# Behavioral Welfare Economics
# and the Quantitative Intentional Stance

by

Glenn W. Harrison and Don Ross [†]

January 2023

Forthcoming in G.W. Harrison and D. Ross (eds.),
*Models of Risk Preferences: Descriptive and Normative Challenges*
(Bingley, UK: Emerald, Research in Experimental Economics, 2023).

ABSTRACT.

Behavioral economics poses a challenge for the welfare evaluation of choices, particularly those that involve risk. It demands that we recognize that the descriptive account of behavior towards those choices might not be the ones we were all taught, and still teach, and that subjective risk perceptions might not accord with expert assessments of probabilities. In addition to these challenges, we are faced with the need to jettison naive notions of revealed preferences, according to which every choice by a subject expresses her objective function, as behavioral evidence forces us to confront pervasive inconsistencies and noise in a typical individual's choice data. A principled account of errant choice must be built into models used for identification and estimation. These challenges demand close attention to the methodological claims often used to justify policy interventions. They also require, we argue, closer attention by economists to relevant contributions from cognitive science. We propose that a quantitative application of the "intentional stance" of Dennett provides a coherent, attractive and general approach to behavioral welfare economics.

[†] Department of Risk Management & Insurance, Robinson College of Business, Georgia State University, USA (Harrison); Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison and Ross); School of Economics, University of Cape Town, South Africa (Harrison and Ross); Research Unit in Behavioural Economics and Neuroeconomics (Ross); School of Society, Politics and Ethics, University College Cork, Ireland (Ross) E-mail contacts: gharrison@gsu.edu and don.ross931@gmail.com. ORCID numbers: Harrison (0000-0003-1837-8020) and Ross (0000-0003-1813-3111). We are grateful to Jimmy Martínez-Correa, Karlijn Morsink, Jia Min Ng and Elisabet Rutström for helpful discussions.

## Table of Contents

One theme lies at the heart of a rigorous evaluation of policy using the insights of behavioral welfare economics: how to judge if some policy is encouraging good choices or bad choices. One approach, which drives the nudge movement and some randomised evaluations, is to assume that judgment away, and simply assert that some change in an *observable* must be good regardless of special features of the preferences and beliefs of individuals. Isn't it obvious that people should save more, eat fewer fatty foods, drink less wine, and take up insurance? The only appropriate answer to this for an economist is "no." Demand for these behaviors depends on preferences and beliefs, and hence the expected consumer surplus they deliver is also conditional on preferences and beliefs. Further, no realistic economics that recognises heterogeneity of preferences and beliefs, and identifies a welfare-enhancing policy conditional on those preferences and beliefs, tells us that such policies would be welfare-enhancing *on average* if applied *unconditionally*.

Revealed preference theory (RPT), the various axiomatisations of which are reviewed by Chambers and Echenique [2016], minimises the inferential gap between agents' observed behavior and ascribed goals by treating preferences as summaries of actual choices. As emphasized by Binmore [2009], under RPT it is simply an *error* to regard preferences as *causes* of choices, or as *explaining* them. The motivation for this is not philosophical commitment to behaviorism about "mental states," but reflects the value to the economist, for the sake of theoretical power, of treating all processes that generate common utility outcomes for an agent as an equivalence class. This raises problems for normative analysis, however, because it seems to leave no room for treating any individual choices as reflecting *error*. In effect, every choice that the agent makes is assumed to make her no worse off. Such an idealization makes it impossible *a priori* for the economist to offer any advice about prospective choices that would be potentially superior with respect to the agent's subjective welfare.

Behavioral welfare economics seeks to bridge this gap. One obvious way to do this, which is common in the literature, is to simply abandon RPT, and directly model preferences as hypothesised psychological states that cause choices. The theorist is then free to suppose that different psychological processes have varying welfare consequences, and that an agent can profit from the economist's advice whenever her choices result from less-than-optimal such processes. But this approach has major costs. It makes welfare economics conditional on strong psychological assumptions that must be expected to vary from case to case, even across choices that are identical with respect to outcomes. Thus it directly undermines the kinds of generalizations economists care about both descriptively and normatively. The psychological assumptions in question are often, necessarily, pure conjectures. Where they are based on psychological experiments, they are hostage to inferential methods in psychology that are heavily criticized, have produced many post-publication retractions by journals, and are rejected by most economists when applied in their own discipline (Ortmann [2021], Yarkoni [2022]). Appeals to hypothesized brain states based on neuroimaging studies involve yet more egregious violations of econometric best practice with respect to both identification and estimation (Harrison [2008], Harrison and Ross [2020]). Worst of all where unified science is concerned, as we discuss, interpretation of preferences (and beliefs) as inner states of agents that cause chosen behavior is now rejected outright by the majority of the relevant experts on the issue, cognitive scientists.

The challenge for behavioral welfare economics, then, is to find a way of remaining within the broad ambit of RPT – that is, deriving preferences (and beliefs) directly from observed behavior – but with a richer set of modeling techniques that avoid idealizing each and every choice as optimal for the agent. These requirements generate a derived demand for thinking carefully about the methodologies of behavioral welfare economics, and that requires economists to think more deeply about the methodology and philosophy of their subject. In this respect thought experiments and laboratory experiments stand as ideal places to begin this long, slippery path. One major risk we face, and that is tragically illustrated by what currently passes for behavioral economic policy, is that the new behavioral economics causes us to forget the old welfare economics: see Atkinson [2001][2009][2011], who eloquently reflects on "the strange disappearance of welfare economics." In the words of Homer Simpson (season 5, episode 22), "every time I learn something new, it pushes some old stuff out of my brain."

We argue that this challenge can be met systematically, generally, and with available tools, by recognizing a body of work in cognitive sciences that seems to have been constructed with economists in mind (though of course it wasn't). We refer to "the intentional stance" developed by Dennett [1971][1987].

In section 1 we briefly review the descriptive evidence on behavior towards risk, as we see it. Section 2 reviews broad approaches towards behavioral welfare economics by economists. Section 3 is then a statement of the proposed application of the intentional stance to behavioral welfare economics, which we refer to as the Quantitative Intentional Stance (QIS). Section 4 reviews several approaches towards behavioral welfare economics by philosophers. Section 5 concludes.

## 1. The Descriptive Evidence, As We See It

Our focus here is on the descriptive evidence about risk preferences, but the general issues that the QIS is intended to address also require attention to time preferences and subjective beliefs, at the very least. And since these interact with the evaluation of risky choices anyway, we also briefly cover them.

### *A. Theoretical Context*

The notion of a risk premium is one of the core concepts that different theories of risk preferences agree on. Expected Utility Theory (EUT) assumes that aversion to variability drives a risk preference, where variability can be much more than just variance. Rank-Dependent Utility (RDU) complicates this assumption with the additional idea that probability optimism or pessimism can augment, positively or negatively, any risk premium due to aversion to variability. And Cumulative Prospect Theory (CPT) is based on a hypothesized psychological mechanism on top of these, where sign dependence relative to some reference point affects risk preferences. All of these models agree on the *concept* of a risk premium, and just decompose it differently. In general, since the utility function under each model will differ, the models will not agree on the *value* of the risk premium. Harrison and Swarthout [2023] review the comparative evidence for these models.

Important extensions to these basic constructions include considerations of downside risk aversion that differs from the loss aversion of CPT, and is related to literature on "higher order risk preferences;" considerations of "regret" or "disappointment" that can arise from contemplation of realized outcomes; intertemporal risk aversion toward variability of outcomes over time; and allowance for multi-variate risk aversion, including connections between foreground and background risk.

Principal theories of time preference include Exponential, Hyperbolic, and Quasi-Hyperbolic discounting. The differences can best be understood by thinking of the lender of money as assigning some cost to not having liquidity for a time period. Exponential discounting assumes a constant variable cost of being without the loaned funds with respect to time and no fixed cost; Hyperbolic discounting assumes a declining variable cost with respect to time and no fixed cost; and Quasi-Hyperbolic discounting assumes a fixed cost and a constant variable cost.[1] An alternative approach from psychology is to view the perception of time horizon as subjective: if the agent perceives time units contracting as the horizon gets longer, declining discount rates will arise. Andersen, Harrison, Lau and Rutström [2014] review these models, and evidence for them.

Virtually all theories of time preference assume an *additive* intertemporal utility function, in which utility over time is a discount-factor-weighted sum of utility for each distinct period. In this respect, the alternative theories behind the discount factor tend to agree. This seemingly technical assumption, however, has dramatic implications for behavior towards risk: it implies that agents are neutral towards risk *over time*, even if they are averse to risk *at a point in time*. In other words, agents might be averse to risk resolved at a point in time, but must then be neutral to risk resolved *over* time. A restrictive corollary of this additivity is that atemporal risk preferences and time preferences are formally "tied at the hip," in the sense that the intertemporal elasticity of substitution *must* be equal to the inverse of relative risk aversion.

[1] The Quasi-Hyperbolic model assumes a rather strange fixed cost, a constant *percentage* of the principal. One can write down models that assume that the fixed cost is a scalar amount of money, or a scalar level of utility.

This corollary sits uncomfortably with everyday observations and the stylized aggregate data, forcing problematic calibrations in macroeconomic models. A simple resolution of this impasse is to allow non-additive intertemporal utility functions, such that interactions between atemporal responses across time periods matter to the agent. Andersen, Harrison, Lau and Rutström [2018] review the theory.

The static theory of subjective beliefs is dominated by Subjective Expected Utility (SEU), which assumes that agents behave as if satisfying the Reduction of Compound Lotteries (ROCL). The effect is that non-degenerate subjective belief distributions can be replaced by the weighted average belief, and then EUT applied as usual. It is noteworthy that SEU does not assume that the subjective belief distributions that agents hold satisfy Bayes' Rule when updated over time, despite Savage [1954][1962] being a staunch advocate for each. Bayes' Rule generates a separate model of (dynamic) risk perception, which may or may not apply with SEU. Relaxations of ROCL that still assume that the agent has a well-defined subjective belief distribution characterize uncertainty, and models of decision-making that do not assume a well-defined subjective belief distribution characterize ambiguity: see Harrison [2011b] for an exposition.

### *B. Experimental Evidence*

There are various methods for eliciting and estimating risk preferences, reviewed by Harrison and Rutström [2008]. Unfortunately some of the methods in use have well-known weaknesses and biases. One of the most flexible is to ask the agent to make a series of unordered binary choices over risky lotteries, where each lottery typically has between 1 and 4 outcomes. This method provides enough flexibility to allow for estimation of risk preferences at the level of the individual. For normative analysis, recognizing the heterogeneity of risk preferences across individuals is critical. Moreover, heterogeneity here refers to much more than the risk premium: it also refers to the *type* of risk preferences. It makes a significant difference for the normative evaluation of insurance products if the agent is an EUT or RDU decision-maker. In general these models will imply different utility functions, and it is the utility function that is used to calculate the Certainty Equivalent (CE) of polices to manage risk (e.g., Harrison and Ng [2016]).

There is overwhelming evidence that laboratory and field samples demonstrate heterogeneity with respect to which models of risk response describe their choice patterns. Harrison and Ross [2016, p. 401] summarize a selection of their own studies, conducted in both developed and developing countries, which find that substantial proportions of subjects in all samples exhibit behavior best characterized by EUT. At the same time, in all of these studies we observe somewhat larger proportions presenting choice patterns better modelled by RDU. This classification refers to estimated models at the level of the individual: comparable classifications arise if one uses mixture models over data that are pooled over individuals, as proposed by Harrison and Rutström [2009]. There is seldom any evidence for Dual Theory, which proposes the special case of RDU in which utility functions are linear, so that the entire risk premium derives from probability weighting.

There is actually very little evidence for CPT in controlled, incentivized experiments. This may come as a shock to some. Harrison and Swarthout [2023] provide an extensive literature review, which finds that most reported evidence for "loss aversion" is actually evidence for probability weighting. Very often there is evidence of probability weighting over choices between two gain-frame lotteries, and that is literally reported as evidence for CPT. The context always makes it clear that it is *evidence for* RDU, and is *consistent with* evidence for CPT, but that distinction is all too often glossed.

Harrison and Swarthout [2023] also report evidence of (at least local) asset integration in the laboratory, which is *fatal* for the empirical adequacy of CPT. In a nutshell, this explains why they find so little support for CPT in a "three horse race" that includes EUT and RDU; Dual Theory never gets out of the gates, by the way. Harrison and Ross [2017] review further evidence, and consider the implications for welfare assessment of the conjecture that the many reported "two horse race" victories of CPT over EUT were really wins for RDU in disguise, where the successes of CPT stemmed from its allowance for probability weighting rather than "utility" loss aversion relative to an idiosyncratic reference point.

Harrison and Swarthout [2023] further stress that there are two pathways for the loss aversion meme that "losses loom larger than gains." The usual way in which CPT models loss aversion comes from Tversky and Kahneman [1992; p. 309], who popularized the functional forms we often see using a CRRA specification of utility over money m: $U(m) = m^{1-\alpha}/(1-\alpha)$ when $m \geq 0$, and $U(m) = -\lambda[(-m)^{1-\beta}/(1-\beta)]$ when $m < 0$, where $\alpha$ and $\beta$ define the curvature of some "basic utility function" that captures the "intrinsic value of outcomes and satisfies usual regularity conditions" (Wakker [2010; p.239]), U is some "overall utility function," and $\lambda$ is the worshiped loss aversion parameter for utility. Here we have the assumption that the degree of utility loss aversion for small unit changes in money is the same as the degree of loss aversion for large unit changes: the same $\lambda$ applies locally to gains and losses of the same monetary magnitude around 0 as it does globally to any size gain or loss of the same magnitude. This is not a criticism, just a restrictive parametric turn in the specification compared to Kahneman and Tversky [1979]. However, even if the basic utility functions for gains and losses are linear, and conventional loss aversion is absent ($\lambda$=1), differences in the *decision weights* for gains and losses could induce the same behavior as if there were utility loss aversion emanating from $\lambda$. This is called "probabilistic loss aversion" by Schmidt and Zank [2008; p.213]. Imagine that there is no probability weighting on the gain domain, so the decision weights are the objective probabilities, but that there is some probability weighting on the loss domain. Then one could easily have losses weighted more than gains, solely from the implied decision weights.

Another critique of EUT that has arisen in experimental settings is the so-called calibration critique popularized by Rabin [2000]. This is the concern that "small stakes risk aversion," supposedly common in lab experiments, implies implausibly large "high stakes risk aversion" under EUT. The point was originally made by Hansson [1988], and has been viewed as an indirect rationale for wanting to consider (utility) loss aversion from CPT as playing an important role in decision-making over low stakes. However, the general experimental literature on risk aversion does not support the theoretical premise of the calibration critique: that premise depends on observing the *same person* facing small stakes lottery choices over a range of wealth levels. Cox and Sadiraj [2006] proposed an elegant design to implement this test, building on the ability to vary "lab wealth" for a given subject. Evidence from university undergraduates in the U.S. indicates that the premise is simply false for that population (Harrison, Lau, Ross and Swarthout [2017]), although evidence from representatives of the adult Danish population shows that the premise is valid for the range of lab wealth considered (Andersen, Cox, Harrison, Lau, Rutström and Sadiraj [2018]). In the latter case there are alternative assumptions about the degree of asset integration between field wealth and lottery prizes that allow one to readily reconcile small stakes risk aversion with plausible high stakes risk aversion, and these assumptions appear to apply to the Danish population.

There is much less evidence for "hyperbolicky" discounting than conventionally assumed. Prior to Coller and Williams [1999], there were very few experiments that provided designs that allowed one to infer monetary discount rates rigorously. This might seem like a simple point, but prior literature typically generated annualized discount rates in the hundreds or thousands of percent (and typically chose not to report them as such, for obvious reasons). Another important insight, often neglected completely, has been to correct for the effect of diminishing marginal utility on inferences about utility discount rates drawn from choices between "smaller, sooner" amounts of money and "larger, later" amounts of money. Modest levels of diminishing marginal utility generate first-order changes in inferred discount rates (Andersen, Harrison, Lau and Rutström [2008a]). Variations in designs allow one to test Exponential discounting of money against all major alternatives, and Exponential discounting clearly characterizes the data best in such settings (Andersen, Harrison, Lau and Rutström [2014]). Nor is there any evidence for the alleged "magnitude effect," whereby elicited discount rates appeared to be lower for higher stakes (Andersen, Harrison, Lau and Rutström [2013]).

There have been important advances in the manner in which subjective beliefs can be elicited. One strand of literature concerns the estimation of subjective *probabilities* over *binary* events, using incentivized scoring rules and corrections for the effect of risk aversion on reports (Andersen, Fountain, Harrison and Rutström [2014]). Another strand tackles the more challenging problem of inferring whole subjective belief *distributions* for *continuous or non-binary* events (Harrison, Martínez-Correa, Swarthout and Ulm [2017]). In the latter case one can directly make statements about the level of "confidence" that individuals have in their beliefs. The application of these *incentivized* methods has not yet been widespread in behavioral economics.

## 2. Methodologies of Behavioral Welfare Economics from Economists

There is a large, general literature on behavioral welfare economics, including Bernheim [2009][2016], Bernheim and Rangel [2008][2009], Bernheim and Taubinksy [2018], Manzini and Mariotti [2012][2014], Rubinstein and Salant [2012], Salant and Rubinstein [2008] and Sugden [2004][2009]. A general concern with this literature is that although it identifies the methodological problem well, it does not yet provide clear guidance to practical, portable, rigorous welfare evaluation with respect to risk preferences as far as we can determine. That is what the approach by Harrison and Ng [2016][2018] and Harrison and Ross [2018], described in §3, seeks to do.

### *A. Nudges and Boosts*

A principal source of interest in behavioral economics has been its advertised contributions to policies aimed at "nudging" people away from allegedly natural but self-defeating behavior toward patterns of response thought more likely to improve their welfare. Leading early promotions of this kind of application of behavioral studies are Camerer, Issacharoff, Loewenstein, O'Donoghue and Rabin [2003] and Sunstein and Thaler [2003a][2003b]. Grüne-Yanoff and Hertwig [2016] have distinguished nudging, which is based on the heuristics-and-biases branch of behavioral economics research associated with Kahneman and Tversky [1982], from policies aimed at "boosting," which apply the simple heuristics research program of Gigerenzer et al. [1999] and Hertwig et al. [2013].

Nudging and boosting are contrasted as follows. Nudges aim to change a decision-maker's ecological context and external cognitive affordances in such a way that the decision-maker will be more likely to choose a welfare-improving option without having to think any differently than before. Boosts aim to supplement cognitive processes with heuristics that are viewed as reliable guides, despite glossing some information and avoiding computationally intensive algorithms, to produce good inferences, choices, and decisions when applied in the appropriate ecological contexts. An alternative way to define boosting builds on the role that "scaffolds" play in aiding cognition (Clark [1998], Dennett [2017]). Access to the internet or experts, for example, might be expected *a priori* to make individuals more literate on facts that affect their cognition, as better inputs to their "cognitive production function." [2]

An additional contrast between nudges and boosts is that a nudge would normally be expected to have effects only on the specific behavior to which it is applied, and only in the setting that the nudge adjusts. A boost, on the other hand, to the extent that it alters standing cognitive capacities and associated behavioral propensities across ranges of structurally similar choice problems, might be hoped to generate "rationality spillovers" discussed by Cherry et al. [2003]. Furthermore, boosting might plausibly capacitate people with defenses against non-benevolent nudging by narrowly self-interested parties, such as marketers and demagogues (de Haan and Linde [2018]).

Nudging is open to the charge that it is manipulative. Its defenders point out that if people are naturally prone to systematic error, then any scaffolding built by any institution unavoidably involves manipulation, so the manipulation in question might as well be benevolent. Of course, as stressed above, what is actually "benevolent" is typically conditional on some unobserved preference or belief ascribed to the decision-maker. How this ascription might occur is the deeper question, addressed in §3.

[2] Hence boosts need not rely on the use of heuristics. The key distinction between an algorithm and a heuristic has to do with the knowledge claim that they each allow one to make. If an algorithm has been applied correctly, then the result will be a solution that we know something about. For example, we may know that it is a local optimum, even if we do not know that it is a global optimum. Heuristics are lesser epistemological beasts: the solution provided by a heuristic has no claim to be a valid solution in the sense of meeting some logical criteria. In the computational literature, if not some parts of the psychological literature, heuristics are akin to "rules of thumb" that simply have good or bad track records for certain classes of problems. The track record may be defined in terms of the speed of arriving at a candidate solution, or the ease of application. Harrison [2008; §4.2] provides more discussion of the role of heuristics in decision-making, particularly their crucial role in "behaviorally plausible" homotopy, or path-following, algorithms.

Boosting, by contrast, involves endowing decision-makers with enhanced cognitive capacities by teaching them more effective decision principles, which they can choose to apply or not once they have been enlightened. Thus boosting avoids manipulating the agents to whom the policies in question are applied, and is to that extent less paternalistic. But boosting unavoidably raises the question of what are more "effective" decision principles.

### *B. Randomized Evaluations in Search of "What Works"*

One of the slogans that burdens most behavioral economic policy studies, and the focus on randomized evaluations, is the claim that they are only interested in "what works."[3] It is hard to imagine a less informative, and more dangerous, slogan. The core problem is that it characterizes approaches that only look at observables.

The problem with just looking at observables is that they tell us nothing about the virtual variables that are of interest in welfare evaluation. For that we need to make inferences about expected Consumer Surplus (CS), and for that we need to know about the preferences that people bring to their choices, such as risk preferences and time preferences. We also need to know about the subjective beliefs that people bring to their choices. The reason for the recurrent focus on observables is easy to discern and openly discussed: a desire to avoid having to take a stand on theoretical constructs as maintained assumptions. We will discuss the general grounds for such abstinence from theory in §3. The same methodological precept guides the choice of statistical methods, but that is another story about modeling costs and

[3] This expression is widely used, but examples might be of value in case one doubts that it is so common. Karlan and Appel [2011; p.5] state that "...at the end of the day, even Sachs and Easterly could agree on the following: Sometimes aid works, and sometimes it does not. That can't be all that controversial a stand! The critical question, then, is which aid works. The debate has been in the sky, but the answers are on the ground. Instead of getting hung up on the extremes, let's zero in on the details. Let's look at a specific challenge or problem that poor people face, try to understand what they're up against, propose a potential solution, and then test to find out whether it works. If that solution works – and if we can demonstrate that it works consistently – then let's scale it up so it can work for more people. If it doesn't work, let's make changes or try something new." And the U.S. Department of Education maintains a "What Works Clearinghouse" at https://ies.ed.gov/ncee/WWC.

benefits. One can fill in these blanks in our knowledge about virtual preferences and beliefs with theories and guessed-at numbers, or with theories and estimated numbers, as stressed by Leamer [2010] and Wolpin [2013]. But one has to use theory to make conceptually coherent statements about preferences and beliefs, and then undertake welfare evaluations.

Advocates of randomized evaluations portray the tradeoff here in overly dramatic fashion: either one uses the methods that avoid these theoretical constructs, or one dives deep into the shoals of full structural, parametric modeling of behavior. This is a false dichotomy. The missing middle ground becomes apparent when empirical puzzles emerge, leading to casual theorizing and even more casual appeal to loose behavioral inferences, documented by case studies in Harrison [2011a].

In any event, randomized evaluations can be wonderful tools for gathering information about the cost-effectiveness of alternative policies towards some given goal, but are silent on the real question of the net welfare consequences of those policies.

### *C. Behavioral Welfare Analysis Using "Frames"*

Bernheim and Rangel [2008][2009] and Bernheim [2009][2016] present an approach to behavioral welfare economics that recognizes the methodological challenge of evaluating welfare when one concerned that choices might refelect mis-framing, weakness of will, or other psychological "disturbance" factors. They propose that one develop two frames with which to ask a question bearing on financial choices, where two conditions are met, and are couched here in terms of a financial literacy application:

1. Each frame is *a priori* presumed to generate actions that have the same welfare consequences for the individuals.
2. One frame is simple and transparent to understand, so *a priori* does not require any significant degree of literacy to assess, while the other frame asks the respondent to bring some degree of such literacy to bear.

Both conditions rely on *a priori* judgments. There is nothing wrong with this, but of course the conditions are open to empirical assessment when one gets to specific applications, and priors may vary on the extent of their validity.[4] The application of these ideas in Ambuehl, Bernheim and Lusardi [2014][2017][2022] and Ambuehl, Bernheim, Ersoy and Harris [2018], reviewed in Bernheim [2019; p.51ff.], provide such an instance.[5]

The application in each case is the same, and tests comprehension of the concept of compound interest as it affects intertemporal choices between a smaller, sooner (SS) amount of money and a larger, later (LL) amount of money. This is a canonical task for the elicitation of time preferences: see Coller and Williams [1999] for an extensive review of the older literature and clean experimental implementation of this task. To illustrate, consider these two statements, which very slightly paraphrase those actually used:

A. You will receive $88 in 72 days.

B. We will invest $22 at 3% interest, compounded daily, for 72 days.

Subjects are then asked, in response to one of these statements, to say "what is the present amount that is equivalent?" Responses are elicited using an Iterative Multiple Price List (iMPL) procedure developed by Andersen, Harrison, Lau and Rutström [2006], and can be assumed for present purposes to lead subjects to reveal their true answer in an incentive compatible manner.

If subjects exhibit financial literacy they "should" give the same answers in response to statements A and B, since we observers know that the amount of money in B will end up being $88 in 72 days. If the

[4] This is the approach adopted in Ambuehl, Bernheim and Lusardi [2014], to view one of the frames as revealing true, virtual valuations. In Ambuehl, Bernheim and Lusardi [2017] this position was qualified, allowing that there might be some normative metric that does not lead one to accept that either frame represents the true, virtual valuation. The example provided is when subjects exhibit Quasi-Hyperbolic discounting in response to both questions, with Exponential discounting *a priori* deemed to be normatively attractive and Quasi-Hyperbolic discounting deemed *a priori* to be normatively unattractive. In this case, they claim, both responses might be "contaminated" by the "passion for the present" one expects from Quasi-Hyperbolic responses. They then present a formal result that essentially says that if the responses to statements A and B are equally contaminated, then as one takes the *limit of the difference between the responses as that difference goes to zero*, a first-order approximation to a valid welfare measure can be obtained. But that says nothing about whether the difference between the responses that are non-zero, or not close to zero, have any valid interpretation, unless one wants to invoke stringent path-independence assumptions from welfare economics (see Boadway and Bruce [1984; p.199] or Harrison, Rutherford and Wooton [1993]). The bulk of responses of interest are decidedly non-zero, and not close to zero, as illustrated in Ambuehl, Bernheim, Ersoy and Harris [2018; Figure 1, p.16].

[5] An application of the same methodology to retirement savings plans is provided by Bernheim, Fradkin and Popov [2015].

answers to A and B differ, then we have identified a financial literacy gap, and it is asserted that we can take the absolute value of the difference in valuations as a measure of the welfare loss from that gap. Since the present value amounts are stated in deterministic form, this welfare loss is in the form of a certainty-equivalent. In effect, here, the observed choice is a willingness to exchange the LL amount mentioned or implied by statement A or B for the SS amount stated in the response elicited by the iMPL procedure.

Now consider whether statements A and B meet the conditions required for inferences about welfare loss due to financial illiteracy.

One immediate concern is that statement B might be interpreted, from a conversational perspective, as already providing the answer: surely it is $22. The interpretation is that you have been asked what amount of money today would generate the implied $88 in 72 days, and this must be a "trick question" because the statement already tells you that it was $22. Of course, we analysts are expecting subjects to tell us the present discounted amount that is equivalent to $88 in 72 days, where the discount rate need not be the same as the interest rate, but that is just one interpretation of the question. One might expect, if inspecting the raw data, to see many respondents simply say $22 in this instance.

Another, more subtle, interpretation issue concerns the information about a 3% interest rate. A subject might reasonably presume that this is taken to be the market (borrowing and lending) interest rate for this question. Then we know from the Fisher Separation Theorem that we cannot recover estimates of discount rates due to censoring: see Coller and Willams [1999]. All that we would recover is the subject's knowledge of the interest rate, which is again included in statement B, hence we would again expect a spike of responses at $22.

Extending this point, the mere mention of interest rates might affect responses differently for statement B compared to statement A. In effect, statement B offers a cognitive scaffold that could be expected to change the response compared to statement A, where there is no such explicit scaffold mentioned. Thus what is claimed to be the welfare effect of literacy might just be the welfare effect of having access to a more specific[6] scaffold, and that is ambiguous as a theoretical matter.

Finally, any difference between responses to statements A and B might simply reflect an inability to apply the principle of compound interest in evaluating statement B, to arrive at the implied $88 correctly. A subject might understand what compound interest is, and just not be able to do the math on the spot, even with a calculator provided. The issue here is whether one labels any difference in present value responses a welfare-significant failure of literacy with respect to the concept of compound interest or a welfare-significant failure of the ability to *apply* the correct concept (recall the earlier distinction between literacy and capability). And the focus throughout Ambuehl et al. [2014][2017][2018][2022] is on the effect of an intervention to improve decision-making, whether or not it is literacy or capability that is driving the effect.

One overarching concern here is that to apply the method of Berneim and Rangel [2008][2009] one must find frames that convince readers that they meet the two conditions noted earlier, and this is not likely to be an easy task across domains. Their method is not, in this sense, a general method.

Our concern about generality in this approach can itself be generalised. All economically interesting choice in humans and other intelligent social animals relies on complex networks of scaffolding.[7] Consequently, the frames methodology makes welfare assessments parasitic on an indefinite number of theories of these myriad exogenous influences that structure choice. We are glad that cognitive anthropologists and others are working hard on such theories and on the fascinating empirical findings that underpin them (see Caporael, Griesemer and Wimsatt [2014]). But importing them into welfare analysis undermines economic generalization in the same way as do flavors of behavioral economics that implicitly collapse the field into the psychology of valuation: they fracture the equivalence classes on which economists and policy-makers depend. The classes in question are selected by reference to *outcomes*,

[6] We say "more specific" to stress that literacy is also scaffolding.

[7] Literacy is itself a general scaffolding architecture, which provides people who can read with access to more specific elements of scaffolding such as technical or popular investment manuals.

not generative mechanisms or enabling conditions. Of course, we want to know about such mechanisms and conditions to be confident that a recommended policy doesn't accidentally interfere with their operation. But we seek welfare specifications that are robust against measurable heterogeneity, not specifications that simply splinter in response to it. The frames approach does not so much rescue welfare economics as supplant it by mixed applications of other sciences. Furthermore, being mixtures, such models tend to lack *any* unifying disciplinary foundations. Every experimental design becomes a fresh adventure in *de novo* identification.

### *D. Identifying the Inner Utility Function*

Many who view RDU or CPT as a better descriptive model of risk preferences nonetheless view EUT as an appropriate normative model of risk preferences. This raises an important practical issue: if all you have before you as an observer is someone exhibiting RDU or CPT behavior, how do you recover the utility function you need to undertake normative evaluations?

One approach is to simply impose EUT on the estimation of risk preferences that are observed, and use the utility function that is then inferred. This approach is used, for purposes of exposition, by Harrison and Ng [2016].

Bleichrodt et al. [2001] maintain that EUT is the appropriate normative model, and correctly note that if an individual is an RDU or CPT decision-maker, then recovering the utility function from observed lottery choices requires allowing for probability weighting and/or sign-dependence. They then implicitly propose using *that* utility function to infer the CE using EUT. This is a radically different normative position than the one proposed by Harrison and Ng [2016; p.116].

Some notation will help. Let RDU(x) denote the evaluation of an insurance policy x in Harrison and Ng [2016] using the RDU risk preferences of the individual, including the probability weighting function. They calculate the CE by solving $U^{RDU}(CE) = RDU(x)$ for CE, where $U^{RDU}$ is the estimated utility function from the RDU model of risk preferences for that individual. But Bleichrodt et al. [2001]

evaluate the CE by solving $U^{RDU}$(CE) = EUT(x) where EUT(x) uses the $U^{RDU}$ utility function in an EUT manner, assuming no probability weighting. This is normatively illogical. The logical approach here would be to estimate the "best fitting EUT risk preferences" for the individual from their observed lottery choices, following Harrison and Ng [2016], and then use the resulting utility function $U^{EUT}$ as the basis for evaluating the CE using $U^{EUT}$(CE) = EUT(x), where EUT(x) uses the same $U^{EUT}$ function used to evaluate the CE.

### *E. Modeling Mistakes*

Another way to undertake normative evaluations is to develop a structural model of mistakes, and consider the effects on behavior of removing those mistakes. The issue here is whether the modeled behavior is indeed reasonably classified as a mistake or not, and from whose perspective. Behavioral economists have not been shy to quickly label any "odd behavior" as due to the first heuristic that comes to mind, rather than dig deeper. Harrison [2019; §5.2] provides a critical review of structural models of this kind applied to health insurance and income annuity choices.

An example in which the modeled "mistake" could obviously be due to a simple risk preference is instructive of the dangers of this approach. Handel [2013] exploits a natural experiment in which a large firm changed health insurance options from an active choice mode to a passive mode where the previously selected choice was the default choice in later years unless action was taken. This change allowed inferences about the role of "inertia" in insurance plan choice. The behavior of new employees, who needed to make an active choice when previous employees were faced with passive choices, provides potential insight into the significance of inertia, assuming comparability of other characteristics between the two employee groups. In addition, some passively enrolled employees faced dominated choices over time as insurance parameters changed, and their sluggishness in the face of these incentives provides indicators of inertia. Atemporal risk preferences are assumed to be distributed randomly over the population sampled, and be consistent with EUT. Individuals know their own risk preferences, but these are unobserved by the analyst.[8] In keeping with other observational studies of health insurance, the distribution of claims was simulated using sophisticated models akin to how an actuary would undertake the task, and individuals were assumed to know the risks they faced exactly.

Since the focus is on "inertia" over time, an important behavioral omission is the implicit assumtion that individuals are *intertemporally risk neutral*. Hence, whatever the implied *atemporal* risk aversion from the random coefficient estimation, individuals in this model are unable to exhibit inertia in choices due to intertemporal risk aversion. This is quite separate from the assumption that "consumers are myopic and do not make dynamic decisions whereby current choices would take into account inertia in future periods" (Handel [2013; p. 2662]). Those assumptions have to do with sophistication with respect to the effect of current consumption on future consumption, akin to "rational addiction" models. Intertemporal risk aversion is just a taste for not having variability in claims risks over time, and that is met simply by choosing the same plan year over year. Just as one is willing to pay a risk premium in terms of expected value to reduce atemporal risk aversion, the willingness to put up with lower *expected value* plans can be seen as a risk premium to reduce intertemporal risk aversion. This has fundamental implications for the resulting welfare analysis (p.2669-2679). The story here is that "consumers enroll in sub-optimal health plans over time, from their perspective, because of inertia. After initially making informed decisions, consumers don't perfectly adjust their choices over time in response to changes to the market environment (e.g., prices) and their own health statuses" (p.2669). Another story, equally consistent with the observed choices and EUT, is that consumers simply have a preference for avoiding intertemporal risk in the health plan lotteries they choose.

Another approach is to use structural models of noisy decision-making to infer a measure of lost welfare by finding the least such noise that can rationalize observed behavior. This exercise should always

[8] This might cause identification problems if the "nonfinancial attributes," to use the expression of Handel and Kolstad [2015], also varied across all plan choices, but three PPO plans had no differences in these attributes: hence their variations in "financial attributes," such as deductible, coinsurance, and out-of-pocket maxima, could be used to identify atemporal risk preferences. The presumption is that individuals do not subjectively believe that these attributes differ across these PPO plans.

be undertaken when one allows for standard errors on preference parameters, since that is another source of noise. However, the popular Fechner or Tremble models of noisy decision-making are a part of the hypothesized economics. The Fechner model, for example, assumes that the agent can evaluate the EU (or RDU) of two lotteries, that the agent can take the difference in the EU (or RDU), but that when making a choice the agent inflates or deflates that difference by some amount. In the extreme the deflation collapses to a zero difference between the EU (or RDU) of the two lotteries, such that the agent is indifferent between the two and always selects one option with a probability of ½. In the extreme the inflation expands the economic significance of any difference in EU (or RDU) such that the agent behaves as if strongly motivated to select the option with the highest EU (or RDU) with near certainty, no matter how small the initial pre-inflation difference in EU (or RDU). This is a story about how sensitive an individual is to differences in EU (or RDU) in terms of how those differences translate into probabilities of choosing the more attractive lottery.[9]

These piecemeal adjustments, however, fail to drill down to a foundational level, for more general re-orientation on the problem raised by choice errors for estimating welfare. In developing our more general approach to behavioral welfare economics in §3, we take advantage of the extensive work in cognitive science, over many decades, of distinguishing functional responses to ecological and strategic problems faced by agents from noisy processes. Much of the work in question has been inspired by practical issues in artificial intelligence (AI). The designer of a programmable robot with a narrow intended application can try to get noise close to zero. The designer of a system intended to be relatively flexible and autonomous cannot; she must accept noise, and learn to theoretically control and contain it,

[9] Measures developed by Alekseev, Harrison, Lau and Ross [2018] to implement this idea for models of risk preferences are, methodologically, similar to the Critical Cost Efficiency Index (CCEI) of Afriat [1972], which is used to evaluate the degree of consistency with the Generalized Axiom of Revealed Preference (GARP). The CCEI relative cost measures is defined on the unit interval, and its complement shows what proportion of monetary value an agent should be allowed to waste in order to rationalize her choices by some utility function. While GARP provides qualitative statements, Alekseev et al. [2018] put more structure on the estimation procedure and provide quantitative evidence of welfare costs from observed choices. This approach can be applied to any model of risk preferences that admits of one or other model of noisy choice.

as a side-cost of making a system able to recognise problems coming from the world as being the same or similar even if they vary significantly in the processes and conditions that generate them. In AI this is known as "the frame problem" (see Pylyshyn [1987] and Ford and Pylyshyn [1996]). It is logically isomorphic to the welfare economist who seeks to discover and design effective policies.

### *F. Reduced Form Inferences*

Townsend [1994] initiated a major stream of research by examining the response of household consumption to income shocks. Examining data from villages in rural India, he found that "household consumptions are not much influenced by contemporaneous own income, sickness, unemployment, or other idiosyncratic shocks, controlling for village consumption (i.e. for village level risk)" (p. 539).[10] Under certain, strong assumptions, evidence that consumption remains "stable" over time in relation to relative volatility of income indicates that there is likely to be small welfare gains, if any, from "social insurance" schemes.

Because of the influence of this approach, it is worth noting the explicit methodological position that motivated it. Townsend [1994] was well aware of the long list of mechanisms and institutions that might provide informal insurance, noting family transfers among villages, informal credit markets, plot and crop diversification, and animal sales. And rigorously documenting this type of long list has occupied him in later work in rural Thailand: see Samphantharak and Townsend [2010]. However, Townsend [1994; p. 540] argues that

> ... in studying one market or institution only, the researcher may miss smoothing possibilities provided by another. For example, transfers may be small or missing, but this may not leave the family vulnerable if credit markets function well. [Hence this study] presents a general equilibrium framework which overcomes the problem of looking at risk-sharing markets or institutions one at a time. Specifically, the general equilibrium model inevitably leads the researcher to focus on outcomes, namely, consumption and labor supply, so that all actual institutions of any kind are jointly evaluated.

[10] This conclusion was qualified in some villages for those that did not own land.

One concern with this position is that a general equilibrium structure is used to generate "reduced form" results which are then empirically evaluated, without the economist being able to go back and verify the structure.

Chetty and Looney [2006; p.2352] note, with citations, that "the presumption that consumption fluctuations give a measure of the welfare costs of risks, and therefore the value of additional insurance, remains prevalent."[11] However, Baily [1978] had much earlier identified an important trade-off between the factors causing benefits from consumption smoothing (higher risk aversion) and the factors causing costs of smoothing consumption in the design of optimal unemployment insurance. Focus on the latter: in a world of complete and perfect markets, these costs are low. Absent these imaginary markets, it is often presumed that private or informal insurance mechanisms at the individual, household, village, state or national level somehow act as if providing "full insurance" against consumption variability. Or in the debate over the roles of social *versus* private insurance, that private insurance serves to do what social insurance proposes doing. However, the logic proposed by Baily [1978] implied that evidence of consumption smoothing in Townsend [1994] might just be evidence of *extremely* high risk aversion and *inefficient* risk management options. As long as the demand for risk reduction is high enough, even wasteful risk management schemes will be tolerated. A review of the vignettes from the *Portfolios of the Poor* financial diaries, by Collins, Morduch, Rutherford and Ruthven [2009], tells of the myriad, costly risk management schemes needed to understand "how the world's poor lives on $2 a day." Chetty [2006] and Chetty and Looney [2006] show precisely how this logic applies to understand the identification issues that plague the conclusions from Townsend [1994] about the potential welfare gains to households from social insurance.

[11] The subsequent literature on full or partial insurance, inferred from such correlations, continues. For example, Blundell, Pistaferri and Preston [2008] conclude from U.S. data that there is "some partial insurance of permanent shocks, especially for the college educated and those nearing retirement [and that there is] full insurance of transitory shocks except among poor households." (p. 1887).

## 3. Behavioral Welfare Economics with the Quantitative Intentional Stance

Economists have long realized that *if one assumes some risk preferences*, it is possible to make normative statements about the observed demand for certain products in which risk plays a critical role. For example, Feldstein [1973] proposed that, on average, U.S. households carried too much health insurance. Armed with estimates of a measure of risk aversion, a price elasticity of demand for health care, the (gross) price change induced by lower insurance coverage, and the decrease in health care quality induced by lower insurance coverage, he estimated that the CE of the EUT loss from reduced insurance coverage would be more than offset by the gain from reduced purchases of lower-priced health care. His estimate of risk aversion was derived (p. 274) from casual introspection about what he regarded as a plausible risk premium for a hypothetical bet of a 50:50 chance of ±$1,000.[12] The challenge is to go beyond just assuming some risk preference and to utilize risk preferences that have some claim to be appropriate for the agent being evaluated.

Harrison and Ng [2016][2018] and Harrison and Ross [2018] estimate the best descriptive model of risk preferences for individuals, and use these estimates to make normative evaluations of the insurance and investment product choices of their laboratory subjects. They justify this method by appeal to a leading interpretation of preferences in cognitive science.

It is unlikely that most people choosing insurance contracts or investment funds attempt to compute internally represented optima, either from EUT or RDU bases, and then make computational errors that could be pointed out to them. This echoes a point made by Infante, Lecouteux and Sugden [2016] when they complain that behavioral welfare economists typically follow Hausman [2011] in "purifying" empirically observed preferences. Infante et al. [2016] argue that purification reflects an implicit philosophy according to which an inner Savage-rational agent is trapped within a psychological, irrational shell from which best policy should try to rescue her. They provide no general theoretical

[12] Feldman and Dowd [1991] updated these calculations with later, improved estimates of the moving behavioral parts. Their estimates of risk aversion came from econometric estimates in the health insurance literature based on comparable observational data and survey questions.

framework within which they motivate their skepticism about "inner rational agents." However, such a framework is available from cognitive science.

Long ago, Aristotle distinguished between four kinds of what translators render as "causation": material, efficient, formal, and final causes. Most contemporary philosophers are reluctant to fracture the concept of causation. But Aristotle's distinctions can readily be updated as referring to different *principles of explanation*. If I explain a change in the price of pizza in Cork by analyzing changes in prices of inputs and demand elasticities from data, then I appeal to efficient "causation." If my explanation adverts to equilibrium in an applied Bertrand model, then Aristotle would see me as giving a formal explanation. And if my story has recourse to a regulator who has intervened to break up a local pizza cartel, I am in the domain of final "causation." Aristotle interpretation is of course not what we are interested in here. We mention him in order to make the point that it has long been recognized that explanations are typically *partial* and *jointly complementary*. An economist might publish a whole study that just focuses on one of the types of explanation and mentions the others only in footnotes. This does not imply that the relegated explanations are false.

Someone might object that we should see the different explanations as stepping-stones along the path to a "complete" explanation. But it is not clear what that might mean in practice. What would a "complete explanation" of pizza prices look like? Would it have to include a general equilibrium model of the global economy? Would it involve a chemical analysis of what makes some cheeses but not others suitable for melting on pizza crust? Economists do not aspire to complete explanations. No sensible scientist does.

The modern extension of Aristotle's general point to the cognitive and behavioral sciences has been fleshed out by Daniel Dennett [1971][1987]. He calls complementary styles of explanation "stances." In cognitive and behavioral science, he argues that we find coherent literatures that make use of three stances, which he calls "physical," "design," and "intentional." If we carry these distinctions into economics, then as in the case with Aristotle we get some semantic awkwardness. For reasons best reserved for a philosophy of science setting, we think that the "physical" stance should more accurately be called the "mechanism" stance, even in its home domain of Dennett's applications. The "design" stance is suitably named for use in biological ecology and evolutionary theory, where Dennett applied it, but this would cause confusion in economics and trip over the fact that economists use the word "design" – in the concept of "mechanism design" – in a way that doesn't map neatly onto what Dennett means. We think that if the "design stance" were renamed the "selection stance," it would still be a good label for Dennett's motivating cases, but also avoid muddying waters in economics.

We will focus here on the intentional stance, which can comfortably keep its name in our context. We mention the others only to help readers understand what a stance is. Instead of defining them, we will illustrate them by reference to economic applications. Suppose an economist wants to predict comparative trade volumes between countries. She would be advised to use a gravity model, which focuses on the micro-scale processes – operational ranges of trucking firms, shared transport infrastructures, cultural and social underpinnings of business networks – through which physical movements of goods and services are implemented. This is an instance of a mechanism stance. But she might be more interested in the strategic and policy-sensitive aspects of international trade. In that case she would control for geographical distance (that is, control for what the gravity model highlights) in a modernized Ricardian, Heckscher-Ohlin, or political economy model. Here her focus is on factors that select for relative dispositions to form trade links implemented by varying transport and business mechanisms across cases treated as equivalence classes in the model. Here she adopts a selection stance. This does not imply rejecting her colleague's gravity modeling; the two imagined economists address different aspects of the same general cluster of phenomena, international trade patterns. The alternative stances are not reflections of subjective hunches about what "really" matters. Each is characterized by rigorously developed families of models and easily identified, distinct, literatures.

Now suppose that further down the corridor in the Economics Department, a third colleague is trying to predict which interest group lobbies in a country will support or reject a newly negotiated trade deal awaiting legislative ratification. She will focus most directly on the utility functions of different firms, labor unions, and sub-national governments. Her explanations will be based on the intentional stance.

### *A. The Intentional Stance*

Dennett [1971][1987] provides a rich account of the relationships between beliefs, preferences and propositional attitudes[13] that provides a rigorous foundation for behavioral welfare economics. He argues[14] that the *attribution of preferences and beliefs involves taking an intentional stance toward understanding the behavior of an agent. This stance consists in assuming that the agent's behavior is guided by goals and is sensitive to information about means to the goals, and about the relative probabilities of achieving the goals given available means.* The intentional stance is a product of cultural evolution. It arose and persists because of the importance of coordinated expectations in an intensely social species with massive behavioural heterogeneity due to large brains that support sophisticated learning. Beliefs, preferences, goals, and other propositional attitudes do not have counterparts at the level of brain states. They instead index relationships between target agents,

[13] Philosophers group statements identifying beliefs, desires, preferences, hopes, fears, wishes, and their boundless counterparts and conjugates across human languages, as all being about *propositional attitudes*. The point of this terminology is that all of these verbs refer to views that a user of language can take to an actual or hypothetical state of affairs in the world. For example, one person might *believe that* Acme stock will crash, while a holder of the stock *fears that* this is so and another investor who has shorted it both *hopes* that the stock will fall and *prefers* to say things that encourage others to affirm the belief. The word "proposition" here signifies that the relevant "state of affairs" (i.e., the fate of Acme stock) is not a *linguistic* item, because a unilingual speaker of English could share the above attitudes with a unilingual speaker of Tagalog or Mandarin. The most difficult and perennial problems in applied logic arise from complexities in propositional attitude *reports*. Metaphysically, their status is bound up with what it is to be or have a mind. That is, a "mind" is often said by philosophers to be a delimited entity or process that can truly be said to express and behaviorally respond to propositional attitudes. This claim is practically central to identifying the theoretical subject matter of cognitive science, and to the goals of artificial intelligence research. A few philosophers doubt that propositional attitudes, and therefore minds, can ever be proper objects of scientific study, and urge that they should ultimately be explained away. The philosopher Alex Rosenberg [1992], for example, argues that economics can never be good science as long as its theory makes ineliminable reference to preferences and beliefs. A main career objective of Dennett, on whose ideas we rely, has been to cleanly integrate study of propositional attitudes and their uses into the general scientific worldview, thereby answering sceptics such as Rosenberg. A critical survey of the debate is provided by Ross [2005].

[14] The origins of debates about propositional attitudes in philosophy, described above, might lead a reader to suppose that Dennett's arguments must be *a priori*. In fact, they are almost exclusively empirical, drawn from details in the uncontroversial success of some scientific research programmes based on the intentional stance, such as the main streams of human developmental psychology, cognitive ethology, and the "deep learning" (or "connectionist") branch of artificial intelligence.

environments, and interpreters trying to explain and anticipate the target agents' behavior (including their communicative behavior). The welfare economist attempting to determine what people regard as subjectively preferable is in the same situation as all people in all social contexts all the time: she seeks accounts of her targets' lattices of propositional attitudes, with particular emphasis on preferences and beliefs about probabilities, that the targets would endorse themselves. She is *not* trying to make inferences about anyone's "latent" states or states that are hidden in brains until someone with a neuroimaging scanner comes along.

Dennett's analysis of propositional attitudes is a scientifically inspired *revision* to everyday "folk psychology." Most people, when they express *their own* preferences and beliefs, take themselves to be publicly reporting private facts about themselves. A major conceptual reform on which cognitive scientists have gradually converged[15] over the past four decades is that this is a *useful illusion*. It arises because people are socially and culturally obliged to take the intentional stance toward *themselves*, and to self-attribute networks of propositional attitudes that others can understand and use as input for coordination. Because this self-interpretation is continuous, except in times of crisis or major life-course discontinuities, its specific elements for a person become habitual and non-reflective. "I am a lover of unspoiled places," reports the environmental activist in response to an open-ended range of social prompts she encounters from day to day, which are made relevant by her various activist activities and arguments. Because no conscious deliberation is involved in these reports, she takes them to reflect direct "inner perception" of feelings and valuations that must be "in her brain," if she lives in a modern scientific culture, or "in her heart," if her culture's metaphysic of the self is traditional or partly mystical. But there is no such literal process as "inner perception" of propositional attitudes (Lyons [1986], Dennett [1991], Schwitzgebel [2011]). The most direct evidence comes from developmental psychology. Children learn to adopt the intentional stance toward others, and *then* apply it to themselves (McGeer [2001]). They can tell plausible

[15] The relevant literature here is vast. Hood [2012] reviews evidence up to a decade ago, and Spivey [2020] draws on more recent developments.

narrative stories about others before they can produce fragments of autobiography. Furthermore, early autobiographical narratives are typically generated by parental encouragement and *correction* (McGeer [2020]), which would make no sense if they were reports of inner perceptions.

The folk psychological view of beliefs and preferences as inner, private states has encouraged the widespread idea that people form expectations about one another's behaviour and communications by "mindreading" (Nichols and Stich [2003]). This idea will be familiar to economists from the standard stylised psychological stories that often accompany models of extensive-form games in which players infer utility functions of other players from observed play. However, Zawidzki [2013] musters evidence that accurate mindreading cannot be carried out in real time except among very closely entangled agents, and that the evolution of mindreading mechanisms could not have been supported by the natural selection of cortex. Zawidzki [2013] convincingly shows that people mainly achieve coordination (to the extent that they do) by *mindshaping* – that is, by influencing one another to conform their preferences and beliefs to narrative models that can easily be socially recycled. Mindshaping is not mere mindless imitation that an economist could model only using something like a replicator dynamics. It is more accurately conceived as high-speed bargaining with a strong strategic dimension. Ross and Stirling [2021] offer a general, formal framework by which game theorists can model it.

The recognition that inner propositional attitudes apprehended by introspection are useful illusions has led some theorists to conclude that they are *mere* illusions, as noted earlier. With application to the design and interpretation of economic lab experiments, this is basically the view urged by Chater [2018], who argues that behavioral scientists should cease to use the concepts of preference and belief in theoretical generalisations. Sugden [2018] also advances some arguments that seem to imply this position. A more common view among both philosophers and cognitive scientists is that Dennett's account of propositional attitudes amounts to instrumentalism about them. This idea emphasises both the "illusory" and the "useful" parts of "useful illusion." "Instrumentalism" about the scientific use of a concept is the word used in philosophy of science when the concept in question is thought to not designate anything that objectively exists in the world, but to be a helpful tool for constructing theory. The original home of the idea was in philosophies of physics according to which there are no real atoms, essential though the concept of an atom is to theory in physics and chemistry.

Instrumentalism is familiar to any economist, even if she is not acquainted with the philosophers' word for it. It is the hugely influential philosophy of economics promoted by Milton Friedman [1953], and many followers over the years. Friedman argued that literal utility and production functions are operated only by economic modelers, not by the actual economic agents they model. Economists use these concepts because they generate good predictions of individual and firm behavior, according to Friedman; but because the concepts have no counterparts in extra-theoretical reality, economists should not regard assumptions cast in terms of them as making any empirical commitments that could be tested against data.

Friedman's instrumentalism finds no support among specialists in economic methodology.[16] One of many bases for objection is that it implies that economists never explain any empirical facts, and should not attempt to do so. Similarly, if the intentional stance is a form of instrumentalism, but is the correct account of beliefs and preferences, then no models in cognitive science that invoke these concepts could ever explain anything. Dennett has thus made various reluctant forays over the years into general philosophy of science (which is not his field) to try to ward off associations with instrumentalism. Ladyman and Ross [2007, chapter 4] show that Dennett's efforts have not been entirely successful, but that Dennett provided a clear pathway that they complete by applying some technical resources from computational information theory.

We avoid getting into these deep philosophical waters here. Instead, we gloss the issue informally as follows. According to the intentional stance, beliefs, preferences, and other propositional attitudes are virtual or interface elements of reality. They denote systematic, recurrent patterns that arise in the relationships between brain-powered organisms and their external environments, particularly including, in

[16] Mäki [2009] reviews the history and critiques of Friedman's instrumentalism.

the case of humans and other highly intelligent animals, social environments.[17] Being virtual is not a way of being fictional; it is a way of being real. Propositional attitudes are scientifically studied, and empirical claims about them are true or false. We need merely acknowledge that these claims are not, in general, decided by facts about brains. It is worth drawing attention here to another example of a type of virtual object, money. There has never been a shortage of pop-philosophical claims to the effect that "money isn't real." But that opinion deserves to be called "the silly stance." Thinking that facts about beliefs and preferences must ultimately boil down to facts about brains is like thinking that facts about a country's money supply must ultimately boil down to facts about its printing presses and mints. They don't; but there are still plenty of facts about money.

The intentional stance implies that propositional attitudes are states of entire intentional systems (e.g., a person), not states of inferred *parts* of systems. Economists still often use the expression "latent states," borrowed from a transition period in psychological theory when post-behaviorists had restored mental states to scientific status, but supposed that they must ultimately all reduce to hidden brain states that awaited identification by a future neuroscience. This use by economists should be discontinued as invoking obsolete psychology.[18] The holistic nature of intentional stance description of agent behavior allows for error, but also complicates it: as stressed by Hey [2005], the "behavioral error" stories that we append to our structural models are part of the economics.[19]

Ross [2014] argues that this marks a fundamental basis for the distinction between economics and

[17] No highly intelligent animals are asocial, with the possible exception of a few species very recently descended from intensively social ancestors.

[18] There are important uses of the word "latent," in statistics, that don't carry these connotations. We are not, for example, advising anyone to stop talking about latent indices. Linguistic theorists use the word in yet another way that isn't threatened by changes in prevailing psychology. Instances of such polysemy abound in the history of science. Wilson [2006] refers to cases where more than one discipline technically co-opts the name of the same pre-scientific concept as "wandering significance." He demonstrates the importance of such histories by documenting instances where unregulated interactions between rival semantic anchors produced unintended consequences for the evolution of theories and experimental traditions.

[19] To add complication, they interact directly with the stochastic specifications that attend to sampling errors in the econometrics, and hence inferences about preferences: see Wilcox [2008][2011] for a masterful review in the case of risk preferences.

psychology. The intentional stance does not deny that brains are information processors, or that such information processing is a crucial *part* of the causal vector that supports attribution of propositional attitudes to people. The intentional stance simply denies that these important brain states should be *identified* with beliefs and preferences, or that brain processes should be modelled as inferences that have beliefs and preferences as conclusions. Psychologists are professionally interested directly in these processes, and in how they influence decisions. Economists, by contrast, are concerned with this only derivatively. If a system of incentives will lead various people, through heterogeneous sets of psychological *along with* social, institutional, financial, and technological processes, to all make the same choices, then the people form, at least for an analysis restricted to that choice, an equivalence class of economic agents. But it is a strictly empirical matter when this heterogeneity with respect to processes will and won't matter economically. Economists, like all scientists, seek generalizations that support out-of-sample predictions. Different data-generating processes tend to produce, sooner or later, different data, including different economic data. Economics is thus crucially informed by psychology in general, as is sociology and anthropology, while not collapsing into the psychology of valuation as some behavioral economists have urged (e.g., Camerer, Loewenstein and Prelec [2005]).

Applying this understanding of mind and agency to the applications to insurance in Harrison and Ng [2016], we assume the intentional stance to make sense of the experimental subjects' overall behavioral patterns, and use the risky lottery choice experiment as a relatively direct source of constraint on the virtual preference structures we assign when we perform welfare assessment of their insurance contract choices. The more precisely we specify the contents of propositional attitudes, especially in quantitative terms, the less weight in identification will rest on "inboard" elements of data generating processes relative to external aspects of the agents' overall behavioral ecologies (i.e., cognitive scaffolds). This somewhat subtle point is of crucial methodological significance. Biological brains, with their dynamic, highly distributed, and essentially analog processing architectures, face strong limits, relative to digital representational technologies, on the extent to which they can stably discriminate between magnitudes. If a theorist conceives of beliefs and preferences as brain states, therefore, then the levels of precision assignable to these states will fall short of the discriminations that matter to theory, particularly in such domains as finance and insurance. But of course people *can* make these finer discriminations, by reading written numbers and using computers. They do not need to store stable quantitative representations "in their heads" (as the common metaphor puts it) when the representations in question are stored on paper or in external data files.

It might be objected here that subjects can only be attributed beliefs and preferences about quantitative and mathematical representations they actually understand. This is basically correct, but the simple statement hides much complexity. People learn to understand mathematical expressions by using them. As everyone who has taught mathematics or statistics to students knows, people can typically respond behaviorally to distinctions before they can correctly write them down or code them. Where normative consulting interactions are concerned, the financial or actuarial advisor can play a crucial pedagogical role. Of course this opens scope for concerns about paternalism, which we address later. The key point for now is that the intentional stance does not require the attributor of preferences and beliefs to stick to those that a naïve subject could apply to themselves.[20] Experimental treatments[21] might provide evidence that attention to certain informational patterns induces a significant number of subjects to act as if they were stochastically closest to being EU optimizers; evidence about other subjects might indicate patterns of probability weighting, as in RDU. These patterns therefore enter into a fully informed analyst's specification of the subjects' beliefs and preferences.

[20] The discussion here is crucial to answering the common objection heard from some behavioral economists, and external critics of economic theory, that people do not or cannot perform the "mental computations" that most microeconomic models specify. The discussion simultaneously shows why the right answer to such critics is *not* to join Gul and Pesendorfer [2008] in asserting that psychological processing is *a priori* irrelevant to economists. It is relevant exactly and only when it makes an economic difference, and this is an empirical matter to be assessed from one application to another.

[21] For example, the informational treatment of Harrison and Ross [2018] with respect to investment decisions, or the various informational treatments of Harrison, Morsink and Schneider [2020] with respect to index insurance decisions.

### *B. An Illustrative Application*

Armed with some rigorous basis for assessing the benefit or harm to an individual from some experimental treatment, how do we make it operational? One general recommendation is to use Bayesian methods. The reason that this recommendation is general is that integrating economic theory with experimental data entails the systematic pooling of priors with data, and that is what Bayesian methods are designed to allow. And, critically, we view the attribution of preferences and beliefs that is central to the QIS as exactly akin to forming priors *about the agent*, and then pooling them with observations *of the agent* to make (normative and descriptive) inferences.

For economists, a canonical illustration of the need to pool priors and data is provided by the evaluation of the expected CS from observed insurance choices. Even if we limit ourselves to EUT, the gains or losses from someone purchasing an insurance product with known actuarial characteristics depend on their (atemporal) risk preferences. If we have priors about those risk preferences, then we can directly infer if the observed purchase choice was the correct one or not, as illustrated by Feldstein [1973]. Here the word "correct" means consistent with the inferred EUT risk preferences for the individual making the choice we evaluate normatively. The same point extends immediately to non-EUT models of risk preferences, such as RDU. From a Bayesian perspective, this inference uses estimates of the posterior distributions of individual risk preferences to make an inference over "different data" than were used to estimate the posterior.[22] Hence these are referred to as *posterior predictive distributions*.

In the simplest possible case, considered by Harrison and Ng [2016], subjects made a binary choice to purchase a full indemnity insurance product or not. The actuarial characteristics of the insurance product were controlled over 24 choices by each subject: the loss probability, the premium, the absence of a deductible, and the absence of non-performance risk. In effect, then, these insurance purchase choices are just re-framed choices over risky lotteries. The risky lottery here is to not purchase insurance and run

[22] The usual application in Bayesian modeling is to additional out-of-sample instances of the same data used to estimate the posterior. A typical example would be to predict choices by one of our subjects if she had been offered a new, different battery of choices over risky lotteries.

the risk of the loss probably reducing income from some known endowment, and the (very) safe lottery is to purchase insurance and deduct the known premium from the known endowment.

The same subjects that made these insurance choices also made choices over a battery of risky lotteries, and a Bayesian model can be used to estimate individual risk preferences for each individual from their risky lottery choices.[23] The task is then to infer the posterior predictive distribution of welfare for each insurance choice of each individual. The predictive distribution is just a distribution of unobserved data (the expected insurance choice given the actuarial parameters offered) conditional on observed data (the actual choices in the risk lottery task). All that is involved is marginalizing the likelihood function for the insurance choices with respect to the posterior distribution of EUT model parameters from the risk lottery choices. The upshot is that we predict a *distribution* of welfare for a given choice by a given individual, rather than a *scalar*.[24] We can then report that distribution as a kernel density, or select some measure of central tendency such as the mean or median.

Figure 1 displays several posterior predictive distributions for insurance purchase choices by one subject. For choice #1 the posterior predictive density shows a clear gain in CS, and for choice #4 a clear loss in CS. In each case, of course, there is a distribution, with a standard deviation of $0.76. The predictive posterior distributions for choice #13 and choice #17 illustrate an important case, where we can only say that there has been a CS gain with some probability.

This example allows us to illustrate how one can undertake *adaptive* welfare evaluation during an experiment, following Gao, Harrison and Tchernis [2022; §3.C].[25] Some of the subjects in this experiment

[23] Details are provided in Gao, Harrison and Tchernis [2022]. A Bayesian hierarchical model was used in which informative priors for the estimation of individual risk preferences were obtained by assuming exhangeability with respect to the risk preferences of other individuals in the sample. A relatively diffuse (weakly informative) prior was employed to estimate the risk preferences of the representative agent, and the posterior distribution from that estimation was used as the informative prior for estimation of individual risk preferences.

[24] If one was using point estimates from a traditional maximum likelihood approach, or even point estimates from one of the descriptive statistics of a posterior distribution (e.g., mean, median or mode), then the inferred welfare measure would be a scalar.

[25] Harrison, Morsink and Schneider [2020] provide a number of examples of the evaluation of *non-adaptive* treatments.

gain from virtually every opportunity to purchase insurance, and sadly some lose with equal persistence over the 24 sequential choices. Armed with posterior predictive estimates of the welfare gain or loss distribution for each subject and each choice, can we adaptively identify *when* to withdraw the insurance product from these persistent losers, and thereby avoid them incurring such large welfare losses? Important recent research by Caria et al. [2020], Hadad et al. [2021] and Kasy and Sautmann [2021] considers this general issue. The challenges are significant, from the effects on inference about confidence intervals, to the implications for optimal sampling intensity, to the weight to be given to multiple treatment arms, and so on.

Assume that the experimenter could have decided to stop offering the insurance product to an individual at the mid-point of their series of 24 choices, so the sole treatment arm was to discontinue the product offering or continue to offer it.[26] The order of insurance products, differentiated by their actuarial parameters, was randomly assigned to each subject when presented to them. Figure 2 displays the sequence of welfare evaluations possible for subject #1, the same subject evaluated in Figure 1. The two solid lines of Figure 2 show measures of the CS: in one case the average gain or loss from the observed choice in that period, and in the other case the cumulative gain or loss over time. Here the average refers to the posterior predictive distribution for this subject and each choice. Since this is a distribution, we can evaluate the Bayesian probability that *each* choice resulted in a gain or no loss, reflecting a qualitative Do No Harm (DNH) metric enshrined in the *Belmont Report* as applied to behavioral research.[27] This probability is presented in Figure 1, in cumulative form, by the dashed line and references the right-hand vertical axis.

Although there are some gains and losses in average CS along the way, and the posterior predictive

[26] A more sophisticated "targeting" policy might use the information from the first 12 insurance choices to adaptively determine the actuarial parameters that might lead each subject to make better decisions in the remaining 12 choices.

[27] See Teele [2014] and Glennerster [2017] for discussion of the *Belmont Report* and some aspects of the ethics of conducting randomized behavioral interventions in economics. Even when randomized clinical trials were not adaptive, or even sequential in terms of stopping rules, it was common to employ termination rules based on extreme, cumulative results (e.g., the "3 standard deviations" rule noted by Peto [1985; p. 33]).

probability of a CS gain declines more or less steadily towards 0.5 over time, the DNH probability is always greater than 0.5 for this subject. And there is a steady, cumulative gain in expected CS over time. These outcomes reflect a common pattern in these data, with small CS losses often being more than offset by larger CS gains. Hence one can, and should, view these as a temporal series of "policy lotteries" which are being offered to the subject, if the policy of offering the insurance contract is in place (Harrison [2011b]). In this spirit, we can think of the probabilities underlying the posterior predictive DNH probability as the probabilities of positive or negative CS outcomes, given the risk preferences of the subject. The fact that the Expected Value (EV) of this series of lotteries is positive, even as the probability approaches 0.5, reflects the asymmetry of CS gains and losses in quantitative terms and the policy importance of such quantification. For now, we might think of the *policy maker* as exhibiting risk neutral preferences over policy lotteries, but recognizing that the evaluation of the purchase lottery by the subject should properly reflect her risk preferences.

Consider comparable evaluations for four individuals from our sample in Figure 3. Subject #5 is a "clear loser," despite the occasional choice that generates an average welfare gain. It is exactly this type of subject one would expect to be better off if not offered the insurance product after period 12 (or, for that matter and with hindsight, at all). Subject #111 is a more challenging case. By period 12 the qualitative DNH metric is around 0.5, and barely gets far above it for the remaining periods. And yet the EV of the policy lottery is positive, as shown by the steadily increasing cumulative CS. This example sharply demonstrates the "policy lottery" point referred to for subject #1 in Figure 2.

The remaining subjects in Figure 3 illustrate different points: that we should also consider the time and intertemporal risk preferences of the agent when evaluating the policy lottery of not offering the insurance product after period 12. Assume that these periods reflect non-trivial time periods, such as a month, a harvesting season, or even a year. In that case the temporal pattern for subject #67 encourages us to worry about how patient subject #67 is: the cumulative CS is positive by the end of period 24, but if later periods are discounted sufficiently, the subjective present value of being offered the insurance product could be negative due to the early CS losses.[28] Similarly, consider the volatility *over time* of the CS gains and losses faced by subject #14, even if the cumulative CS is positive throughout. In this case a complete evaluation of the policy lottery for this subject should take into account the *intertemporal* risk aversion of the subject, which arises if the subject behaves consistently with a non-additive intertemporal utility function over the 24 periods.

Applying the policy of withdrawing the insurance product after period 12 for those individuals with a cumulative CS that is negative results in an aggregate welfare gain of 108%, implicitly assuming a classical utilitarian social welfare function over all 111 subjects.

One general lesson from this example is that we now have the descriptive and normative tools to be able to make adaptive welfare evaluations about treatments during the course of administering the treatment. How one does that optimally is challenging, but largely because we have not paid it much direct attention in economics. Optimality here entails many tradeoffs, and not just those reflecting the preferences of the instant subject.[29] Our focus here is on the partial equilibrium impact on the *welfare* of each and every individual.[30] This is often confused by economists as trying to evaluate social welfare, a different concept altogether, although ideally concepts that are related to each other in subtle ways. Hence, when we report an average of individual welfare effects descriptively, that is not to impose a utilitarian social welfare function, but just to describe our calculations in a familiar manner. The role of formal general equilibrium welfare evaluations is to account for some of the interactions between agents, and second-best constraints, that affect the evaluation of policy. Just as the numerical models evaluating

[28] This point has nothing to do with whether the subject exhibits "present bias" in any form. All that is needed is simple impatience, even with Exponential discounting. Berry and Fristedt [1985; chapter 3] stress the importance of time discounting in sequential "bandit" problems in medical settings.

[29] And this lesson is on top of the lessons from a well-known medical case study, reviewed by Harrison [2021], about the normative basis of experimentation at all based on prior evidence from *non-experimental* environments.

[30] We stress the welfare impact. Many economists confuse impact evaluation with welfare evaluation, arguing that surely the observable impact being measured must matter for welfare. Even when statistical circumstances are ideal, impact evaluation constitutes at best an intermediate input into the welfare evaluation of interventions. That intermediate input is valuable, but should not be confused with the final product, a proper cost-benefit analysis (Harrison [2014]).

general equilibrium welfare effects have been extended over the years to include imperfect competition, scale economies, trade barriers that are not *ad valorem* tariffs, and so on, eventually they could be extended to incorporate richer models of behavior in stochastic policy settings.[31] That is not our immediate focus.

The other general lesson from this case study is the difficulty of making decisions during the instant experiment when the inferences from the experiment have some, presumed welfare implications for *individuals outside the instant experiment*.[32] If we had truncated these experiments adaptively as suggested, would we have been able to draw reliable statistical inferences about the treatment in a way that would influence future applications of the treatment? The only way to evaluate these issues, particularly with multiple treatment arms, is to undertake them in safe laboratory settings in which subjects literally have nothing to lose, and study the implications of "throwing data away" in accordance with such adaptive rules. Then be Bayesian about deciding how much to learn from that for the field.

### *C. Issues in the Application of the Quantitative Intentional Stance*

#### Is Rationality Being Assumed?

Dennett has often[33] characterised the basic posture of the intentional stance as assuming that an agent is rational, assessing her circumstances, and then posing the question, "What should we expect her

[31] See Harrison, Rutherford and Tarr [1997] for discussion of the role of these modeling extensions in the context of a welfare evaluation of the Uruguay Round of multilateral trade reform, and Harrison [2011b] for a review of "policy lotteries" that have been evaluated with computable general equilibrium models. A key feature of those evaluations for practical welfare economics is the ability to explicitly calculate sidepayments required by the Kaldor-Hicks-Scitovsky Compensation Criteria, allowing for the distortionary effects of sidepayments that are not "lump sum." For applications in climate policy, tax policy, and trade policy, respectively, see Harrison and Rutherford [1999], Harrison, Jensen, Lau and Rutherford [2002] and Harrison, Rutherford and Tarr [2003].

[32] This tradeoff has long been felt keenly in the literature on sequential clinical trials in medicine: see Armitage [1985].

[33] For example, in Dennett [1987; p.17], "Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do."

to do in light of her rationality and her circumstances?" Glossed at such a high level of generality, this looks identical to 20th-century "neoclassical" economic methodology as most influentially described by Lionel Robbins [1935]. Is this not precisely the methodology against which behavioral economics is generally taken to be a corrective strategy?

Part of the answer to this implied objection, on which we have been focusing so far, is that the QIS counsels measuring the agent's "circumstances" very carefully and exactly, based on empirical evidence and particularly on experimental manipulation. But every behavioral economist agrees, in contrast to Robbins, that the rationality of any actual agent is necessarily "bounded" in the sense of Simon [1955]. We must assume that the target is rational in the *broad* sense required for agency: that there is a systematic relationship between her behavioral patterns over the time frame about which we seek generalizations, and the information that empirical evidence suggests she registers and tracks. We should *not* assume – and Dennett has never suggested that any inquirer using the IS should assume – that her expectations are "rational" in the ideal sense of reflecting all information that is *in principle* available, or that she makes no errors of inference. These bounds on a given agent's rationality must also be identified empirically.

Crucially, however, the QIS emphasises that we should not attempt such identification by studying the agent individually and in isolation. Her preferences and beliefs will be those that people regarded as approximate equivalence classes by her social reference network would attribute to her and would expect her to attribute to herself, conditional on specifics of the decision setting. According to the IS, *there is no "deeper" fact of the matter about what she prefers and believes.* This strongly licenses the method of identifying response functions – or, more specifically, utility functions and probability weighting functions – by estimating models that include covariates sufficient for specifying the expected *social* model of the target agent. We gain powerful evidence about the information her "society" expects her to attend to when she makes choices by studying the information scaffolds that the society in question affords her. Societies are full of structure, hierarchies, and biases. They expect people with different demographic profiles to respond differently from one another, and these different profiles will consequently tend to rely on different scaffolding packages for many problems of interest to economists. For example, we might hypothesize that it is easier for women than for men, statistically, to get pediatric information and to assess its value. Economists, like other social scientists, tend to be *most* interested in launching studies precisely where they find these variations.

In everyday choices, relevant scaffolds will typically be quite general. For example, an adult in a modern society is expected to be able to understand, and pay at least some attention to, the daily news in the mass media, public notices and signs, and conventions for estimating and applying everyday magnitudes. In a workplace or hobby setting, expectations and the scaffolding that sustains them will be more extensive, and more tightly standardized, within relatively narrow domains. The economist's model of participants in a finance experiment will differ depending on whether her subject pool is drawn from the general population or from the banking and accounting professions. She might be described as modelling the latter as "more rational," within the context of the experiment, than the former. The relevant sense of "rationality" here is what Vernon Smith [2008], drawing on Hayek, calls "ecological rationality." This usage involves substantial transformation of the concept of rationality. The economist does not treat her banker subjects as less likely to make *logical* errors, or less likely to be emotionally biased or cognitively lazy, than her subjects from the general population. She just expects them to be better informed, and to be more likely to accurately apply the information they have. Thus the "rationality" in the concept of ecological rationality is not what the philosophical decision theorist means by that word.

For deep reasons we discuss in §4, we are doubtful about the value to economists of any concept of "rationality" that is intended to be fully general. We regard the phrase "rational agent" as including a redundant adjective. Economists study choices under incentives, so *all* of their subjects (including, sometimes, non-human ones) are agents. The IS, and following it the QIS, apply wherever agency does. Beyond that common threshold, different capacities of different agents to solve problems are mainly a function of the different scaffolding kits they are expected to use. People, but not rats or elephants, can read text; bankers, but not most other people, can compute compound interest rates without special support from the experimenter.

To summarize the answer to the question, then: the economist deploying the QIS *assumes* agency, and *expects* variation in ecological rationality that she must do empirical work to identify and estimate. She makes no special assumptions about "general" rationality, an idea she can and should avoid altogether.

### Are We Assuming Some True Risk Preference?

No, we are instead assuming some prior belief is being formed about the risk preferences of the agent whose behavior is being evaluated. Thinking of these as priors rather than some "assumed truth" has important implications, quite apart from being consistent with the QIS, and also opens the way to developing ways to better inform the choice of priors for behavioral welfare economics.

The value of viewing these QIS attributions as priors, and employing a Bayesian approach derives from the methodological need for normative analysis of risky choices to have estimates of risk preferences from choice tasks *other than the choice task one is making welfare evaluations about.*[34] In settings of this kind, it is natural to want to debate and discuss the appropriateness of the risk preferences being used. In fact, the need for debate and conversation becomes more urgent when, as here, we infer significant losses in expected CS, and significant foregone efficiency. How do we know that the task we used to infer risk preferences, or even the models of risk preference we used, are the right ones? The obvious answer: we don't. We can only hold prior beliefs about those, and related questions. And when it comes to systematically examining the role of alternative priors on posterior-based inference, one wants to be using Bayesian formalisms.

Saying that we view these as priors is not an invitation to then claim that the welfare evaluation is arbitrary. It is recognizing what economists of a wide range of methodological persuasions have been

[34] To be strict, we should say "other than directly, naively inferred from the choice task one is making welfare evaluations about." We discuss the notion of structural models of mistakes in §2.E.

doing for many decades and just formalizing it. The analogy to the nudge literature is apt. Proponents of nudges correctly stress that when we adopt some choice architecture for decision-makers, and have priors over the effect of that architecture on their behavior, we have simply replaced one existing choice architecture with another. That is, some choice architecture is required, and will be used anyway, so why just assume that historical accident has generated a normatively attractive architecture? Another analogy comes from the the classic *Specification Searches* of Leamer [1978]: many of the *ad hoc* methods used by econometricians are clumsy attempts to use priors, so why not recognize that and do it explicitly and elegantly with Bayesian methods?

There are immediate reasons why one would want to use Bayesian estimates of risk preferences for the type of normative exercise illustrated in our extended example in §3.B. One obtains more systematic control of the use of priors over plausible risk preferences, and the ability to make inferences for every individual in a sample.

However, there are also more general reasons for wanting to adopt a Bayesian approach, to make explicit the role for priors when making normative evaluations. A related, general reason for a Bayesian approach derives from the *ethical* need to pool data from randomized evaluations and non-randomized evaluations, discussed by Harrison [2021]. The motivation for randomized control trials in many areas, such as surgical procedures, derives from non-randomized evidence accumulated in widely varying circumstances, such as the health and co-morbidities of the patient. These data are evidently not inferred from "clean beakers," but they are often *completely discarded* when designing a randomized test of the procedure. This practice reflects the notion of "clinical equipoise," which holds that one should initiate and apply the randomized procedure as if none of the prior non-randomized evidence had existed at all. The counter-argument is just to view those prior data as justifying what is actually observed: someone thinking *a priori* that some new procedure is worth testing. That is not, by construction, a completely diffuse prior at work, so one should formally reflect that fact. The ethical issue takes on urgency when patients or parents of patients are being asked to submit to 50:50 chances of a procedure that these priors suggest is inferior. Of course, such equipoise might be justified by a social objective of arriving at a general conclusion more quickly, for the benefit of all potential patients, despite the expected cost to the instant patient; we would disagree with the implied tradeoff, but we see the logic.

### Are We Assuming Stable Risk Preference?

To conduct normative evaluation of insurance decisions in our extended example we needed to make the explicit and necessary assumption that there is a set of risk preferences of an individual that we can identify in a risky lottery task, and that we can apply as priors in an insurance task, so as to infer expected welfare changes from insurance choices. If risk preferences are not stable *over time*, is there a risk of normative evaluation being based on "stale" preferences? If risk preferences elicited in one domain are not stable *across domains*, how do we know that they are appropriate for another domain?

Even though these are relevant concerns, we argue that they are second order, simply because there are no other assumptions that one can make *if the objective is normative evaluation*. Now that we have demonstrated the QIS method based on that assumption, however, it is entirely appropriate to engage in debate over the strength or weakness of our prior and potential alternative priors for risk preferences that might be used. This is where the ongoing discussion of these, and related, descriptive characterisations of risk preferences have a legitimate role: helping us navigate among the various priors we might use. In our first attempt at applying the QIS method in the laboratory the risky lottery choice task and the insurance decisions are made contemporaneously, implying that there is no serious issue of temporal stability that arises in this instance. And the financial-outcome frame of the risky lottery choice task is close to the financial-outcome frame of the insurance purchase task, so we also don't anticipate a serious issue of domain-specificity in this instance. But what can we say in general?

Consider the issues raised by any instability of risk preferences over time. Temporal stability of preferences can mean three things, and can be defined at the aggregate level for pooled samples or for individuals.[35] Our concern is with individuals, and this is arguably a more demanding requirement.[36] One interpretation of temporal preference stability is that risk preferences are *unconditionally stable* over time. This means that the risk preference parameter estimates we obtain for a given individual should predict the risk preference parameters she would use in the future when she makes the decision that we are normatively evaluating, no matter what else happens in her life. This is the strongest version of a "temporal stability of preferences" assumption, and will presumably be rejected for longer and longer gaps between elicitation of the risk preferences and normative evaluation of the decision.[37]

A second interpretation of temporal preference stability is that risk preferences are *conditionally stable*. This interpretation assumes that risk preferences might be state-dependent and a stable function of states over time, but there could be changes in the relevant states over time. This interpretation implies that the risk preference parameter estimates for a subject might depend on her age, for example, and that particular "state" changes in thankfully predictable ways. Of course, this predictability presumes that we have a decent statistical estimate of the effect of age on risk preferences, but it is plausible that this could be obtained. If the states are readily observable, such as age, conditional stability is perhaps a reasonable prior to have for normative evaluation.

A third interpretation is that risk preferences might be state-dependent and the states are not observable, or that the risk preferences are themselves stochastic. In this instance there are stochastic specifications, which in turn embody hyper-priors, that let us say something about stability (e.g., that the unobserved states are fixed for the individual, or that the stochastic variation in preference realisations

[35] The relevant characteristic of stability can also vary with the inferences being made. For some inferences we only care about the ranking of individuals in terms of risk premia, and for some inferences we care about the level of the risk premia for individuals. We assume the latter for our purposes here.

[36] There are very few data collected on any forms of stability at the individual level. Most of the evidence concerns averages or distributions over individuals.

[37] Chuang and Schechter [2015] review the literature and suggest low correlations of risk preferences over time. Harrison et al. [2005] find evidence of unconditional stability over 5 or 6 months for average levels of risk aversion. Andersen, Harrison, Lau and Rutström [2008b, §5.1] similarly find evidence of unconditional stability over 17 months for distributions of risk attitudes.

follows some fixed, parametric distribution).[38]

### What About Source-Dependent Risk Preference?

Domain specificity of risk preferences involves systematically examining the role of alternative priors over the risk preferences from various possible domains on posterior-based inference about the specific domain of the financial decisions we aim to normatively evaluate. This again implies we should use Bayesian techniques.

Imagine one was designing a field experiment, say in rural Ethiopia, in which various interventions for a health insurance product were to be used to improve the welfare of households. Assume that this health insurance product focused on acute conditions, with significant mortality risk. The only priors on risk preferences you have come from university students who participated in laboratory experiments in the United States. Should you go ahead and design interventions that, conditional on the risk preferences of the university students, lead to expected welfare losses for the same students, of the kind we have demonstrated? We suggest that, ethically speaking, you should not.

Now imagine you have been able to conduct comparable incentivised artefactual field experiments with risky lottery choices in Ethiopia with a sample from the target population of rural Ethiopian households that allow you to infer risk preferences over financial outcomes. These are obviously better priors for the risk preferences needed to undertake the eventual normative inference about the health insurance decisions, and should be used. In this case we would completely discard the priors from students in the United States. Next, imagine that you have been able to conduct artefactual field experiments over some risky health outcomes in Ethiopia that allow you to infer risk preferences. Assume

[38] Allowing for unobserved heterogeneity, Harrison, Lau and Yoo [2020] find evidence for temporal stability of distributions of risk preferences over 6 to 12 months, but only when correcting formally for sample selection and attrition. And they infer temporal instability when those corrections are not made. No prior study has corrected for selection or attrition when drawing inferences about temporal stability.

that these health outcomes refer to morbidity risks, not mortality risks, but to real outcomes nonetheless.[39] Clearly the domain of risk preferences here is closer than the risk preferences defined over money because they were elicited in the context of health choices. However, because we know that eliciting risk preferences over health risks is not as reliable, would you now attach *zero weight* to the risk preferences over money by similar Ethiopians? Probably not.

What this discussion shows is that it is simplistic to attempt to make a general statement about the validity of preferences for a reference risk preference elicitation task and a different task that is the target of normative evaluation, since it depends entirely on the timing, context, and domain of the two tasks. However, the statistical methods that we have at our disposal (Bayesian analysis) merely demand that we are able to define a diffuse or an informed prior in one task or domain, to do normative evaluation of choices in another domain, without having to require that the elicited risk preferences are "perfectly valid" for the choice task that we want to make normative statements about.

### But Can't I Just See What Works?

This question comes up a lot, often also in the guise of the question, "does it have to be this hard?" in the sense of requiring so much theory and structural modeling of preferences, beliefs and constraints. The response is to question what can be inferred about welfare directly from observed behavior without some implicit or explicit theory of motivation. Then the structural modeling that some theory requires just has to be done, recognizing that some theories require more or less structure.

A recent debate between Sunstein [2021] and Sugden [2021] illustrates this divide, where the context is commentary by Sunstein [2021] on a critique of the nudge movement in Sugden [2018]:

> Sunstein's response to this critique is one that I have heard from many behavioural economists. It is, as he says, 'brisk' [...]. In rough paraphrase: 'We all know that the criteria we use are conceptually problematic, but the cases we are dealing with are too important

[39] As any experimental economist knows, it is not easy to come up with morbidity outcomes that can be credibly and ethically delivered within the budgets we normally find ourselves working within.

> for us to be held back by abstract problems. It's so obvious that people are making errors that we don't need a definition of error. The effects on people's welfare are so clear-cut that we don't need a definition of welfare. Wake up and smell the coffee.' I have to say that I find this response frustrating. My critique addresses what behavioural welfare economists have actually written about the concepts of preference, welfare, error and bias when explaining how they reach their policy recommendations. Was that not supposed to be taken seriously? (Sugden [2021; p. 420])

We hasten to add that we share some of Sunstein's briskness when it comes to many of the commentaries on behavioral welfare economists by philosophers. They often demand analytically complete theories which, even if we could all eventually be brought to agree on, would be extravagantly demanding of data. But the importance of the "inner rational agent" critique of Sugden [2018] and Infante, Lecouteux and Sugden [2016] cannot be doubted or waved aside because it is inconvenient.

### How Much Structure is Needed?

It is popular to come up with "reduced form" surveys or experiments to elicit risk preferences, and to many researchers it seems like a lot of work to elicit choices over lotteries, and then also to have to engage in structural estimation, before even getting to the task of real interest. A short response might be, "that's life." A longer response recognizes that these are what are often called "nuisance parameters," in the sense that they are not the primary focus of interest, and that there is an opportunity cost of every extra task added to a design. One of the advantages of Bayesian Hierarchical Models (BHM) is that one can make informed inferences at the level of the individual without every lottery choice in a large battery being asked. In our own work it is now common to construct a battery of 80 to 100 lottery pairs, and then randomly select 20 or 30 for each subject to complete, following Gao, Harrison and Tchernis [2022; §3.A]. These efficiencies are particularly important in field applications.

Three qualifications are important. First, one cannot just have some measure of risk preferences that might be (linerarly) correlated with the minimal structure needed to evaluate the expected CS. One needs to know the relevant parameters of some utility function, at a bare minimum, and of course any stochastic error around any point estimates of parameters. Responses to Likert scales that positively correlate with the propensity to bungy-jump need not apply.

Second, there may be settings in which one wants to ensure that the estimated risk preferences allow inferences about specific axioms. A core instance in our work is the ROCL axiom. This axiom is assumed in EUT, RDU and CPT, but is often a central focus of skeptical attention. One example is when there is a compound risk of non-performance of an insurance contract (e.g., Harrison and Ng [2018]), or there is basis risk in index insurance contracts (e.g., Harrison, Morsink and Schneider [2022]). Another general example is when one wants to consider how uncertainty aversion, as distinct from risk aversion, affects inferences whenever there are subjective belief distributions (e.g., Harrison [2011b; §4]). In these cases, we always want to ensure that each subject faces a certain number of paired binary choices that allow direct inference from choice patterns to the extent of ROCL violation.[40]

The third qualification has to do with the difference between *atemporal* risk aversion and *intertemporal* risk aversion. The former refers to risk over outcomes that are to be received by a subject at a point in time, often immediately in controlled experiments.[41] The latter refers to risk over outcome streams that are time-dated. In general there is no theoretical need for one type of risk aversion to imply anything about the other: see Andersen, Harrison, Lau and Rutström [2018] for theoretical results and experimental evidence. There are important field settings in which both are needed to undertake normative evaluations (e.g., annuities and other retirement options). So the general point is that "enough structure" is needed to properly undertake the instant normative evaluation. We appreciate that this entails even more tasks than just one to elicit atemporal risk preferences.

A related point is that one might often need priors over time preferences as well as risk

[40] The idea is to have one binary choice that presents a lottery with a compound risk (C) and some lottery (S) with simple risks. Then we need another binary choice that presents the actuarially-equivalent simple lottery (C′) that corresponds to C and the same lottery S. If the subject selects C (S) in the first choice and C′ (S) in the second choice then there is no violation of ROCL; if not, then there is a ROCL violation. Harrison, Martínez-Correa and Swarthout [2015] develop a battery to test ROCL in this manner. No estimates are needed to determine the fraction of choices in such "pairs of pairs" that reflect ROCL violations.

[41] We can ignore for the moment the fact that many experiments ask for choices over a number of pairs of lotteries and then select one for resolution and payment. The time taken in these choices is minimal, and what is fundamental is that the outcome be received by the subject at one point in time.

preferences, and indeed priors over subjective beliefs as well as preferences.

Our experience is that *all* attempts to find short-cuts to eliciting these preferences and beliefs end in inferential misery. As Mencken once said, "there is always a well-known solution to every human problem – neat, plausible, and wrong." Given the intellectually parlous state of behavioral welfare economics, and its potential value, now is not the time to look for short-cuts.

### Why Just EUT and RDU?

In general we agree that other models of risk preferences should be considered. There seems to be enough support for RDU as the most empirically important alternative to EUT, once the lack of controlled laboratory evidence for CPT is accepted. We view the various "regret theory" models of Bell [1982], Fishburn [1982] and Loomes and Sugden [1982][1987], as well as the "disappointment aversion" models of Bell [1985], Gul [1991] and Routledge and Zin [2010], as particularly worthy of attention.[42]

For now, we take the agnostic view that the risk preferences we have modeled as best characterizing the individual are those that should be used, in the spirit of the "welfarism" axiom of welfare economics. Even though the alternatives to EUT were originally developed to relax axioms of EUT that many consider to be normatively attractive, it does not follow that one is unable to write down axioms that make those alternatives attractive normatively. For instance, consider inverse-S probability weighting in an RDU setting, which leads the decision-maker to place greater weight on the probabilities associated with the best and worst outcomes. This might be a reasoned heuristic for recognizing that "tail probabilities" are known to be inferred less reliably, and are more reliant on parametric forms for probability distributions being correct. In fact, it characterizes one approach to "actuarial prudence" in the calculus of risk management. In terms of decision theory, it may be viewed as one way to extend the reasoning from the "small worlds" of Savage to his "large worlds" (Binmore [2009]).

[42] See Starmer [2000; p.355ff., p.344ff.] for an excellent exposition of the historical context of these models.

### Why Ignore Loss Aversion?

We do not ignore loss aversion. Instead, as noted in §1.B, we just do not find enough evidence in controlled experiments to justify using CPT models that define it as such. In many of our experiments, not all, we have been careful to use a risk lottery battery that allows one to estimate a CPT model if desired, but more recently do not see the point having reviewed the data. Given that, and recognizing that "zombie parameters" walk among us, we do nonetheless have a position on how loss aversion should be viewed from the perspective of the QIS, and review that below in §4.B.

### But There Is No Single, Correct Way to Elicit Risk Preferences

It is true that there are many alternative ways to elcit risk preferences, even if we restrict attention to those that involve incentivized choices. Although we are content, as practicing applied economists, with the methods we use (binary choice over a randomly-ordered battery of lotteries), we see no need for anyone to try to define a single "correct" way to elicit risk preferences, and do not even know how to define a sensible metric to use to undertake that race and seek to determine a winner. More strenuously, we do not see the existence of different risk elicitation methods as a reason to pause using one that meets certain criteria, just because there are others under consideration. These are, after all, just priors. The fact that there are several elicitation methods for these priors does not make the priors *arbitrary*. Rather, it makes them *conditional* on the elicitation methods, and that is all.

### What About Subjective Probabilities?

We have focused attention on "small world" settings with objective probabilities, since there are enough conceptual issues to sort out in that setting. But we completely agree that the very next step must be to extend the approach to when we must also make inferences about subjective probabilities. The reason is that poorly-calibrated subjective probabilities, or failures to update subjective beliefs consistently with Bayes' Rule, are *a priori* likely to be an important source of welfare losses in many policy-relevant settings.

We believe that the next step will be to utilize additional tasks to directly elicit subjective relative likelihoods over events, which can then be used to make inferences about RDU models that separately identify the subjective belief that is then subject to probability weighting. The theoretical framework to do this was developed by Machina and Schmeidler [1992][1995], assuming one could do so without making any assumptions about risk preferences or probability weighting.[43] One would also need to have a task to identify the utility function and probability weighting function, using objective probabilities, as they also propose, but that is by now standard fare for experimental economists.

## What About Social Welfare?

Our approach has been to generate priors about the preferences and, if necessary, subjective beliefs of the individual[44] in order to evaluate two or more risky prospects that the individual has to make a choice over. The evaluation might reflect an EUT or RDU model of risk preferences. Each evaluation results in a comparison of CE, which is the expected CS of choosing one prospect rather than other prospects. The descriptive models of EUT or RDU predict that the individual will choose the prospect with the highest CS. Normative models just tabulate the CS of the observed choice, which may or may not be the prospect with the highest CS. All of this is undertaken for one individual.

Over a sample of individuals we can then build up a distribution of welfare effects of observed choices. Some elements of this distribution may be positive, some might be zero, and some might be negative. If an individual makes a number of choices, it is therefore possible for an individual to accrue some positive welfare effects, some zero welfare effects, and some positive welfare effects of those

[43] This extra step is achieved directly by eliciting subjective beliefs using a binary lottery procedure that, in theory, induces risk neutrality under EUT or RDU. See Allen [1987], McKelvey and Page [1990], Hossain and Okui [2013], Harrison, Martínez-Correa and Swarthout [2014] and Harrison, Martínez-Correa, Swarthout and Ulm [2015].

[44] As in applied economics generally, an "individual" is a single *agent*, not necessarily a single *person*. When a single utility function is assigned to a firm or a labor union or a lobby group, etc., this amounts to modeling the corporate agent as an individual.

choices. A central component of this evaluation of the distribution of welfare effects is to start with priors over the risk preferences and, if necessary, subjective beliefs of each individual. This can be called a "bottoms up" approach to behavioral welfare evaluation.

The final step in this approach is to describe the distribution. One might look at the average, one might look at quartiles, and one might look at measures of variability such as standard deviaton, skewness or even kurtosis. These are just descriptions, and should not be confused with social welfare evaluations. To be sure, some social welfare evaluations might be the same, numerically: for example, taking the unweighted average also reflects a classical utilitarian social welfare function (SWF).

An alternative approach is to start with a SWF defined over *final, non-risky* outcomes, and then allow for risk to be associated with those outcomes. The arguments of the initial SWF in economics are typically utility-based evaluations of final outcomes by individuals. Adler [2012][2019] provides a careful exposition of this approach, which we do not take up here. We have many problems with the treatment of risk being "added on" as a last step, but they take us too far astray for present purposes.

## 4. Distinguishing Welfare from Well-Being

### *A. Well-Being and Reflective Rationality*

An unusual feature of economics, among the academic disciplines, is that it is subject to a swollen, and endlessly replenished, hostile literature that attacks the whole enterprise. This is not a recent phenomenon; it dates from the historical origins of modern economics (Coleman [2002]), and is closely entangled with, though not identical to, hostility to markets (Ross [2012]). Generalised anti-economics often gives pride of place to a range of epistemological arguments, intended to cast doubt on economists' claims to knowledge and to status as scientists. This literature consequently receives persistent attention from philosophers of economics. Notwithstanding these abstract debates, which reach far beyond our scope here, it is obvious that their volume and frequency is fuelled by resentment and anxiety about the role of economists and economics in designing and promoting public and corporate policies. Recent and explicit examples are Applebaum [2019] and Berman [2022].

We set aside debates about whether the basic logic of mainstream normative economics reflects ideology, which one of us has reviewed directly elsewhere (Ross [2012]). A less polarized, but equally dense, body of contestation is about whether normative economics claims an overly expansive domain of influence, implicitly or explicitly promoting a narrow and particular set of values that drive out those championed by non-economists, and professionally explored by moral philosophers, sociologists, and political scientists from outside the "rational-choice" tradition in that discipline. Marglin [2008] provides a sympathetic review of this perspective by an economist. We comment on these debates here because many behavioral welfare economists argue that their special alertness to psychology offers a path for correction of the alleged normative imperialism of traditional welfare economics.

Like much of the economist's special lexicon, the word "welfare" has both everyday and technical meanings that diverge. In professional economics, welfare is an efficiency measure, referring to quantitative comparisons between ratios of marginal gains in preference satisfaction from outputs and marginal sacrifices of preference satisfaction associated with opportunity costs of inputs. Such welfare can be individual or social, and there is of course a vast literature on different approaches to welfare aggregation. Work that compares these methods *can* be purely technical, but very often is not, because alternative ways of defining social welfare are often motivated normatively by reference to implications for distribution across individuals or groups. This creates constant tension between welfare as a technical efficiency concept and an everyday use of "welfare" that treats it as equivalent to "general well-being." Economists frequently fuel this tension by using rhetoric that implies that if there is consensus among economists that policy $X$ delivers more marginal welfare per input unit of cost than alternative policy $Y$, then $X$ should "automatically" be chosen by whichever agents control the choice in question. This suggests that economics should govern policy in general, and triggers backlash and resistance.

We do not begin to have space here to review arguments about the entangled complex of mathematical, philosophical, political, and psychological elements that critics combine and re-combine into comprehensive normative views about the authority and social role of welfare economists. We instead simply summarize conclusions of Ross and Townshend [2021] and Ross [2023], that review arguments we consider most pertinent.[45] We number the claims for subsequent convenience of reference.

1. Comparative states of general human well-being are too multi-faceted, cross-culturally variable, and resistant to full description without use of normatively inflected narratives, to be reduced to or decidable by any quantitative efficiency measure. So welfare in its technical sense is not equivalent to general well-being. The most sophisticated literature on general well-being is produced by some philosophers, such as Tiberius [2010], Williams [1981][2006] and Nussbaum [1997]. This kind of work does not reach scientifically testable conclusions, and does not aspire to do so. Economists have no special expertise, as economists, in assessing the relative merits or soundness of this literature as a whole, or of specific instances of it.
2. Economists best promote clarity of their own work by restricting use of "welfare" to technical interpretations that admit of specifications that can be identified in realistically obtainable empirical data.
3. Because economists are not experts in general human well-being, they should manifest the attitude that Keynes [1936; p. 373] recommended when he famously urged them to adopt the posture of "humble, competent people, on a level with dentists." The meaning of Keynes here is not vague: what he specifically advised economists against was advocating sweeping reforms of individual or social values.[46]
4. In a broadly liberal society, public administrators should be seen as agents who aim to promote the welfare of principals, that is, the body of citizens, whom they are hired to serve. They should act to

[45] In doing this, we do not mean to suggest that these statements are the most important ones. They simply state grounds for our own opinions at greater length and engagement with very large literatures than we can provide here. We also cite some main sources of the most important premises on which the arguments depend.

[46] Some might suppose that when Duflo [2017] compares ideal development economists to "plumbers" she expresses a 21st-century echo of Keynes. This is not so, because she divides the entire labor of ideal policy assessment among different sub-sets of economists. Duflo's program implies that each individual economist might be humble, but economists as a group should rule the policy consulting ecosystem.

promote welfare, not general well-being, because the latter activity would unavoidably impose their special philosophical opinions about what is best for people in general on citizens with conflicting such opinions (Dowding and Taylor [2019]).

5. Economists who are humble experts on welfare are appropriate consultants for public administrators in broadly liberal political and social orders. When an economist's advice is not followed by a government, her response that efficiency is being traded off for something else may be true, but does not justify, in principle, a claim that injustice has been allowed.

These five claims complement a pair of closely linked corollary opinions on the general philosophy of economics.

First, we think that economists should be much more cautious than most are about in rendering judgments about "rationality." One conclusion of Robbins [1935] that has stood the test of time is that what economists *properly* mean by "rationality" in their professional work is consistency of agents' ends with means, and among ends, for some assignment of agency roles and across some time interval. This is a far narrower idea than what anyone other than economists, and certainly the philosophers cited above, mean by "rationality." Just as economists have no special expertise as assessors of well-being in general, they have no special expertise as assessors of rationality "in general." Most economists, we suggest, have always acknowledged this in treating alternative distributional allocations consistent with Kaldor-Hicks-Scitovsky improvements as matters of "ethics," to be modelled by a social choice theory that they have never claimed to monopolise, but share with philosophers and political theorists.

But, second, we also have a sceptical attitude to that literature. Consider one of its best products, the magisterial and technically careful critical review by Adler [2012]. The main literatures on which he relies are technical welfare economics and philosophical decision theory. The goal of the enterprise is a manual for governing a community of members, each of whom is held to, and expects others to be held to, the decision-theorist's standard of general rationality. One representative implication is Adler's conclusion that although most people are, *as a matter of empirical fact*, risk averse, ideally rational agents are risk neutral; therefore the set of best social welfare functions that should be operated by public administrators should be applied *as if* citizens aimed to maximize their expected value. Another implication is that citizen's preferences over "remote" outcomes outside their control space, for example concerning the collective well-being of distant future generations, should be ignored (Adler [2012; p.174-181]). By contrast, we think that economists assessing welfare should be sensitive to all actual preferences they find, including preferences about risk and about "remote" matters. Adler's disagreement with us here can, we suggest, be diagnosed as follows. He assumes that a liberal society is the best kind of society for promoting general well being, and that general well being can be assessed within a framework that assumes general rationality. Thus "illiberal" preferences about people unconnected with the preference-holder, and "irrational" preferences such as risk aversion, should be ruled out of the domain of a social welfare function.

It is a biographical fact about us, the present authors, that we are liberals. As political participants we encourage others to be liberals, or at least to support the liberal policies we prefer. But we do not see this as any of our business *as (humble) economists*. Furthermore, we claim enough expertise in philosophy of economics to hold as a professionally informed opinion that general well-being and general rationality cannot be reduced to a technical decision theory that makes preference-consistency the decisive element of methodology for assessing value. When, on evenings and weekends, we want to contemplate the best forms of private and political life, we find more interesting guidance from writers such as Tiberius, Williams, and Nussbaum than from analysts like Adler. In our view, philosophical decision theorists and architects of "top-down" social welfare functions mis-apply economist's technical tools to the philosopher's discursive, narrative job. In consequence they invite the kind of backlash against excessive ambition to regulate all of society that fuels anti-economics.

In light of these remarks, our suggestion that "welfare" be restricted to "technical interpretations that admit of specifications that can be identified in realistically obtainable empirical data" might be misinterpreted. We certainly do *not* mean by this that changes in welfare should be identified with changes in "material" outcomes. The point of behavioral experimentation is to apply the techniques of quantitative welfare estimation to conditions that otherwise lack metrics for assessment. For example, Harrison, Morsink, and Schneider [2022] criticise other economists, specifically Matsuda, Takahashi and Ikegami [2019], for understanding the "direct" benefits of insurance as consisting in actual payouts received by policy-holders, implying that purchasers of insurance who suffer no redeemable damages have made investments that turned out badly *ex post*. This is clearly confused. Of course people often get convinced to buy insurance policies that reduce their welfare for reasons related to actuarial odds and price. But stable insurance markets exist because reduction of risk is itself welfare enhancing in cases where no claim against a policy is made. Part of the gain is that the policy-holder frees resources for consumption or investment that would otherwise be tied up in self-insurance or self-protection. But another part of the gain is typically psychological: relief from aversive feelings of anxiety. The behavioral welfare economist is not advised by the QIS to try to identify the second element by scanning subjects' brains or measuring their cortisol levels. But it can be estimated from willingness to pay *if* an experiment is administered in which the subject's understanding of the terms of the transaction is effectively promoted by careful design.

So what *do* we have in mind when we say that there are aspects of well-being that are not aspects of welfare, and that can only be addressed by qualitative philosophical reflection? It is crucial to our point that we *cannot* point to specific kinds of circumstances that apply predictably across contexts. Where we *can* point to such targets, a behavioral welfare economist can set out to design an experiment for measuring them. But we can imagine comparing two communities in which preferences are satisfied comparably efficiently, but where almost anyone who is culturally external to both communities would regard one as enjoying a wiser or ethically better way of life. We stress that we do *not* think that such judgments can be expressed on generally agreed metrics – that is exactly half of our point. So we should not expect philosophical reflections on such cases to produce arguments that would settle the case for any soundly thinking observer.

The *other* half of our point, however, is that it is arrogant and churlish of an economist to insist that reflections are useless if they cannot establish such definitive conclusions, that is, cannot turn the judgment about well-being into a judgment about welfare after all. The best substitute for argument here is just to orient skeptics toward some best cases of narrative reflection and hope they see the point. But we can supplement this with a practical argument. If an economist insists that everything of value must be reducible in principle to measurable welfare, then she invites anti-economic cultural backlash. In fact, though we can think of some economists who are inclined to rattle humanistic sensibilities in this way in off-hours pontification, the institutionalized sub-discipline that truly embodies the error is another group of philosophers, those who construct technical theories of general rationality and then insist that a best life must conform to such theories.

Thus we defend a conception of the welfare economist's brief as being identification of feasible Kaldor-Hicks-Scitovsky improvements,[47] given structurally specified subjective, heterogeneous utilities within fields of interacting agents, to be identified empirically from observed choice behavior in the lab and the field, modeled using the QIS.

### *B. Paternalism*

Sugden [2004][2009][2018] develops a framework for normatively evaluating agents' outcomes under alternative institutional arrangements in a way that privileges their autonomy as choosers (i.e., their consumer sovereignty) without depending on their specific preference orderings, and thus without requiring their preferences to even be consistently ordered, let alone fully EUT-compliant. According to Sugden [2004][2009], agents are made better off to the extent that their *opportunity sets* are expanded, and worse off to the extent that their *opportunity sets* are contracted.[48] Against this standard, "pure" boosts will

[47] Allowing for the distortionary effects of sidepayments that are not "lump sum," as noted earlier, and applied to climate policy, tax policy, and trade policy, respectively, by Harrison and Rutherford [1999], Harrison, Jensen, Lau and Rutherford [2002] and Harrison, Rutherford and Tarr [2003].

[48] The reader unfamiliar with Sugden's work might wonder how he considers comparative costs of opportunity set expansion without reference to standard preference orderings. The *technical* answer is that he defines

typically make agents better off and "pure" nudges will typically make them worse off.

This idea is indeed attractive as a way of addressing normative questions in circumstances where welfare analysis in the technical sense is not possible due to substantive preference reversals.[49] Thus, for example, this approach can generate recommendations in cases where the method of Bernheim [2009][2016] and Bernheim and Rangel [2008][2009] would find Pareto indifference and therefore yield no guidance. But we should not abjure ever doing standard welfare analysis merely because it can't be undertaken in every context. In both the Harrison and Ng [2016][2018] case and in the situation presented to Harrison and Ross [2018] by their consulting client, the complications arise from the existence of preferences that violate EUT but are nevertheless well-ordered. Arguably, this is the standard situation where relevant utilities are defined over expected monetary values that are risky.

A pre-behavioral theory of welfare that assumes that all choices by agents optimize their subjective expected utility faces no concerns about paternalism. In that setting, the welfare analyst who gives policy advice is, where each individual's welfare is concerned, merely transmitting the agent's own revealed preferences, and can do this with high confidence because the power of the full set of Savage axioms, when these are applied to a set of observations with reasonable dispersion on the Marschak-Machina triangle (Machina [1987]), typically imposes tight constraints on all unobserved (including future) choices

a "weak dominance" relationship among preferences for opportunity sets that does not depend on consistency of preferences over the members of the sets (Sugden [2018; chapter 5]). Sugden does not do away with the concept of preference, but he can relax the definition because he does not require the standard, more restrictive, one to use for defining welfare. The *substantive* answer, details of which would require undue digression here, lies in Sugden's wider philosophy of economics, which makes markets rather than individual choices fundamental, and distinguishes opportunity sets in terms of available exchanges, both at a time and over time, in markets (Sugden [2018; chapter 6]). We think this wider philosophy has much to recommend it, and that it can be put to valuable work in many contexts. It has particular promise for policy problems in which aggregation of welfare is intractably entangled with incentive effects of regulation, as in Sugden [2018; chapter 7]. We merely resist the idea that Sugden's ideas should crowd out applications of more standard approaches to welfare across the board.

[49] Virtually all (incentivized) evidence for preference reversals over monetary rewards entails individuals foregoing *fractions* of pennies in foregone expected income, due to the "payoff dominance" problem: see Harrison [1992][1994]. So the prevalence of preference reversals should not be *automatically* identified with their welfare significance to the individuals. We emphasize "automatically," since we agree with Ainslie [1992][2001] that when people lack, or fail to use, scaffolding that helps them anchor alternatives to fungible currencies of exchange, they often choose in ways that are inconsistent over time, and that could have significant welfare consequences for the individual. The clinical context which Ainslie uses as his primary source for case studies involves serious addictions that surely involve some behaviors with great welfare consequences.

by the agent.

However, the very point of a behavioral approach is based on *rejecting* the general applicability of the assumption that no observed choices are errors. Once we acknowledge that people often make choices that are inconsistent with one another from the perspective of an SEU model, we seem to face the choice between *either* throwing up our hands and pronouncing the agent's utility function to be undiscoverable, or finding grounds for regarding some of her choices as errors. The behavioral welfare economist is committed to the second approach except in cases where she has good independent evidence that the individual subject has in fact lost her agency.

Loss of agency might occur from extreme cognitive impairment or because she has been completely captured by coercive agents who struggle with one another for control of her. Implications of cases of the first kind are discussed in Ross [2023, pp. 97-99]. The second kind of case may initially look like it could only arise in a *Matrix*-type science fiction setting, but in fact is essentially the view of some traditions in sociological theory that minimize agency and view individuals as captives of power dynamics in networks. The behavioral welfare economist, however, predicates her work on the view that most people at least often exercise agency, but also often make mistaken choices. This confronts her with two problems: she must justify identifying errors in other agents' behaviors that they seem not to have spotted themselves, and she needs acceptable grounds for recommending policy interventions that implicitly correct the choices of people who typically are not explicitly asking to be set straight. In other words, the behavioral welfare economist needs a principled position on paternalism.

The traditional ideal in economics is to avoid paternalism. Referring back to our discussion of welfare and rationality immediately above, this sets the economist apart from philosophers and legal scholars such as Adler [2012] who advocate discarding preferences over remote contingencies, and discarding preferences that fail to meet their tests for general rationality, when selecting objectives for policy designers. We must not give the impression that these theorists enjoy consensus about the standard of general rationality. Buchak [2013] has excited considerable discussion amongst philosophers in arguing that general rationality enjoins conforming to the axioms of RDU, with flexible parameters on risk preferences, rather than EUT.[50] But as economists we prefer to stand aloof from such debates, taking the view that public policy officials should do their best to satisfy the preferences they actually find, rather than try to optimize the general well-being of ideally rational counterparts of actual people.

This gulf with respect to grand objectives should not obscure practical affinities. Our welfare economist who applies the QIS is not motivated to worry about choice consistency because she thinks it is required by general rationality. However, her aim is to infer patterns from observed choices and extend these out of sample, a task which necessarily must gain leverage from such consistency as she finds. Furthermore, her Bayesian inferential methods, relying on demographic and other circumstantial covariates about agents and populations, imply that her efforts to avoid paternalism will lead her to frustration *to the extent that* choices of individual agents are governed by *idiosyncratic* parameters. An important example of such a parameter is the λ parameter for idiosyncratic reference-dependent loss aversion in CPT. We gain illustrative traction on the economist's paternalism problem, under the QIS, by contrasting our attitude to CPT with that of Buchak [2013], who is uninhibited about paternalistically separating the rational from the irrational. Insofar as loss aversion is based on λ, it must undermine consistency across (loss, gain and mixed frame) contexts under any set of decision axioms. Therefore, Buchak argues that policy makers or advisors should seek to correct for CPT preferences, just as they might feel justified in correcting revealed preferences for smoking *if and where* such preferences turned out

[50] For an example of the debate triggered by Buchak [2013], see Pettigrew [2016]. Pettigrew reconstructs the domain of EUT to include utilities over actions as well as outcomes, and argues, as against Buchak, that this allows accommodation of rank-dependent preferences under the standard Savage axioms. From a purely mathematical point of view this is unobjectionable. But here we see an illustrative instance of why economists should keep clear distance from the project of using decision theory to construct analyses of "rationality in general". Pettigrew shows no interest in the fact that his reconstruction of the domain of the axioms would disconnect individual decision theory from game theory: all of the main solution strategies for extensive-form games require that utility functions take only outcomes as arguments. No economist in her right mind would want to make this trade-off. In fact, we think there is ultimately no trade-off to be made. Game theory is core technology for social and behavioral science, whereas the project shared by Buchak and Pettigrew has no sustainable justification, even for philosophers. Philosophers who are followers of Kant will disagree. But on Kantian philosophy we applaud the attitude of Binmore [1994][1998].

to be crucially conditioned on scientifically false beliefs about effects of smoking propagated by tobacco companies and their captured experts.

We argued in §1 that there is little empirical basis for regarding utility functions that include $\lambda$ as descriptively important. But this is strictly a contingent matter. There obviously *might be* some agents who find loss of whatever assets they happen to acquire so painful that they would require extravagant compensation to part with them. Such a case is the most natural psychological analogue to $\lambda$. These kinds of sentiments are *exactly* the kind that an anti-paternalist should take care to respect; "Shake off your pain and be rational!" is as literally paternalistic an attitude as it is possible to have. By contrast, *probabilistic* loss aversion that results from subjective probability weighting, which Buchak [2013] promotes as typically recommended by general rationality, is the kind of case where the welfare economist following the QIS may find her anti-paternalist commitments leading to ambiguity.

We agree with Buchak [2013] that rank-dependent risk preferences make sense as heuristics. But this is not because they can be rendered *technically* "rational" in philosophical decision theory. It is rather because of some features of the social world inhabited by strategic agents. This world abounds with scaffolding built by self-interested parties whose manipulations are obscured. Furthermore, such parties (for example, advertisers, politicians, and clickbait artists) typically exploit a general epistemic limitation: marks who only encounter their ploys from time to time will typically have sparse observations from the tails of distributions of game outcomes.[51] In such a world, it is a sensible policy to behave as if tails of distributions on outcome event spaces are fatter than available direct evidence suggests, particularly on the downside. Advisors should not make it their general policy to steer their clients away from such heuristics.

However, under *some* circumstances clients are poorly served by failure to encourage closer alignment between subjective and objective probabilities. The case reviewed in Harrison and Ross [2018] is a crisp example. In this instance, we as economists had a direct client, a retail investment bank. It aimed

[51] That is why internet phishing ploys get more sophisticated over time.

to discourage its customers from inefficient churning of assets in their portfolios, where "inefficient" was defined by the bank in terms of expected wealth at retirement age. This objective was identified by reference to declared goals of a majority of customers. Of course, it cannot be inferred from this that any *particular* customer's pattern of behavior had been inefficient. The client bank sought to boost customers' knowledge of stock behavior through an educational intervention it would make available to them on a voluntary basis. The bank also wished to avoid wasting customers' time by specifically promoting the boosting intervention to customers who wouldn't be expected to benefit from it. Thus the bank proposed to entangle the boosting intervention with selective nudging for a sub-set of customers. Our task was to identify that subset, based on a population-level model informed by risky (lottery) choice experiments *with individual-level estimations* and a simulated investment task, involving real and salient monetary rewards, conducted with a random sample of customers. The model of course included demographic covariates. Based on the experimental data, we distinguished between subjects whose patterns of lottery choices were best modeled by EUT from subjects whose patterns of lottery choices were best modeled by RDU. We then compared the results of individual-level estimation of the lottery choice data with outcomes from the simulated stock market investment choices. We aimed to identify characteristics of customers who would, statistically, be likely to enjoy welfare improvements if they were nudged to be boosted by the bank's educational intervention.

In deciding how to advise our client, issues of paternalism arose in two places.

First, we could assume or not assume that customers preferred to have more money rather than less at the end of the investment history. We assumed that they preferred more money. Arguably, this was not a genuine "choice" on our part; in the absence of this assumption we would not have been able to address our client's question at all. But the choice could be supported by empirical facts: the portfolio products we simulated in the investment task were modeled on actual portfolios explicitly advertised to individuals and households as devices for maximizing their expected monetary wealth at retirement. This knowledge clearly belonged in the priors for any Bayesian model of the study population. In principle, the prior could have been undone by behavioral observations that our subjects atemporally preferred less money to more, though of course that would have been extremely surprising, to put it mildly.

Second, and more interestingly, we had to decide whether to base welfare estimations on EUT models for EUT-conforming subjects and RDU models for RDU-conforming subjects, or on consumer surplus calculations that "imposed" EUT on all subjects. We opted for the first approach, based on considerations about paternalism. We found that against this rubric RDU-conforming subjects suffered significant welfare losses from their investment choices relative to their EUT-conforming counterparts.[52] Thus we recommended nudges toward boosting for customers whose choice patterns indicated that they were best characterised as RDU-conforming, and no nudging for customers whose choice patterns suggested that they were best characterised as EUT-conforming.

Why did we suppose it would be unacceptably paternalistic to model all subjects' welfare outcomes on the basis of EUT? Again, we defend this decision on empirical grounds reflected in our priors as modelers of welfare. There is empirical evidence that many pension funds underweight stocks in portfolios, given historical equity premia, relative to the weighting they would operate if they thought their representative client was an expected utility maximizer with a utility function linear in money (Gomes and Michaeldes [2005]). RDU preferences over financial investments, in any given sample, might reflect disutility from anxiety some people feel during periods of market retrenchment, and our priors recognise this possibility. It would be paternalistic to over-ride such heuristic preferences. On the other hand, subjects might reveal RDU preferences that reflect concerns about strategic exploitation and mis-apply these in circumstances, such as buying standard portfolios from well-regulated retailers, where incentives of manipulative brokers would work in the *opposite* direction to the intent of the intervention we were studying.[53] Our client's boost intervention, by providing customers with richer information about historical distributions of asset prices than customers might otherwise access, might lead them to detect

[52] The only other covariate that predicted significant welfare differences was gender.
[53] That is, self-interested brokers would encourage more portfolio churning rather than less.

and correct such errors – of course that is precisely the assumption and motivation for the intervention.

We were not advising our client on whether to offer the boosting intervention. We were advising them on who to nudge and who not to nudge. As Sugden [2018] argues, *any* nudge is paternalistic to some extent if it suggests that the nudgee devote time and cognitive resources to attending to a boosting effort before the nudgee has full information about the potential value to them of the boost. This should simply be acknowledged. We do not think it is efficient for a welfare economist to generally try to reduce nudging elements of recommended measures to zero; such a fanatical policy would almost certainly lead to net welfare losses. The key difference between applied behavioral welfare economics following the QIS, and alternative approaches such those we criticized in §2, is that the economist following the QIS does not try to distinguish, *a priori*, behavior that reveals "true" or "purified" or "serious" preferences from behavior that reveals "defective" or "impulsive" or "myopic" preferences, and base welfare estimations only on the former. The fully conscientious behavioral welfare economist includes *all* information about policy clients' behavioral records in her priors. When her posterior settles on a model that identifies some choices as errors, she ideally designs independent tests of the relevant error specification. The QIS tells us to design policy suggestions on the basis of the *best* model of client agency that realistically obtainable data allows. It does not tell us to try to discover the "true" model of the latent rational agent, because there is no such thing.

It is noteworthy in this context that the version of behavioral welfare economics promoted by Bernheim and Rangel [2008][2009] grew from their earlier work on justifying interventions to help people overcome addictions [2004]. Identifying the preferences of addicts from the intentional stance is unusually challenging because the typical addict's pattern of choices is highly ambivalent. Addicts often spend resources to try to establish commitments against future addictive consumption *while also* spending resources on drugs or gambling. Many economic models of addiction, following the lead of Schelling [1978][1980][1984], consequently model addicts as two-agent communities forced to bargain with one another. Paternalism with respect to the single *person* is then unavoidable if the policy-maker sides with the interests of one agent, who is trying to quit, against the other agent, who wants to keep the dopamine spikes rolling.

The way to take this seriously as a problem of paternalism is not to search ingeniously for ways of re-framing the technical model so as to avoid having to say that one is ignoring an internal agent. On any modeling, we have paternalism if an economist ends up telling a person who wants to go the bar that he should not "for your own good." If the economist is a friend of the advisee she might ethically think she should intervene in her role as friend rather than in her role as economist. But if she has studied the social statistics on the kinds of scaffolding regulations that actually work to control relapses in addicts, something that *is* within her professional purview, she will find good empirical reason to emphasize the addict's autonomy and responsibility: only self-management rules that addicts construct for themselves,[54] typically with support from family and friends, tend to successfully produce stable outcomes of either abstinence or controlled consumption (Ross [2020]). The most useful contribution of the welfare economist here is to identify policies that minimize the number of people who become caught in addictive ambivalence in the first place. She should focus on optimal taxation of addictive goods that are allowed to be produced because addiction is a side-consequence of their production (for example, alcohol and painkillers), while studying prospects for banning goods that are profitable to producers only if they are sold to addicts (for example, cigarettes and digital slot machines). Once a person is addicted, the appropriate expert for her to consult is a clinical counselor, not a welfare economist.

Similarly, welfare economists who study consequences of financial behavior should not be viewed as individual investment advisors with the whole population as clients. Investment advisors benefit from understanding economics of markets, but they operate mainly in the role of applied psychologists, not economists (Ross [2023]). In the case we are using as an example, Harrison and Ross [2018], though we

[54] What we mean by "rules" here refers to the work of Ainslie [1992][2001]. He refers to his research program, which is intended mainly for clinicians who seek to guide individuals, as "picoeconomics." This is precisely because it is about strategies by which individuals can frame their internal conflict zones as if they were markets, and thus learn personal rules for constructing stable internal "currencies" that help them exchange sooner and later rewards without undermining their own psychic investments.

estimated agents' risk-preference types at the individual level, we did not design a separate policy for each individual subject. However ideal it might be for the welfare economist's work to be taken up as input by waiting squads of investment counsellors, such an approach is very expensive to scale up in real institutional settings. Our policy advice was based on statistical facts about actual people rather than representative agents, but the basis of the advice *was* statistical. We recommended nudging all RDU-conforming customers. Thus the policy almost certainly involved recommending nudges for some subjects who did not need them.

It is conceptual confusion to regard *that* as representing paternalism. All corporate and government policies are based on population-scale statistics. For example, no doubt there are some drivers who are so careful and whose reaction times are so quick that they would never have accidents if they ignored stop signs. No one is being paternalistic to these drivers by not trying to identify them so they can be exempted from the law that all must do what the signs say. Policies are very seldom intended to optimize the welfare of each individual taken one at a time. They usually have, as a crucial objective, coordinating expectations among people who lack specific individual-scale models of one another.

This is why we resist approaches to welfare, such as that of Adler [2012], based on Social Welfare Functions (SWFs). A SWF is intended to aggregate individual preferences in such a way as to ensure that no individual's welfare is over-ridden simply because her preferences are statistically unusual. This may be ethically admirable, but it is too data-hungry to be a feasible ambition. At the same time, Adler [2012] and other SWF theorists advocate ignoring "remote" preferences, that is, preferences that mainly concern consequences for people other than the preference-holder (for example, people living in the distant future or people living in distantly connected geographical or normative communities). The point here is not that SWF theorists think that policy-makers should ignore the future or the well-being of foreigners; it is that they don't think that policies in these areas should be based on individual preference aggregation. The distinction expresses liberal individualism, indeed atomism: remote preferences are considered unacceptably meddlesome and "bossy." Adler [2012] works hard, and carefully, to identify methods for uncovering potential Pareto improvements based on SWFs, but he is skeptical, indeed scathing, about the value of Kaldor-Hicks-Scitovsky improvements, because in the absence of commitment devices that ensure that compensating transfers will actually be made, the efficiency is merely hypothetical.

Our strongly contrasting view is that identifying potential Kaldor-Hicks-Scitovsky improvements is the distinctive core activity of welfare economists.[55] We suggest that the QIS embeds this way of understanding efficiency in a coherent philosophical and methodological package. The view of preferences and beliefs as constructed from the intentional stance through mindshaping processes[56] directly challenges the atomist view of societies as "adding up" sets of individuals with fully pre-formed and complete propositional attitudes latent in their psychologies. We are highly skeptical of a view of administrators as people hired to promote rigorous liberalism as a best universal ethic for maximising well-being. In complex societies made up of individuals and sub-communities with heterogeneous preferences and beliefs, institutions and rules play vital *coordinating* functions. This activity takes intentional-stance profiles of people as it finds them, but it is highly dynamic; because networks of intentional-stance profiles are constructed to mediate mutual expectations and stabilise bargaining spaces, they adapt to and co-evolve with institutional and legal structures. Underlying welfare economics in every society is continuous cultural evolution of values, as Binmore [2005] emphasizes. Thus the need for efficiency analysis never ends – windows for improvement are passing phenomena, many of which are grasped and many of which are wasted.

The welfare economist's concern to keep paternalism in check is an expression of liberalism. It is a *fact* about the history of mainstream economic theory that it has evolved from the outset within a broadly liberal ethical frame (Ross [2012]). But liberalism is a mansion of many rooms. The version according to which the first duty of the policy-maker is to safeguard the sovereignty of atomic individuals is one

[55] Recall our discussion in §3.B of practical calculations with respect to tax policy, trade policy, and climate change policies that demonstrate how such compensation schemes can be implemented with realistic tax instruments.

[56] These processes are explained in §3.A.

tradition in liberal philosophy, and it expresses itself in programs for welfare management such as Adler's. Another tradition emphasizes instead the role of government as an agent hired by a democratic principal - citizens - that should avoid promoting an independent conception of the good. It does this not by blocking out all information except non-remote individual preferences, but by being open to influence from any coalition of interests. Restricted by a constitution that is hard to amend, it tinkers with rules and institutions so that complex networks of bargainers can evolve by their own dynamics. In effect it uses the rule of law to stabilize a policy market, similar in structure to the markets for goods and services that it also enables. In case some readers are unconvinced that this second conception, the one we defend as a sound ethical basis for behavioral welfare economics, is truly liberal, we will just point out here that its leading visionary theorist is Hayek (Ross [2011]). That is a reliable enough anti-paternalist brand for us. The core of our effective anti-paternalism is the Keynesian humbleness we urged on applied economists in §4A: the economist's job is to identify social inefficiencies for general consideration, not to try to effect ideological reforms. Such reforms are often morally salutary or even morally urgent, but they are not the professional business of the economist.

## 5. Conclusions

The path that leads to the QIS for behavioral welfare economics is summarized in six steps.

First, we begin from the recognition that an agent's welfare cannot be identified by naïve RPT, that is, by assuming that the agent *always* chooses what is best for her. The agents that most interest economists, people, make choices that are inconsistent over time. Sometimes that can be accommodated by recognizing that a natural person who lives a full lifespan is a sequence of agents. With respect to the features economists use to define agency, no one is the same agent at 20 years old and 70 years old. But human inconsistency of choice is more pervasive and "instantaneous" than that. We must allow that many human choices are errors. Estimations of welfare must identify and discount these errors.

Second, one way to reduce proportions of choices to be treated as errors is to impose a strong model of general individual rationality. Then the analyst can count as errors any choices that fall short of this standard. This approach has a deep history in philosophy, which in its modern form has its roots in Kant's conception of practical reason. It is now best represented by the enterprise of analytic decision theory, which economists readily recognize because it shares its technical foundations with EUT. But economists should reject this approach. One reason to reject this approach is that, like the other parts of analytic philosophy, it seeks an unreachable objective. Just as there will always be counterexamples to every attempted philosophical analysis of knowledge, so there will always be counterexamples to every attempted philosophical analysis of general rationality. Welfare economics is a practical enterprise intended to guide policy makers and administrators. We will not make progress in this by endlessly constructing, puncturing, and patching technical specifications of general rationality. The other reason to reject this approach is that, because of its radical individualism, this tradition effectively amounts to a project of trying to work out what a single agent should do if she wants to earn the sobriquet "rational". That is not the welfare economist's project. The welfare economist's typical project is to identify policies and structures that are inefficient at the social, statistical level. Reforms at this level inevitably produce winners and losers. In recognition of this, economists should self-consciously maintain Keynesian humbleness. Our job is to be experts on efficiency, not to re-engineer our societies. If an economist's society needs fundamental reform for moral reasons, she could go out to the barricades with everyone else who shares her convictions, but leaving her economist's hat at home.

Third, another popular way to try to reduce the proportion of choices to use as the "real" evidence for welfare optima is to try to identify psychological mechanisms that tend to generate choices that undermine a chooser's welfare. Then choices that are predominantly produced by these mechanisms can be regarded as "anomalies" to be corrected in theory, and nudged against in practice. This has been the dominant approach in behavioral welfare economics. The deep problem with it, as recognized by a few "insiders" to economics such as Ken Binmore, Nick Chater, and Robert Sugden, is that it rests on a naïve conception of psychology. Most human behavior is not mechanically caused by mechanisms that are latent within, or even "supervene on," their brains. Welfare is characterized in terms of preference satisfaction. Preferences cannot be identified from choice behavior independently of specifying beliefs. The most promising, and increasingly dominant, approach in cognitive science emphasizes that preferences and beliefs are social constructions used to make people mutually comprehensible to one another and to coordinate their expectations so they can make shared policies – sometimes cooperatively, and sometimes competitively, but always within arenas of institutional constraint. Saying that preferences and beliefs are social constructions is not a way of saying that they are unreal. It is rather to say that they are virtual kinds of objects. Modern societies teem with such real objects, money being an example very familiar to economists. Preferences and beliefs are identified in everyday contexts by a "folk" intentional stance. Like all folk models, this incorporates many myths that scientific study does not ratify. Cognitive scientists make progress from a scientific intentional stance, meaning one that survives the filtering processes of collective rigor about observability across cases.

Fourth, economics is fundamentally quantitative. This does not reflect a view to the effect that careful qualitative reflection is uninformative. Economists should have deep respect for an older philosophical tradition, which we associate with Aristotle but which has many superb contemporary practitioners, of articulating more and less wise ways for humans to aim to live. Such reflections gain their power and reasonableness precisely by being highly sensitive to nuances of historically and culturally specific cases. This makes them valuable for individuals or densely connected communities, but less useful for policy makers and administrators who must rely on statistical generalisations.[57] Economists should not be committed to quantitative methods because they want to be as respected as physicists. That motivation unquestionably did motivate many of our iconic predecessors, but humble Keynesians should treat that as an embarrassing quirk in tribal history. Economics is quantitative because what we study are statistical patterns in sets of observed choices. We take a special *quantitative intentional stance* with respect to behavior

[57] A caveat is in order here. Reading the best philosophy may from time to time stop an economist from recommending a policy that is literally stupid, such as institutional discouragement of gift-giving because gift-giving is bound to be inefficient.

that is based on choice. We do not identify "choice" with any psychological mechanism. We regard behavior as "chosen" if it can be influenced by changes in incentives. That is also identifiable by statistical evidence. Blinking is not chosen behavior because people will not stop doing it no matter how much you pay them. Physiologically similar winking is chosen behavior, and we have the statistical evidence to assure us of this. Far more of our fathers and grandfathers winked at their female colleagues at their offices than our contemporaries do. Everyone knows why this changed, without our having to know anything about processing details in any heads.

Fifth, the basic engine for identifying choice patterns is Bayesian inference. Crucial to this methodology is that everything we think we know to be relevant, and everything we observe, is included in priors. In this sense, the approach is the opposite to methods that deliberately exclude choices thought to be irrational on philosophical grounds, or that result from purported mechanisms that the subject would disown when thinking at her best. Of course, we can only include in priors what we can represent in a single model. So models must be structural. The aim is to identify the production function for a subject's welfare, so we can ask under what conditions the function in question outputs the highest *flow* of utility. Best methodology is thus to estimate at the level of the individual, even when our ultimate focus is on social efficiency, that is, potential Kaldor-Hicks-Scitovsky improvements.

Sixth, while rejecting simple-minded RPT that allows for no errant choices, the QIS is very much in the *spirit* of RPT. That is, it understands utility functions as summaries of observed choices, conditioned on beliefs. This "contemporary" RPT has been discussed in the leading economic methodology literature for some years; highlights are Binmore [2009], Hands [2013] and Ross [2014][2023]. Ross [2014] refers to the methodological program as "neo-Samuelsonian": it is "Samuelsonian" because, like the intentional stance, it ascribes preferences in order to describe observed behavior, and it is "neo" because it goes well beyond the original weak axiom of revealed preference and applies to finite sets of choices.

We agree with Leamer [2012] that economics is fundamentally a policy science. Like medical and engineering academics, its practitioners, even when they are not directly serving specific clients, choose the domain to which they apply their theory, mathematics, and statistics, on concerns about efficiency.[58] Nothing is efficient or inefficient in economics except with respect to values of inputs and outputs, and it is the realm of practical pursuits that sets these values. This is not a narrow conception of economics. It applies to many transactions that do not take place in markets mediated directly by exchanges of money. The practical pursuits that establish values might be those of fish, bees, elephants or humans.[59] But at least in the background to everything that economists professionally think about is someone's welfare. Therefore, it is as important as can be that we conceptualize and measure welfare with great care.

[58] In saying this, we agree with Sugden [2018] that we would like to see the end of obligatory sections in economics articles of "policy recommendations" that are not directed to any actual agency that could feasibly implement them. Even if all economics is deeply conditioned by practical concerns, many specific exercises in economics are not practical, and it does no favors to our professional reputation to pretend otherwise.

[59] See Kacelnik and Bateson [1996], Bshary [2001] and Chittka [2022].

## Figure 1: Posterior Predictive Consumer Surplus Distribution for Each of Four Insurance Purchase Choices by One Subject

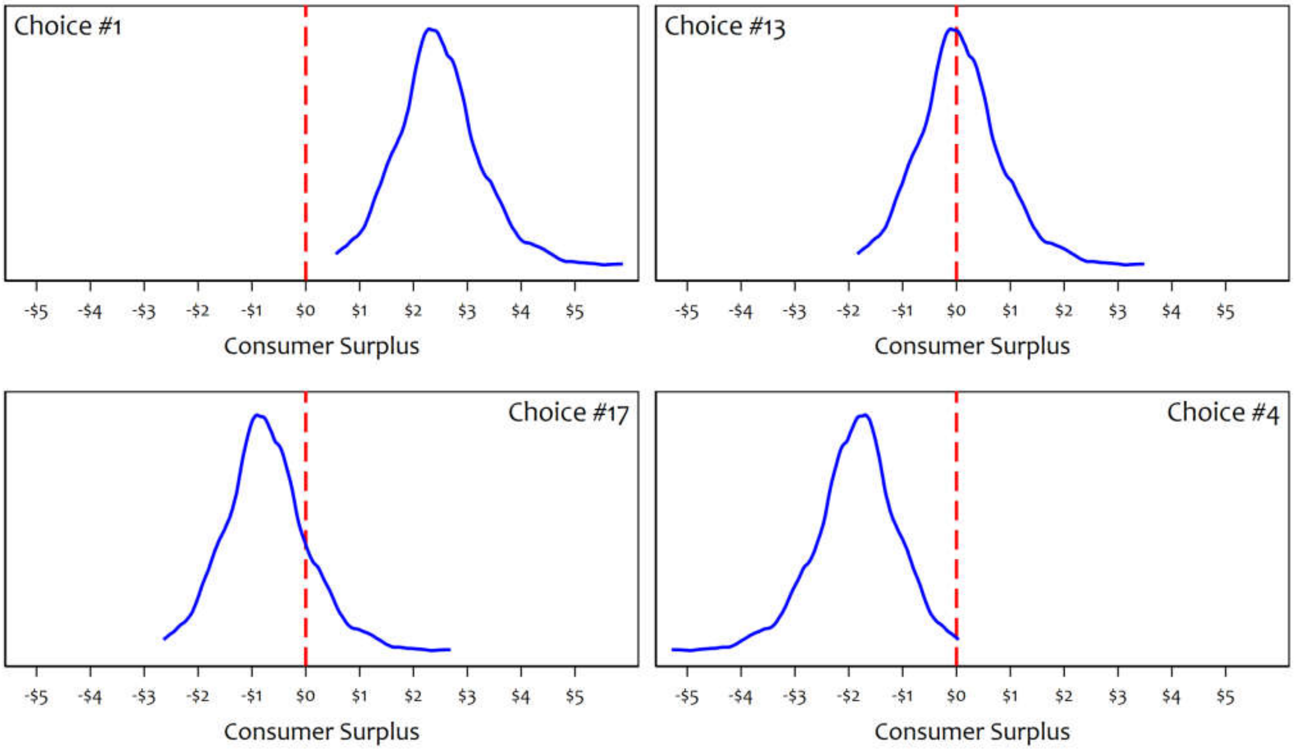


Figure 2: Adaptive Welfare Evaluations for Subject #1

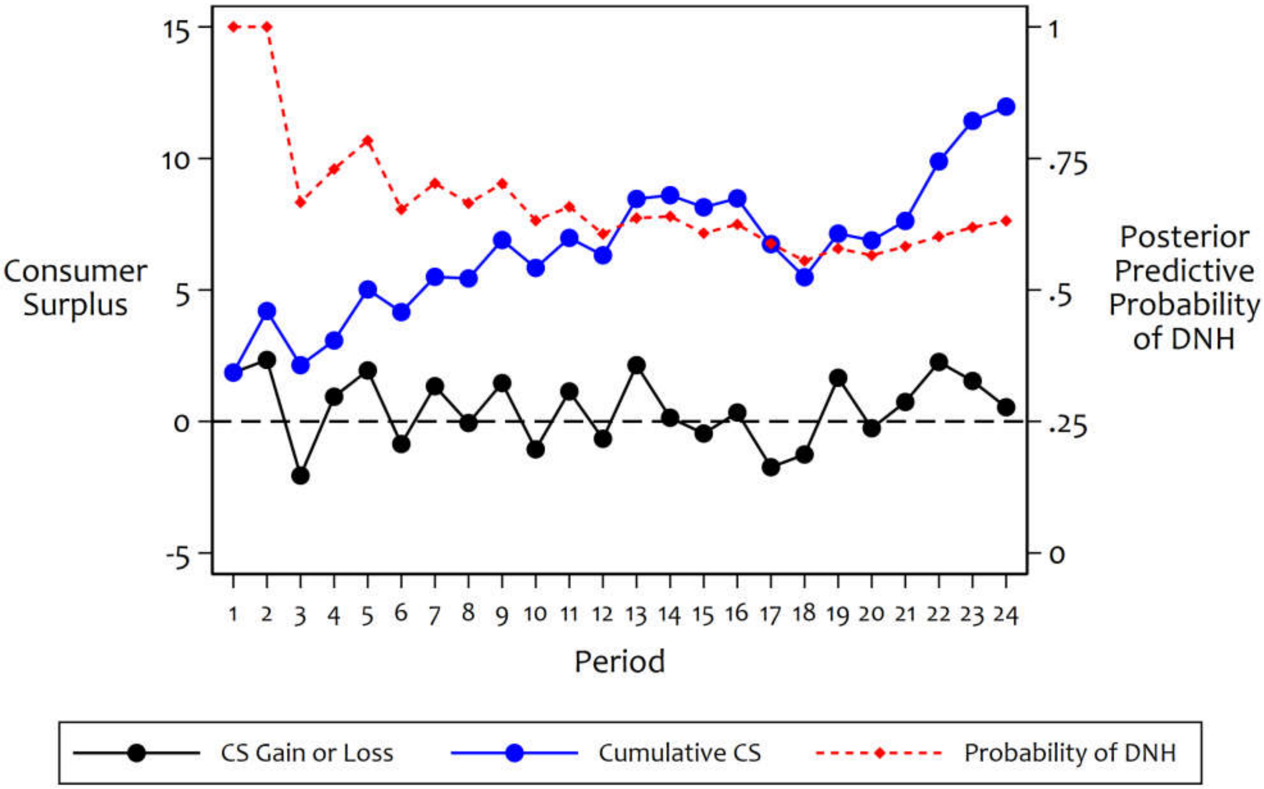
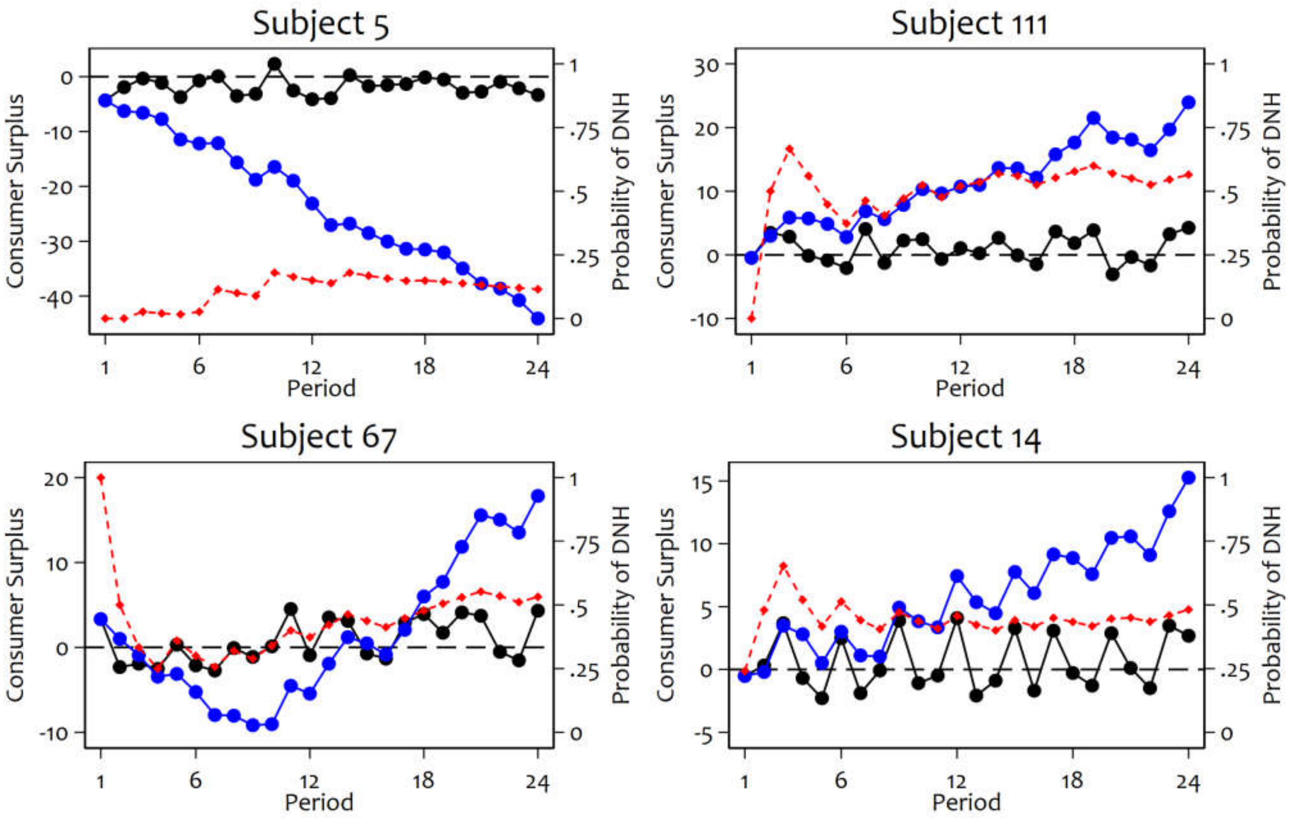


Figure 3: Individual Adaptive Welfare Evaluations for Four Subjects



## References

Adler, Matthew D., *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis* (New York: Oxford University Press, 2012).

Adler, Matthew D., *Measuring Social Welfare: An Introduction* (New York: Oxford University Press, 2019).

Afriat, Sydney N., "Efficiency Estimation of Production Function," *International Economic Review*, 13(3), 1972, 568-598.

Ainslie, George, *Picoeconomics* (Cambridge, UK, Cambridge University Press, 1992).

Ainslie, George, *Breakdown of Will* (Cambridge, UK, Cambridge University Press, 2001).

Alekseev, Alex; Harrison, Glenn; Lau, Morten, and Ross, Don, "Deciphering the Noise: The Welfare Costs of Noisy Behavior," *CEAR Working Paper 2018-0*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2018.

Allen, Franklin, "Discovering Personal Probabilities When Utility Functions Are Unknown," *Management Science*, 33(4), 1987, 452-454.

Ambuehl, Sandro; Bernheim, B. Douglas; Ersoy, Fulya, and Harris, Donna, "Peer Advice on Financial Decisions: A Case of the Blind Leading the Blind?" *Working Paper 25034*, National Bureau of Economic Research, 2018.

Ambuehl, Sandro; Bernheim, B. Douglas, and Lusardi, Annamaria, "The Effect of Financial Education on the Quality of Decision Making," *Working Paper 20618*, National Bureau of Economic Research, 2014.

Ambuehl, Sandro; Bernheim, B. Douglas, and Lusardi, Annamaria, "A Method for Evaluating the Quality of Financial Decision Making, with an Application to  Financial Education," *Working Paper 20618*, National Bureau of Economic Research, 2017.

Ambuehl, Sandro; Bernheim, B. Douglas, and Lusardi, Annamaria, "Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice," *American Economic Review*, 112(11), 2022, 3584-3626.

Andersen, Steffen; Cox, James C.; Harrison, Glenn W.; Lau, Morten I.; Rutström, E. Elisabet, and Sadiraj, Vjollca, "Asset Integration and Attitudes to Risk: Theory and Evidence," *Review of Economics & Statistics*, 100(5), 2018, 816-830.

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Journal of Risk & Uncertainty*, 48, 2014, 207-229.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Elicitation Using Multiple Price List Formats," *Experimental Economics*, 9(4), 2006, 383–405.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008a, 583–618.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Lost in State Space: Are Preferences Stable?" *International Economic Review*, 49(3), 2008b, 1091-1112.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Discounting Behavior and the Magnitude Effect: Evidence from a Field Experiment in Denmark," *Economica*, 80, 2013, 670-697.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Discounting Behavior: A Reconsideration," *European Economic Review*, 71, 2014, 15–33.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion," *International Economic Review*, 59(2), 2018, 537–555.

Applebaum, Binyamin, *The Economists' Hour* (Boston: Little, Brown, 2019).

Armitage, Paul, "The Search for Optimality in Clinical Trials," *International Statistical Review*, 53(1), 1985, 15-24.

Atkinson, Anthony B., "The Strange Disappearance of Welfare Economics," *Kyklos*, 54(2/3), 2001, 193-206.

Atkinson, Anthony B., "Economics as a Moral Science," *Economica*, 76, 2009, 791-804.

Atkinson, Anthony B., "The Restoration of Welfare Economics," *American Economic Review (Papers & Proceedings)*, 101(3), 2011, 157-161.

Baily, Martin Neil, "Some Aspects of Optimal Unemployment Insurance," *Journal of Public Economics*, 10(3), 1978, 379-402.

Bell, David, "Regret in Decision Making under Uncertainty," *Operations Research*, 20, 1982, 961–981.

Bell, David, "Disappointment in Decision Making under Uncertainty," *Operations Research*, 33, 1985, 1–27.

Berman, Elizabeth, *Thinking Like an Economist: How Efficiency Replaced Equality in U.S. Public Policy* (Princeton: Princeton University Press, 2022).

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

Bernheim, B. Douglas, "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics," *Journal of Benefit-Cost Analysis*, 7(1), 2016, 12–68.

Bernheim, B. Douglas; Fradkin, Andrey, and Popov, Igor, "The Welfare Economics of Default Options in 401(k) Plans," *American Economic Review*, 105(9), 2015, 2798–2837.

Bernheim, B. Douglas, and Rangel, Antonio, "Addiction and Cue-Triggered Addiction Processes," *American Economic Review*, 94, 2004, 1558-1590.

Bernheim, B. Douglas, and Rangel, Antonio, "Choice-Theoretic Foundations for Behavioral Welfare Economics," in A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook* (Oxford: Oxford University Press, 2008).

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

Bernheim, B. Douglas, and Taubinksy, Dmitry, "Behavioral Public Economics," in B.D. Bernheim, S. DellaVigna and D. Laibson (eds), *Handbook of Behavioral Economics – Volume 1* (New York, Elsevier, 2018).

Berry, Donald A., and Fristedt, Bert (eds.), *Bandit Problems: Sequential Allocation of Experiments* (New York: Springer, 1985).

Binmore, Ken, *Game Theory and the Social Contract, Volume 1: Playing Fair* (Cambridge MA: MIT Press, 1994).

Binmore, Ken, *Game Theory and the Social Contract, Volume 2: Just Playing* (Cambridge, MA: MIT Press, 1998).

Binmore, Ken, *Natural Justice* (Oxford: Oxford University Press, 2005).

Binmore, Ken, *Rational Decisions* (Princeton: Princeton University Press, 2009).

Bleichrodt, Han; Pinto, J.L, and Wakker, Peter P., "Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility," *Management Science*, 47, 2001, 1498-1514.

Blundell, Richard; Pistaferri, Luigi, and Preston, Ian, "Consumption Inequality and Partial Insurance," *American Economic Review*, 98(5), 2008, 1887-1921.

Boadway, Robin, and Bruce, Neil, *Welfare Economics* (Oxford: Blackwell, 1984).

Buchak, Lara, *Risk and Rationality* (New York: Oxford University Press, 2013).

Bshary, Redouan, "The Cleaner Fish Market," in R. Noë, J. van Hooff and P. Hammerstein (eds.), *Economics in Nature* (Cambridge: Cambridge University Press, 2001).

Camerer, Colin; Issacharoff, Samuel; Loewenstein, George; O'Donoghue, Ted, and Rabin, Matthew, "Regulation for Conservatives: Behavioral Economics and the Case for Asymmetric Paternalism," *University of Pennsylvania Law Review*, 151, 2003, 1211-1254.

Camerer, Colin; Loewenstein, George, and Prelec, Drazen, "Neuroeconomics: How Neuroscience Can Inform Economics," *Journal of Economic Literature*, 43, 2005, 9-64.

Caporael, Linnda; Griesemer, James, and Wimsatt, William (eds.), *Developing Scaffolds in Evolution, Culture, and Cognition* (Cambridge, MA: MIT Press, 2014).

Caria, Stefano; Gordon, Grant; Kasy, Maximilian; Quinn, Simon; Shami, Soha, and Teytelboym, Alexander, "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan," *Draft Working Paper*, Oxford University, May 2020; available at https://maxkasy.github.io/home/research/

Chambers, Christopher, and Echenique, Federico, *Revealed Preference Theory* (New York: Cambridge University Press, 2016).

Chater, Nick, *The Mind is Flat* (New Havn, CT: Yale University Press, 2018).

Cherry, Todd; Crocker, Thomas, and Shogren, Jason, "Rationality Spillovers," *Journal of Environmental Economics and Management*, 45, 2003, 63-84.

Chetty, Raj, "A General Formula for the Optimal Level of Social Insurance," *Journal of Public Economics*, 90(10-11), 2006, 1879-1901.

Chetty, Raj., and Looney, Adam, "Consumption Smoothing and the Welfare Consequences of Social Insurance in Developing Countries," *Journal of Public Economics*, 90, 2006 2351-2356.

Chittka, Lars, *The Mind of a Bee* (Princeton, NJ: Princeton University Press, 2022).

Chuang, Yating, and Schechter, Laura, "Stability of Experimental and Survey Measures of Risk, Time and Social Preferences: A Review and Some New Results," *Journal of Development Economics*, 117, 2015, 151–170.

Clark, Andy, *Being There* (Cambridge, MA: MIT Press, 1998).

Coleman, William, *Economics and its Enemies* (Houndmills, Basingstoke: Palgrave Macmillan, 2002).

Coller, Maribeth, and Williams, Melonie B., "Eliciting Individual Discount Rates," *Experimental Economics*, 2(2), 1999, 107–127.

Collins, Daryl; Morduch, Jonathan; Rutherford, Stuart, and Ruthven, Orlanda, *Portfolios of the Poor: How the World's Poor Live on $2 a Day* (Princeton, NJ: Princeton University Press, 2009)

Cox, James C., and Sadiraj, Vjollca, "Small- and Large-Stakes Risk Aversion: Implications of Concavity Calibration for Decision Theory," *Games and Economic Behavior*, 56, 2006, 45-60.

de Haan, Thomas, and Linde, Jona, "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement," *Economic Journal*, 128(610), 2018, 1180-1206.

Dennett, Daniel, "Intentional Systems," *The Journal of Philosophy*, 68(4), February 1971, 87-106.

Dennett, Daniel, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

Dennett, Daniel, *Consciousness Explained* (Boston: Little, Brown, 1991).

Dennett, Daniel, *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W.W. Norton and Company, 2017).

Dowding, Keith, and Taylor, Brad, *Economic Perspectives on Government* (Houndmills, Basingstoke: Palgrave Macmillan 2019).

Duflo, Esther, "The Economist as Plumber," *American Economic Review*, 107, 2017, 1-26.

Feldman, Roger, and Dowd, Bryan, "A New Estimate of the Welfare Loss of Excess Health Insurance," *American Economic Review*, 81(1), March 1991, 297-301.

Feldstein, Martin S., "The Welfare Loss of Excess Health Insurance," *Journal of Political Economy*, 81(2), 1973, 251-280.

Fishburn, Peter C. "Nontransitive Measurable Utility," *Journal of Mathematical Psychology*, 26, 1982, 31-67.

Ford, Kenneth, and Pylyshyn, Zenon (eds.), *The Robot's Dilemma Revisited* (New York: Praeger, 1996).

Friedman, Milton, *Essays in Positive Economics* (Chicago: University of Chicago Press, 1953).

Gao, Xiaoxue Sherry; Harrison, Glenn W., and Tchernis, Rusty, "Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach," *Experimental Economics*, forthcoming 2022, DOI: https://doi.org/10.1007/s10683-022-09751-0.

Gigerenzer, Gerd; Todd, Peter & the ABC Research Group, *Simple Heuristics that Make Us Smart* (Oxford: Oxford University Press, 1999).

Glennerster, Rachel, "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency," in Banerjee, A. and Duflo, E. (eds.), *Handbook of Field Experiments: Volume One* (Amsterdam: North-Holland, 2017).

Gomes, Francisco, and Michaeldes, Alexander, "Optimal Life-Cycle Asset Allocation: Understanding the Empirical Evidence," *Journal of Finance*, 60, 2005, 869-904.

Grüne-Yanoff, Till, and Hertwig, Ralph, "Nudge *vs.* boost: How coherent are policy and theory?" *Minds and Machines*, 26, 2016, 149–183.

Gul, Faruk, "A Theory of Disappointment Aversion," *Econometrica*, 59, 1991, 667-686.

Gul, Faruk, and Pesendorfer, Wolfgang, "The Case for Mindless Economics," in A. Caplin & A. Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook* (New York: Oxford University Press, 2008).

Hadad, Vitor; Hirshberg, David A.; Zhan, Ruohan; Wager, Stefan, and Athey, Susan, "Confidence Intervals for Policy Evaluation in Adaptive Experiments," *Proceedings of the National Academy of Sciences*, 118(15), 2021, e2014602118; DOI: 10.1073/pnas.2014602118 .

Handel, Benjamin R., "Adverse Selection and Inertia in Health Insurance markets: When Nudging Hurts," *American Economic Review*, 103(7), 2013, 2643-2682.

Handel, Benjamin R., and Kolstad, Jonathan T., "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 105(8), 2015, 2449-2500.

Hands, Wade, "Foundations of Contemporary Revealed Preference Theory," *Erkenntnis*, 78, 2013, 1081-1108.

Hansson, Bengt, "Risk Aversion as a Problem of Conjoint Measurement," in P. Gardenfors and N-E. Sahlin (eds.), *Decisions, Probability, and Utility* (New York: Cambridge University Press, 1988).

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions: Reply," *American Economic Review*, 82, 1992, 1426-1443.

Harrison, Glenn W., "Expected Utility Theory and The Experimentalists," *Empirical Economics*, 19(2), 1994, 223-253.

Harrison, Glenn W., "Neuroeconomics: A Critical Reconsideration," *Economics & Philosophy*, 24(3), 2008, 303-344.

Harrison, Glenn W., "Randomisation and Its Discontents," *Journal of African Economies*, 20(4), 2011a, 626-652.

Harrison, Glenn W, "Experimental Methods and the Welfare Evaluation of Policy Lotteries," *European Review of Agricultural Economics*, 38(3), 2011b, 335-360.

Harrison, Glenn W., "Impact Evaluation and Welfare Evaluation," *European Journal of Development Research*, 26, 2014, 39-45.

Harrison, Glenn W., "The Behavioral Welfare Economics of Insurance," *Geneva Risk & Insurance Review*, 44(2), 2019, 137–175.

Harrison, Glenn W., "Experimental Design and Bayesian Interpretation," H. Kincaid and D. Ross (eds.), *Modern Guide to the Philosophy of Economics* (Cheltenham, UK: Elgar, 2021).

Harrison, Glenn W.; Jensen, Jesper; Lau, Morten, and Rutherford, Thomas F., "Policy Reform Without Tears," in A. Fossati and W. Weigard (eds.), *Policy Evaluation With Computable General Equilibrium Models* (New York: Routledge, 2002).

Harrison, Glenn W.; Lau, Morten I.; Ross, Don, and Swarthout, J. Todd, "Small Stakes Risk Aversion in Experiments: A Reconsideration," *Economics Letters*, 160, 2017, 24-28.

Harrison, Glenn W.; Lau, Morten I., and Yoo, Hong Il, "Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment,"*Review of Economics & Statistics*, 102(3), 2020, 552-568.

Harrison, Glenn W; Martínez-Correa, Jimmy, and Swarthout, J. Todd, "Eliciting Subjective Probabilities with Binary Lotteries," *Journal of Economic Behavior and Organization*, 101, 2014, 128-140.

Harrison, Glenn W; Martínez-Correa, Jimmy, and Swarthout, J. Todd, "Reduction of Compound Lotteries with Objective Probabilities: Theory and Evidence," *Journal of Economic Behavior & Organization*, 119, 2015, 32-55.

Harrison, Glenn W., Martínez-Correa, Jimmy, Swarthout, J. Todd, and Ulm, Eric, "Eliciting Subjective Probability Distributions with Binary Lotteries," *Economics Letters*, 127, 2015, 68-71.

Harrison, Glenn W., Martínez-Correa, Jimmy, Swarthout, J. Todd, and Ulm, Eric "Scoring Rules for Subjective Probability Distributions," *Journal of Economic Behavior & Organization* 134, 2017, 430-448.

Harrison, Glenn W.; Morsink, Karlijn, and Schneider, Mark, "Do No Harm? The Welfare Consequences of Behavioral Interventions," *CEAR Working Paper 2020-12*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2020.

Harrison, Glenn W.; Morsink, Karlijn, and Schneider, Mark, "Literacy and the Quality of Index Insurance Decisions," *Geneva Risk and Insurance Review*, 47, 2022, 66-97.

Harrison, Glenn W., and Ng, Jia Min, "Evaluating the Expected Welfare Gain from Insurance," *Journal of Risk and Insurance*, 83(1), 2016, 91-120.

Harrison, Glenn W., and Ng, Jia Min, "Welfare Effects of Insurance Contract Non-Performance," *Geneva Risk and Insurance Review*, 43(1), 2018, 39-76.

Harrison, Glenn, W., and Ross, Don, "The Psychology of Human Risk Preferences and Vulnerability to Scare-Mongers: Experimental Economic Tools for Hypothesis Formulation and Testing," *Journal of Cognition and Culture*, 16, 2016, 383-414.

Harrison, Glenn W., and Ross, Don A., "The Empirical Adequacy of Cumulative Prospect Theory and its Implications for Normative Assessment," *Journal of Economic Methodology*, 24(2), 2017, 150-165.

Harrison, Glenn W., and Ross, Don A. "Varieties of Paternalism and the Heterogeneity of Utility Structures," *Journal of Economic Methodology*, 25(1), 2018, 42-67.

Harrison, Glenn W., and Ross, Don, "The Methodologies of Neuroeconomics," *Journal of Economic Methodology*, 17, 2020, 185-196.

Harrison, Glenn W., and Rutherford, Thomas F., "Burden Sharing, Joint Implementation, and Carbon Coalitions," in C. Carraro (ed.), *International Environmental Agreements on Climate Change* (Amsterdam: Kluwer, 1999).

Harrison, Glenn W.; Rutherford, Thomas F., and Tarr, David G., "Quantifying the Uruguay Round," *Economic Journal*, 107, 1997, 1405-1430.

Harrison, Glenn W.; Rutherford, Thomas F., and Tarr, David G., "Trade Liberalization, Poverty and Efficient Equity," *Journal of Development Economics*, 71, 2003, 97-128.

Harrison, Glenn W.; Rutherford, Thomas F., and Wooton, Ian, "An Alternative Welfare Decomposition for Customs Unions," *Canadian Journal of Economics*, 26(4), 1993, 961–968.

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Harrison, Glenn W., and Rutström, E. Elisabet, "Expected Utility *And* Prospect Theory: One Wedding and A Decent Funeral," *Experimental Economics*, 12(2), 2009, 133-158.

Harrison, Glenn W., and Swarthout, J. Todd, "Cumulative Prospect Theory in the Laboratory: A Reconsideration," in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2023).

Hausman, Daniel, *Preference, Value, Choice and Welfare* (Cambridge: Cambridge University Press, 2011).

Hertwig, Ralph; Hoffrage, Ulrich, and the ABC Research Group, *Simple Heuristics in a Social World* (Oxford: Oxford University Press, 2013).

Hey, John D., "Why We Should Not Be Silent About Noise," *Experimental Economics*, 8(4), 2005, 325–345.

Hood, Bruce, *The Self Illusion* (Oxford: Oxford University Press, 2012).

Hossain, Tanjim, and Okui, Ryo, "The Binarized Scoring Rule," *Review of Economic Studies*, 2013, 80, 984-991.

Infante, Gerardo; Lecouteux, Guilhem, and Sugden, Robert, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioral Welfare Economics," *Journal of Economic Methodology*, 23, 2016, 1-25.

Kacelnik, Alex, and Bateson, Melissa, "Risky Theories? The Effects of Variance on Foraging Decisions," *American Zoologist*, 36(4), 1996, 402-434.

Kahneman, Daniel; Slovic, Paul, and Tversky, Amos (eds.), *Judgment Under Uncertainty: Heuritics and Biases* (Cambridge: Cambridge University Press, 1982).

Kahneman, Daniel, and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 1979, 263-291.

Karlan, Dean, and Appel, Jacob., *More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy* (New York; Dutton, 2011).

Kasy, Maximilian, and Sautmann, Anja, "Adaptive Treatment Assignment in Experiments for Policy Choice," *Econometrica*, 89(1), 2021, 113-132.

Keynes, John Maynard, "Economic Possibilities for Our Grand-children," re-printed in J.M. Keynes, *Essays in Persuasion* (New York: Norton, 1963, 358-373).

Ladyman, James, and Ross, Don, *Every Thing Must Go: Metaphysics Naturalised* (Oxford: Oxford University Press, 2007).

Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York: Wiley, 1978).

Leamer, Edward E., "Tantalus on the Road to Asymptopia," *Journal of Economic Perspectives*, 24(2), 2010, 31-46.

Leamer, Edward E., *The Craft of Economics* (Cambridge, MA: MIT Press, 2012).

Loomes, Graham, and Sugden, Robert, "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty," *Economic Journal*, 92, 1982, 805-824.

Loomes, Graham, and Sugden, Robert, "Some Implications of a More General Form of Regret Theory," *Journal of Economic Theory*, 41(2), 1987, 270-287.

Lyons, William, *The Disappearance of Introspection* (Cambridge, MA: MIT Press, 1986).

Machina, Mark J., "Choice under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives*, 1(1), 1987, 121-154.

Machina, Mark J., and Schmeidler, David, "A More Robust Definition of Subjective Probability," *Econometrica*, 60(4), 1992, 745–780.

Machina, Mark J., and Schmeidler, David, "Bayes Without Bernoulli: Simple Conditions for Probabilistically Sophisticated Choice," *Journal of Economic Theory*, 67(1), 1995, 106-128.

Manzini, Paola, and Marco Mariotti, "Categorize Then Choose: Boundedly Rational Choice and Welfare," *Journal of the European Economic Association*, 10(5), 2012, 1141-1165.

Manzini, Paola, and Mariotti, Marco, "Welfare Economics and Bounded Rationality: The Case for Model-Based Approaches," *Journal of Economic Methodology*, 21, 2014, 343-360.

Marglin, Stephen, *The Dismal Science: How Thinking Like an Economist Undermines Community* (Cambridge, MA: Harvard University Press, 2008).

Matsuda, Ayako; Takahashi, Kazushi, and Ikegami, Munenobu, "Direct and Indirect Impact of Index-Based Livestock Insurance in Southern Ethiopia," *Geneva Papers on Risk and Insurance: Issues and Practice*, 44, 2019, 481-502.

McKelvey, Richard D., and Page, Talbot, "Public and Private Information: An Experimental Study of Information Pooling," *Econometrica*, 58(6), 1990, 1321-1339.

McGeer, Victoria, "Psycho-practice, Psycho-theory, and the Contrastive Case of Autism: How Processes of Mind Become Second Nature," *Journal of Consciousness Studies*, 8, 2001, 109-132.

McGeer, Victoria, "Enculturating Folk-psychologists," *Synthese*, 199, 2020, 1039-1063.

Mäki, Uskali (ed.), *The Methodology of Positive Economics: Reflections on the Milton Friedman Legacy* (Cambridge: Cambridge University Press, 2009).

Nichols, Shaun, and Stich, Stephen, *Mindreading* (Oxford: Oxford University Press, 2003).

Nussbaum, Martha, *Cultivating Humanity* (Cambridge, UK: Cambridge University Press, 1997).

Ortmann, Andreas, "On the Foundations of Behavioural and Experimental Economics," in Harold Kincaid & Don Ross (eds.), *A Modern Guide to Philosophy of Economics* (Northampton, MA: Edward Elgar, 2021).

Peto, Richard, "Discussion of Papers by J.A. Bather and P. Armitage," *International Statistical Review*, 53(1), 1985, 31-34.

Pettigrew, Richard, "Risk, Rationality, and Expected Utility Theory," *Canadian Journal of Philosophy*, 45, 2016, 798-826.

Pylyshyn, Zenon (ed.), *The Robot's Dilemma* (Norwood, NJ: Ablex, 1987).

Rabin, Matthew, "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica*, 68, 2000, 1281-1292.

Robbins, Lionel, *An Essay on the Nature and Significance of Economic Science* (London: Macmillan, Second Edition, 1935).

Rosenberg, Alex, *Economics: Mathematical Politics or Science of Diminishing Returns?* (Chicago: University of Chicago Press, 1992).

Ross, Don, *Economic Theory and Cognitive Science: Microexplanation* (Cambridge, MA: MIT Press, 2005).

Ross, Don, "Hayek's Speculative Psychology, the Neuroscience of Value Estimation, and the Basis of Normative Individualism," in L. Marsh(ed.), *Hayek in Mind: Hayek's Philosophical Psychology* (Bingley, UK: Emerald, 2011).

Ross, Don, "Economic Theory, Anti-economics and Political Ideology," in U. Mäki (ed.), *Handbook of the Philosophy of Science, Volume 13: Economics* (Amsterdam: Elsevier, 2012).

Ross, Don, *Philosophy of Economics* (London: Palgrave Macmillan, 2014).

Ross, Don, "Addiction is Socially Engineered Exploitation of Natural Biological Vulnerability," *Behavioral Brain Research*, 386, 2020, 1-8.

Ross, Don, "Neo-Samuelsonian Methodology, Normative Economics, and the Quantitative Intentional Stance," in B. Caldwell, J. Davis, U. Mäki and E. Sent (eds.), *Methodology and History of Economics* (New York: Routledge, 2023).

Ross, Don, and Stirling, Wynn, "Economics, Social Neuroscience, and Mindshaping," in J. Harbecke, J. and C. Herrmann-Pillath (eds.), *Social Neuroeconomics* (London: Routledge, 2021).

Ross, Don, and Townshend, Matthew, "Everyday Economics," in H. Kincaid & D. Ross (eds.), *A Modern Guide to Philosophy of Economics* (Cheltenham: Edward Elgar, 2021).

Routledge, Bryan R., and Zin, Stanley E., "Generalized Disappointment Aversion and Asset Prices," *Journal of Finance*, 65(4), 2010, 1303-1332.

Rubinstein, Ariel, and Salant, Yuval, "Eliciting Welfare Preferences from Behavioral Datasets," *Review of Economic Studies*, 79, 2012, 375-387.

Salant, Yuval, and Ariel Rubinstein, "(A,f): Choice with Frames," *Review of Economic Studies*, 75(4), 2008, 1287-1296.

Samphantharak, Krislert, and Townsend, Robert M., *Households as Corporate Firms: An Analysis of Household Finance Using Integrated Household Surveys and Corporate Financial Accounting* (New York: Cambridge University Press, 2010).

Savage, Leonard J., *The Foundations of Statistics* (New York: Wiley, Second Edition, 1954).

Savage, Leonard J., "Subjective Probability and Statistical Practice," and "Discussion," in G.A. Barnard and D.R. Cox (eds.), *The Foundations of Statistical Inference: A Discussion* (New York: Wiley, 1962).

Schelling, Thomas, "Economics, or the Art of Self-Management," *American Economic Review (Papers & Proceedings)*, 68(2), 1978, 290-294.

Schelling, Thomas, "The Intimate Contest for Self-Command," *Public Interest*, 60, 1980, 94-118.

Schelling, Thomas, "Self-Command in Practice, in Policy, and in a Theory of Rational Choice," *American Economic Review (Papers & Proceedings)*, 74(2), 1984, 1-11.

Schmidt, Ulrich, and Zank, Horst, "Risk Aversion in Cumulative Prospect Theory," *Management Science*, 54, 2008, 208–216.

Schwitzgebel, Eric, *Perplexities of Consciousness* (Cambridge, MA: MIT Press 2011).

Simon, Herbert, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69, 1955, 99-118.

Smith, Vernon L., *Rationality in Economics* (New York: Cambridge University Press, 2007).

Spivey, Michael, *Who You Are* (Cambridge, MA: MIT Press, 2020).

Starmer, Chris, "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature*, 38, June 2000, 332–382

Sugden, Robert, "The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences," *American Economic Review*, 94(4), 2004, 1014-1033.

Sugden, Robert, "Market Simulation and the Provision of Public Goods: A Non-Paternalistic Response to Anomalies in Environmental Evaluation," *Journal of Environmental Economics and Management*, 57, 2009, 87-103.

Sugden, Robert, *The Community of Advantage: A Behavioral Economist's Defence of the Market* (New York: Oxford University Press, 2018).

Sugden, Robert, "A Response to Six Comments on *The Community of Advantage*," *Journal of Economic Methodology*, 28(4), 2021, 419-430.

Sunstein, Cass R., "Voluntary Agreements," *Journal of Economic Methodology*, 28(4), 2021, 401-408.

Sunstein, Cass R., and Thaler, Richard H., "Libertarian Paternalism," *American Economic Review (Papers & Proceedings)*, 93, 2003a, 175-179.

Sunstein, Cass R., and Thaler, Richard H., "Libertarian Paternalism is Not an Oxymoron," *University of Chicago Law Review*, 70, 2003b, 1159-1202.

Teele, Dawn Langan, "Reflections on the Ethics of Field Experiments," in Teele, D. (ed.), *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* (New Haven, NJ: Yale University Press, 2014).

Tiberius, Valerie, *The Reflective Life* (Oxford: Oxford University Press, 2010).

Tiberius, Valerie, and Plakias, Alexandra, "Well-Being," in J. Doris and the Moral Psychology Research Group (eds.), *The Moral Psychology Handbook* (Oxford: Oxford University Press, 2010).

Townsend, Robert M., "Risk and Insurance in Village India," *Econometrica*, 62(3), May 1994, 539-591.

Tversky, Amos, and Kahneman, Daniel, "Advances in Prospect Theory: Cumulative Representations of Uncertainty," *Journal of Risk & Uncertainty*, 5, 1992, 297-323.

Wakker, Peter P., *Prospect Theory for Risk and Ambiguity* (New York: Cambridge University Press, 2010).

Wilcox, Nathaniel T., "Stochastic Models for Binary Discrete Choice under Risk: A Critical Primer and Econometric Comparison," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Wilcox, Nathaniel T., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice Under Risk," *Journal of Econometrics*, 162(1), 2011, 89-104.

Williams, Bernard, *Moral Luck* (Cambridge, UK: Cambridge University Press, 1981).

Williams, Bernard, *Philosophy as a Humanistic Discipline* (Princeton: Princeton University Press, 2006).

Wilson, Mark, *Wandering Significance* (New York: Oxford University Press, 2006).

Wolpin, Kenneth I., *The Limits of Inference Without Theory* (Cambride, MA: MIT Press, 2013).

Yarkoni, Tal, "The Generalizability Crisis," *Behavioral and Brain Sciences*, 45, e1, 2022. DOI: https://doi.org/10.1017/S0140525X20001685.

Zawidzki, Tadeusz, *Mindshaping* (Cambridge, MA: MIT Press, 2013).