral welfare economics relying on revealed preferences. For example, Chetty et al. [2009] assume normative preferences are revealed when taxes are completely salient. Other behavioral public economics literature typically selects one or two biases as the main behavioral distortion(s), requiring that decisions purged of this bias reflect normative preferences.

We explicitly allow for uncertainty about which frame is normative, providing a structured approach to characterizing optimal policy under such normative uncertainty. Whereas Bernheim and Rangel [2009]'s dominance criterion alone is often incomplete for practical policy decisions, our approach generates complete characterizations of optimal policy, while allowing the modeler to defer some normative judgments to policymakers. Our approach is nevertheless fundamentally welfarist—defining welfare through revealed preferences and optimizing accordingly—distinguishing it from non-welfarist approaches [e.g. Sugden, 2004].[8]

3 Policy Problems and Normative Criteria

In this section we build on our core assumptions to characterize the planner's problem. We rely entirely on machinery from prior work in decision theory and/or social welfare.

3.1 Behavioral Optimal Policy Problems

Notation. Policies are real-valued vectors denoted $P \in \mathcal{P}$. For a given policy P, the option set the individual chooses from is $X(P)$ and the frame in which the individual makes their choice is $\theta^D(P)$. Where it is irrelevant we suppress the dependence of θ^D on P.

Known Normative Frame. We begin with the case where the normative frame is known to be some $\theta^* \in \Theta$, i.e. there is no normative uncertainty in the model. A benevolent planner's

[7] This is very similar to the concept of a "welfare-relevant domain" in Bernheim and Rangel [2009]. Implementing this extension of our model might require conditioning the welfare-relevant set on the menu as we allow in the extension in Appendix F: the frame in which the agent attends to all their options may depend on the menu.

[8] The opportunity criterion of Sugden [2004] is subject to normative ambiguity in a way that we find intriguing; does changing the default or introducing a default modify one's opportunity set? The answer seems to depend on whether or not opting out of the default requires expending real resources that reflect lost opportunity. We defer a fuller exploration of this idea, and the relationship of opportunity-based normative criteria and our approach to normative ambiguity, to future work.

objective is to choose the policy $P \in \mathcal{P}$ that the individual would choose for themselves according to $\succeq_*$. That is, we should characterize the policy that is optimal subject to the constraint that the individual will choose $x = x(P, \theta^D)$ – the Behavioral Incentive Compatibility (BIC) constraint [Rees-Jones and Taubinsky, 2018; Danz et al., 2022].[9]

RP-Coincidence implies that a benevolent planner should adopt as their the welfare function the utility function that represents $\succeq_{\theta^*}$. The planner's problem under known θ^* is therefore

$$\begin{aligned} &\max_{P \in \mathcal{P}} u(x, \theta^*) \\ &\text{subject to } x = x(X(P), \theta^D(P)). \quad \text{(BIC)} \end{aligned} \tag{3}$$

Many policy problems in the literature on behavioral public economics take the form above, where θ^* is implicitly or explicitly assumed to be known. From this work, we have reduced-form characterizations of the welfare effects of policy variation for known θ^* [e.g. Mullainathan et al., 2012; Allcott and Taubinsky, 2015], and structural characterizations of optimal policies for a variety of structural models [e.g. O'Donoghue and Rabin, 2006].

Unknown Normative Frame. We denote the planner's objective when the normative frame is unknown by the function $w(x)$. The policy problem becomes

$$\begin{aligned} &\max_{P \in \mathcal{P}} w(x) \\ &\text{subject to } x = x(X(P), \theta^D(P)). \quad \text{(BIC)} \end{aligned} \tag{4}$$

One possibility we will consider, for instance, is that the planner has some beliefs about the likelihood that each frame is normative, denoted $\psi \in \Delta(\Theta)$, and maximizes classical expected utility. In this case, w takes an expected utility form like $w(x) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta)$, and the utility function $u(x, \theta)$ is fully comparable across frames (and represents $\succeq_\theta$). Our focus going forward is on axiomatizing such potential forms of $w(x)$.

3.2 Formalizing Behavioral Welfarist Criteria

The planner's preferences over which option the individual consumes are denoted by a relation $\succeq_w$ on $\mathcal{X}$. In writing (4), we are already imposing that $\succeq_w$ has a representation $w : \mathcal{X} \to \mathbb{R}$. We should ensure this representation exists. Traditionally, we say that a planner is *benevolent* if given any x, x',

$$x \succeq_* x' \implies w(x) \geq w(x'). \tag{5}$$

This is insufficient to fully characterize $w(x)$ when the normative frame is unknown. However, under Assumptions 1.1 and 1.3, condition (5) does imply that a benevolent planner should respect BR-dominance. We therefore begin with the following structure on $\succeq_w$:

Assumption 5. ***Basic Structure on Planner's Preferences.***

Assumption 5.1. ***Rationality.*** *$\succeq_w$ is complete and transitive.*

[9] Incorporating an additional constraint like the government budget constraint is straightforward – one can think of this as imposing structure on the set of feasible policies $\mathcal{P}$.

Assumption 5.2. ***Continuity.*** *For any $x \in \mathcal{X}$, the sets $\{x' \in \mathcal{X} : x' \succeq_w x\}$ and $\{x' \in \mathcal{X} : x' \preceq_w x\}$ are closed.*

Assumption 5.3. ***Weak BR-dominance.*** *For any $x, x' \in \mathcal{X}$, if $x \succeq_\theta x'$ for every θ, then $x \succeq_w x'$.*

Our assumptions so far impose that w must depend on information in frame-dependent preferences and it cannot depend on any other information about options.

Proposition 1. *Maintain Assumptions 1.2, 2 and 3. Assumption 5 holds if and only if for any representation of ordinal preferences $u(x, \theta)$, there is a function $\mathcal{W} : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planner's preferences are represented by*

$$w(x) = \mathcal{W}\left(\{u(x, \theta)\}_{\theta \in \Theta}\right), \tag{6}$$

and $\mathcal{W}$ is continuous and weakly increasing in every argument.

Discussion of Proposition 1. As with Pareto dominance, respecting BR-dominance requires that the information in frame-specific preferences $\succeq_\theta$ must be sufficient to evaluate the planner's objective. The proof is an adaptation of an argument in Kaplow and Shavell [2001], which demonstrates that if any other information about the options is used to evaluate the planner's objective, we can find violations of Pareto/BR-dominance – the proof uses continuity and the good x_n from Assumption 3 to construct such violations.

Proposition 1 imposes too little structure on the planner's objective to be of much practical use. Consider, for instance, two options x and x' such that $x \succ_\theta x'$ for some frame θ but $x \prec_{\theta'} x'$ for some other frame θ'. If one policy causes the individual to choose x and another policy causes x', the planner faces a welfare tradeoff between the frames θ and θ'. The assumptions of Proposition 1 require that the planner must find some way to evaluate such tradeoffs, but we do not know how. We introduce two approaches to evaluating tradeoffs given some notion of comparability in the next two sections, and return to the issue of comparability itself in Section 4.

3.2.1 Normative Risk

First, assume the planner approaches these tradeoffs like a subjective expected utility maximizer. We follow Köbberling and Wakker [2003]. Given a utility representation $u(x, \theta)$, a frame θ, an option x, and a real number α such that for some option x', $u(x', \theta) = \alpha$, let $\alpha_\theta x$ denote the option in $\mathcal{X}$ such that $u(\alpha_\theta x, \theta) = \alpha$ and for any $\theta' \neq \theta$, $u(\alpha_\theta x, \theta') = u(x, \theta')$. That such options always exist is ensured by Assumption 4. We denote other options constructed in this fashion using $\beta_\theta x'$, $\gamma_\theta y$, $\delta_\theta y'$, etc. We say that a frame θ is *null* if $\alpha_\theta x \sim_w x$ for every α and every x.[10] Otherwise, the frame is non-null. When the planner's preferences have a subjective expected utility representation, null frames are those with zero probability of being the normative frame.

Assumption 6. ***Tradeoff Consistency.*** *There exists a utility function $u(x, \theta)$ such that $u(x, \theta)$ represents $\succeq_\theta$ for every θ, and for any two non-null frames $\theta_0, \theta_1 \in \Theta$, any four options $x, x', y, y' \in \mathcal{X}$,*

[10] A frame is null if the planner disregards revealed preferences within that frame entirely when evaluating welfare.

and any four real numbers $\alpha, \beta, \gamma, \delta$,

$$\begin{array}{rcl} \alpha_{\theta_0} x \sim_w \beta_{\theta_0} x', & & \alpha_{\theta_1} y \sim_w \beta_{\theta_1} y', \\ \text{and } \gamma_{\theta_0} x \sim_w \delta_{\theta_0} x' & \text{imply} & \gamma_{\theta_1} y \sim_w \delta_{\theta_1} y'. \end{array}$$

Assumption 6 structures how the planner trades off welfare across frames. To unpack the assumption, we find it instructive to think of the utility function whose existence is assumed here as an amount of money that makes the individual in the given frame θ indifferent between option x and being given that amount of money (starting from some baseline situation). In this case, the option $\alpha_{\theta_0} x$ gives α dollars in frame θ_0 and the same payoff as x in other frames. The first two conditions require that in frame θ_0, both the difference between α and β dollars and the difference between γ and δ dollars offset the welfare difference between x and x' in other frames. The planner is willing to make the same tradeoff for both of these differences. Assumption 6 requires the planner to make these types of tradeoffs consistently across frames: if the α-β difference in frame θ_1 offsets the welfare difference between y and y' in other frames, then the γ-δ difference must also do so.

Proposition 2. *Maintain Assumptions 1, 2, 3, and 4. Then Assumptions 5 and 6 hold if and only if there is a probability distribution function* $\psi(\theta)$ *and a utility function* $u : \mathcal{X} \times \Theta \to \mathbb{R}$ *such that* $u(x, \theta)$ *represents* $\succeq_\theta$ *for every* θ*, and planner's preferences* $\succeq_w$ *are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta). \tag{7}$$

Moreover, u *is unique up to positive affine transformation if there are at least two non-null frames.*[11]

Discussion of Proposition 2. Relative to the previous proposition, this result adds that assuming the planner trades off welfare consistently across (non-null) frames, they must be maximizing a subjective expected utility function.[12] In Appendix A.1, we present an alternative approach to deriving an expected utility criterion based on Von Neumann and Morgenstern [1953]. This approach requires introducing the concept of a normative lottery (the equivalent of vNM lotteries over an unknown normative frame) and then imposing the vNM independence axiom over normative lotteries. A unifying analysis of many approaches to expected utility by Wakker and Zank [1999] shows that the key structure imposed by all of them is tradeoff consistency. The approach we take here imposes that the planner finds a way to compare and trade off welfare across frames directly, rather than imposing comparability more indirectly, e.g. via preferences over lotteries.

Application to the Running Example In the running defaults example, assuming that the utility function we wrote down in our initial setup of the example is the one the planner uses

[11] When there is only one non-null frame, we are in the case where the planner knows the normative frame and the utility representation will be unique up to monotonic transformation. When all frames are null, the planner is indifferent to all options, which is a case we generally disregard.

[12] The utility function whose existence is ensured by tradeoff consistency (Assumption 6) is not necessarily identical to the fully comparable utility function whose existence is assured by Proposition 2. However, these two functions must exhibit ordinal level comparability across both options and frames, i.e. they must be monotonic transformations of one another.

to compare utility across frames, Equation (7) becomes

$$w(s,d) = \underbrace{u(x,\theta^A)}_{=v(s)} - \psi(\theta^D)\gamma \mathbb{1}\{x \neq d\}$$

We note that $\psi(\theta^D)$ here is isomorphic to the parameter Goldin and Reck [2022] use to capture normative uncertainty.[13]

3.2.2 Normative Ambiguity

A potential objection to the approach to welfare analysis implied Proposition 2 is that the planner may not know what the probability is that each frame is normative. For decision theory under uncertainty, this critique is usually attributed to Knight [1921] and Ellsberg [1961].

To address this concern, we adapt the theory of ambiguity aversion due to Gilboa and Schmeidler [1989]. In this theory, the planner does not have a unique prior distribution $\psi \in \Delta(\Theta)$ about the likelihood each frame is normative, but rather there is a closed and convex set of distributions $\Psi \subseteq \Delta(\Theta)$ that they find plausible.[14] The planner then maximizes a welfare criterion of the form

$$w(x) = \min_{\psi \in \Psi} \left\{ \sum_{\theta} \psi(\theta) u(x,\theta)) \right\}. \tag{8}$$

Our formal derivation of this objective follows that of Gilboa and Schmeidler [1989] very closely, so we focus on explaining the assumptions intuitively and present formal statements and results in Appendix A.3.

To formally justify the ambiguity averse criterion, we begin with the same setup as the approach using vNM independence discussed above and in Appendix A.1, i.e. we use the concept of normative lotteries and adapt Assumption 5 to preferences over normative lotteries. We directly impose level comparability across frames. Two new assumptions require 1) that when the planner faces known risks, they evaluate tradeoffs in a similar fashion to vNM independence (certainty independence), but 2) when faced with unknown risks (i.e. ambiguity), the planner prefers to *hedge* (uncertainty aversion).[15]

Forms of Ambiguity. We suggest two ways to structure normative ambiguity with the ambiguity averse objective. The first is more global: the planner believes the normative frame comes from a subset of welfare-relevant frames $\Theta^* \subseteq \Theta$, but they have no idea which of these is more likely to be normative, so $\Psi = \Delta(\Theta^*)$. This form suggests that the planner fundamentally lacks a rationale for privileging one potentially normative frame over another, and adopting it leads to a global max-min objective reminiscent of a Rawlsian social welfare function. The second is a more local approach inspired by robust control [Hansen and Sargent, 2001], where the planner begins with a best-guess distribution ψ but accounts for ambiguity by evaluating policies

[13] Goldin and Reck [2022] interpret this parameter, which they denote by π, both in terms of probabilistic beliefs about normative preferences and in terms of models in which opt-out costs are partially but not fully normative; see also the concept of "frame-dependent weighting" in Bernheim et al. [2015]. We explore the relationship between these different potential interpretations of ψ more generally in Appendix D.

[14] We endow the simplex $\Delta(\Theta)$ with a metric suitable for probability distributions e.g. the Wasserstein metric.

[15] We also require a non-degeneracy assumption that prevents the planner from being indifferent to all options/lotteries.

against all distributions in a neighborhood $B(\psi, \kappa)$ for some tolerance parameter $\kappa > 0$. This form seems more appropriate when ψ is identified via a structural model (as in the counterfactual normative consumer approach), but the planner wants to consider robustness because they are concerned about model mis-specification.

All the objectives discussed in this section coincide in the limit under extreme ambiguity or extreme paternalistic risk aversion. Specifically, the planner's objective reduces to the global max-min criterion in three cases: (i) when the tolerance parameter κ is large enough to make $B(\psi, \kappa)$ cover the entire simplex, (ii) when the welfare-relevant subset of frames Θ^* spans all of Θ, and (iii) when the planner is utilitarian but extremely risk averse over a given welfare metric (like the money metric). We state and prove these results formally in Appendix A.3.1.

4 Comparability

A key assumption in Section 3 was that the planner can compare welfare across frames. In this section, we consider a more structured approach to comparability.

4.1 Parallel to Interpersonal Comparisons

We constructed our proposed welfarist criteria by applying decision theory without reference to the theory of social welfare. Fundamentally, however, we view the question of how to compare utility across frames as very similar to the controversy about how to compare utility across individuals in analogous interpersonal problems. An older literature on social welfare derives forms of social welfare function using decision-theoretical axioms, similarly to our work in the previous section.[16] We find axioms in all such prior work that require that the planner can compare utility across individuals, as in our tradeoff consistency assumption. However, there is little consensus on the best way to approach comparability in applied interpersonal problems. Analogously, we do not expect to propose an approach to comparability that will prove universally acceptable to all readers and applicable in all settings.

Instead, we formalize an approach that parallels a common approach to interpersonal comparisons in contemporary public economics [see e.g. Saez and Stantcheva, 2016; Hendren and Sprung-Keyser, 2020; Sher, 2024]. With this approach, one converts all options into money-metric equivalent values, and then treats the value of money to a given individual (and by extension the value of equity) as unknown. In our setting, such an approach views both comparisons of money-metric welfare across frames and the probability that each frame is normative as judgments on the part of the planner, over which we as modelers remain agnostic. This approach requires normative assumptions, whose strength depends on just how agnostic we are about the value of money.

4.2 Money Metric Welfare

We begin with notation and assumptions that discipline equivalent variation. Money is a feature of the choice environment denoted $Z \in \mathbb{R}$; the menu now depends on policy and Z:

[16] Harsanyi [1955] axiomatizes the utilitarian social welfare function, while Hammond [1976] axiomatizes a Rawlsian social welfare function [see also Sen, 1970, 1976; d'Aspremont and Gevers, 1977; Maskin, 1978; Weymark, 1991]. See Sen [1986] for a technical review and see Sen [1997] for a more philosophical perspective. Local maxmin utilitarian objectives are less commonly used to model social welfare, but more recent related papers on this idea include Alon and Gayer [2016] and Mongin and Pivato [2021].

$X(P,Z)$. We use the shorthand notation $x(P,Z,\theta) = x(X(P,Z),\theta)$. To avoid confusion, we observe that $x(P,Z,\theta)$ is not generally the same as the option the individual actually chooses for given (P,Z), which we might express in shorthand as $x(P,Z,\theta^D) = x(X(P,Z),\theta^D(P))$. Under BIC, policy can affect choices by altering both menus and frames, but in thinking about normative willingness to pay, we hold the frame fixed in a potentially normative frame.

Assumption 7. ***Ordinal Equivalent Variation Admissibility.***

Assumption 7.1. ***Strict Monotonicity in Money.*** *For any two Z, Z', any frame θ, and any policy P,*

$$Z > Z' \iff x(P,Z,\theta) \succ_\theta x(P,Z',\theta).$$

Assumption 7.2. ***Continuity over Money.*** *For any Z, any frame θ and any policy P, the sets $\{Z' : x(P,Z',\theta) \succsim_\theta x(P,Z,\theta)\}$ and $\{Z' : x(P,Z',\theta) \precsim_\theta x(P,Z,\theta)\}$ are closed.*

Assumption 7.3. ***Equalizability.*** *For any $x \in \mathcal{X}$, any policy P and any frame θ, there exists Z such that $x(P,Z,\theta) \succ_\theta x$ and Z' such that $x(P,Z',\theta) \prec_\theta x$.*

Discussion of Assumption 7. Combined with RP-Coincidence, Assumption 7.1 ensures that giving the individual more Z improves welfare in the normative frame; imposing strict monotonicity over money allows us to relax strict monotonicity over some good (Assumption 3). Assumption 7.2 requires that welfare is continuous in Z in every frame. Assumption 7.3 ensures that all changes to welfare driven by variation in choices can be fully offset by money, which is obviously key for the existence of equivalent variation. All this is assumed regardless of which frame is normative.

Together, these assumptions discipline the equivalent variation of a given option x in a given frame θ, which is defined as $\zeta \in \mathbb{R}$ such that for a given *baseline* (P_0, Z_0),

$$x \sim_\theta x(P_0, Z_0 + \zeta, \theta). \tag{9}$$

Lemma 1. ***Existence and uniqueness of EV.*** *Under Assumptions 1.2, 2 and 7, for any option x any frame θ and any baseline (P_0, Z_0), equivalent variation ζ exists and is unique. Moreover, $\zeta(x,\theta;P_0,Z_0)$ represents ordinal preferences $\succeq_\theta$ for every θ.*

Lemma 1 implies that for a given baseline, $\zeta(x,\theta;P_0,Z_0)$ is a unique representation of revealed preferences under θ. Combining the representation result in Lemma 1 with the idea in Proposition 1, we find that introducing Assumption 7 yields the following:

Proposition 3. ***Planner's Preferences and Equivalent Variation.*** *Under Assumptions 1.2, 2, 5, and 7, for any baseline P_0, Z_0, there is a function $\mathcal{W}_\zeta : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planner's preferences are represented by $w(x) = \mathcal{W}_\zeta\left(\{\zeta(x,\theta;P_0,Z_0)\}_{\theta\in\Theta}\right)$.*

Perturbation Approach to Welfare Analysis. Proposition 3 suggests that provided that equivalent variation is well-behaved per Assumption 7, we may use equivalent variation welfare metrics to describe the welfare effects of local policy perturbations. Under the assumptions

of the proposition and additionally assuming differentiability of all relevant quantities, the welfare effect of a marginal policy reform (a perturbation) dP that does not modify θ^D will be

$$dw = \sum_{\theta} \frac{\partial \mathcal{W}_z}{\partial \zeta_\theta} d\zeta(x(P, Z, \theta^D), \theta; P, Z). \tag{10}$$

where derivatives are evaluated at the status quo (P, Z, θ^D); $d\zeta(., \theta)$ is the money-metric welfare effect of the marginal reform in frame θ evaluated using the status quo as the baseline.[17] With the addition of tradeoff consistency (Assumption 6), the welfare weight term in the above expression becomes

$$\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta} = \psi(\theta) \frac{\partial u(x(P, Z, \theta^D), \theta)}{\partial Z}. \tag{11}$$

In words, the welfare effect of the perturbation is a weighted mean of its equivalent variation across frames, where the welfare weights $\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta}$ involve two normative judgments: the probability that each frame is normative $\psi(\theta)$, and the value of money in frame θ, $\frac{\partial u^\theta}{\partial z}$.[18] The latter is governed by how the planner compares monetary payoffs across frames.

One could stop here and remain agnostic about these welfare weights, as is common in some prior work on interpersonal welfare weights [e.g. Hendren and Sprung-Keyser, 2020]. Below, we set out some assumptions which ensure the planner can directly compare equivalent variation across frames, which implies welfare weights only depend on $\psi(\theta)$ and the value of money (common across frames). For this to work, tradeoff consistency needs to hold over money-metric utility itself. This in turn requires ordinal level comparability of equivalent variation and the planner's utility function, i.e. for any two options x, x' and any two frames θ, θ', $u(x, \theta) \geq u(x'.\theta') \iff \zeta(x, \theta) \geq \zeta(x'.\theta')$.

Assumption 8. ***Cardinal Equivalent Variation Admissibility.*** *Let $u(x, \theta)$ be a fully comparable utility function from the representation of $w(x)$ in Proposition 2. There is a baseline situation (P_0, Z_0) under which the following conditions hold:*

Assumption 8.1. ***Baseline Indifference.*** *For any θ, θ',*

$$u(x(P_0, Z_0, \theta), \theta) = u(x(P_0, Z_0, \theta'), \theta').$$

Assumption 8.2. ***Comparable Value of Money At Baseline.*** *For any θ, θ' and any $\zeta, \zeta' \in \mathbb{R}$,*

$$u(x(P_0, Z_0 + \zeta, \theta), \theta) - u(x(P_0, Z_0, \theta), \theta) \geq u(x(P_0, Z_0 + \zeta', \theta'), \theta') - u(x(P_0, Z_0, \theta'), \theta') \iff \zeta \geq \zeta'.$$

Assumption 8.1 requires that the level of utility is the same across frames in the baseline situation. Assumption 8.2 requires that starting from the baseline and giving the individual some

[17] We do not consider policies that perturb the frame because we assume the set of frames is finite. We view this mainly as a technical limitation and expect the approach to generalize readily to policies that perturb frames. Note that we refer to the status quo situation as the situation from which we perturb policy, while the baseline situation is the situation from which we construct equivalent variation. In Equation (10) these are the same, but for our next result they may differ.

[18] For example, in models featuring loss aversion that may or may not reflect a bias (Example 1.2 below), the value of a dollar differs across frames because giving the individual money allows them to reduce their losses, which is more valuable when loss aversion reflects a normative preference than when it reflects a bias.

amount of money has the same effect on utility regardless of the frame. One might think of this assumption as a selection criterion for the baseline situation. The assumption seems more intuitive, for instance when the individual makes the same choices in different frames, i.e. when $x(P_0, Z_0, \theta)$ – or even $x(P_0, Z_0 + \zeta, \theta)$ for a range of ζ's – is constant over θ.

Lemma 2. ***Ordinal Level Comparability of Equivalent Variation.*** *Maintain Assumptions 1, 2, 4, 5, 6, and 7. Let $u(x, \theta)$ be a cardinal utility function from the representation in Proposition 2. Assumption 8 holds if and only if there is a baseline (P_0, Z_0) such that $u(x, \theta)$ and $\zeta(x, \theta; P_0, Z_0)$ exhibit ordinal level comparability.*

Proposition 4. *Under Assumptions 1, 2, 4, 5, 6, 7, and 8, there is a probability distribution $\psi(\theta)$, a function $u_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0, Z_0) such that the planner's preferences are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u_\zeta(\zeta(x, \theta; P_0, Z_0)). \tag{12}$$

Moreover, u_ζ is strictly monotonic and unique up to positive affine transformation when more than two frames are non-null.

Discussion of Proposition 4. Under these assumptions, the planner can directly compare money-metric welfare across frames and the only remaining unstructured component of the expression is utility over money itself, u_ζ. The result extends straightforwardly to the representation of the planner's preferences in the ambiguity case from equation (8).

Under equation (12), the effect of a local policy perturbation becomes

$$dw = \psi(\theta) \frac{du_\zeta}{d\zeta} d\zeta(x(P, Z, \theta^D), \theta; P_0, Z_0), \tag{13}$$

where the derivative is again evaluated at the status quo, but unlike before, equivalent variation $d\zeta$ is evaluated relative to the baseline from Assumption 8. When the value of money differs across frames, Proposition 4 suggests that we can compare equivalent variation across frames as long as we specify the baseline situation where across-frame indifference holds. In each example of Section 5, we explain our decision about in which situation we should assume baseline indifference.

Paternalistic Risk Aversion One payoff to this additional structure is that now we can think of paternalistic risk aversion in simple terms. Specifically, the planner's preferences exhibit *paternalistic risk aversion over money* if u_ζ is concave. Paternalistic risk aversion implies a preference not to have too much disagreement across frames: reforms that amplify disagreements across frames are riskier. Under paternalistic risk aversion, the first-order welfare effect of a marginal policy reform is the expected welfare effect plus an adjustment for the change in the variance of welfare across frames. The adjustment term is negative if u_ζ is concave: paternalistic risk aversion introduces a penalty on policies which increase variance. This implies a similar but less extreme preference for robustness than the case of ambiguity aversion.[19]

[19] See Corollary 9.1 in the Appendix. We prove all of the results mentioned in this paragraph formally and further explore a perturbation approach to welfare analysis within our framework in Appendix B.

5 Examples

We now illustrate via examples how prior work on behavioral welfare economics fits within our framework. In each example, we consider what the "expected utility" formulation of the planner's preferences from Equation (7)/Proposition 2 looks like for a given set of beliefs ψ. This is done for expositional clarity to explore which kind of normative uncertainty a planner might face in some concrete examples. We turn to analyses of optimal policy under ambiguity and robustness in Section 6, directly imposing comparability across frames and illustrating how this matters for robust optimality.

5.1 Example 1: Biases Versus Strange Preferences

Let us introduce a general example in which the key intrapersonal question is whether some behavioral phenomenon arises due to a bias or a normative preference. Suppose the decision-making frame is some fixed frame θ^D and there is just one alternative frame denoted θ^A. Our representation of welfare from equation (7) becomes

$$w(x) = \psi(\theta^D)u(x, \theta^D) + [1 - \psi(\theta^D)]u(x, \theta^A). \tag{14}$$

With two frames, we denote disagreements using $V(x) \equiv u(x, \theta^D) - u(x, \theta^A)$. To relate our framework to prior work, let us re-write $w(x)$ using the definition of $V(x)$:

$$w(x) = u(x, \theta^D) - [1 - \psi(\theta^D)]V(x) \tag{15}$$

$$= u(x, \theta^A) + \psi(\theta^D)V(x). \tag{16}$$

$$u(x, \theta^D) = u(x, \theta^A) + V(x). \tag{17}$$

Prior work on behavioral frictions often uses a formulation like equation (16), where we think of $V(x)$ as the "behavioral" component of preferences, while "decision utility" takes a form like equation (17). The behavioral component $V(x)$ is typically a deviation from classical forms of preferences that may or may not be due to a bias. When $\psi(\theta^D) = 1$, for instance, the planner knows with certainty that V is a non-standard but normative preference rather than a bias. When $\psi(\theta^D) = 0$, the planner knows that V reflects a bias.

Next we refine this example by considering specific behavioral frictions from prior literature.

Example 1.1. Defaults. The running example introduced above is obviously nested by Example 1. For $x = (s, d)$ we have

$$V(s, d) = -\mathbb{1}\{s \neq d\}\gamma. \tag{18}$$

Example 1.2. Reference Dependence. Reference dependence is the subject of a rich theoretical and empirical literature [Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Kőszegi and Rabin, 2006; Crawford and Meng, 2011; Thakral and Tô, 2021], including many policy-relevant applications [DellaVigna et al., 2017; Rees-Jones, 2018; Seibold, 2021]. A lack of consensus about whether to regard this phenomenon as a bias or a preference has hindered our ability to evaluate policy in these settings [O'Donoghue and Sprenger, 2018]. Reck and

Seibold [2023] consider a model, nested by Example 1, in which the behavioral component of preferences $V(.)$ is a reference-dependent payoff featuring loss aversion.

We use a similar setup to the running example, but replace the default d with a reference point r; now options are denoted $x = (s, r)$. When researchers model reference-dependent choice, they introduce a utility function over s with classical properties labelled "intrinsic utility" or "consumption utility" [e.g. Kőszegi and Rabin, 2006], which is additively separable from a gain-loss payoff over (s, r). We can nest this in our biases versus strange preferences setup if we posit a naturally occurring frame in which the individual makes choices based on both intrinsic and gain-loss utility, and an alternative frame in which the individual makes choices based on intrinsic utility alone. Following Kőszegi and Rabin [2006], we assume intrinsic utility is additively separable, so that we may write $u(s, r, \theta^A) = \sum_{i=1}^{N} u_i(s_i)$, where s_i is the ith component of the vector s. For parameters $\Lambda_i > 0, \beta \in (0, 1]$, we specify a gain-loss payoff of the form

$$V(s, r) = -\sum_{i=1}^{N} \mathbb{1}\{s_i \leq r_i\} \Lambda_i \left[u_i(r_i) - u_i(s_i)\right]^{\beta}, \tag{19}$$

The individual only incurs a payoff along some dimension if they incur a loss, $s_i \leq r_i$. The parameter Λ_i governs the strength of loss aversion along dimension i, while the parameter β governs diminishing sensitivity. We separately consider the case without diminishing sensitivity ($\beta = 1$) and with it ($\beta < 1$) below.

This is a similar form to that proposed by Kőszegi and Rabin [2006] – gains and losses are evaluated over "utils" rather than units of each good – except that 1) we disregard gain domain payoffs where $x_i > r_i$ along some dimension, and 2) we allow the extent of loss aversion Λ_i to vary across dimensions i rather than being fixed. These choices are motivated by a more detailed analysis of forms of gain-loss utility in Reck and Seibold [2023].[20]

Example 1.1 and Example 1.2 under $\beta = 1$, as well as the salience model of Bordalo et al. [2012], are all cases of the "Affine Categorical Thinking Model" from Ellis and Masatlioglu [2022]. One could adapt the approach we develop here to analyze welfare in this salience model, or another Affine Categorical Thinking Model. Not all such models can be nested within Example 1, but our overall approach is applicable because these models feature a family of intrapersonally comparable utility functions. Our next example does not fall within this class of models.

Example 1.3. Probability Re-Weighting. Starting with Kahneman and Tversky [1979], researchers have modeled deviations from expected utility theory due to the reweighting of objective probabilities [see also Prelec, 1998; Abdellaoui, 2000; Chateauneuf et al., 2007]. In a recent welfare analysis of state-run lotteries, Lockwood et al. [2024] present a model in which the main behavioral friction is probability re-weighting and it is uncertain whether re-weighting reflects a bias or a normative preference. In particular, individuals' revealed preferences – identified empirically using demand responses to changes in lottery prizes – suggest their utility function puts excess weight on the jackpot payoff especially.

[20] Including a gain-domain payoff whose strength is governed an additional parameter, usually denoted by η [Tversky and Kahneman, 1991], would not change the results of interest to us provided that η_i is not too strong along any given dimension i, in a sense formalized in Reck and Seibold [2023], Appendix B6.

To nest their model, we think of each component of $x = (x_1, ..., x_N)$ as the payoff in each realization of an uncertain state variable. Objective probability of each state is $\pi = (\pi_1, ..., \pi_N)$. Individuals re-weight each objective probability according to a function $f(\pi)$ and are endowed with a Bernoulli utility function $\mu(x_n)$. Utility in the fixed decision-making frame is $u(x, \pi, \theta^D) = \sum_n f(\pi_n)\mu(x_n)$.

When $f(\pi) = \pi$ everywhere, we have classical expected utility maximization. If we view the vNM independence axiom as normative, then normative utility should coincide with expected utility. For an alternative frame θ^A in which the individual's choices respect the independence axiom, we have $u(x, \pi, \theta^A) = \sum_n \pi_n \mu(x_n)$. The disagreement between these two views of welfare is then, by our definition,

$$V(x, \pi) = \sum_n [\pi_n - f(\pi_n)]\mu(x_n). \tag{20}$$

Now with our framework, we can think of a planner who is uncertain about whether the excess weight on the jackpot payoff (and the resulting under-weighting of payoffs in other states) is normative. Lockwood et al. [2024] model the extent to which re-weighting reflects a bias with a parameter that is isomorphic to $\psi(\theta^A)$ here; we discuss their strategy for identifying these weights in Section 7 and Appendix E.

5.2 Example 2: Present Focus

Our next example is motivated by prior work on present focus and the notion of intertemporal selves [e.g. Laibson et al., 1998; Caliendo and Findley, 2019]. The options are lifetime consumption plans: $\mathcal{X} = \mathbb{R}_+^T$, where T is the number of time periods. An option is $x = (x_1, ..., x_T)$. The frame in this model is the vantage point from which individuals evaluate a consumption plan. We characterize individuals' preferences under commitment, i.e. we think of the individual selecting a full consumption plan in each period. We assume individuals are quasi-hyperbolic discounters as in Laibson [1997]. We also assume there is a period 0 in which the individual is entirely forward-looking, i.e. they do not consume or receive flow utility. For two parameters $\beta > 0, \delta \leq 1$, a flow utility function $\mu(x_t)$, and a vantage point $\tau = 0, ..., T$ we specify:

$$u(x, \tau) = \mathbb{1}\{\tau > 0\}\delta^\tau \mu(x_\tau) + \beta \sum_{t \neq \tau} \delta^t \mu(x_t). \tag{21}$$

Note that with this formulation, we endow the period τ self with preferences over the prior selves' consumption; this is unconventional and we discuss the rationale for this modeling choice below. Many prior papers adopt the "long-run view" that $\tau = 0$ is the normative frame [e.g. O'Donoghue and Rabin, 2006]. The period 0 self is a classical exponential discounter, and in fact, we find that a planner's welfarist objective based on formulation (21) has a representation along similar lines to the Biases versus Strange Preferences example. By construction, the

planner's welfare function takes the following form:

$$\begin{aligned} w(x) &= \beta \sum_{t=1}^{T} \delta^t \mu(x_t) + \sum_{\tau=0}^{T} \psi(\tau) \mathbb{1}\{\tau > 0\}(1-\beta)\delta^\tau \mu(x_\tau) \\ &= u(x,0) + (1-\psi(0)) \sum_{\tau=1}^{T} \psi(\tau|\tau>0)[u(x,\tau) - u(x,0)] \end{aligned} \tag{22}$$

This formulation for welfare resembles the formulation for welfare from Example 1, Equation (16): the first term is a utility function with classical features and the second is a deviation from classical preferences weighted by the planner's beliefs about the probability the individual has non-standard normative preferences $(1-\psi(0))$. In this case, there is more than one alternative view because each of the period $\tau > 0$ selves could receive different normative weights (see also Example 3 below).

In the theory of social welfare functions, we frequently find an "anonymity" condition requiring that no two individuals should have their utility differently weighted [e.g. Maskin, 1978]. While anonymity is not generally useful for our model, it is intuitive to impose it over the $\tau > 0$ selves. Doing so justifies the planner adopting the long-run view of welfare, which *does not require assuming that present focus is a behavioral bias.*

Proposition 5. ***Intertemporal "Social" Welfare and the Long Run View.*** *In this model, if $\psi(\tau)$ is constant for $\tau > 0$, then for any $\psi(0)$, the planner's preferences coincide with the long-run view of welfare $u(x,0)$.*

Proposition 5 provides a new justification for the long-run view of welfare, contributing to the debate about its normative justification [see Bernheim, 2009; Caliendo and Findley, 2019]. The intuition is that when present-focused payoffs are aggregated across all intertemporal selves, intertemporal tradeoffs over these payoffs become proportional to tradeoffs over long-run utility. This result requires that the extent of present focus β remains constant over time, ruling out e.g. the possibility that individuals are present-focused when young but not when old.

Remark on Intertemporal Preferences Formulation. Holding x_s fixed for every $s < \tau$, the above generates the same choices as the conventional β-δ representation of preferences, i.e. $\tilde{u}(x_{t\geq\tau}, \tau) = \mathbb{1}\{\tau > 0\}\mu(x_\tau) + \beta \sum_{t=\tau+1}^{T} \delta^{t-\tau}\mu(x_t)$. Our formulation differs from prior work by endowing the period τ self with classically discounted preferences over consumption in prior periods $t < \tau$, addressing issues with the intertemporal selves model identified by Bernheim and Rangel [2009].

To specify preferences that cannot be directly inferred by choices over forward-looking consumption plans, we assume the period-τ self agrees with prior selves about intertemporal consumption trade-offs. For any pair $\tau > 0$, $\tau' > 0$, and consumption plans x, x' such that $x_\tau = x'_\tau$ and $x_{\tau'} = x'_{\tau'}$, then $u(x,\tau) \geq u(x',\tau) \iff u(x,\tau') \geq u(x',\tau')$. This ensures the period-τ and period-τ' selves make the same choices, holding fixed consumption in τ and τ'.

Welfare with this formulation accords with BR-Dominance over committed choices. The setup is "rectangular" in that for any frame τ, individuals have frame-dependent rational preferences over the entire option space. Without rectangularity, allocating all resources to the last-

period self would always be an intertemporal-self Pareto optimum, contradicting revealed preference. We address this by adopting rectangular preferences using formulation (21) and considering preferences under commitment.

In Appendix F, we analyze welfare without backward-looking preferences using our relaxation of the rectangularity condition and show that with this setup, uncertainty about whether the long-run or short-run view is normative is equivalent to uncertainty about whether the vantage point τ is a frame, similarly to the next example below. We also note that we consider normative frames constructed from committed choices, implying that uncommitted choices made under naivete [e.g. DellaVigna and Malmendier, 2006] are not normative. Naive agents do not understand the consumption plan that will result from their choices – a case of "characterization failure" [Bernheim and Taubinsky, 2018]. We also assume away welfare-relevant shame experienced by naive agents who fail to adhere to plans. Bernheim et al. [2024] develop tools for relaxing this type of assumption by identifying preferences over emotional experiences, which are compatible with our framework but not nested in our examples.[21]

Comparability of Welfare Across Intertemporal Selves The units of utility are determined by $\mu(x)$, which is cardinal [Montiel Olea and Strzalecki, 2014]. We assume the units of μ are the normative units for welfare analysis. The conventional level normalization, $\mu_\tau(0) = 0$ for every τ, implies that the present-focused self with $\beta < 1$ will always have more utility than the period 0 self, who never experiences the utility benefit of current consumption. Formally:

$$\forall \tau, \mu_\tau(0) = 0 \text{ and } \forall x \geq 0, \frac{d\mu_\tau(x)}{dx} \geq 0 \implies \forall x \geq 0, \min_{\psi \in \Delta(\Theta)} \psi(\tau) u(x, \tau) = u(x, 0). \tag{23}$$

Thus, the globally ambiguity averse planner again adopts the long run view of welfare under this level normalization. This potentially problematic assumption becomes moot if we introduce the anonymity condition from Proposition 5. Otherwise, addressing the problem requires specifying a consumption plan that equalizes utility between $\tau = 0$ and $\tau > 0$ selves.

5.3 Example 3: Is a Feature of the Environment a Frame?

What happens when it is unclear whether a feature of the environment is a frame? In Example 1.1, the default option d itself cannot be a frame because welfare depends directly on d if the opt-out cost is seen as normative, violating frame invariance.[22] When the the opt-out cost is normative, however, we could think of choices made under each default as coming from distinct frames rather than a single frame, e.g. due to an "anchoring effect."

Example 3 uses a similar setup to Example 1.1, but the frame has two components: $\theta = (\theta_1, \theta_2)$. The second component is a factor like the default ($\theta_2 \in X$),[23] and the first component, $\theta_1 \in$

[21] We can envision two sources of normative uncertainty in Bernheim et al. [2024]'s model. First, inconsistencies in observed choices across frames could imply inconsistent rankings of mental states, perhaps due to error in the measurement of mental states themselves. Second, the map from experienced mental states to welfare may be uncertain, due to ambiguity, for instance, in whether individuals should allow emotions to influence their choices or whether normative choices come from a dispassionate frame. For a related discussion of the relationship between negative emotions and normative choices, see Loewenstein and O'Donoghue [2006].

[22] A similar argument can be made about the reference point r itself in Example 1.2.

[23] In this example, there could be continuum of frames, contrary to our initial setup with finite frames. We present

$\{0,1\}$ indicates whether the second component can really be viewed as a frame ($\theta_1 = 1 \implies \theta_2/d$ is a frame).

We express the utility function as $u(x,d,\theta_1,\theta_2)$ and we make two restrictions to capture our intuition. When $\theta_1 = 0$, saying feature d is not a frame requires that $u(x,d,0,\theta_2)$ must be constant over θ_2, which we express with a utility function $u_0(x,d) \equiv u(x,d,0,\theta_2)$ for any θ_2. If $\theta_1 = 1$, feature d is a frame so frame invariance requires $u(x,d,1,\theta_2)$ to be constant over d, which we express with a function $u_1(x,\theta_2) = u(x,d,1,\theta_2)$. Denote disagreements between the $\theta_1 = 0$ and $\theta_1 = 1$ cases by $V(x,d,\theta_2) = u_0(x,d) - u_1(x,\theta_2)$ and $\psi_0 = \sum_{\theta_2 \in D} \psi(0,\theta_2)$ be the total weight on $\theta_1 = 0$ where D is a finite set of default options $D \subset X$. Also, let $\psi(\theta_2|\theta_1)$ the conditional distribution of θ_2 given θ_1. Then, we derive an identity similar to Equation (15):

Observation 2. ***Welfare given uncertainty about whether a feature of the environment is a frame.***

$$w(x) = u_0(x,d) - (1-\psi_0) \sum_{\theta_2 \in D} \psi(\theta_2|1) V(x,d,\theta_2) \tag{24}$$

$$= \psi_0 \cdot u_0(x,d) + (1-\psi_0) \cdot \sum_{\theta_2} \psi(\theta_2|1) \cdot u_1(x,\theta_2) \tag{25}$$

The weight ψ_0 is similar to $\psi(\theta^D)$ in Example 1. With this two-stage setup, we confront more ambiguity than in the biases versus strange preferences case from Example 1. First, the planner must decide whether they can rule out the possibility d is a frame. If so, this corresponds to the case where the planner knows with certainty that the effect of d on choices reflects a normative preference and all ambiguity is resolved ($\psi_0 = 1$ and $w(x) = u_0(x,d)$). But, if $\psi_0 < 1$ in (24), there is a possibility d is a frame and substantial ambiguity in welfare remains due to choice inconsistencies as d varies because the planner needs to decide which d reveals normative preferences.

In the case of default effects with observed default d, letting $x(d)$ denote the choice the individual makes under BIC, Observation 2 implies

$$w(x(d)) = u(x(d)) - \psi_0 \cdot \gamma \mathbb{1}\{x(d) \neq d\} + \sum_{\theta_2} \psi(1,\theta_2) \cdot [u(x(\theta_2)) - u(x(d))]$$

where $x(\theta_2) \coloneqq \arg\max u(x) - \gamma \mathbb{1}\{x \neq \theta_2\}$ and $\psi(1,\theta_2) = \mathbb{P}[\text{frame } \theta_2 \text{ reveals true preferences}]$.[24] Most models of default effects considered in Bernheim et al. [2015] are nested in Example 1.1, but the anchoring model they consider resembles Example 3 under $\psi_0 = 0$. Understanding the difference between these examples clarifies why adopting the anchoring model generates more ambiguous welfare effects.

We do not engage deeply with models like Example 3 in the remainder of this paper (except for Appendix F). This is done in the interest of providing simple illustrations of our robustness

an extension to continuous frames in Appendix D, but proceed by assuming a finite set of default options $D \subset X$ as the welfare relevant domain.

[24] One way in which we could conceive $\psi(1, \text{Active Choice Default}) < 1$ is if opting-out of the default imposes a cognitive tax which operates by undermining choice quality rather than a direct utility cost.

concept rather than our thinking that the approach applied by Example 1 is superior to the one implied by Example 3 for any particular behavioral phenomenon.

6 Robust Optimal Policy in Applications

In this section, we characterize the optimality of policies when there is normative uncertainty. For each case, we start by characterizing the optimal policy given a set of beliefs ψ about which frame is normative and a utilitarian objective like Equation (7). Then, we incorporate the case that the planner has some aversion to uncertainty, i.e. ambiguity aversion from Equation (8).

6.1 Setup

Before we turn to our examples, we translate the different formulations of the planner's preferences from Section 2 to notions of optimal policy:

Definition. A policy P^* is a *ψ-optimum* for $\psi \in \Delta(\Theta)$ if $P^* \in \arg\max_{P \in \mathcal{P}} E_\psi[u(x(P), \theta)]$.

For a given beliefs ψ (and no ambiguity), a ψ-optimum maximizes the planner's preferences.

Definition. For a given set of distributions $\Psi \subseteq \Delta(\Theta)$, a policy P^* is a *robust optimum* if

$$P^* \in \arg\max_{P \in \mathcal{P}} \min_{\psi' \in \Psi} E_{\psi'}[u(x(P), \theta)].$$

If Ψ is the closed and convex set of beliefs which define the planner's preferences in the ambiguity-averse case as in Section 3.2.2 then robustness corresponds to maximizing these preferences. We leave the particular value of ψ under risk or Ψ under ambiguity unspecified and illustrate how they matter for the optimum. In thinking about robust optima, we often adopt the intuition introduced by Hansen and Sargent [2001], thinking of an "evil agent" who, subject to the planner's chosen policy, picks a ψ to minimize welfare; to be a robust optimum, the policy is the best possible policy for welfare given the evil agent's reaction.

Definition. A policy P^* is a *globally robust optimum* if it is a ψ-optimum for all $\psi \in \Delta(\Theta)$.

Obviously, a globally robust optimum will also be a robust optimum for any $\Psi \subseteq \Delta(\Theta)$. Global robustness also has a straightforward relationship to BR-dominance:

Observation 3. ***BR-Optimality and Global Robustness.*** *A policy $P^* \in \mathcal{P}$ is a globally robust optimum if and only if for every $P' \in \mathcal{P}$, $x(P)$ weakly BR-dominates $x(P')$.*

We mainly focus on global characterizations of optimal policies and defer perturbation-based (first-order) characterizations to Appendix B. There, we show that a generalized version of the reduced-form optimality criterion from Mullainathan et al. [2012] continues to hold under normative uncertainty: the welfare effect of a marginal reform equals its expected direct effect minus expected marginal internalities, both averaged across normative frames. We illustrate how this plays out in the examples in this section. A stylized corrective tax example (Section 6.2.3) below illustrates the key intuition and how ambiguity aversion modifies it.

6.2 Examples

Now we explore how the notions of optimality defined in Section 6.1 play out in examples.

6.2.1 Optimal Defaults

Suppose a benevolent planner selects the default in our running example/Example 1.1; this optimal policy problem is studied in Carroll et al. [2009]; Bernheim et al. [2015]; Chesterley [2017]; Goldin and Reck [2022], and others. In this example, the *intrinsic optimum* $s^* \equiv \arg\max_s u(s, \theta^A)$ is assumed to be known to the social planner.[25] We begin there, and then consider the case where the intrinsic optimum is not known, which makes the planner's normative objective equivalent to aggregate welfare in Bernheim et al. [2015] and social welfare in Goldin and Reck [2022]. Note that aggregation over potential intrinsic optima is interpreted in these papers as arising due to unobservable interpersonal heterogeneity rather than intrapersonal concerns.

Our illustrations of this model are simulations built on the assumption that the choice variable is one-dimensional, $s \in \mathbb{R}$. We suppose utility is approximately quadratic: $u(s, d, \theta^A) = -\frac{\alpha}{2}(s - s^*)^2$ for a known parameter $\alpha > 0$ and the intrinsic optimum s^*. For simplicity, we further assume the opt-out cost γ is known and s^* is either known to equal 0 or it follows a Gaussian distribution centered around 0. This pins down the shape of the welfare function, which we illustrate in Figure 1.

Figure 1A depicts welfare as a function of the default (under BIC), given a known intrinsic optimum. We plot welfare for varying weights on the frame θ^D, $\psi(\theta^D)$ from 0 to 1. This case turns out to be degenerate: the intrinsic optimum s^* is the unique globally robust optimum. This clearly stems from the assumption that s^* is known; there is no informational advantage to letting the individual choose for themselves.

Relaxing the assumption that the planner knows s^*, we find the following characterization of robustness in the quadratic/Gaussian case:

Proposition 6. ***Robust Optimal Defaults when the Intrinsic Optimum is Unknown***

- *The ψ-optima are the expected intrinsic optimum and the most extreme default possible in the positive or negative direction (henceforth the* extremum default*).*
- *None of the ψ-optima are globally robust.*
- *If the expected intrinsic optimum is ψ-optimal for some ψ in the interior of Ψ, the expected intrinsic optimum is the unique (locally) robust optimum. If not, the extremum default is the unique (locally) robust optimum.*

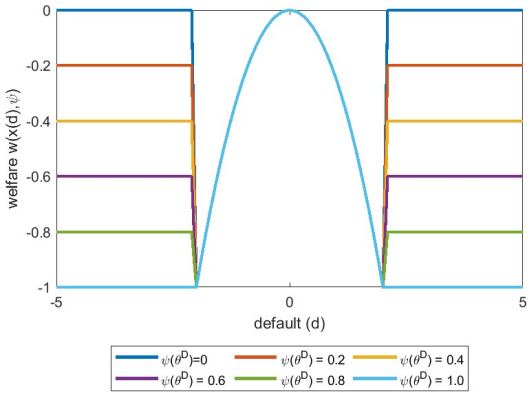
Corollary 6.1. ***Robust Control and the Optimal Default.*** *Suppose $\Psi = B(\psi, \kappa)$ for some $\kappa > 0$ and $\psi \in \Delta(\Theta)$.*

- *If the expected intrinsic optimum is ψ-optimal, then it is a robust optimum for any κ.*

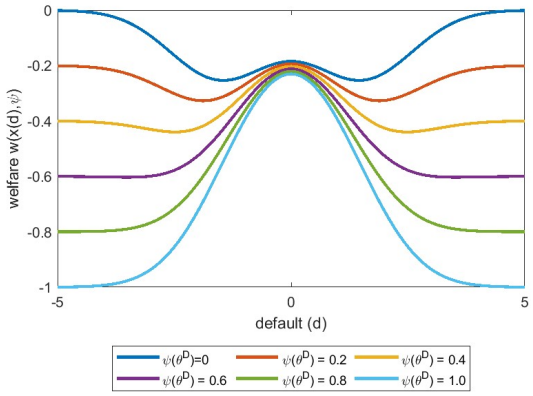
[25] We suppress the default in $u(s, d, \theta^A)$ because utility is constant over d in frame θ^A. The intrinsic optimum s^* is called the "ideal option" in earlier work on defaults [Bernheim et al., 2015; Goldin and Reck, 2022] and the analogous option is called the "intrinsic optimum" in the reference dependence literature [Kőszegi and Rabin, 2006; Reck and Seibold, 2023]. In both cases, s^* does not depend on the default/reference point by construction; it obviously does depend on other aspects of the choice situation, e.g. prices. We suppose s^* is unique for simplicity.

Figure 1: Illustration of the Optimal Default

A. Known Intrinsic Optimum



B. Unknown Intrinsic Optimum



- *If the extremum default is ψ-optimal, then there is a threshold $\bar{\kappa}$ such that it is the unique robust optimum only for $\kappa < \bar{\kappa}$, otherwise the expected intrinsic optimum is the unique robust optimum. $\bar{\kappa}$ is **decreasing** in γ.*

The logic of the proof is illustrated by Figure 1B.[26] When the intrinsic optimum is unknown, the default that maximizes welfare depends on the normative judgment about whether and to what extent the opt-out cost implied by revealed preferences, γ, is normative. The welfare-maximizing default is the expected intrinsic optimum when $\psi(\theta^D)$ is sufficiently large, while the extremum default maximizes welfare when $\psi(\theta^D)$ is sufficiently small. As such, a global robustness criterion like that of Bernheim and Rangel [2009] is inapplicable.

Even so, there is still a sense in which setting the expected intrinsic optimum as the default (i.e. minimizing opt-outs) is a more robust policy recommendation than an extremum default: it is less risky. As we can see in Figure 1B, the expected intrinsic optimum remains a local optimum as we vary normative weights, while the active choice policy becomes strictly worse when we put more normative weight on opt-out costs (because making an active choice requires incurring these costs).

Cautionary remark on comparability The intuition that extremum defaults can be optimal but that they tend not to be robust optima appears in Goldin and Reck [2022]; here we find that formalizing an approach to robustness allows us to capture that intuition.[27] However, understanding the normative foundations of our welfarist criteria also allows us to elucidate the limitations of this intuition. Specifically, throughout this example we have assumed that the default effect arises because of a (potentially irrelevant) *cost of opting out*. Suppose that instead,

[26] Bernheim et al. [2015] and Goldin and Reck [2022] present a similar figure for a fixed opt-out cost model tailored toward their applications to 401(k) contribution defaults.

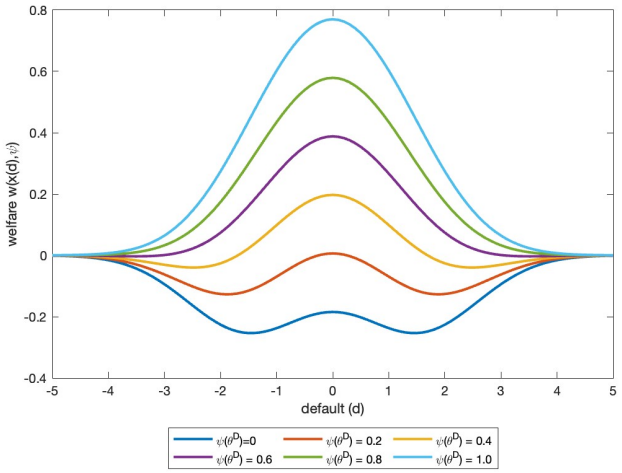
[27] Our proof of this result leverages the simplifying structure of our simulations, but the result generalizes. For less restrictive treatments of the optimal defaults problem, refer to Goldin and Reck [2022]; Bernheim et al. [2015]; Bernheim and Mueller-Gastell [2024].

we modeled default effects as the result of an *opt-in benefit* $\gamma > 0$ with identical behavioral predictions, i.e.

$$u(s, d, \theta^D) = u(s, d, \theta^A) + \gamma \mathbb{1}\{s = d\} \tag{26}$$

Figure 2 illustrates the analogue of Figure 1B when we replace the opt-out cost with an opt-in benefit all else equal. The ψ-optimal defaults are identical to the opt-out costs case, but robust optimality plays out in the opposite fashion: the extremum default is globally robust.

Figure 2: Optimal Defaults under Opt-In Benefits with an Unknown Intrinsic Optimum



Intuitively, with an opt-in benefit that may or may not be normative, disagreements about welfare occur when individuals opt-in rather than when they opt-out. The latter assumption seems more appropriate given the mechanisms behind default effects [e.g. Blumenstock et al., 2018]. For optimal default policy per se, we view this as mostly an academic point. However, researchers who wish to apply our notions of robust optimality in other settings should approach comparability cautiously.[28]

6.2.2 (Unconstrained) Optimal Reference Points

In this example, we employ a two-dimensional version of Example 1.2 and introduce some additional simplifying structure to derive a reduced-form representation of the planner's normative objective.

Suppose options are two-dimensional $s = (s_1, s_2)$, and that intrinsic utility is quasi-linear with x_2 the numeraire. The individual faces a budget constraint for given prices and income with price of p_2 normalized to 1 and $p_1 = p$. The form of loss utility follows from equation (19).

$$u(x_1, x_2, \theta^A) = \log(x_1) + x_2.$$

$$px_1 + x_2 = Z.$$

[28] More subtly, this finding looks to us like another good reason for bringing the cost itself into our definition of the options/outcomes, see also Footnote 5. Implicitly, endogenizing these costs is exactly what we would need to do to model the tests of mechanisms of default effects in Blumenstock et al. [2018].

$$V(x_1, x_2, r_1, r_2) = -\mathbb{1}\{x_1 \leq r_1\}\Lambda_1[\log(r_1) - \log(x_1)]^\beta - \mathbb{1}\{x_2 \leq r_2\}\Lambda_2[r_2 - x_2]^\beta \tag{27}$$

We examine whether planners might wish to induce different reference points. Evidence suggests policy reforms can shift reference points [e.g. Homonoff, 2018; Rees-Jones, 2018; Seibold, 2021]. We consider an environment where the planner can set the reference point anywhere on the budget constraint: $\mathcal{P} = \{(r_x, r_y)|pr_x + r_y = Z\}$.[29] Though this probably grants unrealistic power to the planner, it yields thought-provoking results when we consider robustness.

The model admits a reduced-form representation in terms of s_1. Assuming an interior solution, for fixed p and Z, we can reduce intrinsic utility and loss payoffs as:

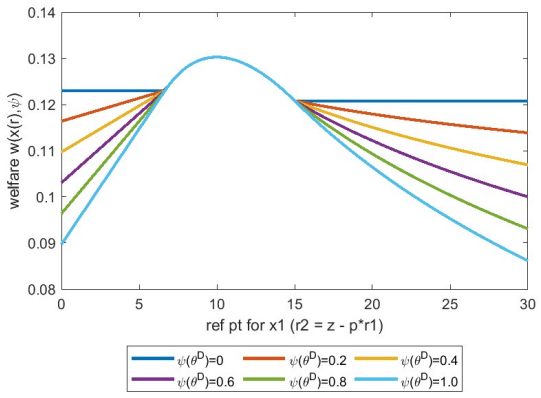
$$u(s_1, \theta^A) = \log(s_1) + Z - ps_1.$$

$$V(s_1, r_1) = \begin{cases} -\Lambda_1[\log(r_1) - \log(s_1)]^\beta, & d_1 \leq r_1 \\ -\Lambda_2[ps_1 - pr_1]^\beta, & s_1 > r_1. \end{cases} \tag{28}$$

The intrinsic optimum is $s_1^* = \frac{1}{p}$. For simulations, we set $p = 0.1 \implies s_1^* = 10$, $\Lambda_1 = \Lambda_2 = 0.5$, $Z = 10$. We express welfare in equivalent variation units relative to $s_1 = r_1$, normalized as a share of income.[30]

Figure 3: Illustration of Optimal Reference Points

A. Without Diminishing Sensitivity ($\beta = 1$)



B. With Diminishing Sensitivty ($\beta = 0.5$)

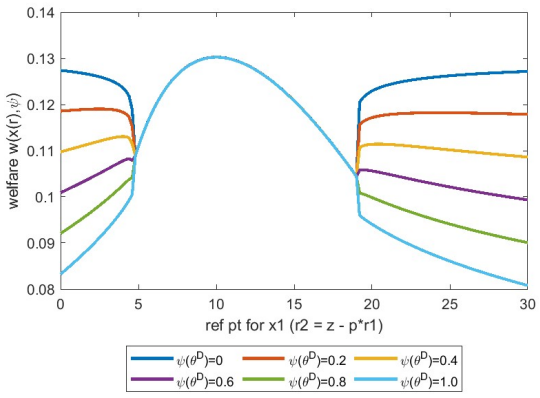


Figure 3 plots welfare as a function of the reference point for s_1. Without diminishing sensitivity ($\beta = 1$), the intrinsic optimum is the globally robust optimum and unique ψ-candidate for any ψ. This aligns with the Preferred Personal Equilibrium concept of Kőszegi and Rabin [2007]: in non-stochastic environments, selecting a PPE equals solving our planner's problem under $\psi(\theta^D) = 1$. For any $\psi(\theta^D)$, the intrinsic optimum is a robust reference point for the plan-

[29] If not constrained to fall on the budget constraint, the globally robust optimum sets the lowest possible reference point along each dimension; see Reck and Seibold [2023] Appendix B.

[30] We plot $\tilde{w}(x, \psi) = \frac{E_\psi[u(x,\theta)] - Z}{Z}$. Using $s = r$ as baseline assumes equal utility levels across frames when the individual chooses the reference point (Assumption 8.1).

ner, or indeed for an individual setting a reference point to maximize welfare of their future reference-dependent self [as in Fudenberg and Levine, 2006].[31]

Welfare behaves differently in three domains in Figure 3A. When the reference point is on the budget constraint, the individual can either consume the reference point itself (no gains or losses, so $V(s,r) = 0$) or accept losses in one good. Near the intrinsic optimum $s_1^* = 10$, the individual chooses the reference point, maximizing welfare. At extreme reference points, loss aversion drives consumption away from s_1^* to minimize losses. Without diminishing sensitivity, welfare is flat in these extremes when $\psi(\theta^D) = 0$ but decreases when $\psi(\theta^D) > 0$ due to the direct welfare cost of losses.

Figure 3B introduces diminishing sensitivity. While welfare behaves similarly where $s = r$, we find non-monotonicity elsewhere: extreme reference points become more desirable for small $\psi(\theta^D)$. The intuition mirrors penalty defaults: with extreme losses, individuals stop avoiding them under diminishing sensitivity. When losses receive little welfare weight ($\psi(\theta^D) \approx 0$), this is desirable; when loss aversion is normative ($\psi(\theta^D) \approx 1$), extreme losses become highly undesirable. As with extremum defaults, extreme reference points' desirability under $\psi(\theta^D) = 0$ is not robust, particularly with uncertainty about the intrinsic optimum (Proposition 6).

Similarity between Default Effects and Reference Dependence. Our notion of robustness plays out very similarly in the defaults and reference points models (compare Figure 1B and Figure 3B). We generalize the common features of this example in Appendix Proposition 10. Intuitively, setting the default or reference point at the (expected) intrinsic optimum minimizes disagreements about welfare across frames. The general proposition shows that disagreement-minimization is a sufficient condition for a ψ-optimum to be robust.

6.2.3 Corrective Taxation

Next we describe optimal corrective taxation in the slightly more general environment of Example 1 (biases versus strange preferences). Suppose for simplicity that we are in the quasi-linear environment from the previous example, and introduce a nonlinear tax on good 1 according to a tax schedule $T(x_1)$, which is fully incident on consumers. Utility under the alternative/classical preferences frame θ^A is expressed as

$$u(x_1, \theta^A) = \mu(x_1) + Z - px_1 - T(x_1) + R$$

where the sub-utility function $\mu(x_1)$ is twice differentiable, increasing, and concave. The variable R is rebated revenue from the corrective tax, which is determined by the simple government budget constraint $R = T(x)$. Suppose further that the tax is unrelated to the behavioral friction, so $V(x)$ is invariant to T; this rules out misperception of tax incentives.[32]

Expressing the disagreement $V(x) = u(x, \theta^D) - u(x, \theta^A)$ as a function of x_1 (leveraging the budget constraint as in the previous example), and assuming paternalistic risk neutrality, we

[31] See Reck and Seibold [2023] for a similar discussion.

[32] This is a common assumption in prior work on corrective taxation, but there are many proposed theories of tax misperception in which the assumption is obviously violated. Integrating the theory of corrective taxes for internalities with the theory of tax misperceptions is beyond the scope of this paper.

write

$$w(x) = u(x_1, \theta^A) + \psi(\theta^D)V(x_1)$$

Following the same logic as Mullainathan et al. [2012], the ψ-optimal corrective tax schedule is

$$T^*(x;\psi) = [1 - \psi(\theta^D)]V(x) + C, \tag{29}$$

where C is a constant pinned down by the government budget constraint. This can be understood by taking a derivative with respect to x_1. Where T^* is differentiable with respect to x_1, we find

$$\frac{\partial T(x_1;\psi)}{\partial x_1} = [1 - \psi(\theta^D)]\frac{dV(x_1)}{dx_1},$$

equating the marginal tax rate with the expected marginal internality.[33] What about robust optima? For a given amount of x_1 chosen by the individual (under BIC in the frame θ^D), we note that welfare is increasing in $\psi(\theta^D)$ if $V(x_1) > 0$ and decreasing if $V(x_1) < 0$. From this observation we can prove the following:

Proposition 7. *Let $\underline{\psi} \equiv \min_{\psi\in\Psi} \psi(\theta^D)$, and let $\overline{\psi} \equiv \max_{\psi\in\Psi} \psi(\theta^D)$. The robust optimal marginal tax rate given Ψ is*

$$\frac{dT^*(x_1)}{dx_1} = \begin{cases} [1 - \underline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) > 0 \\ [1 - \overline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) < 0 \\ 0 & V(x_1) = 0. \end{cases} \tag{30}$$

The intuition is as follows: the ambiguity averse planner wishes to set a tax rate that is optimal in the worst-case scenario for normative preferences. When $V(x_1) > 0$ at some chosen x_1, by construction $u(x_1, \theta^D) > u(x_1, \theta^A)$, so the worst-case scenario places maximal weight on the "biases" frame θ^A and minimal weight on the "strange preferences" frame θ^D. When $V(x_1) < 0$, we find the opposite.

The robust optimal corrective tax therefore depends on whether the level of utility is higher in the biases or strange preferences case. For instance, if we consider the corrective taxation of addictive goods [Gruber and Köszegi, 2001; O'Donoghue and Rabin, 2006; Allcott et al., 2019], ambiguity may arise over the question of whether the individual is rationally or harmfully addicted. In a rational frame, the optimal corrective tax is zero because addiction is not a bias, while in the harmful addiction frame, the optimal corrective tax mitigates over-consumption of the addictive good. If we assume that the level of utility is lower when the individual is harmfully addicted, which seems intuitive, then Proposition 7 suggests the optimal corrective tax will be shaded toward the harmfully addicted frame. In this case, valuing robustness would lead the planner to select a *more* paternalistic policy, in contrast to the two previous examples.

[33] To prove that this describes the optimal tax, observe that with this schedule, the individual's choice of x_1 optimizes the planner's expectation for their welfare over x_1:

$$u(x_1, \theta^D) = \mu(x_1) + Z - px_1 - T(x_1) + R + V(x_1) = u(x_1, \theta^A) + \psi(\theta^D)V(x_1) = w(x_1).$$

See also Equation (45) in the Appendix.

6.2.4 Nudges

Perhaps the most well-known example of behavioral public policy is the use of nudges [Thaler and Sunstein, 2009]—interventions designed to influence behavior without restricting choice or imposing significant costs. In this section, we examine the welfare effect of a nudge under normative ambiguity. We build on a simplified version of the framework of Allcott et al. [2025], seeking to evaluate a nudge intended to "debias" behavior.[34] As a motivating example, we consider the case of a potentially controversial nudge; graphic warning labels on cigarette packets.

Setup. An individual decides whether to buy a product (a cigarette packet). Their value is $v \sim F$, unknown to the government, the price is p and utility is quasi-linear. The individual has bias $\gamma \geq 0$ which shifts demand but not welfare. The policy instrument is a nudge with intensity $\sigma \geq 0$, which *reduces* demand does not affect welfare.[35] We also simplify the setup by assuming exogenous producer prices. Then:

$$\text{Decision Utility: } u(x, \theta^D) = \mathbb{1}\{x = \text{``buy''}\}\{v + \underbrace{\gamma}_{\text{``Bias''}} - \underbrace{\sigma}_{\text{Nudge}} - p\}$$

Demand is $D = \mathbb{P}[v \geq p - \gamma + \sigma]$ with the associated price-response $D'_p = -f(p - \gamma + \sigma)$.

Welfare and Normative Ambiguity: We suppose the planner is uncertain 1) whether the "bias" γ is really a bias or a normatively relevant concern, and 2) whether the nudge operates costlessly via a framing effect, or whether it imposes normatively relevant costs. This generates four potentially normative frames, including decision utility above and the following:

$$\begin{aligned} \text{Classical Utility: } u(\text{buy}, \theta^C) &= v - p \\ \text{Psychic-Cost Utility: } u(\text{buy}, \theta^P) &= v - \sigma - p \\ \text{Unbiased Utility: } u(\text{buy}, \theta^U) &= v + \gamma - p \end{aligned}$$

With probabilistic beliefs about which frame is normative, given linearity, we can express the planner's objective using two marginal probabilities: let $\psi_1 = \mathbb{P}[\gamma \text{ is normative}]$ and $\psi_2 = \mathbb{P}[\sigma \text{ is normative}]$. Then, assuming the (money-metric) utility functions above are comparable across frames in level and unit, under probabilistic beliefs the planner has preferences:

$$w(x) = v + \psi_1 \gamma - \psi_2 \sigma - p \tag{31}$$

We assume that, like intrinsic optimum in prior examples, v is unknown to the government, which makes the problem continuous and allows us to draw demand curves, and we aggregate

[34] The simplifications are that we consider homogeneous bias and no taxes, so as to focus specifically on the analysis of nudges under normative ambiguity in an otherwise simple environment. We incorporate taxes and consider the jointly optimal tax and nudge policy in Appendix C.2 for completeness.

[35] $\gamma \geq 0$ is without loss and consequently the nudge reducing demand is the only non-degenerate case; if the nudge increases demand then clearly the optimal nudge is $\sigma \equiv 0$.

welfare over values of v:[36]

$$\begin{aligned} W &= \mathbb{E}[\text{Consumer Surplus}] = \mathbb{E}[w(\text{buy}) \cdot \mathbb{1}\{\text{buy}\}] \\ &= \int_{v \geq p-\gamma+\sigma} v + \psi_1\gamma - \psi_2\sigma - p \; dF \end{aligned} \tag{32}$$

$\psi-$optima. To build intuition, we start from the case of $\psi_2 = 0$ - e.g. the psychic costs of graphic warning labels are not normatively relevant.

If, additionally $\psi_1 = 0$ (demand for cigarettes is biased) , then $\sigma^* = \gamma$: the nudge perfectly offsets the bias. Relative to this benchmark, normative ambiguity about the initial bias (ψ_1) reduces the optimal nudge intensity, and in the extreme cases where $\psi_1 = 1$ it is optimal not to use a nudge at all. This is because the optimal nudge corrects expected bias; $\sigma = (1-\psi_1)\cdot\gamma$ solves $\frac{dW}{d\sigma}\big|_{\psi_2=0} = 0$.[37]

Now suppose $\psi_2 > 0$. Figure 4 illustrates the different effects on welfare when the government increases the nudge. We illustrate $\frac{dW}{d\sigma}$ under normative ambiguity for small baseline σ in Panel A and for large baseline σ in Panel B. There are three effects on welfare when σ increases.

Internality (Over-)Correction Just as in the case when $\psi_2 = 0$, when $\sigma < (1-\psi_1)\gamma$ at baseline, the individual still overconsumes cigarettes: the nudge has not fully corrected $\mathbb{E}[\text{bias}]$. Therefore, increasing the nudge has an internality correction term equivalent to the reduction in DWL: this is the green trapezoid (1) in Panel 4A.

$$(1) = D'_p\{\sigma - (1-\psi_1)\gamma\} \gtrless 0$$

This is the standard internality correction expression and is increasing in demand sensitivity D'_p and resulting net internality $\sigma - (1-\psi_1)\gamma$.

Analogously, if $\sigma > (1-\psi_1)\gamma$, the government has gone too far. Therefore, increasing σ incurs a welfare cost from increasing DWL (orange trapezoid (1) in Panel 4B). The expression for this term is the same as above, except it is negative and represents an internality over-correction.

Psychic Costs The nudge imposes psychic costs when $\psi_2 > 0$. Therefore, increasing σ imposes a first order cost on inframarginals (the direct effect): this is purple rectangle (2).

$$(2) = -\psi_2 \cdot D < 0$$

Psychic costs are increasing in the share of inframarginals D, and the likelihood they are normatively relevant ψ_2.

Self-Correction At the same time, since $\psi_2 > 0$, the individual internalizes some of the effect of the nudge in their own welfare. This means that the nudge is "self-correcting": increasing σ

[36] Analogous to the case of interpersonal heterogeneity, we will refer to "inframarginals" as $\{v|v \geq p-\gamma+\sigma\}$ which is shorthand for the states of the world in which the individual is inframarginal.

[37] Figure 5 shows an illustration of $\frac{dW}{d\sigma}$ when $\psi_2 = 0$. Increasing the nudge σ changes the area of the DWL triangle due to the expected internality $(1-\psi_1)\gamma$. If $\sigma < (1-\psi_1)\gamma$ then increasing σ will reduce the DWL. If $\sigma > (1-\psi_1)\gamma$) then increasing σ increases DWL; the nudge has gone too far. At the optimum, $\sigma = (1-\psi_1)\gamma$.

reduces demand and hence the psychic cost. This is captured by cyan rectangle (3).

$$(3) = -\psi_2 \cdot D'_p \cdot \sigma > 0$$

The self-correction benefit is increasing in demand sensitivity D'_p, nudge size σ and the likelihood the individual internalizes the costs ψ_2.

Observation 4. *Overall,*

$$\frac{dW}{d\sigma} = \underbrace{D'_p \{\sigma - (1-\psi_1)\gamma\}}_{\textit{Internality Correction}} - \underbrace{\psi_2 \cdot \sigma \cdot D'_p}_{\textit{Self-Correction}} - \underbrace{\psi_2 D}_{\textit{Psychic Costs}} \tag{33}$$

Supposing the nudge-elasticity of demand ($\varepsilon^D_\sigma = \frac{\partial D}{\partial \sigma} \cdot \frac{\sigma}{D}$) is a fixed constant yields a formula for the optimal nudge:[38]

$$\sigma^* = \underbrace{\gamma}_{\textit{Nudge when } \psi_1, \psi_2 = 0} \cdot \underbrace{\frac{(1-\psi_1) \cdot \varepsilon^D_\sigma}{\psi_2 + (1-\psi_2) \cdot \varepsilon^D_\sigma}}_{\textit{Normative ambiguity adjustment}} \tag{34}$$

The optimal nudge is smaller than the case where θ^C is known to be normative ($\psi_1 = \psi_2 = 0$) iff:[39]

$$\overbrace{(\psi_2 - \psi_1) \cdot |D'_p| \cdot \gamma}^{\textit{Net Correction}} < \overbrace{\psi_2 \cdot D}^{\textit{Psychic Costs}} \tag{35}$$

In order to examine when the inequality Equation (35) holds, we discuss the LHS and RHS separately. The RHS is exactly the psychic costs. When psychic costs are large, the government will prefer to reduce the nudge relative to the case where both frictions are known to be biases ($\psi_1 = \psi_2 = 0$).

The LHS measures the net correction of the nudge. If ψ_1 is large (LHS small) then the nudge is likely to be over-correcting (γ not actually a bias) and the optimal nudge becomes less aggressive (relative to $\psi_1 = \psi_2 = 0$). If ψ_2 is large (LHS large), then the nudge is more self-correcting and effectively acts as an uncompensated corrective tax. The LHS scales with γ, the as-if bias and $|D'_p|$, the price response of demand: the nudge's correction only works if demand is elastic.

Overall, whether the optimal nudge is lower than the case where $\psi_1 = \psi_2 = 0$ is determined by whether the self-correction benefit outweighs the psychic costs.

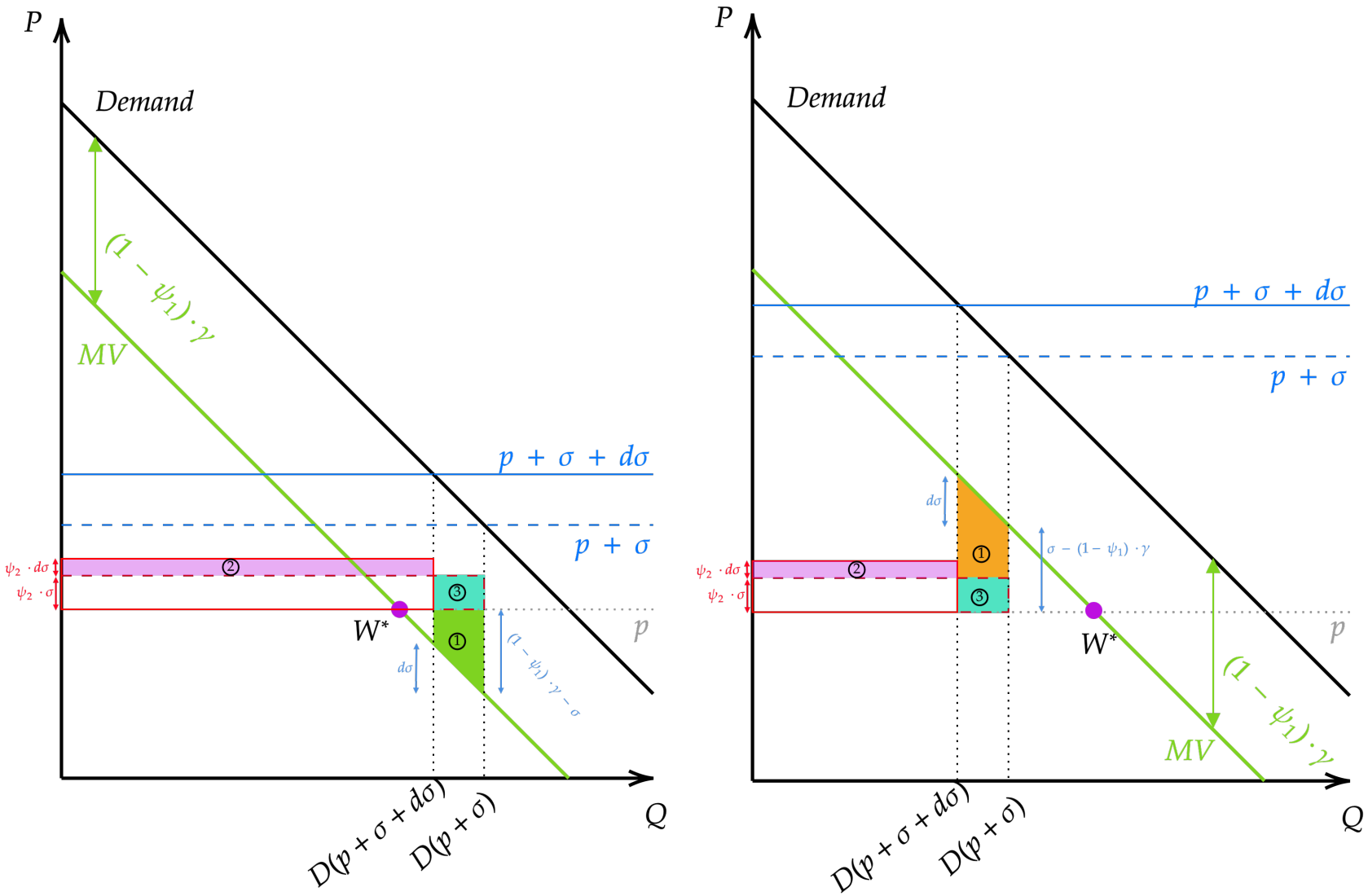
[38] Our model treats nudges exactly like prices. Importantly, we assume σ is continuous. This allows us to define the nudge-elasticity of demand. In practice, nudges might be more naturally thought of as discrete switches. In this case, the σ^* can be written as a slightly less intuitive function of the *price*-elasticity of demand ε^D_p. In our model, $\varepsilon^D_p = \frac{\sigma}{p}\varepsilon^D_\sigma$. In principle, one could estimate ε^D_σ in the example we consider - graphic labels. The exercise would involve estimating how demand responds to making the graphic label more graphic. Quantifying the % change in graphic-ness is straightforward, now, because it involves estimating the % change in *as-if* costs. These can be uncontroversially elicited through WTP experiments similar to Allcott et al. [2025]. Normative ambiguity concerns have been completely disentangled and only operate through ψ_1, ψ_2.

[39] In the full model (see Appendix C.2) with interpersonal heterogeneity and optimal taxation, an analogous inequality determines whether the government shades the nudge towards zero vs the case where $\psi_1, \psi_2 = 0$. However, in that case the LHS measures the *targeting* benefit of the nudge, rather than the average corrective benefit in the simple model.

Robust Optima. For simplicity, we consider the case of global robustness where the planner wishes to solve $\max_{\sigma} \min_{\psi_1,\psi_2} W(\psi_1, \psi_2) = \int_{v \geq p+\sigma-\gamma} v + \psi_1\gamma - \psi_2\sigma - pdF$.

Focusing on the inner minimization leads straightforwardly to $\psi_1^* = 0$ and $\psi_2^* = 1$. Using Equation (34), we know that the optimal nudge with global ambiguity is $\sigma^* = \gamma \cdot \varepsilon_\sigma^D$. Therefore, whether the globally robust nudge σ^* is larger or smaller than the optimal nudge when $\psi_1, \psi_2 = 0$ (i.e. $\sigma = \gamma$) depends on whether $\varepsilon_\sigma^D \gtrless 1$. In general, valuing robustness leads the planner to put minimal weight ψ_1, which increases the optimal nudge, and maximal weight on ψ_2, which increases the optimal nudge if and only if demand is sufficiently elastic.

Figure 4: Welfare effect of increasing the nudge when $\psi_1, \psi_2 \geq 0$



A. Small baseline nudge $\sigma < (1-\psi_1)\cdot\gamma$

B. Large baseline nudge $\sigma > (1-\psi_1)\cdot\gamma$

Notes: MV refers to the marginal value curve defined on expected welfare; $MV(p) = \mathbb{P}[v + \psi_1 \cdot \gamma \geq p]$. W^* represents the social optimum, where MV is equated with p. The dashed red rectangle represents psychic costs before the nudge increase, and solid red rectangle psychic costs after.

Comparability Our formulation assumes warning labels on purchased cigarettes produce weakly negative welfare effects, rather than welfare gains from labels on unpurchased cigarettes. As in the other examples, this assumption matters for our characterization of robust optimality, but it seems consistent with research on the psychological effects of graphic warning labels [Netemeyer et al., 2016].[40] Similarly, we assume over-consumption stems from neglected costs when buying rather than neglected benefits when abstaining. While plausible for cigarettes, these assumptions may not hold for other nudges. Carbon labels on food or energy products might generate positive benefits for non-purchasers through altruism or moral signaling.

7 Discussion: Identifying Normative Weights.

Much of the recent behavioral welfare economics literature attempts to resolve uncertainty about normative preferences—the focus of our model. While a comprehensive account of such identification methods is beyond our scope, we note intriguing similarities between identification strategies for intrapersonal and interpersonal welfare weights.

The simplest approach to pinning down ψ is assuming normative principles, such as the aggregation of interpersonal welfare effects using inverse consumption weights/log utility, or adopting the intrapersonal "long-run view" of welfare from Example 2. A more empirically informed version of this is the "Counterfactual Normative Consumer" approach [e.g. Goldin and Reck, 2020; Allcott et al., 2019; Lockwood et al., 2024], which leverages revealed preferences of "debiased" individuals or experts, assuming: 1) experts lack framing effects, 2) their preferences are observable, and 3) expertise is independent of preferences. We discuss potential failures of these assumptions and the resulting role for robustness in Appendix E. [41]

Alternatively, researchers approximate Harsanyi's [1955] "impartial observer" thought experiment via revealed preference methods. Capozza and Srinivasan [2023] estimate interpersonal welfare weights by eliciting willingness-to-pay for transfers between subjects. Allcott and Kessler [2019] implement an intrapersonal analogue, eliciting willingness-to-pay (WTP) to be nudged and assuming this WTP accounts for subjects' potentially biased future choices. Allcott et al. [2025] identify psychic costs of graphic warning labels with a similar WTP experiment.

Many studies use implemented policies to reveal welfare weights by reverse-engineering which weights would make observed policies optimal [Hendren, 2020; Lockwood and Weinzierl, 2016; Hendren and Sprung-Keyser, 2020]. This approach could apply to behavioral problems: a penalty default reveals policymakers believe default adherence is a bias ($\psi(\theta^D) \approx 0$), while illiquid savings policies reveal judgments about present focus, similar to Beshears et al. [2025].

Crucially, these methods all require untestable normative judgments. Our framework—analyzing how uncertainty about normative judgments affects optimal policy—complements tools for identifying appropriate judgments. When doubt exists about these identification approaches, our framework assesses such doubts' importance for optimal policy.

[40] Alternatively, warning labels might affect everyone regardless of purchase decisions, yielding $\frac{dW}{d\sigma} = D_p' \{\sigma - (1-\psi_1)\gamma\} - \psi_2 < 0$.

[41] The relationship between our $\psi(\theta)$ weights and frame-dependent weights in Bernheim et al. [2015] and Lockwood et al. [2024] is further clarified in Appendix D.

8 Conclusion

Our core argument is that a central problem in behavioral welfare economics–modeling welfare when individuals reveal inconsistent preferences–is the intrapersonal analogue of the interpersonal utility comparison problem. We exploit this parallel to develop criteria for welfare evaluation under uncertainty about individuals' normative preferences.

This parallel could be viewed optimistically or pessimistically, depending on one's stance on interpersonal comparisons. Pragmatically, we provide tools for applied researchers to conduct welfare analysis when facing ambiguity in revealed preferences. Our framework separates empirical quantities (magnitude of internalities, behavioral responses) from normative judgments about optimal policy, allowing exploration of how normative disagreements map to policy disagreements. The approach shares the limitations of interpersonal welfare analysis: restrictions on the set of frames/types and comparability requirements. We aim not for universal consensus but to clarify the merits and limitations of different approaches through axiomatization.

Future theoretical work could extend the framework to models of limited attention (Section 2.2) or choice-process concerns [Bernheim et al., 2024], and develop better approaches to comparability using behavioral decision theory [Ellis and Masatlioglu, 2022]. Applied work exploring what our criteria imply for optimal policy under specific behavioral frictions has only begun to scratch the surface.

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46(11):1497–1512.

Allcott, H., Cohen, D., Morrison, W., and Taubinsky, D. (2025). When do nudgesž increase welfare? Technical Report 5.

Allcott, H. and Kessler, J. B. (2019). The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Allcott, H., Lockwood, B. B., and Taubinsky, D. (2019). Regressive Sin Taxes, with an Application to the Optimal Soda Tax. *Quarterly Journal of Economics*, 23(3):1557–1626.

Allcott, H. and Taubinsky, D. (2015). Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market. *American Economic Review*, 105(8):2501–38.

Alon, S. and Gayer, G. (2016). Utilitarian preferences with multiple priors. *Econometrica*, 84(3):1181–1201.

Benkert, J.-M. and Netzer, N. (2018). Informational requirements of nudging. Technical Report 6.

Bernheim, B. D. (2009). Behavioral Welfare Economics. *Journal of the European Economic Association*, 7(2-3):267–319.

Bernheim, B. D., Fradkin, A., and Popov, I. (2015). The Welfare Economics of Default Options in 401(k) Plans. *American Economic Review*, 105(9):2798–2837.

Bernheim, B. D., Kim, K., and Taubinsky, D. (2024). Welfare and the act of choosing. Working paper, National Bureau of Economic Research.

Bernheim, B. D. and Mueller-Gastell, J. (2024). Optimal default options. Technical report.

Bernheim, B. D. and Rangel, A. (2009). Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *Quarterly Journal of Economics*, 124(1):51–104.

Bernheim, B. D. and Taubinsky, D. (2018). Behavioral Public Economics. In *Handbook of Behavioral Economics: Applications and Foundations*, volume 1, pages 381–516. Elsevier.

Beshears, J., Choi, J. J., Clayton, C., Harris, C., Laibson, D., and Madrian, B. C. (2025). Optimal illiquidity. Technical report.

Blumenstock, J., Callen, M., and Ghani, T. (2018). Why do defaults affect behavior? experimental evidence from afghanistan. Technical Report 10.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *Quarterly Journal of Economics*, 127(3):1243–1285.

Bronchetti, E., Kessler, J., Magenheim, E., Taubinsky, D., and Zwick, E. (2023). Is attention produced optimally? *Econometrica*, 91(2):669–707.

Brot-Goldberg, Z., Layton, T., Vabson, B., and Wang, A. Y. (2023). The behavioral foundations of default effects: Theory and evidence from medicare part d. *American Economic Review*, 113(10):2718–2758.

Caliendo, F. N. and Findley, T. S. (2019). Commitment and welfare. *Journal of Economic Behavior & Organization*, 159:210–234.

Capozza, F. and Srinivasan, K. (2023). Who should get money? estimating welfare weights in the us. Working paper.

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2009). Optimal Defaults and Active Decisions. *Quarterly Journal of Economics*, 124(4):1639–1674.

Chateauneuf, A., Eichberger, J., and Grant, S. (2007). Choice under uncertainty with the best and worst in mind: Neo-additive capacities. *Journal of Economic Theory*, 137(1):538–567.

Chesterley, N. (2017). Defaults, Decision Costs and Welfare in Behavioural Policy Design. *Economica*, 84(333):16–33.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and Taxation: Theory and Evidence. *American Economic Review*, 99(4):1145–1177.

Crawford, V. P. and Meng, J. (2011). New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income. *American Economic Review*, 101(5):1912–32.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

d'Aspremont, C. and Gevers, L. (1977). Equity and the informational basis of collective choice. *Review of Economic Studies*, 44(2):199–209.

Debreu, G. (1959). Topological methods in cardinal utility theory. In Arrow, K. J., Karlin, S., and Suppes, P., editors, *Mathematical Methods in Social Sciences*, pages 16–26. Stanford University Press.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-Dependent Job Search: Evidence from Hungary. *Quarterly Journal of Economics*, 132(4):1969–2018.

DellaVigna, S. and Malmendier, U. (2006). Paying Not to Go to the Gym. *American Economic Review*, 96(3):694–719.

Dietrich, F. and Jabarian, B. (2022). Decision under normative uncertainty. *Economics & Philosophy*, 38(3):372–394.

Ellis, A. and Masatlioglu, Y. (2022). Choice with endogenous categorization. *The Review of Economic Studies*, 89(1):240–278.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, 75(4):643–669.

Fleurbaey, M. and Maniquet, F. (2011). *A Theory of Fairness and Social welfare*, volume 48. Cambridge University Press.

Fudenberg, D. and Levine, D. K. (2006). A Dual-Self Model of Impulse Control. *American Economic Review*, pages 1449–76.

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

Goldin, J. and Reck, D. (2020). Revealed Preference Analysis with Framing Effects. *Journal of Political Economy*, 126(7):2759–95.

Goldin, J. and Reck, D. (2022). Optimal Defaults with Normative Ambiguity. *Review of Eco-*

nomics and Statistics, 104(1):17–33.

Gruber, J. and Köszegi, B. (2001). Is addiction "rational"? theory and evidence. *Quarterly Journal of Economics*, 116(4):1261–1303.

Hammond, P. J. (1976). Equity, arrow's conditions, and rawls' difference principle. *Econometrica*, pages 793–804.

Handel, B. and Schwartzstein, J. (2018). Frictions or mental gaps: what's behind the information we (don't) use and when do we care? *Journal of Economic Perspectives*, 32(1):155–178.

Handel, B. R. (2013). Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts. *American Economic Review*, 103(7):2643–2682.

Hansen, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*, 91(2):60–66.

Hansen, L. P. and Sargent, T. J. (2008). *Robustness*. Princeton university press.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.

Hendren, N. (2020). Measuring economic efficiency using inverse-optimum weights. *Journal of Public Economics*, 187:104198.

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *Quarterly Journal of Economics*, 135(3):1209–1318.

Homonoff, T. A. (2018). Can Small Incentives Have Large Effects? The Impact of Taxes Versus Bonuses on Disposable Bag Use. *American Economic Journal: Economic Policy*, 10(4):177–210.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2):263–92.

Kaplow, L. and Shavell, S. (2001). Any non-welfarist method of policy assessment violates the pareto principle. *Journal of Political Economy*, 109(2):281–286.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.

Köbberling, V. and Wakker, P. P. (2003). Preference foundations for nonexpected utility: A generalized and simplified technique. *Mathematics of Operations Research*, 28(3):395–423.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. *Quarterly Journal of Economics*, 121(4):1133–65.

Kőszegi, B. and Rabin, M. (2007). Reference-Dependent Risk Attitudes. *American Economic Review*, 97(4):1047–73.

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

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, pages 443–477.

Laibson, D. I., Repetto, A., Tobacman, J., Hall, R. E., Gale, W. G., and Akerlof, G. A. (1998). Self-control and saving for retirement. *Brookings Papers on Economic Activity*, 1998(1):91–196.

Lockwood, B. B., Allcott, H., Taubinsky, D., and Sial, A. (2024). What drives demand for state-run lotteries? evidence and welfare implications. Technical report.

Lockwood, B. B., Sial, A., and Weinzierl, M. (2021). Designing, not checking, for policy robustness: An example with optimal taxation. *Tax Policy and the Economy*, 35:1–54.

Lockwood, B. B. and Weinzierl, M. (2016). Positive and normative judgments implicit in us tax policy, and the costs of unequal growth and recessions. *Journal of Monetary Economics*, 77:30–47.

Loewenstein, G. and O'Donoghue, T. (2006). "We Can Do This the Easy Way or the Hard Way": Negative Emotions, Self-Regulation, and the Law. *University of Chicago Law Review*, 73(1):183–206.

MacAskill, M., Bykvist, K., and Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed Attention. *American Economic Review*, 102(5):2183–2205.

Maskin, E. (1978). A theorem on utilitarianism. *The Review of Economic Studies*, 45(1):93–96.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

Mongin, P. and Pivato, M. (2021). Rawls's difference principle and maximin rule of allocation: a new analysis. *Economic Theory*, 71(4):1499–1525.

Montiel Olea, J. L. and Strzalecki, T. (2014). Axiomatization and measurement of quasi-hyperbolic discounting. *Quarterly Journal of Economics*, 129(3):1449–1499.

Mullainathan, S., Schwartzstein, J., and Congdon, W. J. (2012). A Reduced-Form Approach to Behavioral Public Finance. *Annual Review of Economics*, 4:511–540.

Netemeyer, R. G., Burton, S., Andrews, J. C., and Kees, J. (2016). Graphic health warnings on cigarette packages: the role of emotions in affecting adolescent smoking consideration and secondhand smoke beliefs. *Journal of Public Policy & Marketing*, 35(1):124–143.

O'Donoghue, T. and Rabin, M. (2006). Optimal Sin Taxes. *Journal of Public Economics*, 90(10):1825–1849.

O'Donoghue, T. and Sprenger, C. (2018). Reference-Dependent Preferences. In *Handbook of Behavioral Economics: Applications and Foundations*, volume 1, pages 1–77. Elsevier.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Rafkin, C. and Soltas, E. (2024). Eviction as bargaining failure: Hostility and misperceptions in the rental housing market. Working paper.

Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

Reck, D. and Seibold, A. (2023). The welfare economics of reference dependence. Nber working paper no. 31381.

Rees-Jones, A. (2018). Quantifying Loss-Averse Tax Manipulation. *Review of Economic Studies*, 85(2):1251–78.

Rees-Jones, A. and Taubinsky, D. (2018). Taxing humans: Pitfalls of the mechanism design approach and potential resolutions. *Tax Policy and the Economy*, 32(1):107–133.

Saez, E. and Stantcheva, S. (2016). Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review*, 106(1):24–45.

Seibold, A. (2021). Reference Points for Retirement Behavior: Evidence from German Pension Discontinuities. *American Economic Review*, 111(4):1126–65.

Sen, A. (1976). Welfare inequalities and rawlsian axiomatics. *Theory and Decision*, 7(4):243–262.

Sen, A. (1986). Social choice theory. *Handbook of Mathematical Economics*, 3:1073–1181.

Sen, A. (1997). *On economic inequality*. Oxford university press.

Sen, A. K. (1970). *Collective Choice and Social Welfare*. Holden-Day.

Sher, I. (2024). Generalized social marginal welfare weights imply inconsistent comparisons of tax policies. Technical Report 11.

Sugden, R. (2004). The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review*, 94(4):1014–1033.

Thakral, N. and Tô, L. T. (2021). Daily Labor Supply and Adaptive Reference Points. *American Economic Review*, 111(8):2417–43.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics*, 106(4):1039–61.

Von Neumann, J. and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press.

Wakker, P. (1984). Cardinal coordinate independence for expected utility. *Journal of Mathematical Psychology*, 28(1):110–117.

Wakker, P. P. and Zank, H. (1999). A unified derivation of classical subjective expected utility models through cardinal utility. *Journal of Mathematical Economics*, 32(1):1–19.

Weymark, J. A. (1991). A reconsideration of the harsany-sen debate on utilitarianism. In Elster, J. and Roemer, J. E., editors, *Interpersonal Comparisons of Well-Being*, pages 255–320. Cambridge University Press.

Online Appendix

This Online Appendix contains additional analysis, including a perturbation approach to welfare analysis in our framework, technical extensions (continuous sets of frames and "non-rectangular" choice environments), a review of the relationship of our framework to the "counterfactual normative consumer" approach to welfare analysis, and proofs of all results.

A Theory Appendix

In this section we present some assumptions and results discussed informally in Section 3 of the the main text .

A.1 Expected Utility Representation of Planner's Preferences Using Von Neumann & Morgenstern's Theory

Re-defining the Planner's Objective. Adapting von Neumann and Morgenstern's Expected Utility Theory to this setting requires us to conceive of counterfactuals that describe situations in which the planner attaches different weights to each frame. We introduce the notion of an intrapersonal lottery to capture this. The primitive components of such a lottery are an option $x \in \mathcal{X}$, the state space Θ, and a distribution $\psi \in \Delta(\Theta)$. The outcomes of a lottery entail consuming a particular x in a particular state θ. We conceive of a lottery $L(x, \psi)$ in terms of a vector of weights/probabilities $\big(\psi(\theta_1), ..., \psi(\theta_{|\Theta|})\big)$ and a vector of outcomes $\big((x, \theta_1), ..., (x, \theta_{|\Theta|})\big)$. Compound lotteries entail mixtures of weights: for $p \in [0, 1]$ and two distributions ψ_1, ψ_2 we describe these using the notation

$$pL_1(x) + (1-p)L_2(x) = L(x, p\psi_1 + (1-p)\psi_2),$$

where $L_n(x) = L(x, \psi_n)$.

We abuse notation slightly by denoting the planner's preferences over lotteries by $\succeq_w$. Now, the planner's preferences are defined not only over what option the individual consumes but also over the planner's normative beliefs: $\succeq_w$ is a binary relation on the set of lotteries $\mathcal{L}$. We strengthen Assumption 5 as follows:

Assumption 9. ***Expected Utility Assumptions Over Intrapersonal Lotteries.***

Assumption 9.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 9.2. ***Continuity.*** *For any $L \in \mathcal{L}$, the sets $\{L' \in \mathcal{L} : L' \succsim_w L\}$ and $\{L' \in \mathcal{L} : L' \precsim_w L\}$ are closed.*

Assumption 9.3. ***Strong BR-Dominance.*** *For any ψ, x, if $x \succsim_\theta x'$ for every θ, then $L(x, \psi) \succsim_w L(x', \psi)$. If, additionally, there exists θ such that $x \succ_\theta x'$ and $\psi(\theta) > 0$, then $L(x, \psi) \succ_w L(x', \psi)$.*

Assumption 9.4. ***Independence.*** *For any x, any $L_1(x), L_2(x), L_3(x) \in \mathcal{L}$, and any $p \in [0, 1]$*

$$L_1(x) \succsim_w L_2(x) \implies pL_1(x) + (1-p)L_3(x) \succsim_w pL_2(x) + (1-p)L_3(x). \tag{36}$$

Proposition 8. *Maintain Assumptions 1, 2 and 3. Then Assumption 9 holds if and only if there is a function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $u(x, \theta)$ represents individual preferences $\succeq_\theta$ for every θ, and the planner's preferences $\succeq_w$ are represented by*

$$w(x; \psi) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta). \tag{37}$$

Moreover, u is continuous and unique up to positive affine transformation.

Proof. Assumptions 9.1, 9.2, and 9.4 are the axioms of classical expected utility over the outcomes (x, θ). We therefore obtain a payoff function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that the planner's preferences take the expected utility form $w(x, \psi) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta)$. That $u(x, \theta)$ must be a representation of $\succeq_\theta$ follows from 9.3 by the same logic as Proposition 1; the strong form of BR-dominance is required to rule out the degenerate case where u is constant over x. This establishes sufficiency of Assumption 9 for the desired representation of $\succeq_w$; necessity is easily verified. ■

Discussion of Proposition 8 The proof of this proposition is a straightforward adaptation of the Expected Utility Theorem of Von Neumann and Morgenstern [1953]. The way that the planner trades off risk according to the independence assumption 9.4 implies a cardinalization of utility, which is the function $u(x, \theta)$ in (37).

A.2 Paternalistic Risk Aversion

Although we find that some fully comparable utility function that is suitable for evaluation of intrapersonal tradeoffs must exist under the assumptions of Proposition 2 or Proposition 8, these results do not not shed light on which particular representation of frame-dependent preferences we ought to suppose is fully comparable across frames in any given setting or model. The following Corollary to these propositions, in which we consider a specific welfare metric $v(x, \theta)$ that represents ordinal preferences, proves useful for thinking about comparability and the planner's risk preferences. In the main text, the only candidate for the function v we considered was money-metric equivalent variation and we discussed paternalistic risk aversion in terms of risk preferences over monetary payoffs. Here, we take a slightly more general approach.

Definition. We say that two utility functions $u(x, \theta)$ and $v(x, \theta)$ exhibit *ordinal level comparability* if for any (x, θ) and (x', θ'),

$$u(x, \theta) \geq u(x', \theta') \iff v(x, \theta) \geq v(x', \theta').$$

Corollary 8.1. *Consider a utility function $u(x, \theta)$ that gives the representation in Proposition 2 or Proposition 8. Under the assumptions of either proposition, for any function $v(x, \theta)$ that exhibits ordinal level comparability with $u(x, \theta)$, there is a transformation $\omega : \mathbb{R} \to \mathbb{R}$ such that*

$$w(x; \psi) = \sum_{\theta \in \Theta} \psi(\theta) \omega(v(x, \theta)). \tag{38}$$

Moreover, ω is strictly increasing, continuous, and unique up to positive affine transformation.

Proof. This result obviously follows from Proposition 8.1 and the definition of ordinal level comparability. ■

The function ω converts the welfare metric $v(x,\theta)$ into the cardinal units the planner uses to conduct welfare comparisons across frames. In main text Section 4.2, we introduce conditions under which money-metric equivalent variation provides a representation of individual preferences that exhibits ordinal level comparability with cardinal utility, in which case ω should account for any non-linearity of the individual's preference for money. In the main text, we denoted the transformation ω for monetary payoffs by u_ζ.

The Value of Robustness with Probabilistic Uncertainty To understand the value of the robustness of welfare across frames in this case, it is instructive to impose some smoothness the transformation ω from Corollary 8.1.

Corollary 8.2. ***Variance Representation*** *Assume the function ω from representation (38) is twice differentiable. Then up to second-order Taylor approximation of ω, the planner's objective is*

$$w(x,\psi) \approx \omega\Big(E_\psi[v(x,\theta)]\Big) + \frac{\omega''\Big(E_\psi[v(x,\theta)]\Big)}{2} \cdot Var_\psi\Big[v(x,\theta)\Big] \tag{39}$$

where $E_\psi[v(x,\theta)] = \sum_{\theta\in\Theta} \psi(\theta) v(x,\theta)$ and $Var_\psi\Big[v(x,\theta)\Big] = \sum_{\theta\in\Theta} \psi(\theta)\Big[v(x,\theta) - E_\psi[v(x,\theta)]\Big]^2$.

Proof. For ease of notation shorten $v = v(x,\theta)$, a random variable with respect to θ for given x, and $\overline{v} = E_\psi[v(x,\theta)]$, a deterministic number for given x, ψ. Using a Taylor Expansion of ω around $\overline{v}$ we find

$$\begin{aligned} &\omega(v) \approx \omega(\overline{v}) + \omega'(\overline{v}) \cdot (v - \overline{v}) + \omega''(\overline{v}) \cdot (v-\overline{v})^2 \\ \implies &\mathbb{E}_\psi\big[\omega(v)\big] \approx \underbrace{\omega(\overline{v})}_{\text{Fixed Number}} + \omega'(\overline{v}) \cdot \underbrace{\mathbb{E}_\psi[v-\overline{v}]}_{=0} + \frac{\omega''(\overline{v})}{2} \cdot \mathbb{E}_\psi[v-\overline{v}]^2 \end{aligned}$$

The result follows, as $w(x,\psi) = \mathbb{E}_\psi\big[\omega(v(x,\theta))\big]$. ■

We say that the planner's preferences exhibit *paternalistic risk aversion over v* if $\omega'' < 0$ and *paternalistic risk neutrality over v* if $\omega'' = 0$ (this can be converted to statements about primitive preferences over normative lotteries using standard expected utility theory). Corollary 8.2 implies that under paternalistic risk aversion over a welfare metric v with probabilistic uncertainty, the planner values robustness. Under paternalistic risk neutrality over v, the planner's objective coincides with expected welfare according to v, i.e. $E_\psi[v(x,\theta)]$. But under paternalistic risk aversion over v, the variance of the welfare metric v across normative frames begins to matter (up to second-order approximation), and in particular welfare is decreasing in this variance. When, according to the welfare metric v, there is more disagreement in revealed preferences across frames about welfare under some policy P_0 compared to an alternative P_1,

and mean welfare is similar between the two, paternalistic risk aversion over v suggests that P_0 is less desirable than P_1. Unlike the notion of robustness we present in the main text for ambiguity aversion, whether this notion of robustness is relevant is specific to the welfare metric we have in mind. Proposition 8 tells us that under our assumptions, there will always be some measure of welfare $u(x,\theta)$ over which the planner's preferences exhibit paternalistic risk neutrality.

If ω is a homogeneous transformation (i.e. the planner's preferences exhibit *scale invariance* over the welfare metric v), we find a familiar functional form for ω. For a parameter $\eta \in \mathbb{R}$, we have

$$\omega(v) = \begin{cases} \frac{v^{1-\eta}}{1-\eta}, & \eta \neq 1 \\ \log(v) & \eta = 1. \end{cases} \tag{40}$$

Paternalistic risk aversion over v further implies $\eta > 0$, and η is of course the Arrow-Pratt coefficient of relative (paternalistic) risk aversion.

A.3 Formalizing Ambiguity Averse Objectives

We say that a lottery $L(x,\psi)$ is *constant over* u if for the given x, $u(x,\theta) = u(x,\theta')$ for any θ, θ', i.e. if it generates a constant payoff for every normative frame. We note that this obviously imposes comparability of u across frames. Abandoning Assumption 9.4, we introduce conditions on the planner's preferences drawn from Gilboa and Schmeidler [1989].

Assumption 10. ***Ambiguity Aversion Assumptions.***

Assumption 10.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 10.2. ***Continuity.*** *For any $L \in \mathcal{L}$, the sets $\{L' \in \mathcal{L} : L' \succsim_w L\}$ and $\{L' \in \mathcal{L} : L' \precsim_w L\}$ are closed.*

Assumption 10.3. ***Certainty Independence.*** *There is a representation $u(x,\theta)$ such that for any x, any pair $L_1(x), L_2(x) \in \mathcal{L}$, any lottery $L_3^c(x)$ that is constant over u, and any $p \in (0,1)$,*

$$L_1(x) \succsim_w L_2(x) \implies pL_1(x) + (1-p)L_3^c(x) \succsim_w pL_2(x) + (1-p)L_3^c(x).$$

Assumption 10.4. ***Weak BR-dominance.*** *For any $x, x' \in \mathcal{X}$ and any $\psi \in \Delta(\Theta)$, if $x \succsim_\theta x'$ for every θ, then $L(x,\psi) \succsim_w L(x',\psi)$.*

Assumption 10.5. ***Uncertainty Aversion.*** *For any x, any pair $L_1(x), L_2(x)$, and $p \in (0,1)$,*

$$L_1(x) \sim_w L_2(x) \implies pL_1(x) + (1-p)L_2(x) \succsim_w L_1(x).$$

Assumption 10.6. ***Non-degeneracy.*** *There exists $L, L' \in \mathcal{L}$ such that $L \succ_w L'$.*

Comments: Relative to Assumption 9, Assumption 10.3 weakens Assumption 9.4 so that it only holds when we mix a given pair of lotteries with a constant lottery. Assumption 10.6 rules out the degenerate case where the planner is indifferent across all policies/options.

Proposition 9. ***MaxMin Welfare Under Ambiguity Aversion.*** *Maintain Assumptions 1, 2 and 3. Assumption 10 holds if and only if there exist a function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ and a set $\Psi \subseteq \Delta(\Theta)$ such that $u(x, \theta)$ represents $\succsim_\theta$ for every θ, Ψ is closed and convex, and the planner's preferences $\succsim_w$ are represented by*

$$w(x) = \min_{\psi \in \Psi} \left\{ \sum_\theta \psi(\theta) u(x, \theta)) \right\}. \tag{41}$$

Proof. Take the representation of individual preferences whose existence is implied by 10.3 and denote this $\tilde{u}$. Observe that BR-dominance implies Gilboa and Schmiedler's weak monotonicity condition over realizations of $\tilde{u}(x, \theta)$ for this representation. Theorem 1 of Gilboa and Schmeidler [1989] then implies there is a strictly increasing transformation $\omega(\tilde{u})$ such that the planner's preferences are represented by $w(x) = \min_{\psi \in \Psi} \{\sum_\theta \psi(\theta) \omega(\tilde{u}(x, \theta))\}$. The result follows, as $u \equiv \omega(\tilde{u})$ is also a representation of individual preferences by construction. ■

A.3.1 Forms of Ambiguity and Related Alternative Axiomatizations

Gilboa and Schmeidler [1989] defer the structure of the set Ψ to applications, apart from the requirement that Ψ is closed and convex [for a useful discussion, see Hansen and Sargent, 2001]. Following our discussion above about interpretations of the set of frames and the weights themselves, we envision two potential approaches. The first is more global: we could define a subset of the set of frames $\Theta^* \subseteq \Theta$, and let $\Psi = \Delta(\Theta^*)$. This approach is very similar in spirit to the concept of a "welfare-relevant domain" in Bernheim and Rangel [2009], and seems suitable when the planner has no philosophically acceptable way of specifying a unique set of welfare weights. The second approach is more local and drawn from the literature on robust control [e.g. Hansen and Sargent, 2008]: the planner begins with a specific distribution ψ that represents their best guess about the correct normative weights, and accounts for ambiguity in a neighborhood of this distribution. In this case, Ψ could be a ball of distributions around the best guess ψ whose radius is determined by a tolerance parameter $\kappa \geq 0$: $\Psi = B(\psi, \kappa) \equiv \{\psi' \in \Delta(\Theta) \text{ s.t. } ||\psi' - \psi|| \leq \kappa\}$. This approach seems more applicable in the case where the planner uses a statistical model like the "counterfactual normative consumer" approach discussed below to identify welfare weights, but nevertheless confronts ambiguity because the underlying model may be misspecified. We return to this last idea in Section 7 and Appendices D and E.

Global MaxMin Criteria. In the case where normative ambiguity is more globally conceived of over the set $\Psi = \Delta(\Theta^*)$ for a subset of "welfare-relevant" frames Θ^*, the max-min expected welfare criterion from (8) becomes a more global max-min criterion:

$$\min_{\psi \in \Delta(\Theta^*)} \left\{ \sum_\theta \psi(\theta) u(x, \theta)) \right\} = \min_{\theta \in \Theta^*} u(x, \theta) \tag{42}$$

Building on social welfare theory, rather than using Assumption 10, one could derive this criterion this using an analogue of Rawls' [1971] Difference Principle: assume there is some representation of $\succsim_\theta$, u, such that $x \succsim_w x'$ if and only if the individual prefers x to x' in the frame $\theta \in \Theta^*$ in which they are the least well-off according to u. This obviously implies a criterion

like equation (42).[42] Intuitively, with this more global type of ambiguity, endowing the planner with a direct preference for equity across frames (without any notion of intrapersonal lotteries) leads to the same place as giving the planner a preference to hedge in Assumption 10.5 together with a global notion of ambiguity.[43]

All of the objectives discussed above intersect at a global robustness criterion, which obtains under extreme ambiguity, or extreme paternalistic risk aversion with probabilistic uncertainty. Formally, we define the *global max-min* criterion as the one implied by equation (8) for $\Psi = \Delta(\Theta)$. The global max-min criterion is the closest analogue to Rawlsian social welfare in our framework.

Corollary 9.1. ***Intersection of Various Objectives at Global Max-Min***

- *If* $\Psi = B(\psi, \kappa)$*, for any* ψ*, the planner's objective in* (8) *coincides with the global max-min criterion for* $\kappa > 1$*.*
- *If* $\Psi = \Delta(\Theta^*)$*, the planner's objective in* (8) *and/or* (42) *coincides with the global max-min criterion for* $\Theta^* = \Theta$*.*
- *Given a welfare metric* v *under scale invariance over* v *for the parameter* η *and probabilistic uncertainty with* $\psi(\theta) > 0$ *for very* $\theta \in \Theta$*, the planner's objective – Equation* (38) *with the functional form in equation* (40) *– approaches the global max-min criterion as* $\eta \to \infty$*.*[44]

Proof. The first three claims are obvious from equation (8) and (42). The last has a well-known analogue in the nesting of Rawlsian welfare functions in the family of generalized utilitarian welfare functions taking the form in equation (40). ■

Partial Characterization of Robust Optimality. Our next result provides a sufficient condition for a policy that is a ψ-optimum for $\psi \in \Psi$ to also be a robust optimum. Note that this is not a full characterization as we do not obtain necessity; the condition nevertheless builds intuition and proves useful in applications below. To state the condition we use the cardinal *disagreement* in welfare between some frame θ and the decision-making frame θ^D:

$$V(x, \theta, \theta^D) = u(x, \theta^D) - u(x, \theta).$$

Note that in the main text, when there is only one frame besides θ^D, we suppress the second and third inputs as these are always equal to θ^A and θ^D respectively; below, only θ^D is fixed and suppressed.

[42] Respecting strict BR-dominance requires a slight modification of the Difference Principle – the "Equity" axiom from Sen [1970] and Hammond [1976] – to break ties under indifference in the least well-off frame. Then we would obtain a lexicographic max-min criterion.

[43] For a deeper discussion of the theory of ambiguity aversion due to Gilboa and Schmeidler [1989] and Rawlsian social welfare, see Mongin and Pivato [2021].

[44] The interpersonal analogue of this is a well-known result about Rawlsian social welfare functions; see also Lockwood et al. [2021].

Proposition 10. ***Sufficient Condition for a ψ-Optimum to be a Robust Optimum.*** *Let $P^* \in \mathcal{P}$ be a $\psi-$optimum for some $\psi \in \Delta(\theta)$. Then, for any $\Psi \subseteq \Delta(\Theta)$ such that $\psi \in \Psi$, P^* is a robust optimum if*

$$P^* \in \arg\min_{P\in\mathcal{P}} \max_{\psi'\in\Psi} \sum_{\theta\in\Theta} \big(\psi'(\theta) - \psi(\theta)\big) \cdot V\big(x(P, Z, \theta^D), \theta, \theta^D\big). \tag{43}$$

Proof. By supposition,

$$\begin{aligned}
P^* &\in \arg\min_{P\in\mathcal{P}} \max_{\psi'\in\Psi} \sum_{\theta\in\Theta} \big(\psi'(\theta) - \psi(\theta)\big) \cdot V\big(x(P, Z, \theta^D), \theta\big) \\
&= \arg\min_{P\in\mathcal{P}} \left\{ \sum_{\theta\in\Theta} \psi(\theta) \cdot V\big(x(\theta^D, P), \theta, P\big) - \min_{\psi'\in\Psi} \sum_{\theta\in\Theta} \psi'(\theta) \cdot V\big(x(\theta^D, P), \theta, P\big) \right\} \\
&= \arg\min_{P\in\mathcal{P}} \Bigg\{ u(x(P, Z, \theta^D), \theta^D) - \sum_{\theta\in\Theta} \psi(\theta) \cdot V\big(x(P, Z, \theta^D), \theta\big) \\
&\qquad - \min_{\psi'\in\Psi} \Big[u(x(P, Z, \theta^D), \theta^D) - \sum_{\theta\in\Theta} \psi'(\theta) \cdot V\big(x(P, Z, \theta^D), \theta\big) \Big] \Bigg\} \\
&= \arg\min_{P\in\mathcal{P}} \Big\{ W(P, Z, \theta^D; \psi) - \min_{\psi'\in\Psi} W(P, Z, \theta^D; \psi') \Big\} \\
&\iff \forall P \in \mathcal{P}, W(P^*, Z, \theta^D; \psi) - \min_{\psi'\in\Psi} W(P^*, Z, \theta^D; \psi') \leq W(P, Z, \theta^D; \psi) - \min_{\psi'\in\Psi} W(P, Z, \theta^D; \psi')
\end{aligned}$$

However, $W(P^*, Z, \theta^D; \psi) \geq W(P, Z, \theta^D; \psi)$ as P^* is a $\psi-$optimum. We therefore obtain

$$\min_{\psi'\in\Psi} W(P^*, Z, \theta^D; \psi') - \min_{\psi'\in\Psi} W(P, Z, \theta^D; \psi') \geq W(P^*, Z, \theta^D; \psi) - W(P, Z, \theta^D; \psi) \geq 0 \tag{44}$$

So P^* is a robust optimum.

Note that above we suppressed the dependence between θ^D and P in writing out the steps of the proof above. But on inspection, we can see that each step of the proof obtains when θ^D depends non-trivially on P. ■

The condition from Proposition 10 for a given ψ-optimum to be robust is more likely to be met when disagreements about welfare evaluated at that policy are not too large and the set Ψ over which the planner evaluates robustness is a relatively close neighborhood around the relevant distribution ψ. We note that θ^D can depend arbitrarily on P in Equation (43).

B Perturbations

In this section, we explore a perturbation approach to evaluating policy reforms in our framework. We consider one-dimensional policy variation, supposing $\mathcal{P} = \mathbb{R}$ for simplicity. We assume all the derivatives necessary to apply the perturbation approach exist. Expressing the planner's welfare under BIC as $W(P, Z, \theta^D) = w(x(P, Z, \theta^D))$ we are interested in $\frac{\partial W}{\partial P}$, or equivalently the change in welfare dW that results from a marginal policy perturbation dP.

In order to illustrate how the envelope theorem plays out in many applications, we modify our setup slightly. We suppose that in reduced-form we can conceive of every option $\tilde{x} \in \mathcal{X}$ as a component the individual can choose and a fixed feature like a default: $\tilde{x} = (x, P)$. Here, we

assume there is a bijection such that every value of the fixed feature corresponds to a value of the policy P, so we might as well denote the fixed feature by P.

As the set of frames is finite, we suppose the frame θ^D is unaffected by policy; this can be relaxed with the continuous-frames extension developed in D. As θ^D is fixed throughout this section, we express disagreements as $V(x,\theta) = u(x,P,\theta^D) - u(x,P,\theta)$.

We present three characterizations, one leveraging Proposition 2, where we think of u as a utility function that is fully comparable (Paternalistic Risk Neutrality), one based on Proposition 4/Equation (12), where we think of u as a money-metric utility function that is ordinally level comparable to cardinal utility with diminishing utility over money $u''_\zeta < 0$ (Paternalistic Risk Aversion), and one rooted in Proposition 9/Equation (8) (Ambiguity Aversion) given a fully comparable utility function.

B.1 Under Risk Neutrality

We begin with the planner's objective under probabilistic uncertainty about the normative frame and risk-neutrality. Applying the envelope theorem of Milgrom and Segal [2002] under θ^D, we find[45]

$$\frac{\partial W(P,Z,\theta^D)}{\partial P} = \underbrace{E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right]}_{\text{Direct Effect}} - \underbrace{\frac{\partial x(P,Z,\theta^D)}{\partial P}}_{\text{Beh. Resp.}} \cdot \underbrace{(1-\psi(\theta^D))E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial x}\middle|\, \theta \neq \theta^D\right]}_{\text{Marginal Internality}} \tag{45}$$

The term $\frac{\partial u(x,P,\theta)}{\partial P}$ in equation (45) is the partial derivative of $u(x,P,\theta)$ with respect to its second argument – the *direct effect* of varying P. All terms are evaluated at the status quo (P,Z,θ^D), where $x = x(P,Z,\theta^D)$.

In a model in which normative preferences are known – i.e. a model with a singular alternative frame θ^A, as in Example 1, and weight $\psi(\theta^A) = 1$ – this derivation matches the reduced-form characterization of welfare in Mullainathan et al. [2012].[46] Here, we extend this characterization to accommodate an unkown normative frame. Under risk-neutrality, we find similar set of terms as Mullainathan et al. [2012], but we replace the direct effect and marginal internality under a known normative frame with their expected values under an uncertain normative frame. This focuses applied analysis of normative ambiguity on specific questions: how do frames differ in their implied direct effects and marginal internalities?

How do disagreements about welfare shape the effects in (45)? We can see the answer for the internality term in equation (45). For the direct effect term, we observe that

$$E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right] = \frac{\partial u(x,P,\theta^D)}{\partial P} - [1-\psi(\theta^D)]E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial P}\middle|\, \theta \neq \theta^D\right]. \tag{46}$$

[45] Note that P is one-dimensional by assumption here. When x is multidimensional, the second term of this expression should be regarded as a dot product of the vectors $\frac{\partial x}{\partial P}$ and $\frac{\partial V}{\partial x}$.

[46] With a known normative frame it is also not necessary to account for potential differences in the value of a dollar across frames in characterizing when a local perturbation improves welfare, so we could freely use any equivalent variation representation of preferences in the application of the perturbation approach [see also Allcott and Taubinsky, 2015].

B.2 Under Paternalistic Risk Aversion

How do disagreements in money-metric welfare matter for policy evaluation? Assuming ordinal level comparability of equivalent variation $\zeta(x, P, \theta)$ (suppressing the baseline parameters) and diminishing marginal utility of money $u''_\zeta < 0$, we find an intuitive representation that leverages the mean-variance characterization from Corollary 8.2. The variance of welfare equals the variance of the disagreement with θ^D. We express disagreements here as V_ζ to highlight these are in the units of ζ (dollars):

$$Var_\psi\Big[\zeta(x, P, \theta)\Big] = Var_\psi\Big[V_\zeta(x, P, \theta)\Big].$$

Denoting mean indirect utility at (P, Z, θ^D) by $\overline{W}_\zeta(P, Z, \theta^D; \psi) = E_\psi[\zeta(x(P, Z, \theta^D), P)]$, we find that up to second-order approximation of u_ζ,

$$\frac{\partial W(P, Z(P), \theta^D)}{\partial P} \approx u'_\zeta(\overline{W}_\zeta)\frac{\partial \overline{W}_\zeta}{\partial P} + \frac{u''_\zeta(\overline{W}_\zeta)}{2} \cdot \frac{dVar_\psi\Big[V_\zeta(x(P, Z, \theta^D), P, \theta)\Big]}{dP}. \tag{47}$$

By construction, equations (45) and (46) above characterize the effect of dP on mean welfare, $\frac{\partial \overline{W}_\zeta}{\partial P}$ in the first term. The second term then captures how disagreements matter when the planner is risk averse. The characterization is intuitive (and arguably obvious from Corollary 8.2): given $u''_\zeta < 0$ a policy reform that increases the variance of disagreements about money-metric welfare is less desirable, holding the effect on expected welfare $\overline{W}_\zeta$ fixed.

Remark: Setting Aside Money Metrics. The above characterization holds for any welfare metric under ordinal level comparability and paternalistic risk aversion, but we stated it in terms of money-metric utility to emphasize the relationship with prior work (and diminishing marginal utility over money is intuitive). We do not engage with the money-metric welfare concept going forward in this appendix. We simply assume a utility function that is comparable across frames.

B.3 Under Paternalistic Ambiguity Aversion

Now we turn to the ambiguity averse objective from Proposition 9. We let

$$\psi^*(\theta, P) \equiv \arg\min_{\psi \in \Psi} E_\psi[W(P, Z, \theta^D; \theta)].$$

Following Hansen and Sargent [2008], on develop intuition by thinking of ψ^* as being chosen by an "evil agent" who minimizes welfare given the planner's choice of policy. When ψ^* is differentiable in P, we find

$$\frac{\partial W}{\partial P} = \frac{\partial \overline{W}(P, Z, \theta^D; \psi^*)}{\partial P}. \tag{48}$$

This welfare effect is the same as $\frac{\partial \overline{W}}{\partial P}$ above (direct effects and behavioral effects multiplied by marginal internalities) but mean welfare is evaluated over the welfare-minimizing distribution ψ^*. Re-optimization by the evil agent ($\partial\psi^*/\partial P$) does not have a first-order welfare effect as a

consequence of the envelope theorem where it applies.[47]

B.4 Further Possibilities

We could leverage the parallel to prior thinking on the value of money across individuals to go even further. For instance, we could imagine an "impartial observer" with the same risk preferences over money as the individual evaluating the normative risk the planner confronts, and then construct u_ζ from the way the individual trades off risk [as in Harsanyi, 1955]. One problem with this approach is that individuals in behavioral models (and in reality) often do not act like expected-utility maximizers. If it is ambiguous whether we should respect revealed preferences that do not conform with expected utility theory (as in Example 1.3 below), the impartial observer's preferences become ambiguous. Rather than probe this question further, we next turn our focus to more applied questions.

B.5 Examples

Under probabilistic uncertainty in Example 1, we find that the variance of utility over frames is quadratic in the disagreement between the two frames, V:

$$Var_\psi(V) = \psi(\theta^D)(1-\psi(\theta^D))V(x,P)^2. \tag{49}$$

Evaluating the change in welfare from a policy reform due to the change in variance – the second term from equation (47) – we find:

$$\frac{\omega''(\overline{W})}{2} \cdot \frac{dVar_\psi\left[V(x(P,Z,\theta^D),P)\right]}{dP} = \omega''(\overline{W})Var_\psi(V) * \frac{1}{V}\frac{dV}{dP} \tag{50}$$

where V and dV/dP are evaluated at $(x(P,Z,\theta^D),P)$. This is a reduced-form expression that carries some intuition. Note that the last term resembles a semi-elasticity; this term is positive when V moves away from zero following a marginal change in P. The importance of disagreements for policy evaluation depends on 1) the degree of paternalistic risk aversion over our measure of welfare (ω''), 2) the extent of disagreement in the status quo ($Var_\psi(V)$), and 3) the change in the magnitude of disagreement generated by the reform.

The character of the $\frac{1}{V}\frac{dV}{dP}$ term depends on more specific features of the model. Let us illustrate this in Example 1.1. To obtain differentiability we introduce some unobserved heterogeneity (conventional uncertainty about the individual's type) so that instead of $V = -\mathbb{1}\{x \neq d\}\gamma$, we have $V = -Pr[x \neq d]\gamma$.[48] The right-hand side of (50) becomes

$$\omega''(\overline{W})\underbrace{\psi(\theta^D)(1-\psi(\theta^D))Pr[x\neq d]^2\gamma^2}_{Var_\psi(V)}\left\{\frac{1}{Pr[x\neq d]}\frac{\partial Pr[x\neq d]}{\partial d}\right\} \tag{51}$$

The last term in this expression is the semi-elasticity of opt-outs with respect to a change in

[47] Where ψ^* is discontinuous or non-differentiable in P, the envelope theorem does not apply and we require a more global approach to fully characterize optimal policy.

[48] We acknowledge this is informally construed in the interest of avoiding extra notation. We continue to assume that γ is uniform for simplicity, so the unobserved heterogeneity should involve other preference parameters. See Goldin and Reck [2022] for a more thorough treatment of the question of interpersonal heterogeneity in this setting.

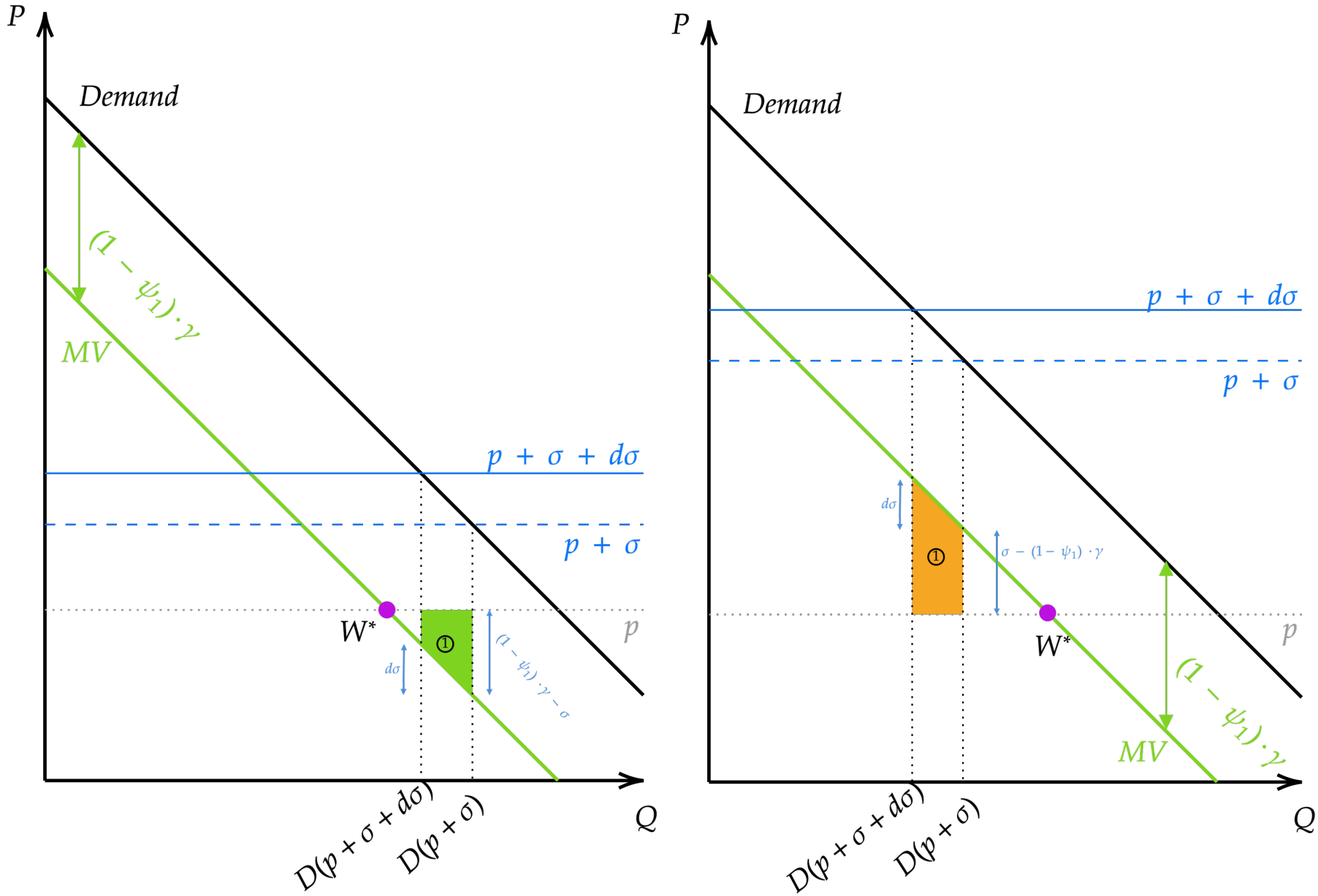
the default [see also Brot-Goldberg et al., 2023]. A reform of the default rule that increases opt-outs will be less desirable when the planner values robustness, to an extent governed by the other terms in the expression. In Example 1.2, the analogous semi-elasticity term is a weighted semi-elasticity of losses across various dimensions, where the weights depend on the strength of loss aversion in each dimension.

Under ambiguity aversion, the evil agent selects the $\psi \in \Psi$ that puts maximal weight on the frame in which welfare is lowest: when $V < 0$, ψ^* places maximal weight on θ^D and where $V > 0$, the evil agent places maximal weight on θ^A. As $V \leq 0$ everywhere in Examples 1.1 and 1.2 – note that this is an implication of the assumption that the level of utility is the same across frames where $x = d$ or $x = r$ – the evil agent always places maximal weight on θ^D in these models. By similar reasoning to the probabilistic uncertainty case, this will make policies where opt-outs are frequent less desirable in Example 1.1, and it will make policies where losses relative to the reference point are larger less desirable in Example 1.2.

C Optimal Nudges

In this section we include some additional material related to the examples we consider in the analysis of optimal nudges in Section 6.2.4.

C.1 Figures



A. Small baseline nudge $\sigma < (1-\psi_1)\cdot\gamma$

B. Large baseline nudge $\sigma > (1-\psi_1)\cdot\gamma$

Figure 5: Welfare effect of increasing the nudge when $\psi_2 = 0$

C.2 More complete model based on Allcott et al. [2025]

Relative to the model of Allcott et al. [2025], our setup in Section 6.2.4 is substantially simplified: we are essentially ignoring inter-personal heterogeneity and we restrict to the case where the government does not have access to a corrective tax instrument. In this section we allow for interpersonal heterogeneity as in Allcott et al. [2025].

The only additional insight which is generated by this fuller treatment is that the inequality which governs whether normative ambiguity leads to a lower nudge than the case where $\psi_1, \psi_2 = 0$, the "benefit" part of the tradeoff between net corrective benefit and psychic cost is a *targeting benefit*, rather than a net corrective benefit. Therefore, we mainly include the below for completeness.

Setup. A unit mass of individuals are deciding whether to buy a product (cigarette packet). Their value is $v \sim F$, the price is p and utility is quasi-linear. All individuals have heterogeneous bias $\gamma \in \mathbb{R}$ which shifts demand but not welfare. The government has access to a nudge (intensity $\sigma \in \mathbb{R}^+$) which reduces demand by sensitivity $\tau \in \mathbb{R}$ but does not affect welfare.[49] We still simplify the setup relative to Allcott et al. [2025] by assuming a perfectly competitive supply-side. The planner can levy a tax t which increases the price paid to $p + t$. In this simplified setup, as in the main text, there are four potentially normative frames:

$$\begin{aligned}
\text{Decision Utility: } u(\text{buy}, \theta^D) &= v + \underbrace{\gamma}_{\text{"Bias"}} - \underbrace{\sigma\tau}_{\text{Nudge}} - p \\
\text{Classical Utility: } u(\text{buy}, \theta^C) &= v - p \\
\text{Psychic-Cost Utility: } u(\text{buy}, \theta^P) &= v - \sigma - p \\
\text{Unbiased Utility: } u(\text{buy}, \theta^U) &= v + \gamma - p
\end{aligned}$$

Again let $\psi_1 = \mathbb{P}[\gamma \text{ is normative}]$ and $\psi_2 = \mathbb{P}[\sigma \text{ is normative}]$. Then comparability and these probabilistic beliefs yield planner preferences:

$$\begin{aligned}
w(\text{buy}) &= v + \psi_1\gamma - \psi_2\sigma\tau - p \text{ for each individual} \\
W &= \text{Consumer Surplus} + \text{Govt Revenue} \\
&= \int_{v \geq p+t+\sigma\tau-\gamma} v + \psi_1\gamma - \psi_2\sigma\tau - p \, dF
\end{aligned}$$

$\psi-$optima. One of the key results of Allcott et al. [2025] is that with heterogeneous bias γ, taxes should correct average bias and nudges should only be used to reduce variance. How do these conclusions change in the case of normative uncertainty / ambiguity? We start with the expected-utility formulation of a risk-neutral planner.

Following an analogous logic to Section 6.2.3, the optimal tax with heterogeneous bias γ and heterogeneous τ indeed corrects average bias (net of nudge effects and psychic costs):

[49] Note, now $\tau \gtrless 0$ are both possible, whereas in the main text we focussed on the case of $\tau > 0$. We make the nudge work against the bias as we view it as price, although it could of course be a negative price.

Observation 5.

$$t^*_\psi = \mathbb{E}_m[(1-\psi_1)\gamma - (1-\psi_2)\sigma\tau] \tag{52}$$

where $\mathbb{E}_m$ integrates across the marginal consumers: $\{(v,\gamma,\tau)|v = p+\sigma\tau-\gamma\}$. Demand is defined as $D(p) = \mathbb{P}[v > p+t+\sigma\tau-\gamma]$, let $D'_p < 0$ be its derivative. Let $\mathbb{E}_I[X] = \mathbb{E}[X(v,\gamma,\tau)|v \geq p+t+\sigma\tau-\gamma]$ i.e. conditional mean of X for the inframarginals.

Intuitively, t^*_ψ corrects γ to the extent that it is actually a bias $(1-\psi_1)$ and accounts for the average effect of the nudge σ to the extent that consumers themselves don't internalize the nudge's effects $(1-\psi_2)$.

Observation 6. *When the tax is set optimally according to Observation 5:*

$$\left.\frac{dW}{d\sigma}\right|_{t^*} = \underbrace{D'_p}_{<0}\{(1-\psi_1)Cov_m[\gamma,\tau] - (1-\psi_2)\sigma Var_m[\tau]\} - \psi_2 \cdot \mathbb{E}_I[\tau]\cdot D(p)$$

If $\psi_2 < 1$, and nudge elasticity is a fixed constant, then we can solve for an interior optimum:

$$\begin{aligned}\left.\frac{dW}{d\sigma}\right|_{t^*} &= 0 \\ \implies \sigma^* &= \frac{(1-\psi_1)\cdot\varepsilon^D_\sigma \cdot Cov_m[\gamma,\tau]}{\psi_2 \cdot \frac{\mathbb{E}_I[\tau]}{\mathbb{E}_m[\tau]} + (1-\psi_2)\cdot\varepsilon^D_\sigma \cdot Var_m[\tau]}\end{aligned} \tag{53}$$

Equation (53) is the analogue to Equation (34) in the case of interpersonal heterogeneity and an optimal tax. Now, average bias is irrelevant for whether the government prefers to nudge (as in Allcott et al. [2025]. Instead, σ^* is increasing in $(1-\psi_1)\cdot Cov_m[\gamma,\tau]$ i.e. if (1) the nudge is well targeted: the people with the largest bias have the largest reduction in demand by the nudge **and** (2) the government thinks it is likely that there is over-consumption in the first place. The optimal nudge is decreasing in psychic costs, proportional to $\psi_2 \cdot \frac{\mathbb{E}_I[\tau]}{\mathbb{E}_m[\tau]}$ and the self-correction benefit of the nudge $(1-\psi_2)\cdot\varepsilon^D_\sigma \cdot Var_m(\tau)$.

In general, the optimal nudge is less aggressive than the case where $\psi_1, \psi_2 = 0$ when:

$$\underbrace{(\psi_2-\psi_1)\cdot|D'_p|\cdot Cov_m[\gamma,\tau]\cdot\mathbb{E}_m[\tau]}_{\text{Net correction}} \leq \underbrace{\psi_2\cdot\mathbb{E}_I[\tau]}_{\text{Psychic costs}} \tag{54}$$

Again, the RHS measures psychic costs and the LHS net correction. Now, however, the corrective benefit of the nudge scales with $Cov_m[\gamma,\tau]$, as in Allcott et al. [2025]. Also, the model weighs the net-correction by the average sensitivity of the marginals (as this is who the correction comes from) to the psychic costs of the inframarginals. In general, it seems reasonable that $\mathbb{E}_m[\tau] \geq \mathbb{E}_I[\tau]$. Therefore, the $\psi-$optimal nudge in the case of interpersonal heterogeneity is likely higher than the case of homogoeneity all else equal.

Ambiguity. For simplicity, we first consider the case of global robustness where the planner wishes to solve $\max_{t,\sigma}\min_{\psi_1,\psi_2} W(\psi_1,\psi_2) = \int_{v\geq p+t+\sigma\tau-\gamma} v + \psi_1\gamma - \psi_2\sigma\tau dF - p\cdot Q^*$.

Focusing on the inner minimization leads straightforwardly to:

$$\psi_1^* = \mathbb{1}\{\mathbb{E}_I[\gamma] \leq 0\} \tag{55}$$
$$\psi_2^* = \mathbb{1}\{\mathbb{E}_I[\tau] \geq 0\} \tag{56}$$

As above, the natural case is $\mathbb{E}_I[\gamma] > 0$ and $\mathbb{E}_I[\tau] > 0$ which means $\psi_1^* = 0$ and $\psi_2^* = 1$. Plugging this into Equation (53), this implies that $\sigma^* = \frac{\varepsilon_\sigma^D \cdot Cov_m[\gamma,\tau] \cdot \mathbb{E}_m[\tau]}{\mathbb{E}_I[\tau]}$.

D Continuous Frames and Interpretation of Normative Weights

Here we discuss an extension of our model in which the frame space is conceived of as the convex hull of a finite set of frames. The extension serves to generalize our results and clarify the relationship of our work to the counterfactual normative consumer approach to behavioral welfare analysis, taken up in the next section of this Appendix.

D.1 Basics

In the main text we have a finite set of frames Θ. We can think of these as parameters of a (cardinal) utility function $u(x, \theta)$. Suppose each $\theta \in \Theta$ corresponds to a set of N real-valued preference parameters $\theta = (\theta_1, ..., \theta_N)$. Let $\tilde{\Theta}$ be the convex hull of Θ – the set of all convex combinations of elements of Θ. We note that if each of the elements of Θ is non-trivial – no component $\theta \in \Theta$ can be expressed as the convex combination of other components – then the dimensionality of $\tilde{\Theta}$ must be equal to the number of elements of Θ. Here we assume non-triviality and note that trivial frames can be thought of as elements of $\tilde{\Theta}$ rather than Θ.

D.2 Linearity and Equivalence to Previous Objectives

By construction, for each $\tilde{\theta} \in \tilde{\Theta}$, there exists a unique weighting function $\psi : \Theta \to \mathbb{R}$ such that $\tilde{\theta} = \sum_{\theta \in \Theta} \psi(\theta)\theta$ and $\sum_{\theta \in \Theta} \psi(\theta) = 1$.

Definition. A utility function is *frame-wise linear* if for any weighting function ψ,

$$u(x, \psi_1\theta_1 + ...\psi_n\theta_N) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta).$$

Framewise linearity is a significant restriction but we find it in applied work. Utility is linear in the χ parameter in Lockwood et al's lottery paper, and the π parameter in Goldin and Reck [2022] and Reck and Seibold [2023]. It is also met in the quasi-hyperbolic discounting model of Laibson [1997] – utility is linear in the present focus parameter β from Example 2. In all of these models, we obtain this equivalence between normative weights on discrete frames and the convex hull of frames under linearity.

If the utility function is frame-wise linear, then for any $\tilde{\theta} \in \tilde{\Theta}$ we have a weighting function ψ such that

$$u(x, \tilde{\theta}) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta).$$

This has an obvious equivalence to the utilitarian representation from Propositions 2 and 8. We discuss this in the main text in reviewing the relationship of our work to the counterfactual

normative consumer model.

We therefore find two potential interpretations of the $\psi(\theta)$.

- If the set of potential utility functions is captured by the discrete set Θ, ψ's are Bayesian beliefs about the probability $\theta \in \Theta$ is normative.
- If the set of potential utility functions is captured by the continuous set $\tilde{\Theta}$, under frame-wise linearity, ψ's may reflect a known normative frame θ^* in the convex hull $\tilde{\Theta}$.

The first of these interpretations is our main focus in the paper, but we can combine the two: if we maintain frame-wise linearity, introducing a pdf over $\tilde{\theta}$ to capture subjective/Bayesian beliefs about which frame is normative leads to a criterion of the same form, but with a more nuanced interpretation of the weighting function. Under the independence assumption, for any pdf over $\tilde{\theta}$, we can find weights on the discrete set Θ such that the planner's objective maximizes expected welfare over Θ given these weights.

E Counterfactual Normative Consumers and Identification of Normative Weights

Continuing with the setup introduced in the previous appendix section, we now discuss the relationship of our work with the counterfactual normative consumer (CNC) approach. The basic idea here is to suppose that the normative weights are implied by the planner's information, $I \in \mathcal{I}$, so let us express the weights as $\psi(\theta, I)$ such that for fixed I, $\psi \in \Delta(\Theta)$. To be clear, we are using the interpretation of the weights in ψ from the previous section here, assuming framewise linearity.

We consider a stylized version of the counterfactual normative consumer to abstract from interpersonal heterogeneity. Suppose the planner knows about the choices of one other individual. Let us call this other individual the expert and let us call the decision-maker for whom the planner is trying to set an optimal policy the main decision-maker or main DM. The CNC research design is justified by the following assumptions:

- **CNC0 (Knowledge of Expert)**: the expert's revealed preferences are constant over frames, and RP-coincidence holds for the expert.
- **CNC1 (Observation of Expert)**: the preferences of the expert are known/observed in at least one frame.
- **CNC2 (Similarity of Expert and Main DM)**: the main DM and the expert have the same preferences in the normative frame.

Discussion. We observe that CNC0 and CNC1 imply that the planner knows the expert's normative preferences, or their choices in the normative frame. We might assume this directly, but stating the assumptions this way emphasizes that we do not need to observe the expert's choices in the normative frame specifically. For instance, the planner might observe choices

in the same decision-making frame θ^D as the main decision-maker and *assume* the expert's choices are constant over frames on the basis of a survey bias proxy. So adopting CNC0 and CNC1 matches how the CNC approach is implemented in practice. In Goldin and Reck [2020], experts are assumed to be those who choose consistently across frames, while in Allcott et al. [2019], experts are identified via a survey bias proxy; both of these require RP-coincidence for the expert – Goldin and Reck label this the consistency principle, Allcott et al. et al impose it when they specify normative benchmarks for their bias proxies. Both CNC0 and CNC2 have untestable normative aspects. Generally, CNC2 can be thought of as the assumption that enables extrapolation from information on the preferences of experts to the preferences of others. Implementations of the CNC approach in practice tend to impose CNC2 via a statistical independence assumption, modeling some interpersonal heterogeneity we do not include here for simplicity, and typically assuming CNC2 conditional on a set of observables.

In any case, the implication of these assumptions is that the planner can infer the normative frame $\tilde{\theta} \in \tilde{\Theta}$ for the main DM by observing the expert's choices, and $\tilde{\theta}$ identifies a weighting function $\psi(\theta)$ on Θ such that the planner's (utilitarian) objective represents normative preferences $\succ^*$ with certainty. That the objective given known preferences takes the same form as the ones we have studied can be viewed as a consequence of the linearity assumption.

This idealized version of CNC is therefore a model in which true/normative preferences are known given the planner's information, which includes CNC0, CNC1, and CNC2. Moreover, because the weighting function implied by the expert's preferences is unique, we can say that knowledge of the expert's preferences *point-identifies* the appropriate normative weights.

The robustness concepts we develop in our paper help us think through policy problems in which the planner believes that CNC0, CNC1, or CNC2 might fail. For the sake of illustration, let us consider how these assumptions might fail in the context of sugary drink consumption in Allcott et al. [2019]. In this model, the expert is an individual with similar characteristics to the main decision-maker; the expert has no self-reported problems with self-control, no present bias according to a survey bias proxy, and good knowledge about the health risks of consuming sugary drinks.

1. **Frame misspecification.** CNC0 could fail if the set of frames is mis-specified, so that normative choices are not constant across what the researchers consider as frames (this can be formalized along the lines of Example 3). For instance, the payoff due to impulsiveness or lack of self-control could be normatively relevant for the main decision-maker even though the "expert" does not experience these payoffs. (In other words, the possibility here is that the "survey bias proxy" is not capturing a bias but a normative preference).

2. **Expert misspecification.** Alternatively, CNC0 might fail because the "expert" is not completely debiased, for instance because they do struggle a little bit with self-control even though they report having no difficulties with it.

3. **Noisy choices.** CNC1 might fail if the experts revealed preferences are not perfectly observed. For example, we might only have a noisy estimate of how much soda the

expert consumes.

4. **Selection Bias.** CNC2 might fail if being an expert is correlated with preferences. This could obviously happen because of frame misspecification above, but setting this aside, the possibility is that experts may be a statistically selected group of individuals, so that they have different preferences from non-expert decision-makers. For instance, those with especially high knowledge about the health consequences of consuming sugary drinks could have gained this knowledge because they especially value good health, and this could lead them to consume less sugary drinks regardless of how well-informed they are. This type of possibility suggests Roy-type selection into expertise that would threaten the (conditional) independence required by CNC2 [Goldin and Reck, 2020].

We do not claim all of these failures are material in the context of corrective taxes on sugary drinks. Allcott et al. [2019] work to defend against many of these concerns, even though the importance of robustness for optimal policy is not formalized in their model.

How we evaluate robustness when employing the CNC identification strategy seems to depend on which of the failures listed above we have in mind. The case of noisy expert choices (Failure 3) seems to suggest thinking about robustness in terms of probabilistic uncertainty; the relevant variance could then be derived from $u(x, \theta)$ for given x and the standard error for our estimate of weights in the normative frame. The possibility of (unstructured) expert misspecification or selection bias suggests a local max-min criterion, using a neighborhood $\Psi = B(\theta^*, \kappa)$ around our estimate of θ^*, as in the robust control literature. The possibility of frame mis-specification, however, suggests a more global robustness concept, as this involves more philosophical questions about whether the influence of some factor like self-control is due to a framing effect (see also Example 3). We emphasize that these are only suggestions for the appropriate robustness concept to apply to entertain potential failures of these assumptions. For instance, one could also think of expert misspecification in probabilistic terms. Ultimately, the question is how we think about uncertainty about the validity of these assumptions, and preferences for how to accommodate it.

F Non-Rectangular Environments: Extension and Application

In this section, we relax the rectangularity assumption maintained throughout the main text and apply the extended theory to intertemporal choice.

Concretely, in the main text we assume the set of choice situations is the Cartesian product $2^{\mathcal{X}} \times \Theta$. Here, we suppose that the set of choice situations is $G \subseteq 2^{\mathcal{X}} \times \Theta$. In other words, every choice situation is described by a menu and a frame, but some menus and frames cannot be combined.

F.1 General Model without Rectangularity

Apart from relaxing rectangularity, our notation is the same as main text Section 2.1. We introduce some definitions that describe which options are relevant for which frames and vice versa.

For a given frame θ, the set of relevant options is

$$\mathcal{X}(\theta) = \{x \in \mathcal{X} | \exists (X, \theta') \in G, x \in X, \theta' = \theta\}.$$

For any two options $x, x' \in \mathcal{X}$, the set of relevant frames is

$$\Theta(x, x') = \{\theta \in \Theta | \exists (X, \theta') \in G, x, x' \in X, \theta' = \theta\}.$$

Comparability Partition Without loss of generality, we can partition the set $\mathcal{X}$ into

$$\mathcal{X} = \cup_n \mathcal{X}_n$$

such that 1) for each $(X, \theta) \in G$, X is contained in exactly one $\mathcal{X}_n$, and 2) for any $(X, \theta), (X', \theta') \in G$ such that $X \in \mathcal{X}_i$, $X' \in \mathcal{X}_j$, we have $X \cap X' \neq \emptyset$ if and only if $i = j$. Intuitively, this partition describes which options the individual compares, directly or indirectly, when making choices in G. If the model is rectangular, ($G = 2^{\mathcal{X}} \times \Theta$), the partition is trivial ($\mathcal{X} = \mathcal{X}_1$). The converse is not true: in some non-rectangular environments, the individual compares all options (see the next subsection for an example).

Now we revise our core assumptions as follows:

Assumption 11. ***Core Assumptions without Rectangularity.***

Assumption 11.1. ***Normative Preferences Exist.*** *There is binary preference relation $\succeq_*$ on $\mathcal{X}$, which is complete and transitive over $\mathcal{X}_n$ for each n. The individual's normative choices are those that maximize $\succeq_*$.*

Assumption 11.2. ***Frame-Dependent Rational Preferences.*** *For every $\theta \in \Theta$, there is a complete and transitive preference relation $\succeq_\theta$ on $\mathcal{X}(\theta)$. Individual choices maximize these preferences.*

Assumption 11.3. ***Revealed Preference Coincidence.*** *For every $x, x' \in \mathcal{X}$ such that there exists $(X, \theta) \in G$ with $x, x' \in X$, there exists $\theta^* \in \Theta(x, x')$ such that*

$$x \succeq_{\theta^*} x' \iff x \succeq_* x'.$$

The difference between Assumption 11.1 and its main text analogue is that we require completeness and transitivity over options the individual compares when making choices rather than all options. If two options x and x' are in distinct partitions of $\mathcal{X}$, the individual will never make a choice that requires that they compare x and x', so we do not assume that they have normative preferences describing whether they should choose x or x'. A social planner forced to choose whether to assign the individual x or x' would face a non-comparability problem like the one described in Kőszegi and Rabin [2008]; Bernheim et al. [2024] proposes a solution but here we will focus on comparisons that could conceivably be grounded in conventional revealed preferences.

Compared to their main text analogues, Assumption 11.2 requires that frame-dependent rational preferences are defined only over relevant choices made within a given frame, and

Assumption 11.3 requires that for any two options the individual chooses from, there is a normative frame that reveals which option is preferred according to normative preferences.

We observe that with this modification to the setup of our model, our core assumption admits the version of BR-dominance as originally proposed by Bernheim and Rangel, without the rectangularity restriction.

Observation 7. ***BR-Dominance without Rectangularity.*** *Under Assumption 11, for any pair of options* $x, x' \in \mathcal{X}$,

$$\forall (X, \theta) \in G \text{ s.t. } x, x' \in X, x \in x(X, \theta), x' \notin x(X, \theta) \implies x \succ_* x'.$$

Proof. Take any situation (X, θ) such that $x, x' \in X$ and suppose the individual chooses x and not x' in this situation. Note by design that it must be the case that $x, x' \in X(\theta)$. By Assumption 11.2, we know that $x \succ_\theta x'$. The precondition of the implication we are trying to prove is therefore equivalent to

$$\forall (X, \theta) \in G, x, x' \in X, x \succ_\theta x'. \tag{57}$$

The result follows θ^* is one of these θ's such that $(X, \theta) \in G, x, x' \in X, x \succ_\theta x'$.

■

Disciplining the Model One likely objection to the version of our core assumptions above is that it could be the case that for some θ, the individual could be endowed by Assumption 11.2 with preferences over options they never compare, i.e. the domain over which $\succ_\theta$ is defined, $\mathcal{X}(\theta)$, could include elements of both $\mathcal{X}_i$ and $\mathcal{X}_j$ for $i \neq j$. Because the individual never makes choices that compare options across these partitions, we find it natural to assume the set of frames that is relevant in each partition is also distinct.

Assumption 12. ***Structural Assumption.*** *The set* Θ *can be partitioned into* $\Theta_1, .., \Theta_N$ *such that for any* $(X, \theta) \in G$ *and any* n, $X \in \mathcal{X}_n \iff \theta \in \Theta_n$.

With this assumption, we ensure preferences under each frame $\succeq_\theta$ can be rooted in revealed preference. The assumption disciplines the set of relevant options/frames as follows:

Observation 8. *Under Assumption 12, for a frame* $\theta \in \Theta_n$, $X(\theta) \subseteq \mathcal{X}_n$. *For two options* $x \in \mathcal{X}_i$, $x \in \mathcal{X}_j$, $\theta(x, x') \subseteq \Theta_i$ *if* $i = j$ *and* $\theta(x, x') = \emptyset$ *if* $i \neq j$.

Another difficulty we wrestle with in building on the core assumptions to generalize our main results is that Assumption 11.3 allows that the normative frame might be different for different pairs of options. This means that to adapt our the main propositions of the main text to this setting, we could need to endow the planner with beliefs about which frame is relevant for given pairs of options, which could get very complicated. The following assumption allows us to articulate a stronger version of RP-coincidence with which the main propositions can more easily be adapted to a non-rectangular environment.

Definition. A frame $\theta \in \Theta$ is a *global frame* on $\mathcal{X}_n$ if $\mathcal{X}(\theta) = \mathcal{X}_n$.

Assumption 13. ***Global Revealed Preference Coincidence*** *For each $\mathcal{X}_n$, there exists $\theta \in \Theta_n$ such that θ is a global frame on $\mathcal{X}_n$ and for any $x, x' \in \mathcal{X}_n$,*

$$x \succeq_\theta x' \iff x \succeq_* x'.$$

Under the global version of RP-coincidence in Assumption 13, the adaptation of main text Propositions 1-4 to a setting without rectangularity becomes straightforward. The planner's objective will be distinct for each n under the structural assumption, and under Assumption 13, the planner forms beliefs over whether each global frame is normative. We defer to future work the question of whether it might be possible to apply our framework to non-rectangular models without Assumption 13; as we discuss in the intertemporal choice application below, attempting to do so naively can generate inconsistencies in the planner's welfarist objective.

F.2 Application: Intertemporal Choice

In Example 2 in the main text, we considered a model of intertemporal choice in which the vantage point from which an individual evaluates a consumption plan is a frame, and we imposed rectangularity on preferences expressed in each vantage point. Although we find the way in which we rectangularized the model intuitive, a much more conventional approach would be to assume that the agent evaluating consumption plans at vantage point τ has preferences over present and future consumption and not over past consumption, because preferences over past consumption cannot be rooted in revealed preference (without a time machine). Here, we use the extension of our results to non-rectangular environments to examine welfare in a model of quasi-hyperbolic discounting with a more conventional setup.

Due to some subtleties that emerge from our analysis, we find it instructive to begin with a simple version of the model in which 1) all options the agent chooses reflect committed choices, and 2) the vantage point can be viewed as a frame. We show that with these simplifications, welfare under our core assumption has a very simple characterization. We cover the relaxation of these simplifying assumptions below.

Setup The setup of the model is equivalent to the one in the main text except that preferences are defined over present and future consumption and not past consumption. There are T periods indexed by t. Options are (committed) consumption plans: $\mathcal{X} = \mathbb{R}^T_{++}$. In each period $\tau = 0, ..., T$, the agent makes choices to maximize utility that takes the quasi-hyperbolic form [Laibson, 1997]:

$$u(x, \tau) = 1\{\tau > 0\}\mu(x_\tau) + \beta \sum_{t=\tau+1}^{T} \delta^{s-\tau}\mu(x_s), \tag{58}$$

where the subutility function μ is continuous, strictly increasing and concave. The agent does not consume in period 0 so under $\tau = 0$ they make entirely forward-looking choices.

The set of choice situations is the set of subsets of consumption plans and vantage points such that consumption in periods prior to the vantage point is held fixed. Formally,

$$G = \{(X, \tau) \subseteq 2^{\mathbb{R}^T_{++}} \times \{0, ..., T\} \mid \forall x, x' \in X, t < \tau \implies x_t = x'_t\}.$$

By additive separability, when consumption is held fixed in prior periods it becomes irrelevant for preferences as of period τ and we may therefore represent the utility function from vantage point τ using Equation (58).

Implications of assuming the vantage point is a a frame Assuming τ is a frame has borderline obvious implications that, combined with our core assumptions, impose a substantial restriction on welfare.

We first observe that under commitment, any consumption plan the agent chooses in $\tau > 1$ could also be chosen in period 1. It follows that for any $X \subseteq \mathcal{X}$, $(X, 1) \in G$, and as a result we have the following observation:

Observation 9. *If the vantage point τ is a frame, all options are comparable, i.e. the comparability partition described in the previous section is trivial.*

Relatedly, there are exactly two global frames in the model where τ is a frame:

Observation 10. *If the vantage point is a frame, $\tau = 0$ and $\tau = 1$ are the only global frames in the model.*

Now we observe that Assumption 11.1 requires that normative preferences cannot depend on the vantage point (frame invariance). As such:

Observation 11. *If the vantage point τ is a frame, 11.1 implies that normative preferences must be time consistent.*

Each of these observations follows straightforwardly from the previous observation and our construction of the model, but the final implication is very strong: if we assume the vantage point is a frame and we insist that the normative frame must be global, we must end up with the long-run view.

Observation 12. *Under Assumptions 11.1, 11.2, Global revealed preference coincidence (Assumption 13) implies that the planner takes the long-run view for all choices made in period $\tau > 1$.*

In fact, slightly less obviously, this must be the case *even without global RP-coincidence.* That is, the long-run view is the only admissible welfare criterion for choices in $\tau > 1$ even under the weaker version of RP-coincidence introduced in Assumption 11.3.

Proposition 11. *Under Assumption 11, normative preferences must accord with the long-run view for all choices made in period $\tau > 1$.*

Proof. Toward a contradiction, suppose that welfare does not accord with the long-run view for some choice made in some period $\tau > 1$. Then we must be able to find two options x, x' such that 1) $x_t = x'_t$ for $t < \tau$, i.e. these options can be compared in period τ, 2) the long-run view self has the preference $x \succeq_0 x'$, 3) the period τ self has the preference $x' \succ_\tau x$, and 4) normative preferences coincide with the non-long-run-view frame:

$$x' \succ_* x.$$

Note that because consumption in periods prior to τ is held fixed across these options, the preferences of prior selves must agree with the long-run self; formally, for $t < \tau$, $x \succ_0 x' \iff x \succ_\tau x'$. We can focus on the period τ frame and the period 0 frame and ignore the others because they are redundant. Relatedly, as we assumed $\tau > 1$, we know that $x \succ_0 x' \iff x \succ_1 x'$, which will be useful in the next step of the proof.

Let x'' be an option that is identical to x in all periods except that we reduce period 1 consumption by $\varepsilon > 0$. Because within-period flow utility μ is continuous, we know that for sufficiently small ε,

$$x \succ_0 x'' \succ_0 x'.$$

and

$$x \succ_1 x'' \succ_1 x'.$$

Observe that because x'' differs from x' and x in period 1, it can only be comapred under the $\tau = 0$ and $\tau = 1$ frames. As such, Assumption 11.3 implies

$$x \succ_* x'' \succ_* x'.$$

This contradicts what we supposed above, requiring $x' \succ_* x$, violating transitivity of $\succ_*$ under Assumption 11.1. ■

Discussion of Proposition 11 This proposition requires that if we assume RP-coincidence, i.e. for any two options one of the frames in which the individual chooses from these options must be normative, and if we assume that the vantage point τ is a frame, the long-run view of preferences must prevail for choices in almost all periods (choices in any period $\tau > 1$). Intuitively, if the vantage point is a frame then normative preferences must be time consistent, and the only time consistent criterion that always respects revealed preferences is the long-run view.

Normative Uncertainty and Present Focus Our central goal with Example 2 in the main text was to consider normative uncertainty over whether the long-run view of welfare is normative, as suggested by Bernheim [2009] and, implicitly, Bernheim et al. [2015]. What we learn from the prior result is that with this setup of the model, which is more conventional than what we did in the main text, such uncertainty cannot be accommodated within our framework if we assume that the vantage point is a frame. Instead, the way to introduce this type of uncertainty in our framework is to introduce *uncertainty over whether the vantage point is a frame*, just as we considered in Example 3 of the main text.

When the vantage point is not a frame Before we consider uncertainty over whether the vantage point is a frame, it is useful to think through the implications of our core assumptions in the case where the vantage point is not a frame.

Observation 13. *If the vantage point is a feature of the options rather than a frame, then the comparability partition is $\mathcal{X} = \cup_{\tau=1}^{T} \mathcal{X}_\tau$ such that for each consumption path x the individual could choose in period τ, the option $(x, \tau) \in \mathcal{X}_\tau$.*

When τ is not a frame, potential consumption paths the individual can choose from in period 2 are distinct from those chosen in period 1. Given this, there is only one potentially normative frame when τ is (with certainty) not a frame, i.e. the revealed preferences of the period τ self. Under RP-coincidence, these preferences must be from the normative frame.

Observation 14. *If the vantage point is not a frame, under Assumption 11, normative preferences over choices in each $\mathcal{X}_\tau$ must coincide with the (present-focused) preferences of the period τ self.*

Intuitively, the only revealed preferences in the model for period τ are the choices the individual makes in period τ, so if at least one such revealed preference must be normative, normative preferences must reflect the "short-run view."

With these conditions, we also encounter a non-comparability problem of the form described by Kőszegi and Rabin [2008]. A social planner setting a policy that affects welfare in multiple periods would have no way to use revealed preferences to aggregate welfare across periods from an ex ante perspective. We set this issue aside for now and return to it shortly.

Uncertainty over whether the vantage point is a frame Now we think through the case where the planner does not know whether the vantage point should be a frame, which we now know to be equivalent to modeling uncertainty over whether the long-run view of welfare is normative. For simplicity, we aside the possibility raised above that when the vantage point is a frame, the short-run view could be normative at $\tau = 1$ and no other vantage point. So we assume that if the vantage point is a frame, the long-run view is normative for all τ.

As we have a distinct partition over every τ when the vantage point is definitely not a frame, we continue to have T partitions of the grand menu set, so the planner will have distinct (time-inconsistent) normative criteria in each period. Under probabilistic uncertainty as in Proposition 2 and related results, the planner has subjective beliefs over this uncertainty. Denoting the probability that the vantage point is not a frame denoted ψ, the planner in period $\tau \geq 1$ maximizes[50]

$$w(x, \tau) = \sum_{t=\tau}^{T} \delta^{s-\tau} \mu(x_s) - \psi(1-\beta)\mu(x_\tau) \tag{59}$$

The second term in this expression is the disagreement between the long-run view and the short-run view of welfare, and this is weighted by the probability that the short-run view is normative.

Aggregation over Periods If we are focused in decision-making in a single period, the expression above seems like a suitable intrapersonal welfare criterion. However, if we broaden our focus to dynamic decision-making, we will need to compare $w(x, \tau)$ over different vantage points τ. If the planner is certain the vantage is a frame, they take the long run view and such aggregation is straightforward. Otherwise, it is not clear how the planner should aggregate welfare across periods. For instance, naively adding up $w(x, \tau)$ above would lead us to put a lot more (T times more) weight on utility over consumption in the final period than the first

[50] This probability could depend on τ but we find it intuitive and simplifying to set aside this possibility.

period. This is related to problems with comparisons across periods in this model discussed in Bernheim and Rangel [2009]. The most intuitive way we can think of to aggregate welfare across periods in an intertemporal selves model is to use the rectangular version of the model we presented in main text Example 2, despite the unconventional setup. The root problem is that agents never make choices in which they compare consumption plans from different vantage points. We defer a fuller consideration of how one might approach this aggregation problem to future work.

Non-Committed Choices In the above analysis of the case where τ is a frame, the period 1 self's revealed preferences over consumption in future periods are recovered implicitly by allowing the agent to commit to consumption in each period. Our approach does not necessarily require that all options the agent could choose feature commitment, but options featuring commitment must be in $\mathcal{X}$, e.g. so that we can ask the agent in period 1 if they prefer x or x'. One can formalize this idea by making commitment an explicit feature of the environment. We find it intuitive to assume that which consumption path an individual prefers does not depend on whether they commit to this path ex ante, so whether a choice features commitment can be regarded as a component of the frame in addition to the vantage point. The essential requirement of our analysis above is that the normative frame must coincide with revealed preferences in a committed frame.

While assuming committed options exist solves the problem of how to embed the long-run view in revealed preferences in the quasi-hyperbolic discounting model, some additional possibilities for the set of normative possibilities (the set of frames) arise if we want to entertain the possibility that choices an agent makes under both lack of commitment and naivete reflect normative preferences. We observe that Bernheim Fradkin, and Popov's application of BR-dominance in a model of present focus with partial naivete allows that such naive choices may reflect normative preferences.

However, the normative assumption that naive, uncommitted choices cannot reflect normative preferences can be derived from the less controversial assumption that choices that exhibit "characterization failure" cannot be normative [see Bernheim and Taubinsky, 2018, and other writing by Bernheim]. To see why, let us use an example with $T = 3$ periods. Suppose the period 1 self is naive about the period 2 self's present focus, so their beliefs about what period 2 self will choose in period 2 are incorrect. As such, the agent in period 1 believes that when they choose between x and x', they're choosing between the consumption paths $y = \{100, 110, 100\}$ and $y' = \{105, 100, 100\}$ (these are amounts of consumption in periods 1, 2, and 3). But given what the agent will actually choose in period 2, by consuming 100 or 105 in period 1, the period 1 self is choosing between $x = \{100, 115, 95\}$ and $x' = \{105, 110, 90\}$. Notice that in reality, the period 2 self consumes more than the naive period-1 agent believes that they will, so we can think of this agent's behavior intuitively as procrastinating the decision to save.

Under committed choices over these consumption paths, suppose the period 1 self expresses $y' \succeq_\theta y$ and $x \succeq_\theta x'$. Without commitment, therefore, the naive agent will choose to consume 105 in period 1 and the agent's consumption path will be x' instead of x. If we view this choice as potentially normative, we must consider the possibility that $x' \succeq_* x$, because that is

what the agent chooses under naivete without commitment. However, the choice "revealing" potentially $x' \succeq_\theta x$ in the uncommitted frame is entirely driven by mistaken beliefs (naivete) about what future selves will choose, so this is a case of characterization failure. It therefore seems reasonable to rule out $x' \succeq_* x$ ex ante, regardless of what we believe about whether the present focus of the period 1 self is normative. (Note we have not considered what the period 0 self would choose).

In summary, throughout our analysis of intertermporal choice, we implicitly assumed that choices under naivete and lack of commitment cannot be normative, which can be derived from the premise that choices under characerization failure cannot be normative. The general approach to welfare analysis we develop in this paper does not require that assumption. Relaxing it could nevertheless lead to further complications involving comparability.

G Proofs for Section 2 (Setup)

Observation 1. ***BR-Dominance.*** *Under Assumption 1, for any $x, x' \in \mathcal{X}$,*

$$\forall \theta \in \Theta,\ x \succeq_\theta x' \implies x \succeq_* x'. \tag{2}$$

Proof. By Assumption 1.3, a normative frame $\theta^* \in \Theta$ exists such that $x \succeq_{\theta^*} x'$. Since $x \succeq_\theta x'$ for all $\theta \in \Theta$, it follows that $x \succeq_{\theta^*} x'$, and thus Assumption 1.3 implies $x \succeq_* x'$. ■

H Proofs for Section 3 (Policy Problems and Normative Criteria)

Proposition 1. *Maintain Assumptions 1.2, 2 and 3. Assumption 5 holds if and only if for any representation of ordinal preferences $u(x, \theta)$, there is a function $\mathcal{W} : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planner's preferences are represented by*

$$w(x) = \mathcal{W}\left(\{u(x, \theta)\}_{\theta \in \Theta}\right), \tag{6}$$

and $\mathcal{W}$ is continuous and weakly increasing in every argument.

Proof. This argument is due to Kaplow and Shavell [2001]. We observe that $\succ_w$ does not have the proposed representation if and only if there are two options x, x' such that for every θ, $x \sim_\theta x'$ but $w(x) \neq w(x')$. Toward a contradiction, suppose we find two such x, x'; without loss of generality $x \succ_w x'$. Starting from x', construct x'' by increasing the good x_n from Assumption 3 by a small amount $\delta > 0$. By continuity (5.2), if δ is sufficiently small we must have $x \succ_w x''$. But for every θ, $x'' \succ_\theta x' \sim_\theta x$, so BR-dominance require $x'' \succsim_w x$. This establishes sufficiency of our assumptions for representation (6); necessity is easily verified. ■

Proposition 2. *Maintain Assumptions 1, 2, 3, and 4. Then Assumptions 5 and 6 hold if and only if there is a probability distribution function $\psi(\theta)$ and a utility function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $u(x, \theta)$ represents $\succeq_\theta$ for every θ, and planner's preferences $\succeq_w$ are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta). \tag{7}$$

Moreover, u is unique up to positive affine transformation if there are at least two non-null frames.[51]

Proof. We use the utility function whose existence is assumed in Assumption 6 to construct an outcome space comprising the output of the utility function. The fact that we obtain a subjective expected utility representation of the planner's preferences then follows from Theorem 5 of Köbberling and Wakker [2003]: their solveability pre-condition is implied by continuity (Assumption 2 and Assumption 5.2) and the richness of our option set (Assumption 4). For the set of axioms that is equivalent to the existence of a subjective expected utility representation with uniqueness up to affine transformation with at least two non-null frames/states in their theorem, Assumption 5.1 and Assumption 5.2 ensure the *weak ordering* condition, *monotonicity* is ensured by weak BR-dominance (Assumption 5.3), *tradeoff consistency* is ensured by Assumption 6. Their Archimedean axiom is satisfied when there is more than one frame by Assumption 4; with more technical work we conjecture that one could weaken 4 to something approaching their Archimedean axiom. Proposition 1 requires that the utility function in our representation of the planner's preferences is a representation of $\succeq_\theta$. ■

I Proofs for Section 4 (Comparability)

Lemma 1. ***Existence and uniqueness of EV.*** *Under Assumptions 1.2, 2 and 7, for any option x any frame θ and any baseline (P_0, Z_0), equivalent variation ζ exists and is unique. Moreover, $\zeta(x, \theta; P_0, Z_0)$ represents ordinal preferences $\succeq_\theta$ for every θ.*

Proof. Suppose first that $x \succ_\theta x(P_0, Z_0, \theta)$, i.e. $u(x, \theta) > u(x(P_0, Z_0, \theta), \theta)$. Assumption 7.1 ensures that $u(x(P_0, Z_0 + \zeta, \theta), \theta)$ is a strictly increasing function in ζ; Assumption 7.2 ensures this function is continuous. Assumption 7.3 implies that there is some $\hat{\zeta}$ such that $u(x(P_0, Z_0 + \hat{\zeta}, \theta), \theta) > u(x, \theta)$. The result follows from the Intermediate Value Theorem – note that $u(x, \theta)$ is continuous by Assumption 2. The same logic applies where $x \prec_\theta x(P_0, Z_0, \theta)$, and in the case of indifference, $\zeta = 0$.

That $\zeta(x, \theta; P_0, Z_0)$ represents $u(x, \theta)$ is easily verified. ■

Proposition 3. ***Planner's Preferences and Equivalent Variation.*** *Under Assumptions 1.2, 2, 5, and 7, for any baseline P_0, Z_0, there is a function $\mathcal{W}_\zeta : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planner's preferences are represented by $w(x) = \mathcal{W}_\zeta\left(\{\zeta(x, \theta; P_0, Z_0)\}_{\theta \in \Theta}\right)$.*

Proof. The result follows the exact same logic as the proof of Proposition 1, but we use small amounts of Z to construct BR-dominant options rather than small amounts of the good described by Assumption 3 (which is no longer required). ■

Lemma 2. ***Ordinal Level Comparability of Equivalent Variation.*** *Maintain Assumptions 1, 2, 4, 5, 6, and 7. Let $u(x, \theta)$ be a cardinal utility function from the representation in Proposition 2. Assumption 8 holds if and only if there is a baseline (P_0, Z_0) such that $u(x, \theta)$ and $\zeta(x, \theta; P_0, Z_0)$ exhibit ordinal level comparability.*

[51] When there is only one non-null frame, we are in the case where the planner knows the normative frame and the utility representation will be unique up to monotonic transformation. When all frames are null, the planner is indifferent to all options, which is a case we generally disregard.

Proof. First we prove that level comparability implies Assumption 8. Assuming ordinal level comparability, Assumption 8.1 follows from the observation that $\zeta(x(P_0, Z_0, \theta), \theta) = \zeta(x(P_0, Z_0, \theta'), \theta') = 0$ by construction. If ζ and u exhibit ordinal level comparability, it must also be the case that $u(x(P_0, Z_0, \theta), \theta) = u(x(P_0, Z_0, \theta'), \theta')$. The second condition then follows from Assumption 7.1 (strict monotonicity over money).

Second we prove that Assumption 8 implies ordinal level comparability. Take any x, x', θ, θ'. Using Assumption 8.1 we have

$$u(x, \theta) \geq u(x', \theta') \iff u(x, \theta) - u(x(P_0, Z_0, \theta), \theta) \geq u(x', \theta') - u(x(P_0, Z_0, \theta'), \theta').$$

Using the definition of equivalent variation, suppressing the baseline input, we have

$$\begin{aligned} & u(x, \theta) - u(x(P_0, Z_0, \theta), \theta) \geq u(x', \theta') - u(x(P_0, Z_0, \theta'), \theta') \\ \iff & u(x(P_0, Z_0 + \zeta(x, \theta)), \theta) - u(x(P_0, Z_0, \theta), \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x(P_0, Z_0, \theta'), \theta'). \end{aligned}$$

Now using Assumption 8.2, we find

$$\begin{aligned} & u(x(P_0, Z_0 + \zeta(x, \theta)), \theta) - u(x(P_0, Z_0, \theta), \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x(P_0, Z_0, \theta'), \theta') \\ \iff & \zeta(x, \theta) \geq \zeta(x, \theta'). \end{aligned}$$

■

Proposition 4. *Under Assumptions 1, 2, 4, 5, 6, 7, and 8, there is a probability distribution $\psi(\theta)$, a function $u_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0, Z_0) such that the planner's preferences are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u_\zeta(\zeta(x, \theta; P_0, Z_0)). \tag{12}$$

Moreover, u_ζ is strictly monotonic and unique up to positive affine transformation when more than two frames are non-null.

Proof. Lemma 1 implies the equivalent variation exists, is unique and represents $\succeq_\theta$ for some baseline. Lemma 2 implies that this representation satisfies ordinal level comparability with the planner's cardinal utility function. Then applying Corollary 8.1 gives the result. ■

J Proofs for Section 5 (Examples)

Proposition 5. ***Intertemporal "Social" Welfare and the Long Run View.*** *In this model, if $\psi(\tau)$ is constant for $\tau > 0$, then for any $\psi(0)$, the planner's preferences coincide with the long-run view of welfare $u(x, 0)$.*

Proof. With constant weights for $\tau > 0$, $\psi(\tau | \tau > 0) = \frac{1}{T}$, and equation (22) simplifies as

follows:

$$w(x) = \beta \sum_{t=1}^{T} \delta^t \mu(x_t) + \frac{1-\psi(0)}{T}(1-\beta) \sum_{\tau=1}^{T} \delta^\tau \mu(x_\tau), \tag{60}$$

$$= \left[\beta + (1-\beta)\frac{1-\psi(0)}{T}\right] \sum_{t=1}^{T} \delta^t \mu(x_t), \tag{61}$$

which is a constant multiple of $u(x, 0)$. ■

Observation 2. ***Welfare given uncertainty about whether a feature of the environment is a frame.***

$$w(x) = u_0(x, d) - (1-\psi_0) \sum_{\theta_2 \in D} \psi(\theta_2|1) V(x, d, \theta_2) \tag{24}$$

$$= \psi_0 \cdot u_0(x, d) + (1-\psi_0) \cdot \sum_{\theta_2} \psi(\theta_2|1) \cdot u_1(x, \theta_2) \tag{25}$$

Proof. Let $\psi(\theta_1, \theta_2)$ be the joint distribution of θ_1, θ_2, $\psi_1(\theta_1)$ the marginal distribution of θ_1 and $\psi(\theta_2|\theta_1)$ the conditional distribution of θ_2 given θ_1. $\psi(\theta_1, \theta_2) = \psi_1(\theta_1) \cdot \psi(\theta_2|\theta_1)$ and $\psi_1(\theta_1) = \sum_{\theta_2} \psi(\theta_1, \theta_2)$

Then,

$$\begin{aligned}
w(x) &= \sum_{\theta_1, \theta_2} \psi(\theta_1, \theta_2) \cdot u(x, d, \theta_1, \theta_2) \\
&= \psi_1(0) \sum_{\theta_2} \psi(\theta_2|0) \cdot \underbrace{u(x, d, 0, \theta_2)}_{\equiv u_0(x,d), \text{ i.e. constant wrt } \theta_2} + [1-\psi_1(0)] \sum_{\theta_2} \psi(\theta_2|1) \cdot \underbrace{u(x, d, 1, \theta_2)}_{=u_1(x,\theta_2)} \\
&= \psi_1(0) \cdot u_0(x, d) \cdot \underbrace{\sum_{\theta_2} \psi(\theta_2|0)}_{=1} + [1-\psi_1(0)] \sum_{\theta_2} \psi(\theta_2|1) \cdot u_1(x, \theta_2) \\
&= u_0(x, d) - [1-\psi_1(0)] \sum_{\theta_2} \psi(\theta_2|1) \cdot V(x, d, \theta_2)
\end{aligned}$$

because $V(x, d, \theta_2) = u_0(x, d) - u_1(x, \theta_2)$. If $\psi_1(0) = 1$ then immediately $w(x) = u_0(x, d)$. ■

K Proofs for Section 6 (Robust Optimal Policy in Applications)

Observation 3. ***BR-Optimality and Global Robustness.*** *A policy $P^* \in \mathcal{P}$ is a globally robust optimum if and only if for every $P' \in \mathcal{P}$, $x(P)$ weakly BR-dominates $x(P')$.*

Proof. Suppose $x(P^*)$ BR-dominates any other $x(P')$. Then global optimality of P^* follows from the monotonicity of expected welfare. For the other direction, suppose P^* does not BR-dominate some P', i.e. there is some θ' strictly better off under P' than P^*. Let $\psi(\theta) = \mathbb{1}\{\theta = \theta'\}$. As P^* is not a ψ-optimum for this ψ, it cannot be globally optimal. ■

Proposition 6. ***Robust Optimal Defaults when the Intrinsic Optimum is Unknown***

- *The ψ-optima are the expected intrinsic optimum and the most extreme default possible in the positive or negative direction (henceforth the* extremum default*).*

- *None of the ψ-optima are globally robust.*
- *If the expected intrinsic optimum is ψ-optimal for some ψ in the interior of Ψ, the expected intrinsic optimum is the unique (locally) robust optimum. If not, the extremum default is the unique (locally) robust optimum.*

Proof. In the defaults case, the policy parameter is one-dimensional: $P = d$, the default option. As discussed previously, this is usually thought of as an example of Bias vs Strange Preferences - where under θ^D, the as-if cost implied by behavior is normative, and under θ^A it is a pure bias. $\psi(\theta^D)$, which we abbreviate to just $\psi = \mathbb{P}[\theta = \theta^D]$.

$$u(s, d, \theta^D) = v(s) - \gamma \cdot \mathbb{1}\{x \neq d\} \tag{62}$$
$$u(s, d, \theta^A, d) = v(s) \tag{63}$$

Therefore, $V(s, d) = -\gamma \cdot \mathbb{1}\{s \neq d\}$ and welfare $W(d) = u\big(s(d)\big) - \psi^D \cdot \gamma \cdot \mathbb{1}\{s(d) \neq d\}$.

As a simple example, let $u(s) = -\frac{\alpha}{2}(s - s^*)^2$ where s^* is unknown. $s, s^* \in S$, the choice set which is $S \subset \mathbb{R}$ and defaults at the max and min of S force the consumer to choose actively. $\psi \in [0, 1]$.

1. First, show that the expected intrinsic optimum d_{min} default is a $\kappa - \psi$ robust optimum for any ψ making d_{min} a candidate optimum. This $\psi \approx 1$. So, $B(\psi, \kappa) = [\psi - \kappa, min(\psi + \kappa, 1)]$. Recall $W(d, \psi') = -\frac{\alpha}{2}(s(d) - s^*)^2 - \psi' \cdot \gamma \cdot \mathbb{1}\{s(d) \neq d\}$. Since $\gamma > 0$...

$$\arg \min_{\psi' \in B(\psi, \kappa)} W(d_{min}, \psi') = min(\psi + \kappa, 1) \tag{64}$$

   The evil agent wants to make the opt-out cost as large as possible so chooses ψ' as large as possible. Therefore, the $\kappa - \psi$ robust optimum is defined by...

$$d^* = \arg\max_d -\frac{\alpha}{2}(s(d) - s^*)^2 - min(\psi + \kappa, 1) \cdot \gamma \cdot \mathbb{1}\{s(d) \neq d\} \tag{65}$$

   Since d_{min} is a candidate optimum for ψ, it is also a candidate optimum for $\psi' > \psi$ since under those judgments the opt-out cost is strictly more likely to be normative - suggesting that minimizing opt-outs will be better. Therefore, d_{min} is a $\kappa - \psi$ robust optimum for any κ.

2. Now show that the penalty default is only a $\kappa - \psi$ robust optimum for small κ. Let ψ be the judgment which makes the penalty default a candidate optimum ($\psi \approx 0$). Then, $B(\psi, \kappa) = [max(0, \psi - \kappa), \psi + \kappa]$. Similarly to the minimizing opt-outs example, the evil agent wants to maximize ψ' and so sets...

$$\arg \min_{\psi' \in B(\psi, \kappa)} W(d_{pen}, \psi') = \psi + \kappa \tag{66}$$

By definition, $s(d_{pen}) = s^*$ and the individual opts-out for sure, therefore...

$$\min_{\psi' \in B(\psi,\kappa)} W(d_{pen}, \psi') = 0 - \gamma(\psi + \kappa) \tag{67}$$

Consider an alternative policy $\bar{d} = \mathbb{E}[s^*]$, i.e. the minimizing opt-out default. From before, we know that...

$$\min_{\psi' \in B(\psi,\kappa)} W(\bar{d}, \psi') = \underbrace{-\frac{\alpha}{2}(\mathbb{E}[s^*] - s^*)^2}_{=-\Lambda \text{ fixed w.r.t. } \kappa} - (\psi + \kappa) \cdot \gamma \cdot \underbrace{\mathbb{P}\{s(\bar{d}) \neq \bar{d}\}}_{=p \text{ small}} \tag{68}$$

Therefore, $\bar{d} \succ d_{pen}$ if

$$\begin{aligned} & -\Lambda - (\psi + \kappa) \cdot \gamma \cdot p > -\gamma(\psi + \kappa) \\ \iff & \gamma \cdot (\psi + \kappa) \cdot (1 - p) > \Lambda \\ & \iff \kappa > \frac{\Lambda}{\gamma \cdot (1 - p)} - \psi \triangleq \bar{\kappa} \end{aligned}$$

where $\bar{\kappa}$ is most likely > 0 given $\psi \approx 0$. I.e. d_{pen} is only a $\kappa - \psi$ robust optimum for at most $\kappa < \bar{\kappa}$. Importantly, note that $\bar{\kappa}$ is **decreasing** in $\gamma = V(s(d_{pen}), d_{pen})$.

■

Corollary 6.1. ***Robust Control and the Optimal Default.*** *Suppose $\Psi = B(\psi, \kappa)$ for some $\kappa > 0$ and $\psi \in \Delta(\Theta)$.*

- *If the expected intrinsic optimum is ψ-optimal, then it is a robust optimum for any κ.*
- *If the extremum default is ψ-optimal, then there is a threshold $\overline{\kappa}$ such that it is the unique robust optimum only for $\kappa < \overline{\kappa}$, otherwise the expected intrinsic optimum is the unique robust optimum. $\bar{\kappa}$ is **decreasing** in γ.*

Proposition 7. *Let $\underline{\psi} \equiv \min_{\psi \in \Psi} \psi(\theta^D)$, and let $\overline{\psi} \equiv \max_{\psi \in \Psi} \psi(\theta^D)$. The robust optimal marginal tax rate given Ψ is*

$$\frac{dT^*(x_1)}{dx_1} = \begin{cases} [1 - \underline{\psi}] \frac{dV(x_1)}{dx_1} & V(x_1) > 0 \\ [1 - \overline{\psi}] \frac{dV(x_1)}{dx_1} & V(x_1) < 0 \\ 0 & V(x_1) = 0. \end{cases} \tag{30}$$

Proof. This result is derived in the discussion in the main text preceding the statement of the proposition. ■

Observation 4. *Overall,*

$$\frac{dW}{d\sigma} = \underbrace{D'_p \{\sigma - (1 - \psi_1)\gamma\}}_{\textit{Internality Correction}} - \underbrace{\psi_2 \cdot \sigma \cdot D'_p}_{\textit{Self-Correction}} - \underbrace{\psi_2 D}_{\textit{Psychic Costs}} \tag{33}$$

Supposing the nudge-elasticity of demand ($\varepsilon^D_\sigma = \frac{\partial D}{\partial \sigma} \cdot \frac{\sigma}{D}$) is a fixed constant yields a formula for the optimal nudge:[52]

$$\sigma^* = \underbrace{\gamma}_{\text{Nudge when } \psi_1, \psi_2 = 0} \cdot \underbrace{\frac{(1-\psi_1)\cdot \varepsilon^D_\sigma}{\psi_2 + (1-\psi_2)\cdot \varepsilon^D_\sigma}}_{\text{Normative ambiguity adjustment}} \tag{34}$$

The optimal nudge is smaller than the case where θ^C is known to be normative ($\psi_1 = \psi_2 = 0$) iff:[53]

$$\overbrace{(\psi_2 - \psi_1)\cdot |D'_p| \cdot \gamma}^{\text{Net Correction}} < \overbrace{\psi_2 \cdot D}^{\text{Psychic Costs}} \tag{35}$$

Proof. Equation (33) follows directly by applying the Leibniz integral rule to Equation (32).

Equation (33), evaluated at the optimal nudge without ambiguity $\sigma^*_{\psi_1,\psi_2=0} = \gamma$ yields

$$\left.\frac{dW}{d\sigma}\right|_{\sigma=\gamma} = -D'_p \cdot \gamma \cdot \{\psi_2 - \psi_1\} - \psi_2 \cdot D < 0$$

Rearranging gives Equation (35). ■

L Proofs for Section C (Optimal Nudges)

Observation 5.

$$t^*_\psi = \mathbb{E}_m[(1-\psi_1)\gamma - (1-\psi_2)\sigma\tau] \tag{52}$$

Proof.

$$\begin{aligned} W &= \int_{v \geq p+t+\sigma\tau-\gamma} v + \psi_1\gamma - \psi_2\sigma\tau - p\, dF \\ \implies \frac{dW}{dt} &= -\mathbb{E}_m[t - \{(1-\psi_1)\gamma - (1-\psi_2)\sigma\tau\}]\cdot D'_p = 0 \\ \implies t^* &= \mathbb{E}_m[(1-\psi_1)\gamma - (1-\psi_2)\sigma\tau] \end{aligned}$$

■

Observation 6. *When the tax is set optimally according to Observation 5:*

$$\left.\frac{dW}{d\sigma}\right|_{t^*} = -\underbrace{D'_p}_{<0}\{(1-\psi_1)Cov_m[\gamma,\tau] - (1-\psi_2)\sigma Var_m[\tau]\} - \psi_2 \cdot \mathbb{E}_I[\tau]\cdot D(p)$$

[52] Our model treats nudges exactly like prices. Importantly, we assume σ is continuous. This allows us to define the nudge-elasticity of demand. In practice, nudges might be more naturally thought of as discrete switches. In this case, the σ^* can be written as a slightly less intuitive function of the *price*-elasticity of demand ε^D_p. In our model, $\varepsilon^D_p = \frac{\sigma}{p}\varepsilon^D_\sigma$. In principle, one could estimate ε^D_σ in the example we consider - graphic labels. The exercise would involve estimating how demand responds to making the graphic label more graphic. Quantifying the % change in graphic-ness is straightforward, now, because it involves estimating the % change in *as-if* costs. These can be uncontroversially elicited through WTP experiments similar to Allcott et al. [2025]. Normative ambiguity concerns have been completely disentangled and only operate through ψ_1, ψ_2.

[53] In the full model (see Appendix C.2) with interpersonal heterogeneity and optimal taxation, an analogous inequality determines whether the government shades the nudge towards zero vs the case where $\psi_1, \psi_2 = 0$. However, in that case the LHS measures the *targeting* benefit of the nudge, rather than the average corrective benefit in the simple model.

If $\psi_2 < 1$, and nudge elasticity is a fixed constant, then we can solve for an interior optimum:

$$\begin{aligned} \left.\frac{dW}{d\sigma}\right|_{t^*} &= 0 \\ \implies \sigma^* &= \frac{(1-\psi_1)\cdot\varepsilon_\sigma^D \cdot Cov_m[\gamma,\tau]}{\psi_2 \cdot \frac{\mathbb{E}_I[\tau]}{\mathbb{E}_m[\tau]} + (1-\psi_2)\cdot\varepsilon_\sigma^D \cdot Var_m[\tau]} \end{aligned} \tag{53}$$

Proof. Again follows directly from Leibniz integral rule and rearranging.

■