Intrapersonal Utility Comparisons as Interpersonal Utility Comparisons:

Welfare, Ambiguity, and Robustness in Behavioral Policy Problems*

Canishk Naik
London School of Economics

Daniel Reck
University of Maryland

January 2, 2025

Abstract

We consider the optimal policy problem of a benevolent planner, who is uncertain about an individual's normative preferences because of inconsistencies in revealed preferences across behavioral frames. We adapt theories of expected utility maximization and ambiguity aversion to characterize the planner's objective, which results in welfarist criteria similar to social welfare functions, with intrapersonal frames replacing interpersonal types. Under paternalistic risk aversion or ambiguity aversion, a policy is less desirable to the planner, holding all else fixed, when it leads to more disagreement about welfare from revealed preferences. We map some examples of behavioral models into our framework and describe how this notion of robustness plays out in applied settings.

*Contact: dreck@umd.edu and c.naik@lse.ac.uk. For valuable discussions and comments, we thank Scott Elliott, Jacob Goldin, Louis Kaplow, Ben Lockwood, Yusufcan Masatlioglu, Emel Filiz-Ozbay, Alex Rees-Jones, Daniel Schaffa, Joel Slemrod, Luminita Stevens, and Dmitry Taubinsky. 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 the real world, Pareto improvements are rare. In evaluating realistic policy reforms, we often confront the question of how to trade off gains to winners and losses to losers. It is an understatement to say that there is no universal consensus among economists, let alone philosophers, about how to approach this problem. But a common practical approach is the use of social welfare functions like

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

Here $\theta \in \Theta$ captures interpersonal heterogeneity, and $x(\theta)$ is the option allocated to type θ, which must be privately optimal under incentive compatibility. In taking this approach to data, we usually think of the utility function $u(x, \theta)$ as being identified by revealed preference from the choices of type θ, while $\psi(\theta)$ reflects a normative judgment on the part of the observer or social planner about the social value of increasing type θ's utility.

A need for normative judgments that cannot be resolved by revealed preference alone also arises in behavioral economics, but here the problem involves different potential views of welfare for a given individual [see e.g. Bernheim and Rangel, 2009]. For example, default options exert influence over choices in many contexts. Revealed preference data from choices under varying defaults suggests that individuals often behave as if there is a cost of opting out of the default [see e.g. Carroll et al., 2009]. But it is unclear whether the "as-if" cost of opting out of the default reflects a normative utility cost or a mistake. In the latter case, we might evaluate welfare using revealed preference information from an "active choice" environment in which default effects are eliminated. A benevolent policymaker seeking to optimize the default therefore faces uncertainty about whether the as-if cost is a normative cost, and this judgment can be pivotal for the optimal policy [Goldin and Reck, 2022].

In this paper, we develop a theoretical approach to the analysis of individual welfare in the presence of behavioral frictions that parallels the conventional approach to the analysis of social welfare with individual heterogeneity. Suppose we reinterpret $\psi(\theta)$ above as a judgment about what utility function the planner should use to evaluate individual welfare, and θ is a *frame*: an aspect of the choice environment which affects choices but not welfare. This paper examines the foundations and practical implications of adopting such an intra-personal welfare criterion to analyze *behavioral policy problems:* optimal policy problems in which we modify incentive compatibility to require that the individual's choices maximize utility in the frame in which the individual chooses [Rees-Jones and Taubinsky, 2018; Danz et al., 2022].

The core assumptions of our framework are as follows. (i) An individual has *normative preferences* that describe what they 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*. With this core setup, the welfare criterion of Bernheim and Rangel [2009], which we label BR-dominance, is formally admissible, and normative preferences must be constant over the frame.

When the normative frame is known, our behavioral policy problem takes a form studied in much prior literature in which normative preferences are known to the planner or point-identified by empirical data [e.g. O'Donoghue and Rabin, 2006; Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019]. We consider problems in which normative preferences are unknown, which we formulate as uncertainty over the normative frame. We provide sets of conditions under which familiar forms of welfare criteria from interpersonal problems represent a benevolent planner's objective. In the first, we endow the planner's preferences over which option the individual consumes with rationality (completeness and transitivity), continuity, and respect for BR-dominance. Noting the similarity of BR-dominance and Pareto dominance, we adapt an argument from Kaplow and Shavell [2001]. So long as there is a good the individual always prefers in strictly larger amounts, the information in revealed preferences in each frame must be sufficient to evaluate the planner's objective, i.e. welfare must take what Kaplow and Shavell [2001] call a "welfarist" form.

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; Sen, 1986; Wakker and Zank, 1999]. For example, if policy A is optimal in one potentially normative frame but policy B is optimal in another, then the planner faces a tradeoff in choosing between A and B, and 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 assume the planner has a unique prior about which frame is normative, or equivalently a unique set of normative weights they attach to each frame, and they trade off welfare under an independence condition as in Von Neumann and Morgenstern [1953]. This assumption implies the planner's objective takes a utilitarian form like equation (1). In another approach, the normative frame is *ambiguous* rather than *probabilistically uncertain* [Knight, 1921; Ellsberg, 1961]. Here, the planner does not have a unique prior about which frame is normative but there is a set of priors they find plausible. Following Gilboa and Schmeidler [1989], we assume the planner's preferences satisfy independence when there is no unresolvable uncertainty (certainty independence), but when faced with unresolvable uncertainty, the planner prefers to hedge (uncertainty aversion). With these conditions, the planner adopts a max-min expected welfare criterion over plausible priors. Leveraging the broader theory on which we draw and the parallel to social welfare, we discuss some additional structure one might impose on the planner's objective, such as paternalistic risk aversion, and alternative ways of axiomatizing welfarist objectives.

We then turn to more practical considerations about how we apply our intrapersonal welfare criteria to policy problems. We examine the usefulness of money-metric equivalent variation, which is often used to cardinalize welfare in applied work. The key questions for us involve whether equivalent variation is sufficiently well-behaved and comparable across frames to be a useful input into our welfare objectives. We provide a set of conditions under which we can use equivalent variation to represent ordinal preferences within frames, and stronger conditions under which equivalent variation from some suitably chosen baseline situation is fully comparable across frames. Intuitively, the key to using equivalent variation for intrapersonal welfare analysis is that in the baseline situation, the individual must have the same utility level across frames and the increase in utility achieved by giving the individual any amount of money is the same across frames.

After analyzing the applicability of equivalent variation, we turn to a broader consideration of comparability, drawing on analogies with prior literature on social welfare. 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].

Next, we examine the perturbation approach to optimal policy analysis under our normative objectives. We compare the first-order welfare effects of policy variation with our approach to the case in which normative preferences are known [e.g. Mullainathan et al., 2012]. One might naively expect that we should simply replace the reduced-form determinants of welfare we find when normative preferences are known with their expected values under unknown normative preferences. We show that these expected values are insufficient to characterize the first-order welfare effect of a policy reform when one of two robustness concepts is at play. Under probabilistic uncertainty, the planner prefers to reduce the variance of realized welfare across normative frames, if they are risk averse over a given welfare metric like money-metric equivalent variation – we clarify below what we mean by paternalistic risk aversion over a given welfare metric. Under ambiguity, meanwhile, welfare effects are expressed as expectations not over a unique prior distribution but over the worst-case scenario, i.e. the distribution that yields minimal welfare over all plausible distributions.

We provide a range of examples illustrating how prior work maps into our framework. Much prior work, including the defaults example above, can be thought of in terms of the question of whether some behavioral phenomenon reflects a bias or a strange preference [e.g. Goldin and Reck, 2022; Reck and Seibold, 2023; Lockwood et al., 2023]. We capture this idea with a two-frame example. We propose a modification of intertemporal selves models with present focus [Laibson, 1997; Laibson et al., 1998] and derive conditions under which we can also think of welfare under present focus in terms of biases versus strange preferences. We show that if the planner assigns equal weight to each present-focused self (c.f. "anonymity" for social welfare functions), then the planner adopts the "long-run view" of welfare regardless of whether they view present focus as a bias. We also present an example in which it is ambiguous whether some feature of the decision-making environment, such as the default itself, is actually a frame.

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 generate normative objectives for behavioral policy problems. Relative to prior work in behavioral welfare economics [reviewed in Bernheim and Taubinsky, 2018], the novelty of our approach primarily lies in the development of robust criteria for selecting policy in the presence of uncertainty or ambiguity about normative preferences. These criteria resolve the incompleteness of the welfare criteria proposed by Bernheim and Rangel [2009] using principled reasoning. They are practically useful because incompleteness makes BR-dominance alone uninformative for a wide range of behavioral policy problems [Benkert and Netzer, 2016].

We illustrate how our criteria and their valuation of robustness play out in three settings: defaults, reference dependence, and corrective taxation with an unknown internality. We show that the first two of these share some common elements in that selecting a default or reference point at the *intrinsic optimum* – the choice the individual would make if they did not care about costs of opting out of the default or gain-loss utility relative to the reference point – is often a robust optimum in both models. Selecting an extreme default or reference point, meanwhile, can be attractive when the behavioral friction is viewed as a bias, but we show that such extreme potential optima are not robust. For corrective taxation, we show that compared to the case where we ignore uncertainty and set the corrective tax rate at the (expected) marginal internality [Mullainathan et al., 2012; Allcott and Taubinsky, 2015], the robust optimal corrective tax is shaded toward the marginal internality according to the worst-case-scenario for welfare.

Outline. In Section 2 we develop our normative criteria from axiomatic foundations. Section 3 discusses comparability, including money-metric welfare. Section 4 describes the first-order welfare effects of policy perturbations. 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. We discuss approaches to resolving the normative uncertainty that is primitive in our theory in Section 7.

2 General Model

2.1 Core Assumptions on Individual Choices and Welfare

Primitives. A choice situation is fully characterized by (σ, θ^D), where $\sigma \in \Sigma$ captures variation in the option set and $\theta^D \in \Theta$ is a decision-making frame drawn from a finite set Θ. For a given σ, the set of options available is $X(\sigma) \in \mathcal{X}$. The map from parameterized situations to menus, $X : \Sigma \to 2^{\mathcal{X}}$, is continuous. We assume $\mathcal{X} \subseteq \mathbb{R}^N$ and that $\mathcal{X}$ is convex. Options are denoted $x = (x_1, ..., x_N)$. To simplify notation, we suppose that for all (σ, θ^D), the individual's choice is unique, so that we can work with choice functions rather than correspondences. The choice function is $x : \Sigma \times \Theta \to \mathcal{X}$ such that $x(\sigma, \theta^D) \in X(\sigma)$. Our model includes two types of preferences, which we describe using standard notation: $\succeq$ for weak preference relations, $\sim$ for indifference, $\succ$ for strict preference, and $\preceq$ for reversed weak preference.

Assumption 1. *Core Assumptions.*

Assumption 1.1. ***Normative Preferences Exist.*** *There is a complete and transitive binary relation $\succeq_*$ defined on $\mathcal{X} \times \Theta$, which represents the individual's normative choice – the choice the individual should make.*

Assumption 1.2. ***Frame-Dependent Rational Preferences.*** *For all $\theta^D \in \Theta$, there is a complete and transitive preference relation $\succeq_{\theta^D}$ on $\mathcal{X}$, such that for any (σ, θ^D), for any $x \in X(\sigma)$, $x(\sigma, \theta^D) \succeq_{\theta^D} x$.*

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

$$x \succeq_{\theta^*} x' \iff (x, \theta^D) \succeq_* (x', \theta^D).$$

Discussion of Core Assumptions. Assumption 1.1 defines the core objective of a benevolent social planner: to maximize welfare according to normative preferences $\succeq^*$. Assumption 1.2 requires that all choice inconsistencies are driven by framing effects: holding the decision-making frame θ^D fixed, revealed preferences are consistent with maximization of a complete and transitive preference $\succeq_{\theta^D}$; this is a restriction on the model of Bernheim and Rangel [2009], which we discuss further below.[1]

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.*

"Behavioral" Characterizations. If we think in terms of a normative choice rule describing what the individual should choose, Assumption 1.1 requires that the normative choice rule satisfies the Generalized Axiom of Revealed Preference (GARP). Assumption 1.2 requires that holding the frame θ^D fixed, the choice function $x(\sigma, \theta^D)$ satisfies GARP [see also Salant and Rubinstein, 2008]. Assumption 1.3 requires that choices in the normative frame from any menu/σ reveal what the individual should choose in any environment (σ, θ^D).

Implications of Core Assumptions. These initial assumptions have two straightforward but useful implications. As we assume the implication of Revealed Preference Coincidence (henceforth RP-Coincidence) holds for any decision-making frame (θ^D), an obvious requirement of RP-Coincidence is that $\succeq^*$ is constant over frames:

Lemma 1. ***Frame Exclusion.*** *Under Assumption 1, for any $x, x' \in \mathcal{X}$ and any $\theta, \theta' \in \Theta$,*

$$(x, \theta) \succeq_* (x', \theta) \implies (x, \theta') \succeq_* (x', \theta').$$

The one-line proof is in the Appendix, as are subsequent proofs. Given Lemma 1, the initial setup of Assumption 1.1 might seem like an odd detour – why not immediately define $\succeq^*$ on $\mathcal{X}$? We find this detour useful for thinking about the set of frames in practice. Lemma 1 defines the property that distinguishes frames from all other features of options: frames are irrelevant for normative preferences. We return to this in Example 1.3, where it is ambiguous whether some aspect of the choice situation is a frame. In such an application, if we tried to define $\succeq^*$ on $\mathcal{X}$ alone, then it would be ambiguous how to do so. Going forward, we suppress the frame θ when describing normative preferences.

Our core assumptions also admit the welfare criterion of Bernheim and Rangel [2009]:

[1] For a more formal articulation of the revealed preference basis for this assumption, see Salant and Rubinstein [2008], Proposition 1. These authors label our Assumption 1.2 "Salient Consideration." A precondition of making Assumption 1.2 is that the set of all possible choice situations is a "rectangular" cross-product, i.e. that any menu may be associated with any frame. Rectangularity is a restriction on Bernheim and Rangel [2009]; it is maintained in Salant and Rubinstein [2008] and many applied papers. We return to this issue when considering inter-temporal choice in Example 1.2.

Lemma 2. ***BR-Dominance.*** *Under Assumptions 1-3, given any $x, x' \in \mathcal{X}$,*

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

Technical Assumptions. We make two more assumptions that could be relaxed in principle, but which 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 $\theta^D \in \Theta$, the sets $\{x \in \mathcal{X} : x \succeq_{\theta^D} x_0\}$ and $\{x \in \mathcal{X} : x \preceq_{\theta^D} x_0\}$ are closed.*

Assumptions 1.2 and 2 give us an ordinal utility function denoted $u(x, \theta^D)$, which represents frame-dependent preferences $\succeq_{\theta^D}$ for given θ^D.

Finally, we assume there is one good that the individual prefers to consume in strictly increasing amounts regardless of the frame. This allows us to employ an argument based on Kaplow and Shavell [2001] below.

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

2.2 Discussion of Setup and Relationship to Prior Literature

Remark on Observability. The setup of our model supposes that frame-dependent preferences $\succeq_\theta$ are known. Some prior literature focuses on identifying $\succeq_\theta$ with limited choice data. In contrast, our focus is normative: how should we map the information in $\succeq_\theta$ to welfare? We do not specify whether knowledge of $\succeq_\theta$ comes from direct observation of choices in every frame, or from an identification condition, by which observed choices in some frames imply choices in others, under decision-theoretical assumptions and/or interpersonal extrapolation conditions [as in Goldin and Reck, 2020; Allcott et al., 2019, discused further below].[2] Relatedly, the essential structure we require of the set of frames Θ is that it contains all the frames in which individuals choose for a given policy (introduced below), and, per Assumption 1.3, all the frames that could be normative. The essential structure we require of the option set $\mathcal{X}$ is that it comprises all information relevant for the evaluation of both welfare and (conditional on the frame) choices.[3]

Core Assumptions and Prior Literature. One feature of our core assumptions compared to prior literature is that we provide a formal definition of a frame. Bernheim and Rangel [2009] verbally resist the assumption that a normative preference exists, but they impose the Frame

[2] For an example in which decision-theoretical axioms facilitate empirical identification, refer to Masatlioglu and Ok [2005]. Their model is not a normative one, but they present assumptions under which behavior in a potentially normative frame – in their case one where status quo bias is eliminated – can be inferred from behavior in more naturally occurring / observed frames. Intuitively, the idea is that taking the view that status quo bias is in fact a bias, non-status-quo (i.e. active) choices reveal normative preferences. This idea underpins identification strategies used in the literature on default effects, see Example 1.1.

[3] If the process of choosing matters for welfare due to e.g. emotions experienced while choosing, what we label revealed preferences would account for preferences over choice processes (or the emotional outcomes they generate). Bernheim et al. [2024] develop an extension to traditional revealed preference analysis that accommodates this possibility. Our general model may be applied in this type of environment, but some of our examples preclude the idea; see further discussion in Example 2.

Exclusion condition from Lemma 1 informally when defining frames (which they originally labelled *ancillary conditions*). Without a notion of well-being rooted in the concept of a normative preference ($\succeq_*$), we have difficulty formalizing the premise that by definition, frames are features of choice situations that "have no direct bearing on well-being, but instead impact biases" [Bernheim and Taubinsky, 2018].

Does assuming a normative preference exists contradict theories of choice commonly adopted in psychology? Psychological theories of choice suggest that preferences are constructed endogenously at the moment of choice [see e.g. Lichtenstein and Slovic, 2006]. This is one reason we might resist the assumption that normative preferences exist [see also Bernheim, 2016; Bernheim and Taubinsky, 2018]. Assuming θ^* is known to an observer or social planner, as in prior work assuming knowledge of normaive preferences, seems vulnerable to this critique. Assuming θ^* *exists but is unknown*, however, allows for a deeper analysis of normative ambiguity and judgments, which in turn helps us understand potential sources of disagreement about whether and how we might resolve normative ambiguity. Moreover, the possibility that the normative frame might be fundamentally ambiguous rather than just uncertain in a probabilistic sense motivates our consideration of ambiguity below in the tradition of Knight [1921]. This perspective allows us to derive principled solutions to the "incomplete ordering" problem of Bernheim and Rangel's welfare criteria, as discussed in the introduction.

Assumption 1.2 is a restriction on the model of Bernheim and Rangel [2009] requiring that all inconsistencies in choice are driven by framing. Relatedly, without Assumption 1.2, the dominance criterion in Lemma 2 is not equivalent to Bernheim and Rangel's criterion. This restriction imposes some useful discipline. Assumption 1.2 rules out limited attention/consideration, where inconsistencies arise across menus within a frame because the individual does not always consider all available options. Masatlioglu et al. [2012] show that Bernheim and Rangel's welfare criterion is not generally admissible in models of limited attention: it could be the case that in all choice situations where two options x and x' are available, the individual might only pay attention to x and choose it even though they prefer x'. Limited attention therefore threatens not only Assumption 1.2, but also RP-coincidence (1.3). Ruling out limited attention by Assumption 1.2 frees us from this problem, and more generally our core assumptions clarify when BR-dominance is generally admissible, in response to the critique of BR-dominance by Masatlioglu et al. [2012]. We conjecture that one could nevertheless modify the core setup of our model to accommodate welfare analysis in some models featuring limited attention.[4]

We argue that Assumption 1.3, RP-Coincidence, is implicitly or explicitly assumed in all literature from behavioral welfare economics that seeks to use choices to inform welfare. For example, in Chetty et al. [2009], RP-coincidence is implied by the assumption that when a tax

[4] One route could be to follow Bernheim and Rangel [2009] and relax Assumption 1.2 to hold within a"welfare-relevant" subset of situations in which the individual understands/considers all their options. But our objective is to analyze judgments within the model, so we do not wish to require judgments overturning revealed preferences ex ante. A more attractive solution could be to introduce an attention cost that rationalizes apparent inconsistencies due to limited attention. Whether to respect this cost or to consider welfare under an alternative, perfect-attention frame would then align with the type of question we take up in Example 1 below [see also Goldin and Reck, 2022; Bronchetti et al., 2023]. However, to formalize such an approach to welfare analysis, the attention cost would need to depend on the menu $X(\sigma)$. We are not confident about how allowing $u(x, \theta)$ to depend flexibly on σ would interact with other ideas we confront below (e.g. comparability), so we defer this to future work.

is salient, the individual chooses optimally. Many prior studies adopt a similar assumption, ruling out behavioral biases other than the bias that is the focus of the model, implying that RP-coincidence obtains in a frame in which the bias of interest is fully alleviated.

2.3 Behavioral Optimal Policy Problems

Notation. We now introduce some notation that helps us think about policy variation and welfare. One of the components of $\sigma = (P, \tilde{\sigma})$ is a policy $P \in \mathcal{P} \subseteq \mathbb{R}^{N_P}$ affecting the individual's option set; the other components $\tilde{\sigma}$ are suppressed where they are obviously held fixed. The decision-making frame may also depend on policy; where this is dependence is relevant we write $\theta^D(P)$, but otherwise we suppress this input.

2.3.1 Known Normative Frame

We begin, for instructive purposes, with the case where the normative frame θ^* is known, i.e. there is no normative uncertainty/ambiguity in the model. A benevolent planner's 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].[5]

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

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

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 more specific structural models [e.g. O'Donoghue and Rabin, 2006].

2.3.2 Unknown Normative Frame

Now we wish to characterize a benevolent planner's objective when the normative preference is unknown. Our policy problem is re-written as

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

Unlike problem (3), here we do not impose that $w(x)$ coincides with utility under a particular normative frame. The indirect utility function under BIC is now denoted $W(P) \equiv w(x(P, \theta^D(P)))$.

[5] 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 $\mathcal{P}$.

What structure should we impose on the planner's objective? The planner's preferences over which option the individual consumes are denoted by a relation $\succeq_w$ on $\mathcal{X}$.[6] In writing (4), we are already imposing that $\succeq_w$ has a representation $w : \mathbb{R}^N \to \mathbb{R}$. 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, property (5) does imply that a benevolent planner should respect BR-dominance. We therefore begin with the following structure on $\succsim_w$:

Assumption 4. ***Basic Structure on Planner's Preferences.***

Assumption 4.1. ***Rationality.*** *$\succeq_w$ is complete and transitive.*

Assumption 4.2. ***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 4.3. ***Weak BR-dominance.*** *For any $x, x' \in \mathcal{X}$, if $x \succeq_\theta x'$ for every θ, then $x \succeq_w x'$.*

Assumptions 1.1 and 1.3 justify BR-dominance per Lemma 2; we now impose this directly via Assumption 4.3. How does the planner make policy decisions when BR-dominance does not apply and the planner does not have perfect knowledge of which choices are normative? Turning to the parallel with social welfare, adapting an argument from Kaplow and Shavell [2001], we find that our assumptions so far require that $w(x)$ takes what they call a "welfarist" form [see also Sher, 2023].[7]

Proposition 1. *Maintain Assumptions 1.2, 2 and 3. Assumption 4 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 by Kaplow and Shavell [2001] 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. However, our assumptions so far place 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'. In this case, the planner faces a tradeoff between welfare under θ and welfare under θ'. All that we know given the conditions of Proposition 1 is that the planner must find some way to evaluate such tradeoffs, but we do not know how.

[6] Note we are using the implication of Lemma 1 in positing that $w(x)/\succeq_w$ is independent of the frame.

[7] There is some disagreement about whether the concept of a "welfarist" criterion requires additional conditions like the independence assumption introduced below; the way Kaplow and Shavell use the term it does not, but see also the comment by Fleurbaey et al. [2003]. This is a purely semantic issue for our purposes.

2.4 Probabilistic Uncertainty versus Ambiguity

In this section we explore conditions under which we can move beyond Proposition 1 and characterize how the planner weighs intrapersonal tradeoffs. But first, we review the concpetual challenges involved.

Comparability. One question we must now confront, as with the theory of social welfare functions [see e.g. Sen, 1986], is comparability, in our case comparability of utility across normative frames. We have rich economic theory of optimality in the presence of tradeoffs that seems useful for analyzing such policy problems, but in order to use this theory we require a notion of cardinal comparability [Debreu, 1959]. We introduce conditions that imply the existence of a utility function that is fully comparable (in level and unit) across normative frames below, and we return to the conceptual question of how we might approach comparability in intrapersonal problems in practice in Section 3 and in our examples.

Which Decision Theory to Adapt? Given a practical resolution to the comparability question, we confront a related modelling choice. There are multiple distinct theories from normative decision theory that we might apply here. A unifying analysis of many different approaches by Wakker and Zank [1999] offers the insight that in any model we might use, the key structure will indeed be how the planner trades off realized welfare across potentially normative frames. In our view, the most important distinction between potential approaches is whether we can think of the planner as having well-formed beliefs about the weight they ought to attach to each frame. We therefore seek to develop two approaches: one in which the planner does have such beliefs and one in which they do not. For the first approach, we adapt classical expected utility theory from Von Neumann and Morgenstern [1953], and for the other we adapt the model of ambiguity aversion from Gilboa and Schmeidler [1989]. Approaches in which normative weights/probabilities emerge as a consequence of assumptions about more abstract primitives are available [e.g. Savage, 1954; Maskin, 1978, for decision theory and social welfare, respectively]; we use simpler assumptions that are familiar to a broader audience.

An assumption, implicit or explicit, in canonical models of decision-making under uncertainty is that exactly one state is the true state. In our model where the state corresponds to the normative frame, this is explicitly assured under Assumptions 1.1 and 1.3, so we re-introduce these assumptions for the remaining results.

2.5 Known Normative Weights/Probabilistic Uncertainty

Primitives and their Interpretation. Here, we assume the planner weighs potential realizations of welfare according to a distribution over Θ, denoted $\psi \in \Delta(\Theta)$. For example, in the case where θ^* is known, only one view of welfare receives positive weight so ψ is degenerate. We assume here that ψ is unique, which makes the next result a characterization of welfare under *known normative weights*, or *probabilistic uncertainty*. In the next section, we relax this assumption and consider ambiguity.

The set of normative weights has multiple potential interpretations. One interpretation is to think of the weights as Bayesian beliefs: $\psi(\theta)$ is the probability θ is normative given the planner's information. Another interpretation is to think of normative weights as philosophical judgments about how to weigh the different views of welfare implied by revealed preferences under different frames. A third interpretation, which we discuss further in Appendix A, involves thinking of $\Delta(\Theta)$ as the convex hull of the set of discrete normative preference parameters implied by Θ. In this last case, the Independence assumption introduced below can be interpreted as a linearity condition. Which interpretation of the frames we use makes no difference for our purposes here, but would matter when considering how to infer normative weights from empirical data (see Section 7).

Re-defining the Planner's Objective. Adapting classical 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 $\left(\psi(\theta_1), ..., \psi(\theta_{|\Theta|})\right)$ and a vector of outcomes $\left((x, \theta_1), ..., (x, \theta_{|\Theta|})\right)$. 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 4 as follows:

Assumption 5. ***Expected Utility Assumptions Over Intrapersonal Lotteries.***

Assumption 5.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 5.2. ***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 5.3. ***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 5.4. ***Independence.*** *For any x, any $L_1(x), L_2(x), L_3(x) \in \mathcal{L}$, and any $p \in [0, 1]$*

$$L_1(x) \succsim_w L_2(x) \implies pL_1(x) + (1-p)L_3(x) \succsim_w pL_2(x) + (1-p)L_3(x). \tag{7}$$

Proposition 2. *Maintain Assumptions 1, 2 and 3. Then Assumption 5 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*

$$w(x; \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) u(x, \theta^*). \tag{8}$$

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

Discussion of Proposition 2 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 5.4 implies a cardinalization of utility, which is the function $u(x, \theta)$ in (8). Although we find that some fully comparable utility function that is suitable for evaluation of intrapersonal tradeoffs must exist under Assumption 5, Proposition 2 still does not shed any 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 Proposition 2, in which we consider a specific welfare metric $v(x, \theta)$ that represents ordinal preferences, proves useful for thinking about comparability below.

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 2.1. *Maintain Assumptions 1, 2, 3,and 5, and consider a utility function $u(x, \theta)$ that gives the representation in Proposition 2. 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*

$$w(x; \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) \omega(v(x, \theta)). \tag{9}$$

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

The function ω converts the welfare metric $v(x, \theta)$ into the cardinal units the planner uses to conduct welfare comparisons across frames. Below, 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.

2.5.1 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 2.1.

Corollary 2.2. ***Variance Representation*** *Assume the function ω from representation (9) is twice differentiable. Then up to second-order Taylor approximation of ω, the planner's objective is*

$$w(x, \psi) \approx \omega\Big(E_\psi[v(x, \theta)]\Big) + \frac{\omega''\Big(E_\psi[v(x, \theta)]\Big)}{2} \cdot Var_\psi\Big[v(x, \theta)\Big] \tag{10}$$

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

We say that the planner's preferences exhibit *paternalistic risk aversion over v* if $\omega'' < 0$ and *paternalistic risk neutrality over v* if $\omega'' = 0$. Corollary 2.2 motivates our first notion of robustness, which arises under paternalistic risk aversion over a welfare metric v under probabilistic uncertainty. 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 second notion of robustness below, whether this notion of robustness is relevant is specific to the welfare metric we have in mind. Proposition 2 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

$$\omega(v) = \begin{cases} \frac{v^{1-\eta}}{1-\eta}, & \eta \neq 1 \\ \log(v) & \eta = 1. \end{cases} \tag{11}$$

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

2.6 Ambiguity

A potential objection to the approach to welfare analysis implied by our results in the previous section is that the planner may not know what weights they should attach to each frame. With the Bayesian interpretation, the planner may not have a unique prior about which frame is normative; with a more philosophical interpretation, the planner may not have a principled way of specifying a unique set of normative weights. Here we adapt the model of ambiguity aversion due to Gilboa and Schmeidler [1989] – which in turn was motivated by the seminal contributions of Knight [1921] and Ellsberg [1961] – to propose a welfare criterion that accommodates this possibility.

Primitives. Rather than representing the planner's beliefs by a unique distribution $\psi \in \Delta(\Theta)$, here we present conditions under which the planner is endowed with a set of distributions $\Psi^* \subseteq \Delta(\Theta)$, which we interpret as a set of weights/probabilities the planner finds acceptable/plausible.[8]

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. Abandoning Assumption 5.4, we introduce conditions on the planner's preferences drawn from Gilboa and Schmeidler [1989].

Assumption 6. ***Ambiguity Aversion Assumptions.***

Assumption 6.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 6.2. ***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.*

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

Assumption 6.3. ***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 6.4. ***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 6.5. ***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 6.6. ***Non-degeneracy.*** *There exists $L, L' \in \mathcal{L}$ such that $L \succ_w L'$.*

The six conditions we present here correspond to the six axioms of Gilboa and Schmeidler [1989], except we replace their weak monotonicity assumption with weak BR-dominance and our definition of a constant lottery involves constant a constant payoff $u(x,\theta)$ across frames θ, which is the analogue in our model of a "constant act" in their framework. We note that this obviously imposes comparability of u across frames, in a more direct fashion than Assumption 5.4 above. Relative to Assumption 5, Assumption 6.3 weakens Assumption 5.4 so that it only holds when we mix a given pair of lotteries with a constant lottery. Assumption 6.6 rules out the degenerate case where the planner is indifferent across all policies/options.

Crucially, Assumption 6.5 implies that when the planner is weighing welfare under ambiguity over two plausible distributions $\psi_1, \psi_2 \in \Psi^*$, the planner prefers to *hedge*.

Proposition 3. ***MaxMin Welfare Under Ambiguity Aversion.*** *Maintain Assumptions 1, 2 and 3. Assumption 6 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*

$$w(x) = \min_{\psi \in \Psi^*} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x,\theta^*)) \right\}. \tag{12}$$

2.6.1 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 A and B.

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 (12) becomes a more global max-min criterion:

$$\min_{\psi \in \Delta(\Theta^*)} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x, \theta^*)) \right\} = \min_{\theta \in \Theta^*} u(x, \theta) \tag{13}$$

Building on social welfare theory, rather than using Assumption 6, one could derive this criterion this using an analogue of Rawls' [1971] Difference Principle: assume there is some representation of $\succeq_\theta$, u, such that $x \succeq_w x'$ 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 (13).[9] 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 6.5 together with a global notion of ambiguity.[10]

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 (12) for $\Psi^* = \Delta(\Theta)$. The global max-min criterion is the closest analogue to Rawlsian social welfare in our framework.

Corollary 3.1. ***Intersection of Various Objectives at Global Max-Min***

- *If $\Psi^* = B(\kappa, \psi)$, for any ψ, the planner's objective in (12) coincides with the global max-min criterion for $\kappa > 1$.*
- *If $\Psi^* = \Delta(\Theta^*)$, the planner's objective in (12) and/or (13) 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 (9) with the functional form in equation (11) – approaches the global max-min criterion as $\eta \to \infty$.*[11]

[9] 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.

[10] 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].

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

2.6.2 Robust Optimality under Ambiguity Aversion

Next we make some observations about the notion of robust optimality we find when the planner is ambiguity averse. It is useful here and in applications below to define three notions of optimality:

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

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

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

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:

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

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 introduce the cardinal *disagreement* in welfare between some frame θ and the decision-making frame θ^D:

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

Proposition 4. ***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*

$$P^* \in \arg\min_{P \in \mathcal{P}} \max_{\psi' \in \Psi^*} \sum_{\theta \in \Theta} \left(\psi'(\theta) - \psi(\theta)\right) \cdot V\left(x(P, Z, \theta^D), \theta, \theta^D\right). \tag{14}$$

The condition from Proposition 4 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 (14).

3 Approaches to Comparability

In this section, we discuss comparisons of welfare across normative frames. Comparability is famously controversial in the theory of social welfare functions, so we do not expect to propose an approach that will be universally acceptable. Our point is that the same difficulties we wrestle with in writing down a social welfare function also arise in behavioral policy problems. To think through this, we consider one common approach and its applicability in intrapersonal problems, and then we broaden our perspective and consider the parallel to debates about comparability involving interpersonal welfare.

3.1 Money Metric Welfare

In this section, we explore the relationship between money-metric equivalent variation and the welfarist criteria we developed above.

We begin with some notation and assumptions that discipline equivalent variation. We introduce a new feature of the choice environment denoted $Z \in \mathbb{R}$; choice environments are now described using $\sigma = (P, Z, \tilde{\sigma})$ and the decision frame θ^D.

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

Assumption 7.1. ***Strict Monotonicity in Money.*** *For any two Z, Z' any two θ^D, θ, and any $P, \tilde{\sigma}$,*

$$Z > Z' \iff x(P, Z, \tilde{\sigma}, \theta^D) \succ_\theta x(P, Z', \tilde{\sigma}, \theta^D).$$

Assumption 7.2. ***Continuity over Money.*** *For any $(P, Z, \tilde{\sigma})$ and any two $\theta^D, \theta \in \Theta$, the sets $\{Z' : x(P, Z', \tilde{\sigma}, \theta^D) \succsim_\theta x(P, Z, \tilde{\sigma}, \theta^D)\}$ and $\{Z' : x(P, Z', \tilde{\sigma}, \theta^D) \precsim_\theta x(P, Z, \tilde{\sigma}, \theta^D)\}$ are closed.*

Assumption 7.3. ***Equalizability.*** *For any $x \in \mathcal{X}$, any $P, \theta^D, \tilde{\sigma}$ and any θ^*, the sets $\{Z' : x(P, Z', \tilde{\sigma}, \theta^D) \succ_{\theta^*} x\}$ and $\{Z' : x(P, Z', \tilde{\sigma}, \theta^D) \prec_{\theta^*} x\}$ are non-empty.*

Discussion of Assumption 7. Combined with RP-Coincidence, Assumption 7.1 ensures that giving the individual more Z always improves welfare in the normative frame. One reason Assumption 7.1 might fail is if, for example, in an "addicted" frame θ^D, the individual spends all their money on an addictive substance that is not a "good" but a "bad" from the perspective of some other potentially normative frame.[12] Assumption 7.2 implies 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 θ^* is normative.

Together, these assumptions discipline *Equivalent Variation* (EV) in any (potentially normative) frame θ, which is defined as ζ such that for a given *baseline* (P_0, Z_0, θ_0^D),

$$x \sim_\theta x(P_0, Z_0 + \zeta, \theta_0^D). \tag{15}$$

Lemma 4. ***Existence and uniqueness of EV.*** *Under Assumptions 1.2, 2 and 7, for any x, any $\theta \in \Theta$ and any (P_0, Z_0, θ_0^D), equivalent variation ζ exists and is unique. Moreover, $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ represents the individual's ordinal preferences $\succeq_\theta$.*

Lemma 4 implies that for a given baseline, $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ is a unique representation of revealed preferences under θ, $u(x, \theta)$. Combining the representation result in Lemma 4 with the idea in Proposition 1, we find that introducing Assumption 7 yields the following:

Proposition 5. ***Planner's Preferences and Equivalent Variation.*** *Under Assumptions 1.2, 2, 4, and 7, for any baseline P_0, Z_0, θ_0^D, there is a function $\mathcal{W}_\zeta : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planners preferences are represented by $w(x) = \mathcal{W}_\zeta \left(\{\zeta(x; \theta^*, P_0, Z_0, \theta_0^D)\}_{\theta^* \in \Theta}\right)$.*

[12] If we are willing to assume the "addicted" frame cannot be normative, is straightforward to relax this assumption to accommodate this possibility and rescue the validity of money-metric welfare.

Proposition 5 suggests that provided that equivalent variation is well-behaved per Assumption 7, the model will allow us to use equivalent variation welfare metrics to describe the welfare effects of local policy perturbations. Practical applications of such approaches in interpersonal problems often leave the marginal value of a dollar under a given type/frame θ, $\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta}$, unspecified [e.g. Hendren and Sprung-Keyser, 2020]. Imposing additional structure on this aspect of the planner's objective allows us to go further in describing such a welfare weight. With the structure we impose under known normative weights (Assumption 5), we observe that the application of Corollary 2.1 to a representation rooted in equivalent variation requires ordinal level comparability between normative, cardinal utility $u(x,\theta)$ and $\zeta(x,\theta;P_0,Z_0,\theta_0^D)$. Our next result characterizes when ordinal level comparability obtains for some baseline situation. To economize on notation we express the option the individual chooses in a given baseline as $x_0 \equiv x(P_0,Z_0,\theta_0^D)$.

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 or 3. There is a baseline situation (P_0,Z_0,θ_0^D) under which the following conditions hold:*

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

$$u(x_0,\theta) = u(x_0,\theta').$$

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

$$u(x(P_0,Z_0+\zeta,\theta_0^D),\theta) - u(x_0,\theta) \geq u(x(P_0,Z_0+\zeta',\theta_0^D),\theta') - u(x_0,\theta') \iff \zeta \geq \zeta'.$$

Lemma 5. ***Ordinal Level Comparability of Equivalent Variation.*** *Maintain Assumptions 1, 2 and 7. Let $u(x,\theta)$ be a cardinal utility function from the representation in Proposition 2 or 3. Assumption 8 holds if and only if there is some baseline (P_0,Z_0,θ_0^D) such that $u(x,\theta)$ and $\zeta(x,\theta;P_0,Z_0,\theta_0^D)$ exhibit ordinal level comparability.*

Combining the previous Lemma with Corollary 2.1, we obtain the following proposition:

Proposition 6. *Under Assumptions 1, 2, 5, 7 and 8, there is a function $\omega_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0,Z_0,θ_0^D) such that the planner's preferences are represented by*

$$w(x,\psi) = \sum_{\theta\in\Theta} \psi(\theta)\omega_\zeta(\zeta(x,\theta;P_0,Z_0,\theta_0^D)). \tag{16}$$

Under Assumptions 1, 2, 6, 7 and 8, there is a function $\omega_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0,Z_0,θ_0^D) such that the planner's preferences are represented by

$$w(x,\Psi) = \min_{\psi\in\Psi^*} \sum_{\theta\in\Theta} \psi(\theta)\omega_\zeta(\zeta(x,\theta;P_0,Z_0,\theta_0^D)). \tag{17}$$

Discussion of Proposition 6. With the additional structure on the planner's objective we impose in Propositions 2 and 3, the structure imposed on the welfare weights is governed by normative beliefs ψ/Ψ^* and the (cardinal) value of money, ω_ζ for an admissible baseline. We turn to the structure of ω_ζ in the next section, building on the parallel to interpersonal differences in the value of money.

Bernheim and Rangel [2009] derive bounds on equivalent variation from BR-dominance alone. Proposition 6 provides stronger conditions under which the lower bound identified by their approach is in fact a sufficient statistic for optimal policy: under the structure we impose on equivalent variation (Assumption 7 and Assumption 8 for a suitably chosen baseline), with a global maxmin criterion per Corollary 3.1, maximizing welfare is equivalent to maximizing the minimum of equivalent variation.

Happiness-Metric Equivalent Variation? Assumptions 7 and 8 describes the key properties of the variable Z in this approach to welfare analysis. These assumptions do not require that the variable Z be money, and they are distinct from the structure usually used to derive equivalent variation in practice (budget constraints, expenditure functions etc). The variable Z could be identified with something in the real world other than money, such as a mental state.

3.2 Relationship to Interpersonal Comparability

What structure might we impose the function in ω_ζ? If we endow the individual with Von Neumann-Morgenstern preferences over intrapersonal lotteries, then we could show that the individual's preferences over lotteries will also have an expected utility representation like (8), and if Assumptions 7 and 8 hold, we could then formulate their Bernoulli utility function as a transformation of the monetary equivalent payoff in any situation, $\mu(\zeta)$. Then respecting BR-dominance would require $\omega = \mu$, individual risk aversion would imply $\omega'' < 0$, and more fundamentally we would be required to use the intrapersonal utilitarian criterion from Proposition 2 rather than the ambiguity averse criterion from Proposition 3. This line of reasoning echoes a classic argument for utilitarianism due to Harsanyi [1955], and the intuition behind it resembles a "veil of ignorance" or "impartial observer" thought experiment for intrapersonal problems. Related arguments not relying on notions of objective probabilities/lotteries are found in d'Aspremont and Gevers [1977] and Maskin [1978].

However, endowing individuals with vNM preferences over intrapersonal lotteries and using this as the basis for welfare analysis could be controversial, as argued for interpersonal problems in Sen [1976] and Weymark [1991]. Behind the (intrapersonal) veil of ignorance, an individual may be willing to sacrifice any amount of utility while well-off to get a little more utility while in the worst-case scenario. From a more Bayesian perspective, an impartial observer may have no idea which frame is normative (or what is the probability that each frame θ is normative) and therefore prefer to hedge their bets. These lines of reasoning lead to criteria like our maxmin criteria. Implementing this criterion using equivalent variation requires comparability of equivalent variation in the worst-case scenario, and sufficient information to determine what is the worst-case scenario (Assumptions 7 and 8 ensure both).

There are two differences between intrapersonal problems and interpersonal analogues we find noteworthy. First, in some behavioral models, we find ordinal level comparability plausible because there is a set of situations where the revealed preferences in different frames agree about the value of money. If we assume that the level of utility is also comparable across frames in these situations, then if we use one of these situations as our baseline, we obtain Assumption 8. For money-metric equivalent variation (or another welfare measure with similar properties), the comparability problem is then simplified to specifying ω_ζ. However, this insight is model-specific: it applies in some of the examples below but not all of them. Second, even if we do adopt the intrapersonal utilitarian perspective and wish to specify structure on ω_ζ, assuming individuals are endowed with vNM preferences in order to resolve the comparability question following Harsanyi [1955] could be controversial, because in some behavioral models, departures from Expected Utility Theory are core to the behavioral phenomenon of interest. If we wish to entertain that these departures reflect normative preferences rather than biases (e.g. Example 1.3 below), the rationale for an approach like Harsanyi's fails.

Leveraging insights about specific classes of models from behavioral decision theory might reveal more and better ways to approach comparability. For example, Ellis and Masatlioglu [2022] present a model in which an individual evaluates options by partitioning the menu space into categories based on a particular option – they call this option the reference point but it could also be interpreted as a default or status quo – and then uses a different utility function to evaluate options within each category. Theorem 2 of their paper presents conditions under which cardinal comparisons of differences in utility *across categories* are well-defined, which implies that if we consider the various categories as frames, intrapersonal welfare comparisons of interest to us will be well-behaved. Moreover, we obtain full comparability in their Affine Categorical Thinking Model if we are willing to assume the level of utility can be assumed to be the same across all categories when the individual chooses the reference point.

4 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.

4.1 Unstructured Welfare Weights

We begin with the least restrictive representation of the planner's objective from Proposition 1: $w(x) = \mathcal{W}(\{u(x, \theta)\}_\theta)$, where $u(x, \theta)$ can be any representation of frame-dependent ordinal preferences. We also suppose equivalent variation representations of preferences are well-behaved so that we can think in terms of willingnesses to pay (Assumption 7); this is common in prior work employing the perturbation approach but not essential. Let us express the indirect utility function under a given normative frame θ^*, imposing BIC, as $W^*(P, Z, \theta^D; \theta^*) = u(x(P, Z, \theta^D); \theta^*)$. Assuming all the necessary differentiability requirements are met in some status quo situation (P, Z, θ^D), for a policy perturbation dP, we have:

$$dW(P, Z, \theta^D) = \sum_{\theta^* \in \Theta} \frac{\partial \mathcal{W}(\{W^*(\cdot; \theta^*)\}_{\theta^*})}{\partial W^*(\cdot; \theta^*)} \frac{\partial W^*(P, Z, \theta^D; \theta^*)}{\partial Z} d\zeta(x, \theta^*; P, Z, \theta^D). \tag{18}$$

where $d\zeta(x,\theta^*;P,Z,\theta^D)$ is the equivalent variation of dP given a normative frame θ^* and the baseline (P,Z,θ^D).

$$d\zeta = \zeta(x(P+dP,Z,\theta^D),\theta^*;P,Z,\theta^D) = \frac{\frac{\partial W^*}{\partial P}}{\frac{\partial W^*}{\partial Z}} dP. \tag{19}$$

In other words, each $d\zeta$ term is the money-metric willingness to pay for the marginal reform when the normative frame is θ^*.

The product $\frac{\partial \mathcal{W}}{\partial W^*(\cdot;\theta^*)} \frac{\partial W^*(\cdot;\theta^*)}{\partial Z}$ from equation (18) resembles a welfare weight like those used often in contemporary optimal tax theory for interpersonal problems [e.g. Saez, 2001; Hendren and Sprung-Keyser, 2020]. As alluded to above, essentially all we know about this term with the structure of Proposition 1 is that this weight is weakly positive. The additional structure introduced in subsequent results above imposes more structure on this term. Under Assumption 5/Proposition 2, we have $\frac{\partial \mathcal{W}}{\partial W^*(\cdot;\theta^*)} = \psi(\theta^*)$. The second part of the weight is the marginal value of a dollar under frame θ. If we further impose paternalistic risk neutrality, i.e. $u = \zeta$, the second part of the welfare weight becomes $\frac{\partial W^*_\zeta(\cdot;\theta^*)}{\partial Z} = 1$ by construction and evaluating whether the derivative in equation (18) is positive is analogous to a Kaldor-Hicks criterion (using equivalent rather than compensating variation). But this requires Assumption 8 obtains for the status quo policy as the baseline. More generally, under Assumption 8 for a baseline (P_0, Z_0, θ_0^D), we have $\frac{\partial W^*_\zeta(\cdot;\theta^*)}{\partial Z} = \frac{d\omega_\zeta}{d\zeta}$, the derivative of ω_ζ with respect to ζ, evaluated at the status quo $\zeta(x(P,Z,\theta^D),\theta^*;P_0,Z_0,\theta^D)$. In this sense the differential value of a dollar accross frames is accounted for when we specify an admissible baseline to obtain Assumption 8; going from a local to a global optimality condition rooted in equivalent variation further requires accounting for potential nonlinearity in the value of money ω_ζ.

Technical Aside. We do not apply equations (18) and (19) to marginal policy reforms that vary Z itself (e.g. monetary transfers) or the decision-making frame θ^D. Allowing Z to depend on P is a simple extension. Allowing θ^D to depend on P is more complicated because our set of frames is discrete. We present an extension with a continuous set of frames in Appendix A; there, one could use a perturbation approach to examine the welfare effects of marginal variation in the frame.

4.2 Perturbations with Fully Specified Objectives

With the structure we impose on intrapersonal tradeoffs in Corollary 2.1 and Proposition 3, we can use the envelope theorem to further characterize the welfare effect of marginal policy changes. Doing so requires a slight modification of our setup. 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. Disagreements are now denoted using $V(x,\theta) = u(x,P,\theta^D) - u(x,P,\theta)$.

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 Corollary 2.1/Equation (16), where we think of u as a money-metric utility function that is ordinally level comparable to cardinal utility with diminishing utility over money $\omega_\zeta'' < 0$ (Paternalistic Risk Aversion), and one rooted in Proposition 3 (Ambiguity Aversion) given a fully comparable utility function.

4.2.1 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[13]

$$\frac{\partial W(P,Z,\theta^D)}{\partial P} = \underbrace{E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right]}_{\text{Direct Effect}} - \underbrace{\frac{\partial x(P,Z,\theta^D)}{\partial P}}_{\text{Beh. Resp.}} \cdot \underbrace{(1-\psi(\theta^D))E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial x}\middle|\theta \neq \theta^D\right]}_{\text{Marginal Internality}} \tag{20}$$

The term $\frac{\partial u(x,P,\theta)}{\partial P}$ in equation (20) 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].[14] 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 (20)? We can see the answer for the internality term in equation (20). For the direct effect term, we observe that

$$E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right] = \frac{\partial u(x,P,\theta^D)}{\partial P} - [1-\psi(\theta^D)]E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial P}\middle|\theta \neq \theta^D\right]. \tag{21}$$

4.2.2 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 $\omega_\zeta'' < 0$, we find an intuitive representation that leverages the mean-variance characterization from Corrollary 2.2. 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):

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

[13] 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 $\frac{\partial V}{\partial x}$.

[14] 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].

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 ω_ζ,

$$\frac{\partial W(P, Z(P), \theta^D)}{\partial P} \approx \omega'(\overline{W}_\zeta)\frac{\partial \overline{W}_\zeta}{\partial P} + \frac{\omega''(\overline{W}_\zeta)}{2} \cdot \frac{dVar_\psi\left[V_\zeta(x(P, Z, \theta^D), P, \theta)\right]}{dP}. \tag{22}$$

By construction, equations (20) and (21) 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 Corrollary 2.2): given $\omega''_\zeta < 0$ 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. We simply assume a utility function that is comparable across frames. We do consider the variance of a given utility function over frames and typically interpret this under paternalistic risk aversion. This can be interpreted in terms of money-metric utility as a welfare metric, or more broadly as any utility function rooted in Corollary 2.1.

4.2.3 Under Paternalistic Ambiguity Aversion

Now we turn to the ambiguity averse objective from Proposition 3. 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

$$\frac{\partial W}{\partial P} = \frac{\partial \overline{W}(P, Z, \theta^D; \psi^*)}{\partial P}. \tag{23}$$

This welfare effect is the same as $\frac{\partial \overline{W}}{\partial P}$ above (direct effects and behavoiral 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.[15]

5 Examples

We next illustrate via examples how prior work on behavioral welfare economics fits within our framework.

5.1 Example 1: Biases Versus Strange Preferences

Let us introduce a running example in which the key intrapersonal question is whether some behavioral phenomenon arises due to a bias or a normative preference. Suppose the decision-

[15] 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.

making frame is some fixed frame θ^D and there is just one alternative frame denoted θ^A. Our representation of welfare from equation (9) becomes

$$w(x) = \psi(\theta^D)u(x,\theta^D) + [1-\psi(\theta^D)]u(x,\theta^A). \tag{24}$$

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

$$\begin{align} w(x) &= u(x,\theta^D) - [1-\psi(\theta^D)]V(x) \tag{25}\\ &= u(x,\theta^A) + \psi(\theta^D)V(x). \tag{26}\\ u(x,\theta^D) &= u(x,\theta^A) + V(x). \tag{27} \end{align}$$

Prior work on behavioral frictions often uses a formulation like equation (26), where we think of $V(x)$ as the "behavioral" component of preferences, while "decision utility" takes a form like equation (27). 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. We now denote elements of $\mathcal{X}$ by (x,d), where the first element is a choice object and the second is the default, a fixed feature as in the previous section. In any situation $\sigma = (d,\tilde{\sigma})$, both of these are drawn from a set of available options: $d, x \in X(\tilde{\sigma})$. To nest the fixed cost model of default effects in Example 1 we specify

$$V(x,d) = -1\{x \neq d\}\gamma. \tag{28}$$

where γ is the fixed cost of choosing some option other than the default [see e.g. Carroll et al., 2009; Bernheim et al., 2015]. The fixed opt-out cost structure matches key empirical aspects of default effects, and this structure nests a variety of mechanisms by which default effects might influence behavior [Goldin and Reck, 2022].[16] But depending on the mechanisms and our normative interpretation of them, the fixed cost may or may not be a normative cost.

The alternative frame implied by (28) is the utility function individuals maximize when they opt out of the default and make an active choice. Drawing parallels between this example and the next, we label the utility function $u(x,\theta^A)$ "intrinsic utility." When $\psi(\theta^D) = 1$, we impose that γ reflects a welfare-relevant cost; when $\psi(\theta^D) = 0$, γ reflects a bias.[17]

In treating default adherence as a "biases versus strange preferences" question, we do not al-

[16] This empirical pattern is observed in widely varied contexts [Choi et al., 2006; Haggag and Paci, 2014; Altmann et al., 2013; Brown et al., 2013] but not everywhere [Brot-Goldberg et al., 2023]. 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. One could employ our overall normative approach to evaluate defaults while allowing for this possibility, but this is not nested in Example 1.1.

[17] When a real cost is inflated above its true value, for instance due to present bias, we capture this possibility by $0 < \psi(\theta^D) < 1$. Convexifying the possible views of welfare in this way also effectively captures views of welfare according to which fixed costs are partially but not fully normative, e.g. models of present bias.

low active choosers to make mistakes. This restriction is relaxed in the more general version of the model in Goldin and Reck [2022], but doing so obviously requires introducing more frames than the two we posit here. An analogous limitation to Example 1 applies in general: in models of biases versus strange preferences, the modeller picks one behavioral factor to consider as a bias or a strange preference, and assumes away deviations from individual welfare maximization due to any other behavioral factor to obtain RP-coincidence (1.3).

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 previous example, but replace the default d with a reference point $r \in X(\tilde{\sigma})$. When researchers model reference-dependent choice, they introduce a utility function over x 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 (x, r). We can nest this in our biases versus strange preferences setup if we posit a naturally occuring 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.[18] Following Kőszegi and Rabin [2006], we assume intrinsic utility is additively separable, so that we may write $u(x, r, \theta^A) = \sum_{i=1}^N u_i(x_i)$. For parameters $\Lambda_i > 0, \beta \in (0, 1]$, we specify a gain-loss payoff of the form

$$V(x, r) = -\sum_{i=1}^{N} 1\{x_i \leq r_i\}\Lambda_i \left[u_i(r_i) - u_i(x_i)\right]^{\beta}, \tag{29}$$

The individual only incurs a payoff along some dimension if they incur a loss, $x_i \leq r_i$. 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].[19]

Example 1.1 and Example 1.2 under $\beta = 1$ are both cases of the "Affine Categorical Thinking Model" from Ellis and Masatlioglu [2022]. The salience model of Bordalo et al. [2012] is also

[18] We borrow the term "naturally occuring frame" from Bernheim et al. [2015].

[19] 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.

an Affine Categorical Thinking Model. 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 to these models because they 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 modelled 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. [2023] present a model in which the main behavioral friction is probability re-weighting and it is ambiguous 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, i.e. more weight than expected utility requires given the extremely low probability of winning a jackpot. This finding suggests a particular form of probability re-weighting, and the main question they confront for welfare analysis is whether this jackpot payoff effect reflects a bias or a normative preference.

To nest their model in Example 1, we think of each component of $x = (x_1, ..., x_N)$ as the payoff that is realized for each realization of an uncertain state variable. Objective probability of each realized state is $\pi = (\pi_1, ..., \pi_N)$; this is the main aspect of situations (σ) that is relevant for the behavioral friction. Individuals re-weight each objective probability according to a function $f(\pi)$. Individuals 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,

$$V(x, \pi) = \sum_n [\pi_n - f(\pi_n)]\mu(x_n). \tag{30}$$

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. [2023] model the extent to which re-weighting reflects a bias with a parameter that is isomorphic to $\psi(\theta^A)$ here.

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 now $x = (x_1, ..., x_T)$. 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:

$$u(x, \tau) = \mathbb{1}\{\tau > 0\}\delta^\tau \mu(x_\tau) + \beta \sum_{t \neq \tau} \delta^t \mu(x_t). \tag{31}$$

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 modelling choice below. A commonly adopted approach to welfare analysis in models like this is to respect the revealed preferences of the period 0 self, sometimes called the "long-run view." The period 0 self is a classical exponential discounter, and in fact we find that a planner's welfarist objective based on formulation (31) has a representation along similar lines to the Biases versus Strange Preferences example. By construction, the planner's welfare function takes the following form:

$$\begin{aligned} w(x) &= \beta \sum_{t=1}^{T} \delta^t \mu(x_t) + \sum_{\tau=0}^{T} \psi(\tau)\mathbb{1}\{\tau > 0\}(1-\beta)\delta^\tau \mu(x_\tau) \\ &= u(x,0) + (1-\psi(0)) \sum_{\tau=1}^{T} \psi(\tau|\tau > 0)[u(x,\tau) - u(x,0)] \end{aligned} \tag{32}$$

This formulation for welfare resembles the formulation for welfare from Example 1, Equation (26): 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 imposed on social welfare functions, which requires that given a fully no two individuals should have their utility differently weighed [e.g. Maskin, 1978]. Anonymity is not a useful assumption for our model in general because we do not have an objective distribution of types (a set of individuals) over which to impose it, but it is intuitive to impose such a condition over the $\tau > 0$ selves. Doing so, we find a justification for the planner's adopting the long-run view of welfare, which *does not require assuming that present focus is a behavioral bias.*

Proposition 7. *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 7 presents a new justification for the long-run view of welfare, which contributes to a debate in the literature about the normative justification (or lack thereof) for adopting the long-run view in welfare analysis [see e.g. Caliendo and Findley, 2019]. The assumption that the extent of present focus β is constant over time seems noteworthy. This assumption rules out, for example, that individuals are present focused while they are young but not when they are old. In that case, the planner must make a material judgment about how to weigh the present focus payoff that is only present for the younger selves.

Remark on Intertemporal Selves 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. $\tilde{u}(x_{t \geq \tau}, \tau) = \mathbb{1}\{\tau > 0\}\mu(x_\tau) + \beta \sum_{t=\tau+1}^{T} \delta^{t-\tau}\mu(x_t)$. The formulation differs from most prior work on intertemporal selves in that the period τ self is endowed with classically discounted preferences over consumption in periods $t < \tau$. This approach appears to fix multiple related issues with the intertemporal selves model identified by Bernheim and Rangel [2009], at the cost of being, admittedly, philosophically confusing.

What does it mean for the period τ self, who cannot go back in time to choose a different amount of consumption, to have preferences over past consumption? Loosely speaking, we address this question by assuming that the period-τ self agrees with most of their prior selves about intertemporal consumption tradeoffs, so that endowing this self with preferences over past consumption does not generate any new choice inconsistencies relative to those we find for observable, forward-looking choices. More formally, we assume that for any pair $\tau > 0$, $\tau' > 0$, if we consider two consumption plans x, x' such that $x_\tau = x'_\tau$ and $x_{\tau'} = x'_{\tau'}$ then we have $u(x, \tau) \geq u(x', \tau) \iff u(x, \tau') \geq u(x', \tau')$. Behaviorally, the period-τ and period-τ' selves make the same choices when we hold consumption in τ and τ' fixed in the menu. This approach works well for the β-δ model but appears to be less well-suited to more generic models of non-classical discounting.

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; Bernheim and Rangel show that a lack of rectangularity in the naive application of "intertemporal-self Pareto optimality" leads to conceptual problems. If we naively write down a utility function in which the selves care only about current and future consumption, allocating all resources to the last-period self will always be an intertemporal-self Pareto optimum. But revealed preference does not suggest that the individual robustly prefers options that defer all consumption to the final period. We address this problem by adopting rectangular preferences using formulation (31) and considering preferences under commitment. Players in conventional games have deep structural preferences over outcomes influenced by the actions of prior movers even when they cannot practically go back in time and change another player's actions; we make a similar assumption for the intrapersonal game here.

Why focus on revealed preferences under commitment to define welfare? Here, we are making an ex ante normative assumption that the consumption plan and not the process of choice (and the associated emotions, etc.) is sufficient for the evaluation of normative preferences. For example, we require that a naive agent who makes a plan and fails to adhere to it will not experience shame that alters their normative choices. This can also be viewed as an assumption about the set of potentially normative frames, which precludes revealed preferences in situations featuring non-commitment from being normative. This type of assumption, a version of which we make generally in defining normative preferences over options rather than menus, is criticized in Bernheim et al. [2024]. They also develop tools for relaxing it using additional information to identify the normative import of emotions. We do not know how our argument on the robustness of the long-run view would be altered by accounting for additional normative concerns due to choice processes and emotions. We defer a fuller treatment of this question to future work.

Finally, we remark on comparability of welfare across intertemporal selves. The units of utility with our formulation are determined by $\mu(x)$, which is cardinal (so that the individual can evaluate intertemporal tradeoffs) [Montiel Olea and Strzalecki, 2014]. We require an assumption that that the units of μ are the normative units for welfare analysis, but then comparisons of utility across τ with the formulation above are well-defined. For $\tau > 0$, level comparability also does not seem to be a problem: evaluating equation (31), we find that a constant consumption growth path generates the same level of utility for any $\tau > 0$. But when we compare $\tau = 0$ versus $\tau > 0$, examining equation (31), we find that 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 due to the present-focused payoff that only the $\tau > 0$ self receives. Formally,

$$\forall \tau, \mu_\tau(0) = 0 \text{ and } \forall x \geq 0, \frac{d\mu_\tau(x)}{dx} \geq 0 \implies \forall x \geq 0, \min_{\psi \in \Delta(\Theta)} \psi(\tau) u(x, \tau) = u(x, 0). \tag{33}$$

This logic implies that the globally ambiguity averse planner adopts the long run view of welfare. Whether this is another novel rationale for the long-run view or an artefact of a dubious level comparability assumption is debatable. This is a moot point when we introduce the anonymity condition from Proposition 7. If it is a problem, the solution requires specifying a consumption path in which the level of utility is equal between $\tau = 0$ and $\tau > 0$ selves.

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

In Example 1.1, the default cannot be a frame by construction; the same is true of the reference point in Example 1.2. If we treat default adherence as a normative preference then by definition the default should not be a frame,[20] but if we do not, then it could be one and we might think of choices made given each default as coming from distinct frames rather than a unitary, naturally occurring frame. In the present example, there is a component of the situation σ, labelled d and drawn from a finite set D,[21] and the frame has two components: $\theta = (\theta_1, \theta_2)$. The first component, $\theta_1 \in \{0, 1\}$ indicates whether the second component, $\theta_2 \in D$, can really be viewed as a frame ($\theta_1 = 1 \implies \theta_2/d$ is a frame), i.e. whether we obtain the frame exclusion condition from Lemma 1.

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

[20] To see this, suppose two options and the individual chooses x_1 when x_1 is the default and x_2 when x_2 is the default (x_1 and x_2 are in the menu for both of these choices). When θ^D is the normative frame, this implies the individual's normative preference depends on which option is the default, which violates Frame Exclusion (Lemma 1) if we regard the default itself as a frame.

[21] We require a finite set of frames so D must be finite to nest this example in our general model, but we view this as a technical issue and do not expect much to change e.g. when D is a continuum.

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 exclusion requires $u(x,d,1,\theta_2)$ to be constant over d, which we express with a function $u_1(x,\theta_2) = u(x,d,1,\theta_2)$. Denoting 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 letting $\psi^0 = \sum_{\theta_2} \psi(0,\theta_2)$ be the total weight on $\theta_1 = 0$, we derive an identity similar to equation (25):

$$w(x) = u_0(x,d) - (1-\psi^0) \sum_{\theta^2 \in D} \frac{\psi(1,\theta_2)}{1-\psi^0} V(x,d,\theta_2). \tag{34}$$

The weight ψ_0 is similar to $\psi(\theta^D)$ in Example 1. With this setup, we confront more ambiguity than in the biases versus strange preferences case from Example 1. For instance, if the planner knew with certainty that the effect of d on choices reflects a bias, then this resolves all ambiguity in Example 1 ($\psi(\theta^D) = 0$), but if analogously $\psi_0 = 0$ in (34), substantial ambiguity in welfare remains due to choice inconsistencies as d varies. We note that the 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, but this is done in the interest of providing simple illustrations of our robustness 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 Optimality

In this section, we explore how the robustness concepts we developed above play out in some of our examples, focusing mainly on models nested by Example 1.

6.1 Robustness and Perturbations

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

$$Var_\psi(V) = \psi(\theta^D)(1-\psi(\theta^D))V(x,P)^2. \tag{35}$$

Evaluating the change in welfare from a policy reform due to the change in variance – the second term from equation (22) – we find:

$$\frac{\omega''(\overline{W})}{2} \cdot \frac{dVar_\psi\left[V(x(P,Z,\theta^D),P)\right]}{dP} = \omega''(\overline{W})Var_\psi(V) * \frac{1}{V}\frac{dV}{dP} \tag{36}$$

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 $V = -1\{x \neq d\}\gamma$, we have $V = -Pr[x \neq d]\gamma$.[22] The right-hand side of (36) becomes

$$\omega''(\overline{W})\underbrace{\psi(\theta^D)(1-\psi(\theta^D))Pr[x \neq d]^2\gamma^2}_{Var_\psi(V)}\left\{\frac{1}{Pr[x \neq d]}\frac{\partial Pr[x \neq d]}{\partial d}\right\} \tag{37}$$

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

6.2 Robust Optimality Under Ambiguity Aversion

Now we return to the notions of optimality defined in Section 2.6.2 and explore how these play out in some of these examples.[23]

6.2.1 Optimal Defaults

We begin with the optimal defaults problem studied in Carroll et al. [2009]; Bernheim et al. [2015]; Chesterley [2017]; Goldin and Reck [2022], and others. In the model we introduced in Example 1.1, the *intrinsic optimum* $x^* \equiv \arg\max_x u(x, \theta^A)$ is assumed to be known to the social planner.[24] We begin there, and then consider the case where the intrinsic optimum is not known, which which makes the planner's normative objective equivalent to aggregate welfare in Bernheim et al. [2015] and social welfare in Goldin and Reck [2022]. Aggregation

[22] 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.

[23] Deriving characterizations of robust optimality under probabilistic uncertainty is straightforward but not very instructive beyond what we do here. Note that our work on perturbations has essentially already characterized the first-order condition for an interior (local) optimum.

[24] The intrinsic optimum x^* 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, x^* does not depend on the default/reference point by construction; it obviously does depend on other aspects of the choice situation encoded in σ, e.g. prices. We suppress dependence of x^* on σ for convenience. We suppose x^* is unique for simplicity.

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, $x \in \mathbb{R}$. We suppose utility is approximately quadratic: $u(x, \theta^A) = -\frac{\alpha}{2}(x - x^*)^2$ for a known parameter $\alpha > 0$ and the intrinsic optimum x^*. For simplicity, we further assume the opt-out cost γ is known and x^* 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. Applying our definitions of optimality, we observe the following, which turn out to be true in full generality, i.e. without any of the restrictions introduced in the previous paragraph:

- The ψ-optimal defaults are the intrinsic optimum x^* and any default under which the individual chooses actively.[25]
- The intrinsic optimum x^* is the unique globally robust optimum.
- An active choice optimum is robust if and only if there is no ambiguity (Ψ^* is singleton) and $\psi(\theta^D) = 0$.

The fact that we find a globally robust optimum in this case clearly depends on the assumption that x^* is known, i.e. that the policymaker can set as the default the option that the individual would choose if they opt out. In this case, setting that option as the default is ensured to give the individual the best possible option and avoid any potentially normative opt-out cost. Relaxing the assumption that the planner knows x^*, we find the following characterization of robustness in the quadratic/Gaussian case:

Proposition 8. ***Robust Optimal Defaults when the Intrinsic Optimum is Unknown***

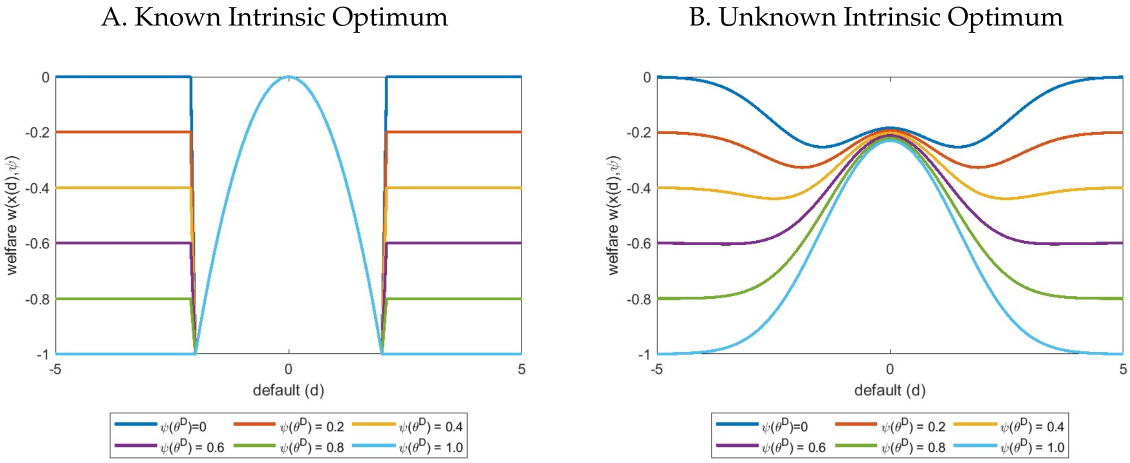
- *The ψ-optima are the expected intrinsic optimum and the most extreme default possible in the positive or negative direction (henceforth the* extremum default*).*
- *None of the ψ-optima are globally robust.*
- *If the expected intrinsic optimum is ψ-optimal for some ψ in the interior of Ψ^*, the expected intrinsic optimum is the unique robust optimum.*
- *If the expected intrinsic optimum is not ψ-optimal for any $\psi \in \Psi^*$, the extremum default is the unique robust optimum.*
- *If the expected intrinsic optimum is ψ-optimal for some ψ on the boundary of Ψ^* but not the interior, both the expected intrinsic optimum and the extremum default are robust optima.*

Corollary 8.1. ***Robust Control and the Optimal Default.*** *Suppose $\Psi^* = B(\kappa, \psi)$ for some $\kappa > 0$ and some $\psi \in \Delta(\Theta)$.*

If the extremum defualt is ψ-optimal, there is a threshold $\overline{\kappa}$ such that

[25] Formally, the set of ψ-optima is $\{d | d = x^* \text{ or } d < \underline{x} \text{ or } d > \overline{x}\}$, where $\underline{x}$ and $\overline{x}$ are the thresholds for active choice. These thresholds are equal to -2 and +2 in the illustration in Figure 1A.

Figure 1: Illustration of the Optimal Default


A. Known Intrinsic Optimum

B. Unknown Intrinsic Optimum

- *the extremum default is the unique robust optimum for* $\kappa < \overline{\kappa}$, *but*
- *the expected intrinsic optimum is the unique robust optimum for* $\kappa > \overline{\kappa}$.[26]

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

The logic of the proof is illustrated by Figure 1B. 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 Bernheim and Rangel [2009] is inapplicable, provided that sufficiently extreme defaults (i.e. those where sufficiently many individuals make active choices) are feasible. Even so, there is still a sense in which the setting the expected intrinsic optimum as the default (i.e. minimizing opt-outs) is a more robust policy recommendation than an extremum default. 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). The intuition that this makes the the extremum default a less robust optimum appears in Goldin and Reck [2022]; here we find that formalizing an approach to robustness allows us to capture that intuition.[27]

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 nor-

[26] In the knife-edge case $\kappa = \overline{\kappa}$, both the extremum default and the expected intrinsic optimum are κ-ψ robust.

[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 Gastell [2021].

mative objective with some interesting commonalities to the previous example.

We suppose options are two-dimensional $x = (x_1, x_2)$, and that intrinsic utility is quasi-linear with x_2 the numeraire. The individual faces a budget constraint for given prices and income (components of σ), with price of p_2 normalized to 1 and $p_1 = p$. The form of gain-loss utility follows from equation (29).

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

$$px_1 + x_2 = Z.$$

$$V(x_1, x_2, r_1, r_2) = -1\{x_1 \leq r_1\}\Lambda_1[\log(r_1) - \log(x_1)]^\beta - 1\{x_2 \leq r_2\}\Lambda_2[r_2 - x_2]^\beta \tag{38}$$

We are interested in whether and when the planner might wish to induce the individual to use a different reference point. Evidence from the lab and the field suggests that policy reforms can indeed *shift reference points to some extent* [e.g. Homonoff, 2018; Rees-Jones, 2018; Seibold, 2021]. However, the full policy space $\mathcal{P}$ is difficult to characterize in applied settings where reference dependence appears to matter, because there is little consensus about how to model the formation of reference points. Here, we consider an environment with fixed prices and incomes, and we suppose that the planner can set the reference point at any point on the budget constraint: $\mathcal{P} = \{(r_x, r_y) | pr_x + r_y = Z\}$.[28] This confers a likely unrealistic amount of power on the planner to shape the individual's reference point, but with this approach we nevertheless find a thought-provoking characterization of robustness. To see why, first note that the model admits a reduced-form representation for welfare in terms of a single dimension of choice, x_1, that has some common features with the previous example. Assuming an interior solution, for fixed p and Z, we can re-write intrinsic utility as:

$$u(x_1, \theta^A) = \log(x_1) + Z - px_1.$$

$$V(x_1, r_1) = \begin{cases} -\Lambda_1[\log(r_1) - \log(x_1)]^\beta, & x_1 \leq r_1 \\ -\Lambda_2[px_1 - pr_1]^\beta, & x_1 > r_1. \end{cases} \tag{39}$$

The *intrinsic optimum* is here characterized by $x_1^* = \frac{1}{p}$; for numeric illustration here we simply set $p = 0.1 \implies x_1^* = 10$.[29] To simulate welfare in the model, we suppose $\Lambda_1 = \Lambda_2 = 0.5$, we set $Z = 10$. We express welfare in equivalent variation units relative to a baseline where $x_1 = r_1$,[30] and further normalize this as a share of income for interpretability.[31]

Figure 2 plots welfare as a function of the reference point for x_1, where the reference point for good 2 is then pinned down by the budget constraint. In the first panel, we rule out di-

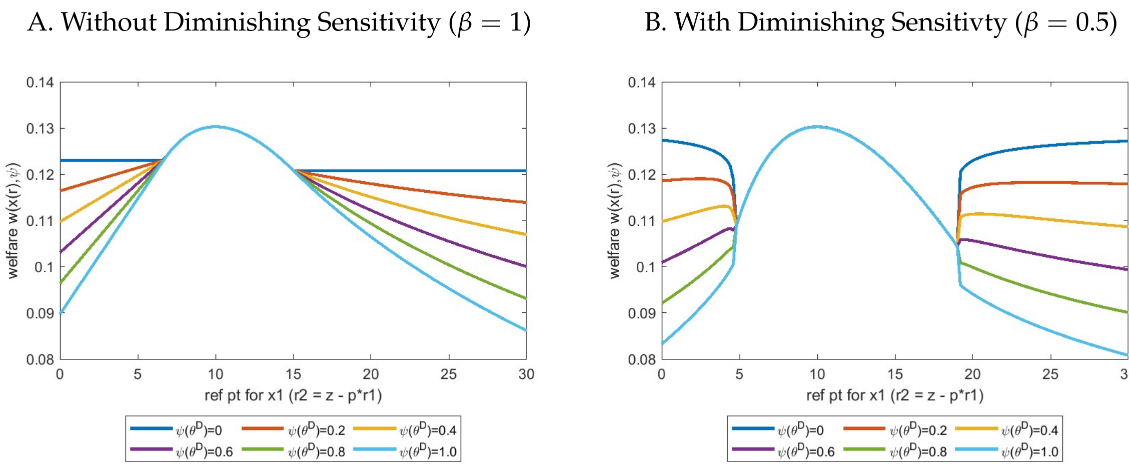
[28] If we relax the restriction that (r_1, r_2) must lie on the budget constraint, the globally robust optimum is for the planner to set the lowest reference point possible along each dimension; see Reck and Seibold [2023] Appendix B for further discussion.

[29] Varying prices is a straightforward extension building on our work in the next subsection. Introducing price variation requires specifying how such variation affects the reference point.

[30] Under our quasilinariy assumption using any baseline where $x_2 \geq r_2$ ($\iff$ $x_1 \leq r_1$) would give the same quantitative expressions for welfare, but outside the quasilinear case, using $x = r$ as the baseline ensures we obtain Assumption 8. Intuitively, because loss aversion modifies marginal value of a dollar according to revealed preferences under θ^D but not θ^A, it seems most sensible to compare money-metric welfare under these two frames from a baseline where the individual incurs no losses.

[31] In other words, we plot $\tilde{w}(x, \psi) = \frac{E_\psi[u(x,\theta)] - Z}{Z}$.

Figure 2: Illustration of Optimal Reference Points


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

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

minishing sensitivity ($\beta = 1$), and in the second we include it, supposing $\beta = 0.5$. Without diminishing sensitivity, we find that the intrinsic optimum is the globally robust optimum, as in the defaults model under known x^*, and in this case it is also the unique ψ-candidate optimum for any ψ. That the optimal reference point is the intrinsic optimum when $\psi(\theta^D) = 1$ is key to the Preferred Personal Equilibrium concept of Kőszegi and Rabin [2007]. In a non-stochastic environment, selecting a Preferred Personal Equilibrium from the set of Personal Equilibria is equivalent to solving our planner's problem under $\psi(\theta^D) = 1$. In fact, we observe here that the planner – or an individual setting a reference point to maximize the welfare of their future, reference-dependent self as in Fudenberg and Levine [2006] – would also want to set the intrinsic optimum as the reference point for any $\psi(\theta^D)$ [see Reck and Seibold, 2023, for further discussion].

We observe that welfare behaves differently in three domains in both panels of Figure 2A. To understand why, first observe that when the reference point is on the budget constraint, the individual can either consume the reference point itself, a bundle with $x_1 < r_1$ and $x_2 > r_2$ (a loss over good 1) or a budle with $x_1 > r_1$ and $x_2 < r_2$ (a loss over good 2). When the reference point falls around the intrinsic optimum of $x_1^* = 10$, the individual chooses the reference point, so $V(x, r) = 0$ because there are no gains or losses, and their welfare is peaked around 10 because this is the intrinsic optimum. At a very high reference point for good 1, the individual chooses to consume some $x_1 > x_1^*$ to reduce their losses over good 1 due to Loss Aversion. Similarly at a very low reference point for good 1, the individual consumes more x_2 to reduce losses in x_2 and therefore consumes less of good 1 than x_1^*. Without diminishing sensitivity x_1 is constant over r in the latter two cases, so under $\psi(\theta^D) = 0$, welfare is flat. When $\psi(\theta^D) > 0$, changing the reference point has direct welfare effects, by increasing the magnitude of losses, and consequently, welfare falls further as r moves to further extremes and the losses grow.

We introduce diminishing sensitivity in Figure 2B. Here, welfare unsurprisingly behaves similarly in the domain where $x = r$ but we find non-monotonicity in welfare outside this domain:

more extreme reference points appear to be more desirable for small $\psi(\theta^D)$. The intuition here is similar to penalty defaults above: under diminishing sensitivity, as losses grow to an extreme, the individual stops trying to avoid losses. When the losses themselves receive little welfare weight (($\psi(\theta^D)$ is near zero), this is desirable, because the planner believes that the individual *should* stop trying to avoid losses. However, when loss aversion is viewed more as a normative preference ($\psi(\theta^D)$ is near 1) the direct negative welfare effect of imposing extreme losses on the individual makes extreme reference points highly undesirable. Like extremum defaults, reference points that generate extreme losses are desirable under $\psi(\theta^D) = 0$ but this desirability is *not robust*. Based on what we found in Figure 1, it is obvious that if we introduced some uncertainty about the intrinsic optimum, we could even get an extreme reference point to be ψ-optimal for some sufficiently small $\psi(\theta^D)$, but this will tend not to be robust just as in Proposition 8.

That our notion of robustness plays out very similarly in the defaults and reference points models (compare Figure 1B and Figure 2B) can be understood as an implication of Proposition 4. In both examples, setting the default or reference point at the (expected) intrinsic optimum minimizes disagreements about welfare across frames.

6.2.3 Corrective Taxation

Let us consider optimal corrective taxation in the biases versus strange preferences example. 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

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

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 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 is

$$T^*(x; \psi) = [1 - \psi(\theta^D)]V(x) + C, \tag{40}$$

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,

[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.

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 (see equation 20).[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 9. *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*

$$\frac{dT^*(x_1)}{dx_1} = \begin{cases} [1-\underline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) > 0 \\ [1-\overline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) < 0 \\ 0 & V(x_1) = 0. \end{cases} \tag{41}$$

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. While we do not examine in detail the joint optimality of a two-dimensional policy involving e.g. a default and a corrective tax, we observe that at a robust optimum in which the sufficient condition from Proposition 4 obtains (e.g. opt-out minimization for defaults), $V(x_1)$ will be small or zero. This in turn suggests a small or zero corrective tax at the joint optimum.

7 Discussion: Identifying Normative Weights.

Much of the recent literature on behavioral welfare economics, including work we drew on in our examples, contains analysis that attempts to resolve the uncertainty about normative preferences that is primitive in our model. Using the model in Example 1.3, for instance, Lockwood et al. [2023] seek to identify the normative weight on the probability-weighting term ($\psi(\theta^D)$) with additional data analysis using a Counterfactual Normative Consumer identification strategy. A formal account of all of the different ways one might justify and identification strategy for normative weights is beyond the scope of this paper. Instead we discuss some intriguing similarities between these prior approaches to intrapersonal welfare analysis and attempts to identify interpersonal welfare weights.

The simplest approach to pinning down ψ is to assume the answer from some normative principle, e.g. by aggregating welfare using inverse-consumption weights in interpersonal problems. Such a treatment of intrapersonal welfare is found, for example, in the assumption by O'Donoghue and Rabin [2006] that the "long-run view" of welfare is normative, which is derived from the normative principle that inter-temporal preferences should be time-consistent

[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:

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

– Bernheim [2009] provides a contrary perspective. A more sophisticated version of this approach is the "Counterfactual Normative Consumer" approach [Allcott and Taubinsky, 2015; Goldin and Reck, 2020; Allcott et al., 2019; Lockwood et al., 2023]. This approach leverages information on the revealed preferences of "debiased" individuals or experts, assuming 1) experts are not subject to framing effects, 2) we can observe the experts' revealed preferences, and 3) expertise is independent of preferences. We discuss this approach and how entertaining failures of these assumptions might lead the planner to employ our proposed approaches to robustness in Appendix B.

Second, researchers use revealed preference methods that approximate the "veil of ignorance" or "impartial observer" thought experiment. In the ideal version of this experiment, an individual, being aware of framing effects but not subject to them (being aware of interpersonal inequality but having not been assigned a type), makes choices that require them to trade off welfare under various frames (types). Under some assumptions, choices in such an ideal experiment are implied by choices in feasible experiments. For example, Saez and Stantcheva [2016] and Capozza and Srinivasan [2023] experimentally estimate interpersonal welfare weights by having participants make choices to reveal their willingness-to-pay to transfer income from person A to person B, varying the incomes of A and B. In Allcott and Kessler [2019], the authors implement a meta-choice approach by eliciting the willingness to pay to be nudged and interpreting this elicitation as if individuals know how to evaluate the potentially biased choices they will make after getting the nudge or not.

Lastly, many recent studies use implemented policies such as tax schedules [Hendren, 2020; Lockwood and Weinzierl, 2016] and transfer policies [Hendren and Sprung-Keyser, 2020] to reveal social welfare weights. The idea is to use the chosen policy to reverse-engineer the weights which would have meant that policy was optimal. This is an interesting approach which could be applied to behavioral problems. For example, if a social planner sets a "penalty-default" in our defaults example, they reveal that their intrapersonal welfare weight sets $\psi(\theta^D) \approx 0$. More ambitiously, we can imagine inferring policymakers' normative judgments about present focus from the design of illiquid/mandated savings policies, which is similar to the central exercise in Beshears et al. [2020].

The important point for us is that these methods all require untestable normative judgments. As such, we view welfare analysis in our framework – analyzing how uncertainty about normative judgments matters for optimal policy using our notions of robustness – as a useful complement to tools that help us identify what the appropriate normative judgment might be. If there is even a little room for doubt about the validity of the approaches summarized here, our framework provides a way to assess the importance of such doubts for optimal policy.

8 Conclusion

The core argument of our paper is that a primary obstacle to the development of behavioral welfare economics – the question of how to do welfare analysis when we get conflicting information from revealed preferences – is the intrapersonal analogue of an older and more familiar problem: interpersonal comparisons of utility. We exploit the parallel between interpersonal and intrapersonal problems to develop criteria for the welfarist evaluation of policy in the presence of uncertainty about an individuals' normative preferences, and we explored what insights the resulting criteria might provide, in general and in the context of specific examples.

Showing that welfare in intrapersonal and intrapersonal problems can be modelled very similarly could be interpreted optimistically or pessimistically, depending on one's views about how economists typically approach interpersonal comparisons. From a pragmatic perspective, our results give applied researchers the tools to conduct welfare analysis when they wish to respect some revealed preferences but are unsure how to resolve ambiguity in revealed preferences. Specifically, we provide applied researchers the conceptual tools to separate empirical quantities that are informative for policy (e.g. what is the magnitude of potential internalities, how does behavior respond to a policy reform, etc) from normative judgments about how exactly these quantities map to an optimal policy. We require some potentially objectionable normative assumptions like those surrounding comparability, but after having made these assumptions we can ask how disagreements about normative judgments and ambiguity map to disagreements or ambiguity in the optimal policy. While such an approach is far from universally accepted in interpersonal problems, it has become popular in recent decades because it allows economists to inform optimal policy using empirical data to an extent, without taking a stand on difficult normative questions like the value of equity.

We identify some opportunities for future work. From a theoretical perspective, a few generalizations of our results are available, upon overcoming some technical challenges. One could formalize the derivation of a normative criteria without endowing the planner with primitive beliefs, e.g. using the approach like that of Savage [1954] or Maskin [1978]. More ambitiously, it might be possible to extend our framework to accommodate some models of limited attention, as discussed in Section 2.2, or models in which the process of choosing introduces potentially normative concerns [Bernheim et al., 2024], as discussed in our present focus example. One could also derive more and better ways to think about comparability and/or to discipline the set of frames using behavioral decision theory, like Ellis and Masatlioglu [2022]. From a more applied perspective, our work on how our notions of robustness play out in specific models barely scratches the surface of what is possible. Future work could examine what more we can learn by applying our approach to optimality and robustness to additional behavioral policy problems.

References

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

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.

Altmann, S., Falk, A., and Grunewald, A. (2013). Incentives and Information as Driving Forces of Default Effects. Working paper.

Benkert, J.-M. and Netzer, N. (2016). Informational Requirements of Nudging. Working paper.

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

Bernheim, B. D. (2016). The good, the bad, and the ugly: A unified approach to behavioral welfare economics1. *Journal of Benefit-Cost Analysis*, 7(1):12–68.

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. and Gastell, J. M. (2021). Optimal default options: The case for opt-out minimization. Nber working paper 28254.

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 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. (2020). Optimal illiquidity. Nber working paper no. 27459.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *The 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.

Brown, Z., Johnstone, N., Haščič, I., Vong, L., and Barascud, F. (2013). Testing the Effect of Defaults on the Thermostat Settings of OECD Employees. *Energy Economics*, 39:128–134.

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.

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.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2006). Saving for Retirement on the Path of Least Resistance. In McCaffery, E. J. and Slemrod, J., editors, *Behavioral Public Finance: Toward a New Agenda*. Russell Sage Foundation.

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.

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.

Fleurbaey, M., Tungodden, B., and Chang, H. F. (2003). Any non-welfarist method of policy assessment violates the pareto principle: A comment. *Journal of Political Economy*, 111(6):1382–1385.

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 Economics and Statistics*, 104(1):17–33.

Haggag, K. and Paci, G. (2014). Default Tips. *American Economic Journal: Applied Economics*, 6(3):1–19.

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

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. *The 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.

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.

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.

Lichtenstein, S. and Slovic, P. (2006). *The Construction of Preference*. Cambridge University Press.

Lockwood, B. B., Allcott, H., Taubinsky, D., and Sial, A. (2023). What drives demand for state-run lotteries? evidence and welfare implications. Nber working paper no. 28975.

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.

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

Masatlioglu, Y. and Ok, E. A. (2005). Rational Choice with Status Quo Bias. *Journal of Economic Theory*, 121(1):1–29.

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). Rawlsâs 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.

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.

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. (2001). Using Elasticities to Derive Optimal Income Tax Rates. *Review of Economic Studies*, 68(1):205–29.

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

Salant, Y. and Rubinstein, A. (2008). (A, f): Choice with Frames. *Review of Economic Studies*, 75(4):1287–1296.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

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

Sher, I. (2023). Generalized social marginal welfare weights imply inconsistent comparisons of tax policies. Working paper, arxiv:2102.07702.

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

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. In *Theory of Bames and Economic Behavior*. Princeton University Press.

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. pages 255–320. Cambridge University Press.

A Convex Sets of Frames and Interpretation of Normative Weights

A.1 Continuous frames and convex hulls

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.

A.1.1 Basics

In the main body we have a set of N frames $\Theta = \{\theta_1, ..., \theta_N\}$. We can think of these as parameters of a (cardinal) utility function $u(x, \theta)$. Suppose each θ_n implies a set of N real-valued preference parameters $\theta_n = (\theta_{n1}, ..., \theta_N)$. 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 θ_n 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 by the logic below.

A.1.2 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 arguably a strong 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 Proposition 2. We discuss this in the main text along the lines of an alternative interpretation of normative weights. What we have shown here justifies these two interpretations of normative weights:

- 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 are weights derived from the planner's Bayesian beliefs over a convex and continuous set of potential preference parameters.

We could combine the two concepts: if we maintain frame-wise linearity, further 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.

B 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 second interpretation of the weights in ψ above 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 the 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, modelling 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 prefernece s 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 the normative frame ($\theta^* \in \tilde{\theta}$). 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 what kind of uncertainty the planner has about the validity of these assumptions, and their preferences for how to accommodate this uncertainty.

A Proofs for Section 2 (General Model)

Lemma 1. ***Frame Exclusion.*** *Under Assumption 1, for any $x, x' \in \mathcal{X}$ and any $\theta, \theta' \in \Theta$,*

$$(x,\theta) \succeq_* (x',\theta) \implies (x,\theta') \succeq_* (x',\theta').$$

Proof. $(x,\theta) \succeq_* (x',\theta) \implies x \succeq_{\theta^*} x'$ by Assumption 1.3 (“ $\Longleftarrow$ ”). Then, $x \succeq_{\theta^*} x' \implies (x,\theta') \succeq_* (x',\theta')$ by Assumption 1.3 (“ $\implies$ ”). ■

Lemma 2. ***BR-Dominance.*** *Under Assumptions 1-3, given any $x, x' \in \mathcal{X}$,*

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

Proof. This follows straightforwardly from Assumption 1.3. Suppose that $\forall\theta \in \Theta$, $x \succeq_\theta x'$. Then for the normative frame $\theta^* \in \Theta$ whose existence is assured by RP-Coincidence, it must be that $x \succeq_{\theta^*} x'$, and then RP-Coincidence implies $x \succeq_* x'$. ■

Proposition 1. *Maintain Assumptions 1.2, 2 and 3. Assumption 4 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 (4.2), 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 requires $x'' \succsim_w x$. This establishes sufficiency of our assumptions for representation (6); necessity is easily verified. ■

Proposition 2. *Maintain Assumptions 1, 2 and 3. Then Assumption 5 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*

$$w(x;\psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) u(x,\theta^*). \tag{8}$$

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

Proof. Assumptions 5.1, 5.2, and 5.4 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 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 5 for the desired representation of $\succeq_w$; necessity is easily verified. ■

Corollary 2.1. *Maintain Assumptions 1, 2, 3,and 5, and consider a utility function $u(x,\theta)$ that gives the representation in Proposition 2. 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*

$$w(x;\psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*)\omega(v(x,\theta)). \tag{9}$$

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

Proof. This result obviously follows from Proposition 2.1 and the definition of ordinal level comparability. ■

Corollary 2.2. ***Variance Representation*** *Assume the function ω from representation (9) is twice differentiable. Then up to second-order Taylor approximation of ω, the planner's objective is*

$$w(x,\psi) \approx \omega\Big(E_\psi[v(x,\theta)]\Big) + \frac{\omega''\Big(E_\psi[v(x,\theta)]\Big)}{2} \cdot Var_\psi\Big[v(x,\theta)\Big] \tag{10}$$

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

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

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

Proposition 3. ***MaxMin Welfare Under Ambiguity Aversion.*** *Maintain Assumptions 1, 2 and 3. Assumption 6 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*

$$w(x) = \min_{\psi \in \Psi^*} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x,\theta^*)) \right\}. \tag{12}$$

Proof. Take the representation of individual preferences whose existence is implied by 6.3 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. ■

Corollary 3.1. ***Intersection of Various Objectives at Global Max-Min***

- *If $\Psi^* = B(\kappa, \psi)$, for any ψ, the planner's objective in (12) coincides with the global max-min criterion for $\kappa > 1$.*
- *If $\Psi^* = \Delta(\Theta^*)$, the planner's objective in (12) and/or (13) 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 (9) with the functional form in equation (11) – approaches the global max-min criterion as $\eta \to \infty$.*[34]

Proof. The first three claims are obvious from equation (12) and (13). 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 (11). ■

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

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 $\psi(\theta) = 1\{\theta = \theta'\}$. As P^* is not a ψ-optimum for this ψ, it cannot be globally optimal. ■

Proposition 4. ***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*

$$P^* \in \arg\min_{P\in\mathcal{P}} \max_{\psi'\in\Psi^*} \sum_{\theta\in\Theta} \left(\psi'(\theta) - \psi(\theta)\right) \cdot V\left(x(P, Z, \theta^D), \theta, \theta^D\right). \tag{14}$$

Proof. By supposition,

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

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

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

$$\min_{\psi' \in \Psi^*} W(P^*, Z, \theta^D; \psi') - \min_{\psi' \in \Psi^*} W(P, Z, \theta^D; \psi') \geq W(P^*, Z, \theta^D; \psi) - W(P, Z, \theta^D; \psi) \geq 0 \tag{42}$$

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. ■

B Proofs for Section 3 (Approaches to Comparability)

Lemma 4. ***Existence and uniqueness of EV.*** *Under Assumptions 1.2, 2 and 7, for any x, any $\theta \in \Theta$ and any (P_0, Z_0, θ_0^D), equivalent variation ζ exists and is unique. Moreover, $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ represents the individual's ordinal preferences $\succeq_\theta$.*

Proof. Suppose first that $x \succ_{\theta^*} x(P_0, Z_0, \theta_0^D)$, i.e. $u(x, \theta^*) > u(x(P_0, Z_0, \theta_0^D), \theta^*)$. Assumption 7.1 ensures that $u(x(P_0, Z_0 + \zeta, \theta_0^D), \theta^*)$ 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 $u(x(P_0, Z_0 + \hat{\zeta}, \theta_0^D), \theta^*) > u(x, \theta^*)$. The result follows from the Intermediate Value Theorem – note that $u(x, \theta^*)$ is continuous by Assumption 2. The same logic applies where $x \prec_{\theta^*} x(P_0, Z_0, \theta_0^D)$, and in the case of indifference, $\zeta = 0$.

Having established existence and uniqueness, that $\zeta(x, \theta^*)$ represents $u(x, \theta^*)$ follows from

$$\begin{aligned} x \succeq_{\theta^*} x' &\iff x(P_0, Z_0 + \zeta(x; \theta^*), \theta_0^D) \succeq_{\theta^*} x(P_0, Z_0 + \zeta(x'; \theta^*), \theta_0^D) \text{ by definition \& transitivity} \\ &\iff \zeta(x; \theta^*) \geq \zeta(x'; \theta^*) \text{ by Assumption 7.1.} \end{aligned}$$

■

Proposition 5. ***Planner's Preferences and Equivalent Variation.*** *Under Assumptions 1.2, 2, 4, and 7, for any baseline P_0, Z_0, θ_0^D, there is a function $\mathcal{W}_\zeta : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planners preferences are represented by $w(x) = \mathcal{W}_\zeta\left(\{\zeta(x; \theta^*, P_0, Z_0, \theta_0^D)\}_{\theta^* \in \Theta}\right)$.*

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 5. ***Ordinal Level Comparability of Equivalent Variation.*** *Maintain Assumptions 1, 2 and 7. Let $u(x, \theta)$ be a cardinal utility function from the representation in Proposition 2 or 3. Assumption 8 holds if and only if there is some baseline (P_0, Z_0, θ_0^D) such that $u(x, \theta)$ and $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ exhibit ordinal level comparability.*

Proof. First we prove that level comparability implies Assumption 8. Assuming ordinal level comparability, Assumption 8.1 follows from the observation that $\zeta(x_0, \theta) = \zeta(x_0, \theta') = 0$ by construction, so then ordinal level comparability implies $u(x_0, \theta) = u(x_0, \theta')$. 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

$$u(x, \theta) \geq u(x', \theta') \iff u(x, \theta) - u(x_0, \theta) \geq u(x', \theta') - u(x_0, \theta').$$

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

$$\begin{aligned} & u(x, \theta) - u(x_0, \theta) \geq u(x', \theta') - u(x_0, \theta') \\ \iff & u(x(P_0, Z_0 + \zeta(x, \theta)), \theta) - u(x_0, \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x_0, \theta'). \end{aligned}$$

Now using Assumption 8.2, we find

$$\begin{aligned} & u(x(P_0, Z_0 + \zeta(x, \theta)), \theta) - u(x_0, \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x_0, \theta') \\ \iff & \zeta(x, \theta) \geq \zeta(x, \theta'). \end{aligned}$$

■

Proposition 6. *Under Assumptions 1, 2, 5, 7 and 8, there is a function $\omega_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0, Z_0, θ_0^D) such that the planner's preferences are represented by*

$$w(x, \psi) = \sum_{\theta \in \Theta} \psi(\theta) \omega_\zeta(\zeta(x, \theta; P_0, Z_0, \theta_0^D)). \tag{16}$$

Under Assumptions 1, 2, 6, 7 and 8, there is a function $\omega_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0, Z_0, θ_0^D) such that the planner's preferences are represented by

$$w(x, \Psi) = \min_{\psi \in \Psi^*} \sum_{\theta \in \Theta} \psi(\theta) \omega_\zeta(\zeta(x, \theta; P_0, Z_0, \theta_0^D)). \tag{17}$$

Proof. Lemma 4 implies the equivalent variation exists, is unique and represents $\succeq_\theta$ for some baseline. Lemma 5 implies that this representation satisfies ordinal level comparability with the planner's cardinal utility function. Then applying Corollary 2.1 gives the result. ■

C Proofs for Section 5 (Examples)

Proposition 7. ***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 (32) simplifies as follows:

$$w(x) = \beta \sum_{t=1}^{T} \delta^t \mu(x_t) + \frac{1-\psi(0)}{T}(1-\beta) \sum_{\tau=1}^{T} \delta^\tau \mu(x_\tau), \quad (43)$$

$$= \left[\beta + (1-\beta)\frac{1-\psi(0)}{T}\right] \sum_{t=1}^{T} \delta^t \mu(x_t), \quad (44)$$

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

D Proofs for Section 6 (Robust Optimality)

Proposition 8. ***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 robust optimum.*
- *If the expected intrinsic optimum is not ψ-optimal for any $\psi \in \Psi^*$, the extremum default is the unique robust optimum.*
- *If the expected intrinsic optimum is ψ-optimal for some ψ on the boundary of Ψ^* but not the interior, both the expected intrinsic optimum and the extremum default are robust optima.*

Proof. In the defaults case, the policy parameter is 1-d $\sigma = d$, 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 behaviour is normative, and under θ^A it is a pure bias. $\psi(\theta^D)$, which we abbreviate to just $\psi = \mathbb{P}[\theta = \theta^D]$.

$$u(x, \theta^D, d) = u(x) - \gamma \cdot \mathbb{1}\{x \neq d\} \quad (45)$$

$$u(x, \theta^A, d) = u(x) \quad (46)$$

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

As a simple example, let $u(x) = -\frac{\alpha}{2}(x - x^*)^2$ where x^* is unknown. $x, x^* \in X$, the choice set which is $X \subset \mathbb{R}$ and defaults at the max and min of X 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, $B(\kappa, \psi) = [\psi - \kappa, min(\psi + \kappa, 1)]$. Recall $W(d, \psi') = -\frac{\alpha}{2}(x(d) - x^*)^2 - \psi' \cdot \gamma \cdot \mathbb{1}\{x(d) \neq d\}$. Since $\gamma > 0$...

$$\arg \min_{\psi' \in B(\kappa,\psi)} W(d_{min}, \psi') = min(\psi + \kappa, 1) \quad (47)$$

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...

$$d^* = \arg\max_d -\frac{\alpha}{2}(x(d) - x^*)^2 - min(\psi + \kappa, 1) \cdot \gamma \cdot \mathbb{1}\{x(d) \neq d\} \tag{48}$$

Since d_{min} is a candidate optimum for ψ, it is also a candidate optimum for $\psi' > \psi$ since under those judgements 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 judgement which makes the penalty default a candidate optimum ($\psi \approx 0$). Then, $B(\kappa, \psi) = [max(0, \psi - \kappa), \psi + \kappa]$. Similarly to the minimizing opt-outs example, the evil agent wants to maximise ψ' and so sets...

$$\arg\min_{\psi' \in B(\kappa,\psi)} W(d_{pen}, \psi') = \psi + \kappa \tag{49}$$

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

$$\min_{\psi' \in B(\kappa,\psi)} W(d_{pen}, \psi') = 0 - \gamma(\psi + \kappa) \tag{50}$$

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

$$\min_{\psi' \in B(\kappa,\psi)} W(\bar{d}, \psi') = \underbrace{-\frac{\alpha}{2}(\mathbb{E}[x^*] - x^*)^2}_{=-\Lambda \text{ fixed w.r.t. } \kappa} - (\psi + \kappa) \cdot \gamma \cdot \underbrace{\mathbb{P}\{x(\bar{d}) \neq \bar{d}\}}_{=p \text{ small}} \tag{51}$$

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 $\gamma = V(x(d_{pen}), d_{pen})$.

■

Corollary 8.1. *Robust Control and the Optimal Default.* *Suppose $\Psi^* = B(\kappa, \psi)$ for some $\kappa > 0$ and some $\psi \in \Delta(\Theta)$.*

If the extremum defualt is ψ-optimal, there is a threshold $\overline{\kappa}$ such that

- *the extremum default is the unique robust optimum for $\kappa < \overline{\kappa}$, but*

- *the expected intrinsic optimum is the unique robust optimum for $\kappa > \overline{\kappa}$.*[35]

If the expected intrinsic optimum is ψ-optimal, the expected intrinsic optimum is the robust optimum for any κ.

Proposition 9. *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*

$$\frac{dT^*(x_1)}{dx_1} = \begin{cases} [1-\underline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) > 0 \\ [1-\overline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) < 0 \\ 0 & V(x_1) = 0. \end{cases} \tag{41}$$

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

[35] In the knife-edge case $\kappa = \overline{\kappa}$, both the extremum default and the expected intrinsic optimum are κ-ψ robust.