

What You Don't Know May Hurt You: Preferences Over Mental And External States^{*}

Gonzalo Arrieta[†] Lukas Bolte[‡]

September 24, 2024

Abstract

The dominant approach to welfare, revealed preference, is restricted to settings where the individual knows their preferences have been fulfilled. We use a choosing-for-others framework to experimentally study welfare when what the individual believes differs from what is actually true. 42% of participants see welfare as independent of beliefs; 22% see welfare as exclusively determined by beliefs; and 29% care about both beliefs and reality. Furthermore, the average participant values accurate beliefs. While there is large heterogeneity, our results suggest most people support the idea that welfare goes beyond beliefs, which can inform media regulation, informational policies, and government communication.

JEL CLASSIFICATION CODES: C91, D01, D60, I31, I38

KEYWORDS: Welfare, paternalism, revealed preferences, utilitarianism, mental states, beliefs, information policy, experience machine.

1. INTRODUCTION

Welfare considerations are at the core of economic policy evaluations: a policy is good if it enhances welfare. The dominant approach to understanding what enhances welfare is to

*We are especially grateful to B. Douglas Bernheim for his guidance and encouragement and to Paul Milgrom and Alvin Roth for collaborating in the implementation of our experimental design. We also thank Sandro Ambuehl, Maxim Bakhtin, Ori Heffetz, David Huffman, Muriel Niederle, Kirby Nielsen, Peter Schwardmann, and seminar participants at Carnegie Mellon University, FAIR NHH, SBEERC, SITE Psychology, Stanford University, Tilburg University, and the University of Pittsburgh for their helpful comments. Halley Hirokawa, Isha Patel, and Xiang Qing Wang provided excellent research assistance. This study is covered under Stanford University's IRB Protocol 44866 and Carnegie Mellon University's IRB Protocol #IRB-STUDY2015_00000482. The study was registered on the AEA RCT registry under ID AEARCTR-0011851 with the title "Red or Blue Pill? A Positive Welfare Analysis."

[†]University of Zurich. E-mail: gonzalo.arrieta@econ.uzh.ch.

[‡]Carnegie Mellon University. E-mail: lukas.bolte@outlook.com.

defer to choice and make welfare assessments based on revealed preferences. According to this criterion, an alternative x , such as a policy, enhances the welfare of an individual relative to alternative y if the individual chooses x over y .

However, the reliance on choice data limits revealed preference as a welfare criterion to cases in which choice data exists. In particular, using choice data to assess individual welfare essentially restricts us to settings in which the individual *knows* that, say, alternative x is chosen; this is because an individual cannot consciously choose between alternative x and y without, in tandem, adjusting their beliefs about which alternative has been chosen.

This reliance on choice data poses real challenges for evaluating policies as “The parable of the oblivious altruist” (Bernheim and Taubinsky, 2018) illustrates:

A small town in Arkansas experiences massive flooding, leaving many families homeless. To provide financial assistance for the impacted families, the government raises taxes, including a \$100 levy on Norman. As a general matter, Norman thinks government spending is wasteful, but he is also an altruist and would gladly contribute \$100 to the fund if he knew about it. However, he never learns about the flood or the relief effort. Does the government’s policy make him better off or worse off?

It is impossible for Norman’s choices to inform the planner in this setting: if Norman chooses the government policy, he would also learn that the government has such policy, contrary to the thought experiment’s premise. Hence, revealed preferences cannot be used as a welfare criterion. How, then, can we assess what the welfare effect of the policy is when Norman does not learn about it?

More generally, government agencies often hold more information than the population they serve. Benevolent policy-makers use this information to satisfy the preferences of their constituents. In this context, understanding the welfare effects of information provision that go beyond its instrumental value, as well as the welfare consequences of satisfying constituents’ preferences in situations in which they remain oblivious, is of first-order importance to the design of optimal policies. Moreover, such questions arise not only in the context of policy-making, where individuals may be unaware of the chosen policy, but also in individual decision-making, where individuals may misunderstand the consequences of their actions. For example, does the individual’s welfare depend on what they believe the consequences of their actions are, on the actual consequences, or some combination? The traditional revealed preferences criterion is, again, not applicable; how, then, can we learn about the welfare consequences of such policies?

This paper experimentally studies the welfare consequences of different alternatives when

the individual remains unaware of which alternative is chosen. To overcome the aforementioned choice data limitation, we resort to a choosing-for-others framework: an altruistic surrogate chooses between alternatives affecting another individual while keeping fixed that individual's beliefs about the chosen alternative. The surrogate's choices then reveal the effect on the individual's welfare as perceived by the surrogate, which teaches us how people think about welfare in these situations. Furthermore, to the extent that surrogates choose for others as they would choose for themselves, we can apply the standard revealed preferences paradigm to interpret their choices as reflective of their welfare.

We refer to our approach as the “altruistically revealed preference paradigm.” The challenge is in creating a situation in which to apply the paradigm. For the surrogates' choices to reveal how they think about others' welfare in these situations, three requirements must be satisfied:

Welfare relevance: The alternatives x and y need to be perceived by the surrogate to matter to the individual, i.e., the individual would choose x over y in a standard choice problem.

Pure altruism: The surrogate only cares about the chosen alternative in so far it affects the individual's welfare.

Obliviousness: The individual remains oblivious about the chosen alternative unless we inform them; otherwise, we are back in the standard paradigm where preference satisfaction and beliefs about preference satisfaction move in tandem.

To illustrate, a choice between eating an apple or an orange may satisfy *welfare relevance*, but it is impossible to have the individual eat the fruit while keeping them oblivious about what they are eating, which violates *obliviousness*. A choice between which charity to donate to can satisfy *welfare relevance* and *obliviousness*. However, it fails to satisfy *pure altruism*: likely, the surrogate making the choice would directly care about which charity receives the donation.

We design a situation that satisfies the three requirements. One of our experimental participants (henceforth “Receiver”) takes the role of the individual on whose behalf choices are made and whose welfare is assessed. We purchased four books by two Nobel laureates in economics. Two books have “original” notes handwritten by the authors and dedicated to the Receiver, and two have “fake” notes, which are copies of the original ones. We select the Receiver such that they have a strict preference over receiving the original notes, which we convey to the main group of participants (henceforth simply “participants”), to satisfy *welfare relevance*. We satisfy *obliviousness* by making the original and fake notes indistinguishable from each other. Lastly, we satisfy *pure altruism*: it is unlikely that participants

care about whether a stranger receives books with original or fake notes for reasons other than the stranger’s welfare (we do not find evidence against this assumption in our experiment, as we discuss in Section 3).

Our altruistically revealed preference paradigm consists of asking surrogate participants to trade off a monetary bonus given to the Receiver, and the Receiver getting the books with the original notes over those with the fake notes. The bonus amount is a surprise to the Receiver to minimize concerns that they use it to deduce which books they got (i.e., to maintain *obliviousness*). Our experimental design allows us to interpret the bonus amount that leaves participants indifferent between giving the original and fake notes as a measure of the change in the Receiver’s welfare. As a benchmark, we also elicit the welfare effect when the Receiver does learn which notes they get.

Our main findings are these. Consider participants who strictly prefer giving the original notes in our benchmark case, when the Receiver learns about it. When, instead, the Receiver does not learn, the welfare impact of getting the original notes is zero for around 22% of these participants, positive but less than the impact when the Receiver learns for around 29%, and the same as the impact when the Receiver learns for 42%.¹² To illustrate, our results suggest that, in “The parable of the oblivious altruist,” 42% of individuals would judge Norman to be better off by the relief effort, while the remaining individuals would see a benefit from Norman being informed about the use of their taxes. We then investigate how the welfare assessments depend on what the Receiver believes—they could think the notes are most likely original or most likely fake. By exogenously varying the Receiver’s belief, we show that some participants act as if having accurate beliefs is important for welfare; that is, the welfare gain from having original notes is greater if the Receiver believes they have original notes.

We then ask to what extent welfare notions depend on the choice environment. To this end, we elicit two related welfare assessments in other contexts: the impact of a desired policy being implemented without the individual knowing about it (specifically, “The parable of the oblivious altruist” by Bernheim and Taubinsky, 2018) and the impact of receiving an original art piece over an indistinguishable fake (see details in Section 3.4). We find that the incentivized welfare measures are correlated with answers to the impact of receiving an original art piece but not with answers to “The parable of the oblivious altruist,” presenting mixed evidence regarding the context-dependence of welfare notions. We also test whether participants exhibit different welfare notions when choosing for the self rather

¹As previewed by Adams (1979).

²Less than 1% see a welfare loss in giving the original books, and around 5% see a larger welfare gain in giving the original books when the Receiver does not learn compared to when they do.

than for others, which helps us in applying the revealed preferences paradigm to interpret the surrogate’s choices as reflective of their own welfare, as we explain in Section 2.4. To this end, we ask participants if they would plug into the Experience Machine (Nozick, 1974), assessing their own (hypothetical) welfare. We find a correlation between our incentivized choosing-for-others welfare assessment and the answers to the Experience Machine, suggesting welfare notions for the self are related to those for others.

Our study contributes to a millennia-old question: “What is welfare?” Our focus is on two broad classes of theories conceptualizing welfare.³ The first is welfare hedonism, which proposes that “well-being consists solely in the presence of pleasure and the absence of pain” (e.g., Bentham, 1789; Mill, 1879). A common variant of this notion of welfare is mental statism, which postulates that well-being is exclusively a reflection of mental states. The second is preference (or desire) theory, which postulates that well-being consists of having one’s preferences satisfied; it is the dominant approach to welfare in economics. In its simplest form, this welfare notion asks whether the world is as the individual would like it to be and ignores what the individual believes about the world. A more general form of preference theory allows for the possibility that the individual’s preferences encompass their own mental states. In this sense, it allows for both mental states and states outside the individual’s awareness (which we call external states) to affect welfare. By testing the role of mental states in surrogate participants’ welfare assessments, our experiment allows us to shed light on which welfare notions participants adhere to and, in particular, the roles of mental and external states.

We contribute to the experimental literature that discusses the measurement of well-being. Although there is a growing interest in resorting to subjective self-reported measures (Benjamin et al., 2012, 2014, 2023), in economics, the gold standard of welfare measures is the revealed preference paradigm. This approach comes with limitations. For instance, Bernheim et al. (2024) point out that welfare may not be recoverable from standard choice data when the decision-maker cares about the experience of choosing. We explore a different way in which welfare is not recoverable from standard choice data: when mental states are not a degenerate distribution on the associated external state. While driven by different features, both limitations stem from the social planner being incapable of presenting individuals with the alternatives the planner themselves is deciding between (i.e., the planner is incapable of gathering the relevant choice data).

In experimental psychology and philosophy, there is a specific focus on the elicitation of

³A third class is “Objective theories” that we do not directly connect to, by which welfare is maximized when some objective criteria are met, irrespective of whether the individual prefers them (Aristotle, 2011; Sen, 1985).

welfare notions. For example, a predominant thought experiment is the Experience Machine (Nozick, 1974), where researchers examine whether participants plug into a hypothetical machine that can simulate mental states. Ignoring many subtleties, mental statists should plug themselves in, while those valuing “real” experiences should not.⁴ We contribute to this experimental literature by constructing an incentivized measure of a participant’s welfare notion and showing that it correlates with the hypothetical Experience Machine question.

Our paper also contributes to a growing literature on positive welfare economics, especially to the branch which aims to determine how people evaluate the welfare of other individuals and groups, mainly from a paternalistic point of view (e.g., Uhl, 2011; Ambuehl et al., 2021, 2023; Bartling et al., 2023). In particular, this paper improves our understanding of how individuals think about others’ welfare, focusing on the welfare relevance of mental and external states.⁵ The paternalism literature finds that individuals intervene in others’ decision problems as if they seek to align others’ choices with their own aspirations (Ambuehl et al., 2021). This finding supports an interpretation of our altruistically revealed preference paradigm as reflective of the traditional revealed preference paradigm—how individuals think of their own welfare—applied to an extended domain in which individuals are oblivious about whether their preferences have been satisfied. To the extent that surrogate participants choose as if they were the Receiver, their choices directly speak to their own preferences over the mental and external states.⁶

Lastly, a large literature documents that individuals may be motivated to hold particular beliefs (i.e., particular mental states) that may be incompatible with Bayesian reasoning given available evidence (i.e., the external state); see Bénabou (2015) for a review. For instance, individuals may be motivated to think highly of themselves as a source of ego utility (Köszegi, 2006), or about a state of the world to derive utility through anticipation (Caplin and Leahy, 2001; Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005). Some of these models consider these utilities as welfare-relevant, while others are silent on whether the desire to manage one’s mental state is welfare-enhancing or a mistake. Our results suggest people might have heterogeneous responses to this question. On the one hand, for a large

⁴The experimental work has progressively refined experiments that try to account for potential biases when asking participants what they would do if facing the machine, with inconclusive evidence about the role of mental states (Baber, 2008; De Brigard, 2010; Smith, 2011; Weijers, 2013; Rowland, 2017; Hindriks and Douven, 2018).

⁵We run a survey on the Social Science Predictions Platform (DellaVigna et al. 2019; Public Study ID sspp-2023-0032-v1 at www.socialedgeprediction.org) to capture our current understanding about the role of external and mental states for welfare. Predictions, while heterogeneous themselves, broadly capture the heterogeneity in welfare notions that we find in our data. Predictors underestimate, however, the degree to which participants behave as if external states drive welfare.

⁶Using preference theory as the welfare notion—in particular in the traditional revealed preference paradigm—the individual’s preferences, in turn, reflect their own welfare.

portion of participants, external states are not all that drives welfare; this opens the door for beliefs in general and biased ones in particular to have direct welfare implications. On the other hand, more than a third of participants behave as if external states are all that matter, suggesting beliefs do not have direct welfare implications for them.⁷

The rest of the paper proceeds as follows. Section 2 presents a conceptual framework that highlights the limitations of the traditional revealed preference paradigm as an underpinning of welfare economics and formally defines which type of welfare assessments we study. We present our experimental design in Section 3. Section 4 gives the results, and Section 5 concludes.

2. CONCEPTUAL FRAMEWORK

Revealed preference as a welfare paradigm is only applicable when there are, in fact, choice problems that reveal preferences. However, some problems are impossible to implement: one cannot elicit someone’s choice between alternatives, *while holding their belief about which alternative is chosen fixed*.⁸ In this section, we consider a simple framework of choice problems where alternatives consist of pairs of “mental states” (what the person believes) and “external states” (what is actually true). We first use this framework to highlight the limits of the traditional revealed preference paradigm as a foundation of welfare and show how an altruistically revealed preference paradigm can overcome them, and under which conditions. We then illustrate how preferences over pairs of mental and external states map into welfare notions, which we use to interpret our empirical findings.

2.1. *The revealed preference paradigm and its limitations as a welfare criterion*

We assess the welfare of an individual. Let X , with typical element x , be a set of external states and $\mu \in \Delta(X)$ a distribution over X , which we refer to as a mental state (of the individual). Crucially, the individual does not necessarily observe the external state x ; their mental state, μ , represents their beliefs over what the external state is. While in many settings, x and μ move in tandem, our interest lies in those where they do not. Specifically, we want to know how the bundle (x, μ) affects the individual’s welfare, which we represent via a function $\mathcal{W} : X \times \Delta(X) \rightarrow \mathbb{R}$.

⁷Our results also shed light on the welfare consequences of mental illness. Two common mental illnesses—anxiety and depression—have been linked to negative economic consequences, like reductions in employment and poverty (Ridley et al., 2020; Bhat et al., 2022). Our results speak to the direct welfare consequences of such extreme negative mental states beyond the economic repercussions.

⁸For example, in the “The parable of the oblivious altruist” presented in the introduction, we cannot have Norman choose between the government offering disaster relief or not while holding Norman’s belief about whether there is disaster relief fixed.

Under the premise that individuals know what is best for them, and that these judgments guide their choices, the revealed preference paradigm underlying much of welfare economics estimates \mathcal{W} by giving the individual choice problems. A typical problem involves choosing from a set $A' \equiv \{(x, \delta_x) \in A \times \Delta(A)\}$, for some $A \subseteq X$, where δ_x places all weight on x ; that is, the individual chooses their preferred external state (e.g., a good or a policy), and whatever they choose, they must believe.⁹ We can use these choice problems to estimate \mathcal{W} —but only for a restricted domain. To illustrate, suppose $X = \{x, y\}$ and consider choices over $\{(x, \mu), (y, \mu)\}$. Giving this choice problem to the individual is infeasible: when the individual chooses (x, μ) , they know that the external state is x , and so it must be that $\mu = \delta_x$; but then choosing (y, μ) is not possible because, by the same token, upon choosing (y, μ) they know that the external state is y , and so it must be that $\mu = \delta_y$.

We overcome this missing-data problem using a choosing-for-others framework and what we call the altruistically revealed preference paradigm. Suppose a purely altruistic surrogate chooses on behalf of the individual. In particular, assume that they, too, maximize \mathcal{W} as they perceive it. Then, we can study the aforementioned choice problems; that is, the surrogate can choose (for the individual) from $\{(x, \mu), (y, \mu)\}$.

2.2. *The altruistically revealed preference paradigm and its extended domain of application*

Consider an individual (henceforth, surrogate) choosing on behalf of another individual (henceforth, Receiver). The welfare of the Receiver is given by function \mathcal{W} ; the surrogate chooses an external and mental state bundle (x, μ) to maximize

$$U(x, \mu) = u(x, \mu) + \alpha \hat{\mathcal{W}}(x, \mu),$$

where u describes the surrogate's selfish preferences over the external state and mental state (of the Receiver), and $\alpha > 0$ implies the surrogate also values the external and mental state via their impact on the Receiver's welfare, which the surrogate perceives as $\hat{\mathcal{W}}$. In taking such a consequentialist approach, we assume the individual's decisions are not guided by Kantian categorical imperatives (e.g., “do not lie”—an assumption that we briefly discuss in Section 3.1).¹⁰

To apply the altruistically revealed preference paradigm to an extended domain—in particular, one that allows choices beyond (x, δ_x) -bundles—the following three requirements need to be satisfied:

⁹To introduce uncertainty over external states, the object of choice can be defined to be a lottery.

¹⁰In fact, the more general formulation $U(x, \mu) = u(x, \mu) + F(\hat{\mathcal{W}}(x, \mu))$, where F is strictly increasing, suffices. Thus, this formulation does allow for preferences such as “warm-glow” and “ethical obligation” as long as they operate via the Receiver's welfare.

Welfare relevance: $\hat{\mathcal{W}}$ is not constant.

Pure altruism: u is constant.

Obliviousness: Some bundles (x, μ) , where $\mu \neq \delta_x$, can be chosen.

In a choice problem that satisfies *welfare relevance* and *pure altruism*, the bundle (x, μ) improves the Receiver's welfare—as perceived by the surrogate—relative to the bundle (y, μ') if the surrogate chooses (x, μ) over (y, μ') . *Obliviousness* is an empirical requirement for the paradigm to be applied to an extended domain, in particular, one that allows for choice problems between bundles where external and mental states “do not match.” The altruistically revealed preference paradigm can then be used to study preferences over (non-degenerate) mental and external state bundles of others. Importantly, note that whether the surrogate and the Receiver agree on how (x, μ) affects the Receiver's welfare is irrelevant. As long as the three requirements hold, the surrogate's choices reveal how (x, μ) enter $\hat{\mathcal{W}}$, which is what we are interested in.

2.3. Preferences over mental and external states

We consider the following types of preferences over mental and external state bundles. Mental statism implies that all welfare gains come from the change in mental state. Conversely, it may be that only external states matter. Lastly, it could be that both mental and external states determine preferences.

Mental states are all that matter— $\mathcal{W}(x, \mu)$ is independent of x . This functional form corresponds to what we call “pure” mental statism. The intuition is simple: what you don't know can't hurt you.

External states are all that matter— $\mathcal{W}(x, \mu)$ is independent of μ . This functional form corresponds to a simple form of preference theory, which we call pure external statism, and, in some sense, is the opposite of mental statism. Here, the individual's welfare is affected by what is actually true and not what the individual believes to be true. In this case, what you don't know *can* hurt you.

Mental and external states matter— $\mathcal{W}(x, \mu)$ depends on both x and μ . Mental and external states could jointly affect welfare in many ways. Here are two. First, it may be that mental and external states matter independently— $\mathcal{W}(x, \mu) = ES(x) + MS(\mu)$, for some functions ES and MS . Second, external and mental states may be complements— \mathcal{W}

has increasing differences.¹¹ One interpretation is that inaccurate beliefs (i.e., believing the external state is x when it is actually y) decrease welfare. According to such \mathcal{W} , it is inherently bad *to live a lie*. In this case, choosing y (the “bad” external state) can actually improve welfare if μ is low enough.¹²

Our experiment, described in the next section, allows us to study the shape of \mathcal{W} . In particular, we go beyond the usual elicitation of $\mathcal{W}(x, \delta_x) - \mathcal{W}(y, \delta_y)$ and use the altruistically revealed preference paradigm to assess $\mathcal{W}(x, \mu) - \mathcal{W}(y, \mu)$ and individuals’ perception thereof, $\hat{\mathcal{W}}(x, \mu) - \hat{\mathcal{W}}(y, \mu)$.

2.4. From welfare perceptions to welfare

Satisfaction of the three requirements detailed in section 2.2 allows us to interpret the surrogate’s choices as revealing their preferences over mental and external states when choosing for others. These preferences teach us how people think about welfare, as interpreted in section 2.3, which is the primary goal of this paper. Can we do more? In particular, how could we go from the surrogate’s adherence to welfare notions when choosing for others to saying something about what welfare really is?

To say anything about true welfare, we need a welfare criterion. The standard welfare criterion used in economics is the revealed preferences paradigm, according to which an alternative x enhances the welfare of an individual relative to alternative y if the individual chooses x over y . To apply it, we need the surrogate’s preferences—over x and y for themselves—to be revealed. To interpret the surrogate’s choices for others as revealing their preferences for themselves, one additional requirement must be satisfied:

Ideals projection: Surrogates choose for others as they would choose for themselves.

In Section 3.4, we discuss survey questions that test for *ideals projection*, and in subsection 4.3 we show the results. If *ideals projection* is satisfied (Ambuehl et al., 2021 provide evidence supporting it), we can then apply the revealed preferences paradigm in the standard way to make normative statements interpreting the surrogate’s choices as reflective of their welfare. For example, it would allow us to say that only external states matter for the welfare of surrogates whose preferences are such that external states are all that matter when choosing for others.

¹¹Suppose $X = \{x, y\}$ and consider choices over $\{(x, \mu), (y, \mu)\}$; order external states by the individual’s preferences over $\{(x, \delta_x), (y, \delta_y)\}$, and mental states by the likelihood assigned to the higher-ordered external state. Increasing differences then means: For $x > y$ and $\mu' > \mu$, $\mathcal{W}(x, \mu') + \mathcal{W}(y, \mu) \geq \mathcal{W}(x, \mu) + \mathcal{W}(y, \mu')$.

¹²The welfare impact of a mismatch between external and mental state could be asymmetric. For example, overly optimistic beliefs could be more or less harmful than pessimistic beliefs relative to accurate ones.

3. EXPERIMENTAL DESIGN

We conduct an incentivized online experiment to measure how individuals perceive external and mental states to determine welfare. All but one participant make surrogate choices about an external state for a single other participant—the Receiver. The controlled nature of the experiment allows us to create an external state that affects the Receiver (*welfare relevance*), while all other participants only care about it via their altruism toward the Receiver (*pure altruism*), and that can be changed without the Receiver knowing (*obliviousness*), as we detail in Section 3.1. Crucially, participants make choices when the Receiver does not learn the external state—i.e., the Receiver’s mental state is fixed, and only the external state changes (the *NotLearns* case). As a benchmark, participants also make choices when the Receiver does learn about the external state, and so mental and external states move in tandem (the *Learns* case—a more standard elicitation). Section 3.2 provides details on these elicitations.

We also vary the Receiver’s mental state to study how the welfare effect of changing the external state depends on the mental state. In Section 3.3, we describe two additional treatments: one in which the Receiver’s mental state is relatively optimistic about the external state and one in which they are relatively pessimistic.

Lastly, we ask participants additional questions related to welfare notions, allowing us to assess evidence for *ideals projection*, i.e., whether individuals’ perceptions of welfare notions for others correlate with welfare notions for themselves, and study context-dependence of welfare notions more generally; see Section 3.4.

3.1. *The external state*

The key challenge is to construct an external state that satisfies the three requirements detailed in Section 2.2: *welfare relevance*, *pure altruism* and *obliviousness*. We overcome this challenge as follows. We purchased four books by two Nobel laureates in economics: two copies of “Discovering Prices” by Paul Milgrom and two copies of “Who Gets What and Why” by Alvin Roth.¹³ Each book has a handwritten note dedicated to the Receiver.¹⁴ For one copy of each of the books, the note was handwritten by the author, and we refer to these two copies as the “books with the original notes.” For the other two copies, the handwritten notes were copied from the original note, and we refer to these two copies as the “books with

¹³We use both “Discovering Prices” and “Who Gets What and Why,” instead of just one of the two, to make the stakes higher for the Receiver.

¹⁴The participant in the role of the Receiver is called Alex, and the notes say, “To Alex: I hope you enjoy reading about auctions! Paul Milgrom” and “For Alex: I hope you enjoy reading about market design. Alvin E. Roth,” respectively.

the fake notes.” The books with the fake notes are indistinguishable from the books with the original notes, which we ensured by having Paul Milgrom and Alvin Roth themselves copy their respective notes; neither the participants nor the Receiver know how the fake notes came to be.¹⁵ The Receiver will receive two books for sure, which may be either the books with the original or the fake notes.

The external state is defined by whether the notes are original or not.¹⁶ Below, we discuss how our choice of external state satisfies the first three requirements laid out in Section 2.2, which allow us to implement the altruistically revealed preference paradigm to study the perceived welfare effects of external and mental states.

Welfare relevance. We ensure *welfare relevance* by suggesting to the participants that the Receiver prefers the original notes (“[The Receiver] has already read some of their work and told us he has great admiration for them.” See the experimental instructions in Appendix B).¹⁷

Pure altruism. Our design choices minimize concerns that any individual other than the Receiver has strict preferences over the external state that do not operate via the Receiver’s welfare. For example, participants could have been concerned with the costs to the experimenters of creating the book with the notes, and hence not wanting them to go to waste; we addressed this by producing both sets of books. They could also have been concerned with what would happen to the books not gifted to the Receiver; we minimize this concern by returning those books to the authors, who can always create new notes. Other non-altruistic considerations determining participants’ choices, such as those derived from Kantian ethics, would lead us to overestimate external statism. While rigorously estimating the extent of such preferences is an interesting and important direction for future work, in this study, we can assess the relevance of this confounder using participants’ open-ended responses (see the end of Section 3.2 for details); we do not find evidence for such non-altruistic considerations. Furthermore,

¹⁵The fake notes are fake in the sense that the authors were not tasked to write another set of notes (as they were when creating the original notes) but rather to copy their old notes. That is, the fake notes are a copy, by the authors, of the original notes. This makes them fake in the same way that a copy of the Mona Lisa is not the original Mona Lisa, even if copied by Da Vinci himself.

¹⁶The Receiver also receives a surprise monetary bonus that is determined by the participants’ choices, which is also part of the Receiver’s external state and mental state. However, for simplicity in exposition, we define the Receiver’s external and mental state as solely about the notes.

¹⁷In conveying *welfare relevance* to the participants, we were careful not to suggest whether the Receiver preferences were driven by external or mental states, which, as discussed in Section 2 is difficult to interpret anyway. Moreover, as explained in Section 2.2, note that whether the participants would also want the book with the original notes if they were in the Receiver’s shoes is irrelevant. As long as the participants believe the Receiver prefers the book with the original notes, *welfare relevance* holds, which is what we need.

considerations such as warm-glow or ethical obligations that operate via the Receiver's welfare do not invalidate *pure altruism*.

Obliviousness. Since the original notes are indistinguishable from the fake ones, *obliviousness* holds (i.e., external and mental states can be independently manipulated).¹⁸

3.2. Preference elicitation

We measure the welfare effect of different mental and external state bundles with money-to-the-Receiver as the numéraire. Specifically, the Receiver gets a surprise bonus determined by the participants' choices. Our primary outcome is the size of the surprise bonus added to the books with the fake notes such that the participant is indifferent between giving the Receiver the books with the original and the books with the fake notes.¹⁹

We use money-to-the-Receiver instead of money-to-the-participant for three reasons. First, with money-to-the-Receiver as the numéraire, the elicited welfare effect is invariant to varying levels of participants' altruism, lowering noise. Second, altruism levels in an online experiment may be low, and so few participants may be willing to reduce their own payoffs (see Section 3.5 for details about the participant population). Third, with their own money on the line, participants may bias their choices towards mental statism as an excuse not to reduce their monetary payoff; see Exley (2016) for an example of selfishly malleable preferences. On the flip side, a concern with using the Receiver's monetary payoff as the numéraire is that the Receiver could make an inference about the external state from the payment they receive, leading to a violation of *obliviousness*. To minimize this concern, we frame the monetary bonus the Receiver receives as a surprise bonus that they always receive, but for an unknown amount. This makes it implausible that the Receiver can update about the external state from observing their bonus.

We elicit the size of the surprise bonus that leaves a participant indifferent in a modified version of a multiple-price list. In particular, our procedure consists of three steps:

1. We ask participants whether they would rather give the Receiver the books with original notes, the ones with fake notes, or whether they are indifferent.
2. Participants who do not select indifference choose between the notes again, but, this

¹⁸Making the original and fake notes indistinguishable could question the validity of the Receiver's preference over them, which could translate into a violation of *welfare relevance*. We find that about 1 in 5 participants violate *welfare relevance*—i.e., are indifferent between giving the Receiver one or the other even when the Receiver learns about which notes they get.

¹⁹The Receiver's welfare as perceived by the participant is thus given by $\hat{W}(x, \mu) + v(m)$ —or $v(m, \delta_m)$ —where x is the type of notes the Receiver gets, μ the Receiver's belief about the type of notes, and $v(m)$ the welfare effect of having (and knowing about having) monetary amount m . Participants choose among bundles (x, μ, m) .

time, the notes they did not choose in the first step come with a \$1 monetary payment to the Receiver.

3. Participants who chose the books with the original notes in both previous steps are presented with a multiple-price list where the monetary payment to the Receiver when given the books with the fake notes spans \$2, \$3, \$5, \$7, \$10, \$15, \$25, \$45, \$70, \$100, \$140, \$200.

Each participant goes through this three-step procedure for two cases, presented in random order: in *Learns*, we tell the Receiver whether the notes in the book they get are original or fake; in *NotLearns*, we do not tell them.

While we could have used a single multiple-price list, the three-step procedure seems easier for participants to understand and, hence, may produce less noisy responses. Although only participants who choose the books with the original notes in steps 1 and 2 face the multiple-price list, participants know that we use their responses to determine their preferences and ultimately implement the multiple-price list for all participants, regardless of whether they reach step 3 or not. This makes the procedures incentive-compatible. Furthermore, prefacing the multiple-price list with two relatively simple questions allows us to identify participants who are indifferent, those who prefer to give the original notes and those who, for some reason, prefer to give the fake notes. With this typification, we can already answer which share of participants sees welfare effects from changing the external state while holding the mental state fixed—i.e., are not pure mental statists (see Section 2.3).

The monetary value elicited in *NotLearns* allows us to further quantify the welfare gain from changing the external state while holding the Receiver's mental state fixed. The monetary value elicited in *Learns* serves two purposes: first, it allows us to identify participants who do not value giving the Receiver their preferred external state even when the Receiver learns about it—for example, because of a lack of altruism, confusion, or because these participants do not agree with the Receiver's preferences. We exclude such participants because we cannot study welfare notions if there is no welfare change to decompose. Second, it serves as a benchmark for the *NotLearns* outcome: we can ask how much of the welfare gain from changing both the external and mental state comes from changing the external state only.

After participants confirm their choices, we elicit open-ended responses to a question about why they responded the same or differently across cases (i.e., *Learns* and *NotLearns*). We then use their responses to assess data quality, for example, by flagging those who exhibit an evident misunderstanding of instructions and gauging participants' reasoning.

3.3. Varying the Receiver's initial mental state

In our baseline treatment, in the *NotLearns* case, the Receiver has some belief about whether they received the books with the original or fake notes. Varying this belief—i.e., their mental state—in a controlled manner allows us to test whether participants' welfare assessments of changing the external state depend on the (fixed) mental state. For instance, it allows us to test whether participants assign a welfare gain to external and mental states coinciding.²⁰

Therefore, we create two treatments that exogenously vary the Receiver's belief about whether they received the books with original or fake notes. In *HighMS*, we tell participants that the Receiver believes they likely got the original notes; in *LowMS*, we tell participants that the Receiver believes they likely got the fake notes.²¹

3.4. Varying the choice environment

We also want to study the dependence of welfare notions on choice domains. To do so, we ask two questions similar to our incentivized welfare assessments but in more naturalistic settings. The first question directly relates to policy preferences by asking participants whether the policy described in “The parable of the oblivious altruist” by Bernheim and Taubinsky (2018)—presented in the introduction—leaves the protagonist better or worse off. The second question takes advantage of a real scenario where 1000 individuals received a drawing by Andy Warhol, knowing only one got the original copy and the rest had indistinguishable fakes; we ask participants whether the person who has the original drawing is better off getting the original one instead of a fake, even if they—and everyone else—will never know which one they have.²²

To further assess the context dependence of welfare notions, we assess the evidence for *ideals projection*—i.e., do participants choose for others as they would for themselves? If so, then one could apply the standard revealed preferences paradigm as a welfare criterion to interpret the adherence to welfare notions for others as reflective of the participants' welfare.

To assess the relationship between welfare notions for the self and others, we present

²⁰Formally, we measure $\mathcal{W}(x, \mu) - \mathcal{W}(y, \mu)$ as a function of μ .

²¹Specifically, in *HighMS*, we say, “[the Receiver] knows that if we don't tell [them] which books [they] got, there is at least a 75% chance that they are the ones with the original notes” and in *LowMS*, we say, “[the Receiver] knows that if we don't tell [them] which books [they] got, there is at least a 75% chance that they are the ones with the fake notes.” We randomly select between this and the baseline treatment for implementation before approaching the Receiver. We truthfully inform the Receiver that with, e.g., 75% chance, they will have gotten the books with the original notes if we do not reveal it, and with complementary probability, which books they will have gotten will have been determined in some other way.

²²See www.moforgesies.org (accessed 2023/08/23) for a detailed description by the organizers of this project.

participants with the Experience Machine, a canonical thought experiment long discussed in the psychology and philosophy literature on welfare notions and the role of mental states (Nozick, 1974; see the instructions in Appendix Section B for the exact wording of this and the other questions). Evidence that participants choose for others as they would choose for themselves (as Ambuehl et al., 2021 show) allows us to provide evidence for *ideals projection*. Unlike our experimental setup, the Experience Machine studies the welfare impact of improving the mental state while worsening the external state. Nevertheless, answers to it can still be related to our incentivized experimental measure: for example, those subscribing to mental statism should plug into the machine.

3.5. Implementation and recruitment details

We recruited 1477 participants on Prolific, an online platform frequently used for research studies, and randomized them into one of our three treatments. We recruited all participants on August 23, 2023. Participants had to be located in the USA and have a minimum of 100 prior submissions, with a perfect approval rate. We implemented the experiment using the oTree platform (Chen et al., 2016). Each participant received a \$3 completion payment, and the median completion time was around 16.2 minutes. We pre-registered this study in the AEA RCT Registry (AEARCTR-0011851).

We recruited the participant taking the role of Receiver on August 22, 2023. One of the authors had a prior private acquaintance with the participant, and we approached them based on their characteristics, which matched the description given to the other participants.

Participants received extensive instructions and needed to correctly answer understanding questions before proceeding to the main parts of our study. See Section B in the Appendix for full experimental instructions.

4. RESULTS

In Section 4.1, we show the distribution of welfare assessments for the *NotLearns* and the *Learns* cases; we find strong support for a welfare effect of changing the external state even when the mental state is fixed. In Section 4.2, we discuss the results from two additional treatments, *HighMS* and *LowMS*, that vary the Receiver’s fixed mental state; we find evidence of a preference to match mental and external states. Lastly, in Section 4.3, we correlate the incentivized welfare assessments with welfare-related questions, shedding light on the stability and external validity of our results.

4.1. Welfare assessments

Consider *Baseline*—i.e., Table 1’s top row. Denote the welfare assessments by W_L and W_{NL} in the *Learns* and *NotLearns* case, respectively (i.e., the dollar amounts added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes together with the money, and them getting the original notes; we omit hats in our notation, although these welfare effects are as perceived by the participants).²³

	$W_L < 0$			$W_L = 0$			$W_L > 0$				
	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} < W_L$	$W_L = W_{NL}$	$W_{NL} > W_L$
<i>Baseline</i> $N = 492$	0.20%	0.20%	0.20%	0.41%	14.23%	3.05%	0.81%	18.29%	23.58%	33.94%	5.08%
<i>LowMS</i> $N = 497$	0.80%	0.00%	1.01%	0.80%	17.51%	2.21%	3.02%	16.70%	20.12%	32.19%	5.63%
<i>HighMS</i> $N = 488$	0.20%	0.00%	0.41%	0.20%	12.50%	3.48%	2.46%	15.16%	21.72%	35.86%	7.99%

Table 1: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively.

We first focus on the 81.71% of participants who assign a strictly positive welfare gain to the Receiver getting the books with the original notes in the *Learns* case (the $W_L > 0$ column).²⁴ We find that 18.29% of our participants do not see a welfare gain in the *NotLearns* case (while seeing one in the *Learns* case), and 33.94% see the same (positive) welfare gains in both cases. Assuming external and mental states affect welfare independently (i.e., $\mathcal{W}(x, \mu)$ can be written as $\mathcal{W}(x, \mu) = ES(x) + MS(\mu)$), the former may be interpreted as pure mental statist, where $\mathcal{W}(x, \mu) = MS(\mu)$, and the latter as pure external statist, where $\mathcal{W}(x, \mu) = ES(x)$.

Beyond pure mental statist and pure external statist, 23.58% of participants assign strictly positive welfare gains in both cases but a lower amount in *NotLearns*. For this group, both mental and external states matter for welfare.

We summarize these results below.

Result 1. *The modal participant adheres to pure external statism, by which the welfare gain from getting the books with the original instead of the fake notes is independent of whether the Receiver learns which notes they are getting.*

²³As noted at the end of Section 3.2, we assess whether W_L and W_{NL} are driven by non-altruistic considerations by eliciting, in open-ended questions, the reasons behind participants’ responses across *Learns* and *NotLearns*. We find no evidence of non-altruistic considerations, supporting *pure altruism*.

²⁴We classify as $W = 0$ participants for whom the bonus that leaves them indifferent is less than \$1, which is the smallest bonus amount that we elicit.

Result 2. *We find heterogeneity in welfare notions: beyond pure external statism, smaller groups of participants adhere to (i) pure mental statism, by which there is no welfare gain unless the Receiver learns about the books, and (ii) a mix between external and mental statism, by which they report a partial reduction in welfare gains when the Receiver does not learn about the books.*

There are two smaller groups of participants with $W_L > 0$. The larger of these two groups, with 5.08% of participants, sees welfare gains in giving the books with original notes that are larger when the Receiver does not learn about it (*NotLearns*) than when they do (*Learns*). We interpret these participants as seeing welfare loss in an inaccurately optimistic mental state. As a result, participants value giving the Receiver's preferred external state more when the Receiver does not learn about it compared to when they do. The smaller group, with 0.81% of participants, sees welfare gains in giving the Receiver the books with fake notes when they do not learn about it. The size of the group exhibiting this unexpected behavior is negligible, and it may be interpreted as noise.

The remaining participants, those with $W_L \leq 0$, cannot be matched to a welfare notion. Most of them, 78.89%, do not care about the external state (i.e., they assign a zero welfare effect to the Receiver getting the books with the original notes in the *Learns* case—the $W_L = 0$ column). Essentially, they do not see a difference between an original and an indistinguishable fake. Most of these participants also see no welfare gain in the *NotLearns* case.

Noise, inherent in any experimental elicitation, can affect the interpretation of our results. While we believe that the three-step elicitation procedure described in Section 3.2 reduces the likelihood of noise affecting the signs of W_L and W_{NL} , our results are not immune to measurement errors. For example, if some participants choose randomly, only a few participants would have $W_L = W_{NL} > 0$, leading us to underestimate the true share of pure external statists. To address this concern, we filtered the participants for quality to assess whether restricting to high-quality participants—i.e., those with presumably less noisy answers—affects our results.²⁵ While this section considers all participants, in Section A in Appendix A, we report the same analysis for the quality-restricted sample. The results are similar to those when using the full sample.

In the next section, we report the results of two additional treatments, which vary the

²⁵We used participants' open-ended responses detailed in Section 3.2 to flag those who exhibit an evident misunderstanding of instructions (in particular, a misunderstanding of the difference between *Learns* and *NotLearns*), those who seem to be using artificial intelligence tools (e.g., ChatGPT), and those whose quantitative answers are inconsistent with their written reasoning. Three research assistants independently categorized the data. We categorize each participant based on what the majority of research assistants say and filter our main sample using this categorization.

Receiver's mental state, allowing us to gauge the prevalence of preferences for matching external states to mental states.

4.2. Testing mental and external state independence

We test how welfare assessments depend on the Receiver's mental state—i.e., the Receiver's belief that they will get the books with the original notes. In *NotLearns*, varying the mental state allows us to test for a preference for mental and external states coinciding; in *Learns*, varying the mental state can be interpreted as an expectation-based reference point manipulation.

First, we repeat the classification exercise in Section 4.1. We report the results for *LowMS* and *HighMS* in rows 2 and 3 of Table 1. However, since these treatments vary the (initial) mental state in both *NotLearns* and *Learns*, changes across Table 1's rows could be driven by either W_{NL} or W_L , making them difficult to interpret.²⁶ Thus, we instead focus on the marginal distributions of W_{NL} and W_L .

Figure 1 displays the cumulative distributions of the welfare gains by treatment (*Baseline*, *LowMS*, *HighMS*) and by case (*NotLearns* in the left panel and *Learns* in the right panel). For both cases, comparing *HighMS* and *LowMS*, the distribution of welfare gains are increasing (in the first-order stochastic dominance sense) in the Receiver's mental state (a two-sided Kolmogorov–Smirnov test yields p-values of 0.005 and of 0.225 for *NotLearns* and *Learns*, respectively). We refrain from analyzing differences to the *Baseline* treatment since the Receiver's mental state as perceived by the participants is ambiguous.

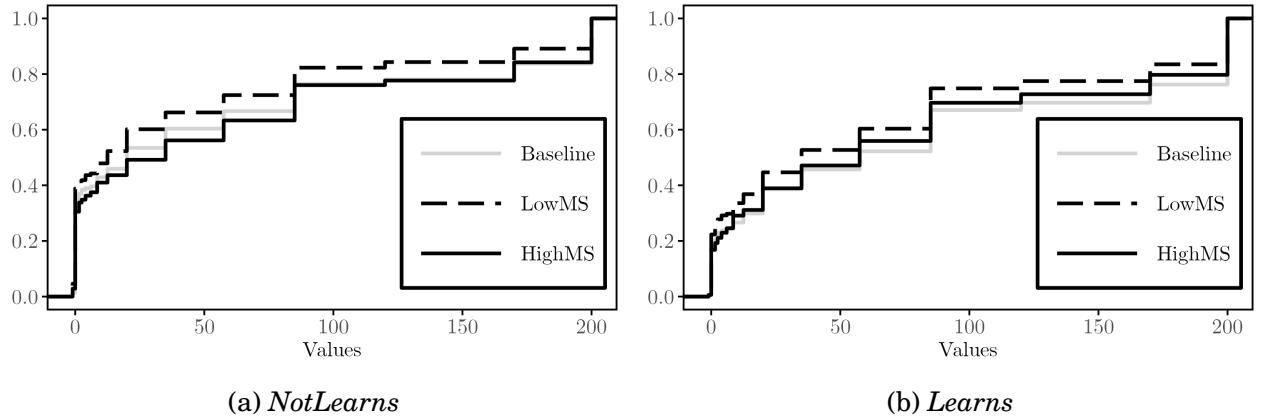


Figure 1: The two panels display the cumulative distributions of the Receiver's welfare gain by treatment for the *NotLearns* and *Learns* cases.

²⁶For example, in *HighMS*, 83.20% of participants have $W_L > 0$, as opposed to only 77.67% of participants (who need not be the same) in *LowMS*. As a result, the set of participants for whom there is welfare to decompose changes, leading to selection issues.

In *NotLearns*, this pattern implies a preference to match the external state to the mental state. In other words, participants assign a welfare gain to the Receiver having (more) accurate beliefs.²⁷ An implication of this finding is illustrated by going back to “The parable of the oblivious altruist” presented in the introduction. Suppose that in one scenario, Norman expects the government to provide relief effort; in another, he does not. The elicited welfare gain from the Receiver having accurate beliefs implies that individuals may perceive the welfare impact of the government providing the relief effort, including if Norman never learns about the effort or what made it necessary, to be larger in the first scenario. Furthermore, if individuals’ assessments of others’ welfare are informative of their own welfare (i.e., *ideals projection* holds), then Norman’s actual welfare gain, not just as perceived by others, is larger in the first scenario.

Result 3. *On average, participants value belief accuracy: they value an external state more if the Receiver’s mental state assigns a higher probability to it.*

In *Learns*, the participants’ decisions do not affect the (mis)match between the final mental and external states since the Receiver will always be informed about the external state, so there is a perfect match by design. Instead, while not statistically significant, the observed pattern can be interpreted as participants assigning some (welfare-relevant) loss aversion to the Receiver (Tversky and Kahneman, 1991): losses loom larger than gains, in a welfare sense. In particular, when the initial mental state is high, the welfare gain for the Receiver from getting the books with original notes is large since getting the fake notes would be perceived as a large loss.

Result 4. *On average, participants assign loss aversion to the Receiver: they value an external state and its corresponding mental state more if the Receiver’s ex-ante mental state assigns a higher probability to it (although the difference is not statistically significant).*

4.3. Testing the stability of welfare notions

We correlate our main incentivized welfare assessments, detailed in Section 4.1, with related unincentivized questions, detailed in Section 3.4, that similarly elicit the welfare notion participants subscribe to. In particular, we ask whether individuals hold different welfare notions for the self than for others, assessing *ideals projection*, and to what extent the

²⁷The welfare costs of a mismatch between mental and external state, together with a general preference for the Receiver to get the books with the original notes, could imply a net welfare gain of 0 if the participant assigns the Receiver a relatively low initial mental state. Thus, some participants who in the *NotLearns* case do not see a welfare gain might, in fact, not be pure mental statist but rather have more nuanced preferences that, in particular, depend on the external state. However, this case is non-generic.

welfare notions individuals adhere to depend on the choice domain, thus speaking to the external validity of our results.

To answer the first question, we present participants with the Experience Machine thought experiment (Nozick, 1974) and ask them “Would you go into the machine?” 51% of participants respond “Yes,” the answer that aligns more with mental statism, and 49% respond “No,” the answer that aligns more with external statism. We are interested in whether the answer to the Experience Machine predicts a participant’s type of welfare assessment—is it more aligned with mental or external statism? Participants who respond differently to our unincentivized questions could have different distributions of W_L , which makes direct comparisons of W_{NL} hard to interpret. Thus, to test whether these two groups’ elicited welfare assessments differ, we consider W_{NL}/W_L , which is the welfare effect from changing the external state normalized by the total welfare effect, W_L .²⁸

Panel (a) in Figure 2 displays the cumulative distributions of the ratio of W_{NL} to W_L , grouped by participants’ answers to the Experience Machine question. The distribution of W_{NL}/W_L for those who give the external-statism answer (i.e., they do not want to plug into the machine) first-order stochastically dominates that of those who give the mental-statism answer (a two-sided K-S test yields a p-value of 0.088). Moreover, the share of pure external statists ($W_{NL}/W_L = 1$) is larger in the former group (44.07% vs. 40.07%; a two-sided t-test yields p-value of 0.122), and the share of pure mental statists ($W_{NL}/W_L = 0$) is larger in the latter group (17.80% vs. 23.51%; p-value of 0.015). Thus, participants giving the external-statism answer indeed put relatively more weight on external states.

Result 5. *There is a positive correlation between the incentivized choosing-for-others welfare assessment and the unincentivized choosing-for-self response.*

We interpret this result as suggestive that participants’ surrogate choices directly speak to their own preferences over the mental and external states—i.e., in support of *ideals projection*—in line with Ambuehl et al. 2021. In turn, in the traditional revealed preference paradigm, this means the welfare notions their choices adhere to reflect their own welfare.

To assess the extent to which welfare notions individuals adhere to depend on the choice domain, we ask participants to make welfare assessments in two more naturalistic settings. In particular, we ask whether the policy described in “The parable of the oblivious altruist” by Bernheim and Taubinsky (2018) makes the protagonist better or worse off. Similarly, we ask participants whether the receiver of the original painting in the Andy Warhol scenario

²⁸For completeness, we also replicate Table 1 by participants’ answers to our unincentivized questions for the whole (Table 3) and the quality-restricted samples (Table 4), respectively. However, our preferred analyses consider the distributions of W_{NL}/W_L .

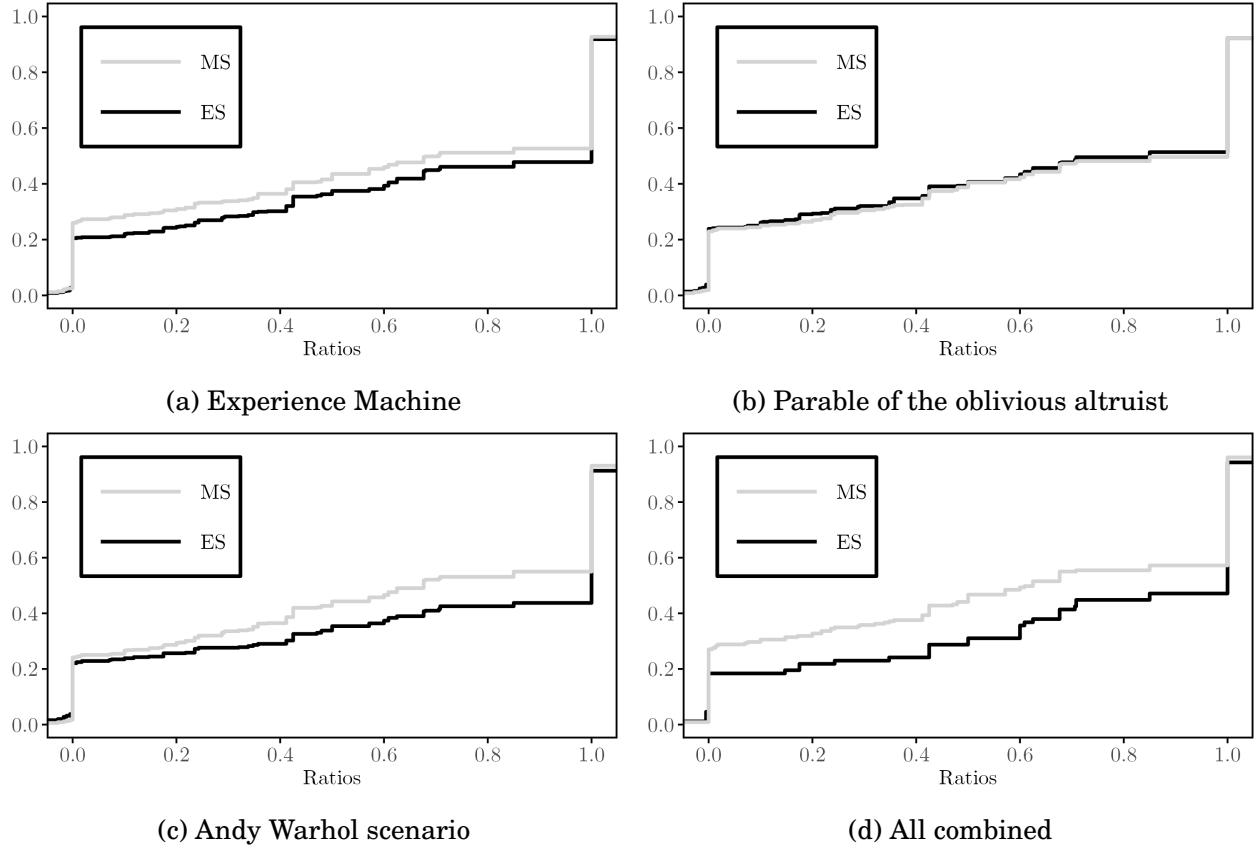


Figure 2: The four panels display the cumulative distributions of the ratio of W_{NL} to W_L , conditional on $W_L > 0$, for participants who give responses aligned with mental statism or external statism. Panel (d) displays participants whose responses to all three questions are consistent with each other (28.2%).

is better off. In both cases, “better off” suggests the participant adheres to external statism; 63% and 57% of participants choose this option,²⁹

Panels (b) and (c) in Figure 2 display the cumulative distributions of the ratio of W_{NL} to W_L for these two questions, again grouped by participants’ answers. While responses to “The parable of the oblivious altruist” do not seem to correlate with our incentivized welfare assessments (a two-sided K-S test yields a p-value of 0.988), in the Andy Warhol scenario, the values of W_{NL}/W_L for participants who gave the external-statism answer mostly first-order stochastically dominates the values for those who gave the mental-statism answer (p-value of 0.001).³⁰ In particular, the share of pure external statists is larger in the former

²⁹Precisely, we ask: “Does the government raising taxes to provide financial relief make John better or worse off?” for the “The parable of the oblivious altruist,” and “Is this person better off by getting the original one instead of a copy?” for the Andy Warhol scenario.

³⁰While “The parable of the oblivious altruist” is a thought experiment that captures the role of welfare notions for policymaking, its relation to policy can introduce confounding concerns. Participants could care

group (47.51% vs. 38.06%; p-value of 0.001), and the share of pure mental statists is larger in the latter group (18.29% vs. 22.42%; p-value of 0.078).

Result 6. *There is a positive correlation between the incentivized welfare assessment and the unincentivized response to the Andy Warhol scenario but not to “The parable of the oblivious altruist.”*

This result suggests that the welfare notions that individuals adhere to are consistent across domains; however, there is also some domain dependence. The extent to which welfare notions depend on the choice domain is an interesting avenue for future research.

The empirical mapping from welfare notions to responses in the three unincentivized questions is noisy. However, we can expect participants who unambiguously adhere to external statism in the unincentivized questions to be more likely to be external statists than those who unambiguously adhere to mental statism. Panel (d) in Figure 2 displays the cumulative distributions of W_{NL}/W_L for participants whose responses to all three questions align with mental statism ($N = 229$) and for those whose three responses align with external statism ($N = 87$)—again, restricted to those with $W_L > 0$. Indeed, we observe that the distribution of W_{NL}/W_L for participants who give the external statism responses first-order stochastically dominates the distribution for those who give the mental statism responses (p-value of 0.043). Similarly to before, the share of pure external statists is larger among those who give the external-statism responses (47.13% vs. 38.86%; p-value of 0.190), and the share of pure mental statists is larger among those who give the mental-statism responses (26.20% vs. 13.79%; p-value of 0.009). This result opens the door for developing a cheap method to assess welfare notions using a composite measure constructed from unincentivized questions.

5. CONCLUSION

Arguably, the very goal of policy-making is to maximize welfare—but what is welfare? We shed light on this question. In particular, we ask how reality, beyond a person’s beliefs, matters for welfare. The revealed preference paradigm cannot be used for this purpose: an individual cannot consciously choose between two realities without also changing their beliefs.

To overcome this limitation, we introduce what we call the altruistically revealed preference paradigm, wherein a surrogate chooses on behalf of the individual, maximizing what they perceive to be the individual’s welfare. This paradigm, relative to the revealed pref-

about how the policy might have “indirect” material consequences over Norman, for example, via equilibrium effects.

erence approach, extends the domain of applications; in particular, it allows us to experimentally study how external states—i.e., “reality”—affect welfare, beyond the individual’s mental state—i.e., their beliefs.

We find that the modal participant adheres to pure external statism, by which they assign welfare gain from changes in the external state independent of what happens to the mental state. We also find substantial heterogeneity, with sizable groups of participants adhering to pure mental statism, assigning welfare gains only from changes in the mental state, and a mix between external and mental statism, reporting a partial reduction in welfare gains when the mental state is fixed. Moreover, on average, participants assign welfare gains to having accurate beliefs. Our results suggest some stability of welfare notions by showing a positive correlation between the incentivized choosing-for-others welfare assessment and responses to three welfare-related unincentivized questions.

Understanding the welfare role of mental states in general, and beliefs in particular, is critical to understanding the value of non-instrumental information and, hence, may prove important for media regulation, informational policies, and government communication. For example, our findings allow for a re-interpretation of the welfare impact of surveillance programs: while the discussion of surveillance versus privacy usually focuses on its instrumental implications, our results suggest that individuals unaware of being surveilled may suffer “direct” welfare costs. The same point applies to a broader set of government policies that rely on an oblivious population—where government agencies hold more information than the population they serve—like those practiced by intelligence agencies and the military. Understanding the welfare effects of information provision that go beyond its instrumental value, as well as the welfare consequences of satisfying people’s preferences in situations in which they remain oblivious, is of first-order importance to the design of optimal policies.

One thought-provoking policy implication has to do with taxation. Chetty et al. (2009) documents that obfuscated taxes can lead to higher spending, which means they are less distortionary. This suggests that the government should use obfuscation as a tool. However, our findings imply that external states (e.g., what the taxes actually are) can matter per se, independent of the perceived taxes (beyond budgetary concerns)—there is a welfare cost from the mismatch of mental and external states. Future work could more directly identify preferences to reveal or obfuscate external states.

Our study also has normative implications for judging individual decision-making. For example, researchers have developed models of belief-based utility and empirically documented their usefulness in predicting behavior (Loewenstein and Molnar, 2018; Brunnermeier and Parker, 2005). In these models, actors choose beliefs, either via information or directly, strategically, trading off the instrumental value of correct beliefs against the utility

they draw from, say, an overly optimistic belief. While being useful positive models, these approaches typically leave important normative questions open. For instance, consider an individual who, say, has an overly optimistic belief about their intelligence to enhance their ego (Kőszegi, 2006). Whether this individual should be informed about their incorrect beliefs crucially depends on whether the utility they draw from them is actually welfare-relevant. If it is not, as is the case in the eyes of many of our participants, correcting the individual's beliefs is unambiguously good, leading to better decisions and a worse mental state, but one that is welfare-irrelevant.

Lastly, the heterogeneity we observe in our results provokes questions about the stability of welfare notions and their determinants. Future work should also explore such questions, for example, considering how culture and norms determine support for different welfare notions.

REFERENCES

- Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Pan Books.
- Ambuehl, S., Bernheim, B. D., and Ockenfels, A. (2021). What motivates paternalism? an experimental study. *American economic review*, 111(3):787–830.
- Ambuehl, S., Blesse, S., Doerrenberg, P., Feldhaus, C., and Ockenfels, A. (2023). Politicians' social welfare criteria: An experiment with german legislators. Technical report, CESifo Working Paper.
- Aristotle (2011). *Aristotle's Nicomachean ethics*. Translation by Bartlett, Robert C and Collins, Susan D and others. University of Chicago Press.
- Baber, H. (2008). The experience machine deconstructed. *Philosophy in the Contemporary World*, 15(1):132–137.
- Bartling, B., Cappelen, A. W., Hermes, H., and Tungodden, B. (2023). Free to fail? paternalistic preferences in the united states. Technical report, NHH Dept. of Economics Discussion Paper.
- Benjamin, D. J., Debnam Guzman, J., Fleurbaey, M., Heffetz, O., and Kimball, M. (2023). What do happiness data mean? theory and survey evidence. *Journal of the European Economic Association*, 21(6):2377–2412.
- Benjamin, D. J., Heffetz, O., Kimball, M. S., and Rees-Jones, A. (2012). What do you think would make you happier? what do you think you would choose? *American Economic Review*, 102(5):2083–2110.
- Benjamin, D. J., Heffetz, O., Kimball, M. S., and Rees-Jones, A. (2014). Can marginal rates of substitution be inferred from happiness data? evidence from residency choices. *American Economic Review*, 104(11):3498–3528.
- Bentham, J. (1789). Pml. *An Introduction to the Principles of Morals and Legislation*.
- Bernheim, B. D., Kim, K., and Taubinsky, D. (2024). Welfare and the act of choosing. Available at SSRN 4739442.
- Bernheim, B. D. and Taubinsky, D. (2018). Behavioral public economics. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 381–516. Elsevier.

- Bhat, B., De Quidt, J., Haushofer, J., Patel, V. H., Rao, G., Schilbach, F., and Vautrey, P.-L. P. (2022). The long-run effects of psychotherapy on depression, beliefs, and economic outcomes. Technical report, National Bureau of Economic Research.
- Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.
- Bénabou, R. (2015). The Economics of Motivated Beliefs. *Revue d'économie politique*, 125(5):665–685.
- Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.
- Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, 116(1):55–79.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–77.
- De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology*, 23(1):43–57.
- DellaVigna, S., Pope, D., and Vivaldi, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.
- Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.
- Hindriks, F. and Douven, I. (2018). Nozick's experience machine: An empirical study. *Philosophical Psychology*, 31(2):278–298.
- Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.
- Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.
- Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3):166–167.

- Mill, J. (1879). *Utilitarianism*. Longmans, Green and Company.
- Nozick, R. (1974). *Anarchy, state, and utopia*. John Wiley & Sons.
- Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, 370(6522):eaay0214.
- Rowland, R. (2017). Our intuitions about the experience machine. *J. Ethics & Soc. Phil.*, 12:110.
- Sen, A. (1985). *Commodities and Capabilities*. North-Holland, Amsterdam. New Delhi: Oxford University Press, 1987; Italian translation: Giuffre Editore, 1988; Japanese translation: Iwanami, 1988.
- Smith, B. (2011). Can we test the experience machine? *Ethical Perspectives*, 18(1):29–51.
- Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4):1039–1061.
- Uhl, M. (2011). Do self-committers mind other-imposed commitment? an experiment on weak paternalism. *Rationality, markets, and morals*.
- Weijers, D. (2013). Intuitive biases in judgements about thought experiments: The experience machine revisited. *Philosophical Writings*, 41(1).

A. ADDITIONAL TABLES

	W _L < 0			W _L = 0			W _L > 0				
	W _{NL} < 0	W _{NL} = 0	W _{NL} > 0	W _{NL} < 0	W _{NL} = 0	W _{NL} > 0	W _{NL} < 0	W _{NL} = 0	W _{NL} < W _L	W _L = W _{NL}	W _{NL} > W _L
	High-quality data										
Baseline N = 393	0.00%	0.00%	0.00%	0.00%	12.21%	1.27%	0.76%	17.30%	24.68%	38.42%	5.34%
LowMS N = 396	1.01%	0.00%	0.76%	1.01%	15.40%	2.02%	2.27%	15.66%	21.21%	35.86%	4.80%
HighMS N = 375	0.27%	0.00%	0.00%	0.27%	10.40%	2.67%	1.87%	13.87%	21.87%	42.40%	6.40%

Table 2: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively.

	W _L < 0			W _L = 0			W _L > 0				
	W _{NL} < 0	W _{NL} = 0	W _{NL} > 0	W _{NL} < 0	W _{NL} = 0	W _{NL} > 0	W _{NL} < 0	W _{NL} = 0	W _{NL} < W _L	W _L = W _{NL}	W _{NL} > W _L
Warhol scenario: ES (N = 641)	0.16%	0.00%	0.62%	0.94%	16.85%	2.96%	2.96%	14.35%	17.00%	37.29%	6.86%
Warhol scenario: MS (N = 836)	0.60%	0.12%	0.48%	0.12%	13.16%	2.87%	1.44%	18.54%	25.48%	31.46%	5.74%
Oblivious altruist: ES (N = 547)	0.18%	0.00%	0.73%	0.73%	14.81%	3.11%	3.11 %	16.09%	22.12%	32.91%	6.22%
Oblivious altruist: MS (N = 930)	0.54%	0.11%	0.43%	0.32%	14.73%	2.80%	1.51%	17.10%	21.61%	34.62%	6.24%
Experience Machine: ES (N = 718)	0.56%	0.14%	0.42%	0.42%	13.51%	2.79%	2.23%	14.62%	22.42%	36.21%	6.69%
Experience Machine: MS (N = 759)	0.26%	0.00%	0.66%	0.53%	15.94%	3.03%	1.98%	18.71%	21.21%	31.88%	5.80%
All combined: ES (N = 114)	0.88%	0.00%	1.75%	1.75%	14.91%	4.39%	3.51%	10.53%	21.93%	35.96%	4.39%
All combined: MS (N = 276)	0.72%	0.00%	0.36%	0.00%	12.68%	3.26%	0.72%	21.74%	25.00%	32.25%	3.26%

Table 3: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively.

	$W_L < 0$			$W_L = 0$			$W_L > 0$				
	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} < W_L$	$W_L = W_{NL}$	$W_{NL} > W_L$
High-quality data											
Warhol scenario: ES ($N = 493$)	0.20%	0.00%	0.20%	0.81%	14.40%	3.04%	1.62%	12.78%	17.65%	44.02%	5.27%
Warhol scenario: MS ($N = 671$)	0.60%	0.00%	0.30%	0.15%	11.48%	1.19%	1.64%	17.73%	26.23%	35.02%	5.66%
Oblivious altruist: ES ($N = 432$)	0.23%	0.00%	0.69%	0.69%	13.43%	2.78%	3.01 %	14.81%	22.69%	36.81%	4.86%
Oblivious altruist: MS ($N = 732$)	0.55%	0.00%	0.00%	0.27%	12.30%	1.50%	0.82%	16.12%	22.54%	40.03%	5.87%
Experience Machine: ES ($N = 568$)	0.70%	0.00%	0.18%	0.35%	11.62%	1.41%	1.76%	12.68%	23.59%	41.02%	6.69%
Experience Machine: MS ($N = 596$)	0.17%	0.00%	0.34%	0.50%	13.76%	2.52%	1.51%	18.46%	21.64%	36.74%	4.36%
All combined: ES ($N = 92$)	1.09%	0.00%	1.09%	1.09%	11.96%	5.43%	4.35%	8.70%	22.83%	39.13%	4.35%
All combined: MS ($N = 229$)	0.44%	0.00%	0.00%	0.00%	10.04%	2.62%	0.44%	21.83%	24.89%	36.24%	3.49%

Table 4: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively.

B. EXPERIMENTAL INSTRUCTIONS

We are giving a present to someone!

The questions we will ask you to answer involve another person. His name is Alex.

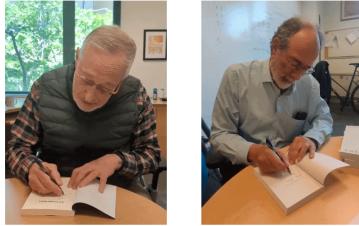
Alex loves economics, and we are going to give him a present!

He is going to get two books by two Nobel laureates in economics—Professors [Paul Milgrom](#) and [Alvin Roth](#) (you can click each of their names to open their Wikipedia page). Professors Milgrom and Roth are professors at our university and have agreed to help with the study.

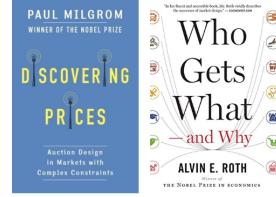
Alex has already read some of their work and told us he has great admiration for them.

The two books come with handwritten notes. But here is the twist! **We have two copies of each of these books.** One with **original notes** from the famous authors themselves (Profs. Milgrom and Roth), and one with **false notes** written by someone excellent at copying their handwriting.

Here are videos of the professors writing the notes



Here are the two books



The fake versions of the handwritten notes are indistinguishable from the original ones. Professors Milgrom and Roth themselves could not tell which is which!

Alex will receive two books, either the two with the original handwritten notes or the two with the fake ones. We will return the two books that we do not give to Alex back to Professors Milgrom and Roth.

When you are ready, click "Next."

[Next](#)

Figure 3

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, he may have got the ones with the original or the fake notes.

When ready, click "Next."

[Next](#)

(a) *Baseline* treatment

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the fake notes, and with a 25% chance, **you will determine which books he gets.**

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the fake notes, and with a 25% chance, **you will determine which books he gets.**

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, there is at least a 75% chance that they are the ones with the fake notes.

When ready, click "Next."

[Next](#)

(b) *LowMS* treatment

Figure 4

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes
With a 75% chance, Alex will get the books with the original notes, and with a 25% chance, **you will determine which books he gets.**

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes
With a 75% chance, Alex will get the books with the original notes, and with a 25% chance, **you will determine which books he gets.**

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, there is at least a 75% chance that they are the ones with the original notes.

When ready, click "Next."

[Next](#)

(c) HighMS treatment

Figure 4

Well done!

On the next pages, we will ask you 15 questions to determine which books Alex gets.

For example, one of them will be: *which books do you prefer Alex to receive? The ones with the original or fake notes?* There are other questions where we add a bonus for Alex to one of the options. You can click on this button to see all questions: [Questions](#)

We will randomly pick one of the questions and implement whatever option you choose. This means that any question can be the one that determines what Alex gets, so please answer them carefully.

When you are ready, click "Next."

[Next](#)

Figure 5

Which books should Alex get?

[Review Instructions](#)

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Fake notes I am indifferent Original notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Fake notes I am indifferent Original notes

[Next](#)

(a) *Baseline* treatment

Which books should Alex get?

[Review Instructions](#)

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Original notes I am indifferent Fake notes

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with fake notes; you now determine which books he gets otherwise.

It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Original notes I am indifferent Fake notes

[Next](#)

(b) *LowMS* treatment

Figure 6

Which books should Alex get?

[Review Instructions](#)

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes
He knows that with 75% chance, he will get the books with original notes; you now determine which books he gets otherwise.

It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?
 Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

I am indifferent Original notes Fake notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes
It is important that you understand which case we are in. Please answer the question below.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?
 Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

I am indifferent Original notes Fake notes

[Next](#)

(c) *HighMS* treatment

Figure 6

Which books and bonus should Alex get?

[Review Instructions](#)

On this page, the options involve Alex's bonus.

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

Fake notes + \$1 Original notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

Fake notes Original notes + \$1

[Next](#)

(a) *Baseline* treatment

Which books and bonus should Alex get?

[Review Instructions](#)

On this page, the options involve Alex's bonus.

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

Original notes Fake notes + \$1

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with fake notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

Original notes Fake notes + \$1

[Next](#)

(b) *LowMS* treatment

Figure 7

Which books and bonus should Alex get?

[Review Instructions](#)

On this page, the options involve Alex's bonus.

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with original notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Original notes Fake notes + \$1

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

- Yes, he will learn No, he will not learn

After correctly answering the question above, you will be able to decide which books Alex receives below.

Which books do you prefer Alex to receive in this case?

- Original notes Fake notes + \$1

[Next](#)

(c) HighMS treatment

Figure 7

Which books and bonus should Alex get?

You make several choices again involving Alex's bonus, filling in a table like the one below. For each row, we ask you to choose between the original notes and fake notes with a bonus for Alex.

Fake notes and...	OR	Original notes and...
...\$2	OR	...\$0
...\$3	OR	...\$0
...\$5	OR	...\$0
...\$7	OR	...\$0
...\$10	OR	...\$0
...\$15	OR	...\$0
...\$25	OR	...\$0
...\$45	OR	...\$0
...\$70	OR	...\$0
...\$100	OR	...\$0
...\$140	OR	...\$0
...\$200	OR	...\$0

We assume that once you choose the fake notes for one row, you will choose the fake notes for all rows below because the rows below simply make the fake notes better by increasing the bonus. You will only need to choose the row in which you switch from preferring the original notes to preferring the fake notes. You do that by clicking on the row.

Once you are confident that you understand how the table works, continue to give your answers.

[Next](#)

Figure 8

(a) *Baseline* treatment

Figure 9

Which books and bonus should Alex get?

[Review Instructions](#)

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...	OR	Fake notes and...
...\$0	OR	...\$2
...\$0	OR	...\$3
...\$0	OR	...\$5
...\$0	OR	...\$7
...\$0	OR	...\$10
...\$0	OR	...\$15
...\$0	OR	...\$25
...\$0	OR	...\$45
...\$0	OR	...\$70
...\$0	OR	...\$100
...\$0	OR	...\$140
...\$0	OR	...\$200

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with fake notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...	OR	Fake notes and...
...\$0	OR	...\$2
...\$0	OR	...\$3
...\$0	OR	...\$5
...\$0	OR	...\$7
...\$0	OR	...\$10
...\$0	OR	...\$15
...\$0	OR	...\$25
...\$0	OR	...\$45
...\$0	OR	...\$70
...\$0	OR	...\$100
...\$0	OR	...\$140
...\$0	OR	...\$200

[Next](#)

(b) *LowMS* treatment

Figure 9

Which books and bonus should Alex get?

[Review Instructions](#)

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes
 He knows that with 75% chance, he will get the books with original notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...	OR	Fake notes and...
...\$0	OR	...\$2
...\$0	OR	...\$3
...\$0	OR	...\$5
...\$0	OR	...\$7
...\$0	OR	...\$10
...\$0	OR	...\$15
...\$0	OR	...\$25
...\$0	OR	...\$45
...\$0	OR	...\$70
...\$0	OR	...\$100
...\$0	OR	...\$140
...\$0	OR	...\$200

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

Yes, he will learn No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...	OR	Fake notes and...
...\$0	OR	...\$2
...\$0	OR	...\$3
...\$0	OR	...\$5
...\$0	OR	...\$7
...\$0	OR	...\$10
...\$0	OR	...\$15
...\$0	OR	...\$25
...\$0	OR	...\$45
...\$0	OR	...\$70
...\$0	OR	...\$100
...\$0	OR	...\$140
...\$0	OR	...\$200

[Next](#)

(c) *HighMS* treatment

Figure 9

Please review your responses for the two cases

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

1. I prefer Alex to receive the ones with the **original notes** and \$1 over the ones with the **false notes**
2. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes**
3. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$1
4. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$2
5. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$3
6. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$5
7. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$7
8. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$10
9. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$15
10. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$25
11. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$45
12. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$70
13. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$100
14. I prefer Alex to receive the ones with the **false notes** and \$140 over the ones with the **original notes**
15. I prefer Alex to receive the ones with the **false notes** and \$200 over the ones with the **original notes**

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

1. I prefer Alex to receive the ones with the **original notes** and \$1 over the ones with the **false notes**
2. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes**
3. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$1
4. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$2
5. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$3
6. I prefer Alex to receive the ones with the **original notes** over the ones with the **false notes** and \$5
7. I prefer Alex to receive the ones with the **false notes** and \$7 over the ones with the **original notes**
8. I prefer Alex to receive the ones with the **false notes** and \$10 over the ones with the **original notes**
9. I prefer Alex to receive the ones with the **false notes** and \$15 over the ones with the **original notes**
10. I prefer Alex to receive the ones with the **false notes** and \$25 over the ones with the **original notes**
11. I prefer Alex to receive the ones with the **false notes** and \$45 over the ones with the **original notes**
12. I prefer Alex to receive the ones with the **false notes** and \$70 over the ones with the **original notes**
13. I prefer Alex to receive the ones with the **false notes** and \$100 over the ones with the **original notes**
14. I prefer Alex to receive the ones with the **false notes** and \$140 over the ones with the **original notes**
15. I prefer Alex to receive the ones with the **false notes** and \$200 over the ones with the **original notes**

Do the above answers reflect what you intended to answer, or do you want to give your answers again?

- Yes, the answers above reflect what I intended to answer
 No, I want to give my answers again

Please click "Next" to proceed to the next page.

[Next](#)

Figure 10

Thank you for your responses

Reminder:

- Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes
- Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

You gave different responses in Case 1 and Case 2. Why? Please tell us in approximately 1-3 sentences.
(There is nothing wrong with your answers! We are just interested in your reasoning)

In the remaining pages, we will present you with scenarios and ask you questions about them. Your responses are very important for our study; please think about the scenarios and answer carefully.

[Next](#)

Figure 11

The Experience Machine

If given the option, would you choose to plug into an experience machine that could provide you with an entirely immersive, simulated reality where you can experience any desirable scenario, despite not being real? Keep in mind that while plugged in, you would never be aware that you are in the experience machine and would believe that the simulated reality is real.

Suppose there was an experience machine that would give you *any* experience you desired (eating good food, having a successful career, making meaningful connections, etc.). While in the machine, you would not know that you are in it; you would think that what you are experiencing is actually happening.

Would you go into the machine?

- Yes
- No

Why? Answer in approximately 1-2 sentences.

[Next](#)

Figure 12

Flood in Arkansas

A small town in Arkansas experiences massive flooding, leaving many families homeless. To provide financial relief to the impacted families, the government temporarily increases taxes, including a \$100 levy on John. John lives far away and *will never learn about the flooding or the relief effort*. However, he cares about helping others and would gladly contribute \$100 to the relief effort if he knew about the flood.

Please tell us what you think using the information provided above. This is not a trick question; we want to understand what you think about the impact that the policy has on John.

Does the government raising taxes to provide financial relief make John better or worse off?

- Better off
- Worse off

Why? Answer in approximately 1-2 sentences.

[Next](#)

Figure 13

Andy Warhol

Hundreds of Andy Warhol fakes, and one original drawing worth \$20k, sold for \$250 each. An art collective purchased an original Warhol drawing and copied it 999 times. The copies are carefully created so that not even their creators can tell them apart from the original drawing. They then mixed the original together with the copies and sold the 1000 drawings.

Please watch the video (1min 1sec) the art collective made (audio is not needed). You can read more about this story [here](#).



Someone got the original Andy Warhol drawing. Since the original drawing and the copies are indistinguishable, please assume that neither the person who got it nor anyone else will ever know which is the original drawing or who has it.

Is this person better off by getting the original one instead of a copy?

- Yes
- No

Why? Answer in approximately 1-2 sentences.

Next

Figure 14: Click [here](#) to see the video: <https://moforgesies.org/images/showcase.mp4> (accessed 2023/08/23).