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

Gonzalo Arrieta† Lukas Bolte‡

October 24, 2025

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, participants' welfare assessments suggest a value of truth. While there is large heterogeneity, our results suggest that most people support the idea that welfare goes beyond beliefs, which can inform media regulation, informational policies, and government communication. We illustrate how our results inform welfare analysis that jointly considers reality and beliefs by revisiting the welfare effects of biased subjective expectations in a quantitative life-cycle model.

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

KEYWORDS: Welfare, subjective expectations, paternalism, revealed preferences, utilitarianism, mental states, beliefs, information policy.

*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, John Conlon, Ori Heffetz, David Huffman, Muriel Niederle, Kirby Nielsen, Charlie Rafkin, Peter Schwardmann, Florian Zimmermann, and seminar participants at Carnegie Mellon University, FAIR NHH, Princeton University, SBEERC, SITE Psychology, Stanford University, Tilburg University, the University of Pittsburgh, and the University of Zurich for their helpful comments. Haley 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 #IRBSTUDY2015_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.

‡Tilburg University. E-mail: lukas.bolte@outlook.com.

> "We are frequently concerned with settings in which people may misunderstand the consequences of their choices. In those cases, does well-being depend on the imagined state of affairs, the real state of affairs, or both? The answer to this question fundamentally shapes the conclusions that follow from normative economic analyses."
>
> —Bernheim and Taubinsky (2018)

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 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, relying on choice data restricts revealed preference as a welfare criterion to cases where such data is available. In particular, assessing welfare through choices restricts us to settings in which the individual knows, for example, that alternative x has been chosen. This is because one cannot consciously choose between x and y without simultaneously updating one's beliefs about which alternative is actually 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.

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 (Nunn and Sierra, 2017), or when they might hold motivated beliefs (Bénabou and Tirole, 2016). 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? Should we always correct wrong beliefs, like thinking too highly of oneself, or can these beliefs increase welfare, beyond any instrumental role? The traditional revealed preferences criterion is, again, not applicable; how, then, can we learn about the welfare consequences of such actions?

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. The elicited welfare perceptions can be used to guide policy-making, not only because people's welfare views may reflect true welfare, but also because democratic principles imply that policy ought to reflect citizens' preferences.[1]

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, 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 insofar as it affects the individual's welfare.
>
> *Obliviousness*: The individual remains oblivious about the chosen alternative unless

[1]Formally, Eden and Piacquadio (2025) characterize conditions that tightly restrict the extent to which a Paretian social welfare function can deviate from citizens' attitudes.

we inform them; otherwise, we are back in the standard paradigm where preference satisfaction and beliefs about preference satisfaction move in tandem.

To illustrate the challenges in implementing the paradigm, consider a surrogate's choice between an individual eating an apple or an orange. This choice may satisfy *welfare relevance*; however, it is impossible to have the individual eat the fruit while keeping them oblivious to what they are eating, which violates *obliviousness*. Similarly, a surrogate choice between which charity the individual donates 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. Furthermore, if the individual were to choose for themselves—as in standard choice data—*welfare relevance* and *pure altruism* would be easily satisfied. It is then the third requirement, *obliviousness*, that is an insurmountable obstacle since it is impossible for a chooser to make a choice while remaining oblivious about the chosen alternative.

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—made by the authors—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%.[2] 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 test whether participants exhibit different welfare notions when choosing for the self rather than for others. While this result is not crucial for the main insights of our study—which focuses on welfare perceptions—assessing the extent to which we are capturing welfare notions for the self sheds light on the possibility of applying the revealed preferences paradigm to interpret the surrogate's choices as reflective of their own welfare, as we explain in Section 3.5. We find evidence suggesting that welfare notions for the self are related to those for others. We further examine whether these welfare notions depend on the choice environment, and find that although they display some context dependence, they also appear to reflect a stable, person-specific component.

Finally, we demonstrate how jointly considering reality and beliefs can inform normative economic analysis by revisiting a quantitative life-cycle model in which agents hold biased expectations about their labor-market prospects (Balleer et al., 2025). We show how incorporating anticipatory utility into the welfare assessment can overturn the normative conclusions drawn from such models under a broad set of parameters. This exercise serves as a proof of concept for embedding our empirically elicited welfare notions into structural frameworks. While it shows that policies aimed at correcting misperceptions need not always enhance welfare, our experimental findings also caution against normative approaches that rely solely on subjective realities, such as beliefs and emotions (Bernheim et al., 2024).

Literature review. Our study tests the dominant approach to welfare in economics, and more broadly contributes to a millennia-old question: "What is welfare?" Our focus is on two broad classes of theories conceptualizing welfare.[3] The first is preference (or desire)

[2]Less 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.

[3]A 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).

theory, the dominant approach to welfare in economics, which postulates that well-being consists of having one's preferences satisfied. 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. The second 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. 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.

Testing the role of mental states for welfare relates us to a large literature documenting 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 and Tirole (2016) 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 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.[4]

We also 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 differ-

[4]Our 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.

ent 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 themself is deciding between (i.e., the planner is incapable of gathering the relevant choice data).

Our choosing-for-others approach situates us in a branch of positive welfare economics 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 the altruistically revealed preference paradigm that is more ambitious than an account of how people evaluate others' welfare: 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. Using preference theory as the welfare notion—in particular in the traditional revealed preference paradigm—the individual's preferences, in turn, reflect their own welfare.

Outside economics, in experimental psychology and philosophy, there has been a persistent interest in the elicitation of 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.[6] 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.

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 the type of welfare assessment we study.

[5]We run a survey on the Social Science Predictions Platform (DellaVigna et al. 2019; Public Study ID sspp-2023-0032-v1 at www.socialscienceprediction.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.

[6]The 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).

We present our experimental design in Section 3. Section 4 gives the results, and Section 5 illustrates how these findings inform welfare analysis by revisiting the welfare effects of biased subjective expectations in a quantitative life-cycle model. Section 6 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*. For example, in "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. 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, not a group. 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 $\mathscr{W} : X \times \Delta(X) \to \mathbb{R}$.

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 $\mathscr{W}$ 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.[7] We can use these choice problems to estimate $\mathscr{W}$—but

[7]To introduce uncertainty over external states, choices can be over lotteries. This does not resolve the limitations of the revealed preferences paradigm. To illustrate, suppose a coin has been tossed and has landed,

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 $\mu = \delta_y$.[8]

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

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

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

where F is strictly increasing in both arguments, u describes the surrogate's selfish preferences over the external state and mental state (of the Receiver), and 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 that the individual's decisions are not guided by selfish Kantian categorical imperatives (e.g., "do not lie")—an assumption that we briefly discuss in Section 3.1. Our formulation allows for preferences such as "warm-glow" and "ethical obligation" as long as they operate via the Receiver's welfare.

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

---

but the outcome is covered. Suppose the chooser chooses a lottery that pays \$1 if the coin is heads over a certain payment of \$0.40. This choice reveals a preference for the lottery over the certain payment, regardless of the fact that the uncertainty has been realized because, as perceived by the chooser, it has not yet been realized. Their beliefs are then, also, that the chosen object is a lottery and not the factually realized coin's outcome. More generally, we argue that choices reveal preferences over the object chosen as perceived by the chooser, making a detachment between the revealed preferences and beliefs over the chosen object impossible to obtain.

[8]Misperception of the external state allows for choosing (x,μ) while believing $\mu \neq \delta_x$, yet it introduces more severe limitations for the use of the revealed preferences paradigm by not revealing information about the preferences of interest in the first place—because of the misperception.

Welfare relevance: $\hat{\mathscr{W}}$ 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 our domain of interest, 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 we do not require the surrogate to be accurate in their assessment of the Receiver's preferences over (x,μ), or their welfare, $\hat{\mathscr{W}}$. Whether the surrogate and the Receiver agree is irrelevant. As long as the three requirements hold, the surrogate's choices reveal how (x,μ) enters $\hat{\mathscr{W}}$, which is what we are interested in. In Section 3.5, we discuss conditions under which our approach reveals actual—and not just perceived—welfare.

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—$\mathscr{W}(x,\mu)$ 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—$\mathscr{W}(x,\mu)$ 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—$\mathscr{W}(x,\mu)$ 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—$\mathscr{W}(x,\mu) = ES(x) + MS(\mu)$, for some

functions ES and MS. Second, external and mental states may be complements—$\mathcal{W}$ has increasing differences.[9] 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.

Our experiment, described in the next section, allows us to study the shape of $\hat{\mathcal{W}}$ (and, under further assumptions, actual and not just perceived welfare). 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 individuals' perceptions of others' welfare, $\hat{\mathcal{W}}(x,\mu)-\hat{\mathcal{W}}(y,\mu)$.

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 study context-dependence of welfare notions; see Section 3.4.

[9]Suppose $X=\{x,y\}$ 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 $\mu'>\mu$, $\mathcal{W}(x,\mu')+\mathcal{W}(y,\mu)\geq\mathcal{W}(x,\mu)+\mathcal{W}(y,\mu')$.

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. We 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. Each book has a handwritten note dedicated to the Receiver.[10] For one copy of each of the books, the note was handwritten by the respective 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 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.[11] 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.[12] Below, we discuss how our choice of external state satisfies the 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 C).[13]
>
> *Pure altruism*. Our design choices minimize concerns that any individual other than

[10]The 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.

[11]The 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.

[12]The 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.

[13]In conveying *welfare relevance* to the participants, we were careful not to suggest whether the Receiver's 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.

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.[14] 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, considerations such as warm-glow or ethical obligations that operate via the Receiver's welfare do not invalidate *pure altruism*.[15]

Obliviousness. Since the original notes are indistinguishable from the fake ones, *obliviousness* holds (i.e., external and mental states can be independently manipulated).[16]

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.[17]

[14]For example, it is likely that participants would prefer avoiding the destruction of the books with the original notes, which would have been a violation of *pure altruism*.

[15]Satisfaction of *pure altruism* permits the revelation of the surrogate's preferences over (x, μ) operating via the Receiver's perceived welfare, $\hat{W}$. However, if *pure altruism* was not satisfied, the surrogate's choices would still reveal preferences over (x, μ), which can be intrinsically relevant to guide policymaking in a welfarist framework (Eden and Piacquadio, 2025). Hence, while *pure altruism* seems to be largely satisfied in our setting, and it facilitates the interpretation of the results, these do not strictly rely on the absence of selfish motives for our findings to be relevant for policymaking by a welfarist planner.

[16]Participants' open-ended responses (see the end of Section 3.2 for details) show no evidence that they question the fact that the notes are indistinguishable. On the flip side, making the original and fake notes indistinguishable could lead participants to 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.

[17]The 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).

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 such as ours may be low, and so few participants may be willing to reduce their own payoffs (see Section 3.6 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 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. This constitutes a *welfare relevance* violation. 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. Formally, we measure $\mathscr{W}(x,\mu) - \mathscr{W}(y,\mu)$ as a function of μ.[18]

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.[19]

[18] In all treatments, we are explicit about the universe of external states—the Receiver knows the books' notes can be either fake or original. Future work could explore whether our results vary in a case in which the Receiver is unaware of the possibility of an external state mismatch.

[19] Specifically, 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

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 1,000 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.[20]

3.5. 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 actual 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. Thus, we need to know something about the surrogate's preferences over x and y *for themselves*. To interpret the surrogate's choices for others as informative about their preferences for themselves, one additional requirement must be satisfied:

> *Ideals projection*: Surrogates choose for others using the same welfare notion as they would use for themselves.

If *ideals projection* is satisfied, 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

the ones with the fake notes." We randomly select between these 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.

[20] See www.moforgeries.org (accessed 2023/08/23) for a detailed description by the organizers of this project.

matter for the welfare of surrogates whose preferences are such that external states are all that matter when choosing for others.

Ambuehl et al., 2021 provide evidence supporting *ideals projection* in a time-preference context. In our context, to assess the relationship between welfare notions for the self and others, we present 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 C 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.[21]

3.6. Implementation and recruitment details

We recruited 1,477 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 C in the appendix for full experimental instructions.

[21]While not incentivizing this question might produce noisier responses, it is precisely the fact that this is an unincentivized thought experiment that allows us to elicit welfare notions for the self (i.e., elicit choices over deluded mental states without killing the delusion). However, we are skeptical of using unincentivized questions as a *main* strategy to study welfare notions for the self (as opposed to the main experiment we propose in this paper) since it requires asking respondents to imagine *obliviousness*, which can be undeniably hard for them, especially given existing evidence on the difficulties people have with contingent thinking.

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* and 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). Figure 1 displays the raw data, plotting W_{NL} against W_L. This figure visually displays our main types described in Section 2.3: Pure external statists, those with $W_{NL} = W_L$, are on the 45-degree line, and pure mental statists are those with $W_{NL} = 0$. The data below the 45-degrees line and above the $W_{NL} = 0$ horizontal line represent participants who value both external and mental states.

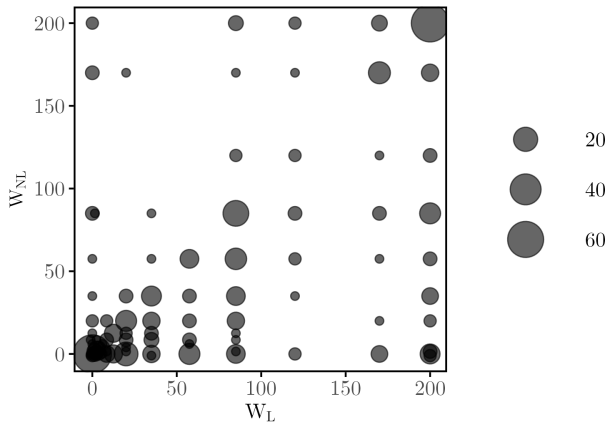


Figure 1: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively. The data is displayed for the *Baseline* treatment.

Table 1's top row classifies participants. 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).[22] 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., $\mathscr{W}(x,\mu)$ can be written as $\mathscr{W}(x,\mu) = ES(x) + MS(\mu)$), the former may be interpreted as pure mental statists, where $\mathscr{W}(x,\mu) = MS(\mu)$, and the latter as pure external statists, where $\mathscr{W}(x,\mu) = ES(x)$.

	$W_L < 0$			$W_L = 0$			$W_L > 0$				
	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} > 0$	$W_{NL} < 0$	$W_{NL} = 0$	$W_{NL} < W_L$	$W_L = W_{NL}$	$W_{NL} > W_L$
Baseline $N = 492$	0.20%	0.20%	0.20%	0.41%	14.23%	3.05%	0.81%	18.29%	23.58%	33.94%	5.08%
LowMS $N = 497$	0.80%	0.00%	1.01%	0.80%	17.51%	2.21%	3.02%	16.70%	20.12%	32.19%	5.63%
HighMS $N = 488$	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. The data is given for the *Baseline*, *LowMS*, and *HighMS* treatments.

Beyond pure mental statists and pure external statists, 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.*

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

[22] We 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.

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} (which are of first-order importance for our proof-of-concept exercise), 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.[23] 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 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

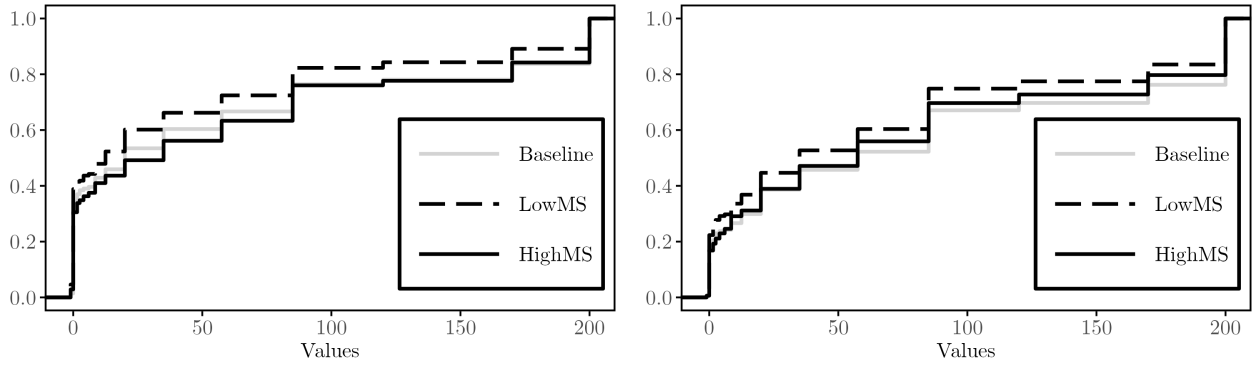
We test how welfare assessments depend on the Receiver's mental state—i.e., the Receiver's belief that they will get the books with the original notes. In *NotLearns*, varying the mental state allows us to test for a preference for mental and external states coinciding; in *Learns*,

[23] We 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.

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.[24] Thus, we instead focus on the marginal distributions of W_{NL} and W_L.

Figure 2 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 0.225 for *NotLearns* and *Learns*, respectively).[25] We refrain from analyzing differences to *Baseline* since the Receiver's mental state as perceived by the participants is ambiguous.



(a) *NotLearns* (b) *Learns*

Figure 2: The two panels display the cumulative distributions of the Receiver's welfare gain by treatment for the *NotLearns* and *Learns* cases.

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.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 wel-

[24]For 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.

[25]One could have been worried that participants do not acknowledge or that they distrust the mental state manipulation we intend to induce in *LowMS* and *HighMS*. The observed difference in the distributions of welfare gains shows that participants respond to the manipulation, minimizing this concern.

fare 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 Sections 3.4 and 3.5, that similarly elicit the welfare notion participants subscribe to. In particular, we ask to what extent the welfare notions individuals adhere to depend on the choice domain, thus speaking to the external validity of our results.

We begin by asking 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 is better off. In both cases, "better off" suggests the participant adheres to external statism; 63% and 57% of participants choose this option, respectively.[26]

[26]Precisely, 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.

We are interested in whether the answers to these questions predict 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 groups' 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.[27]

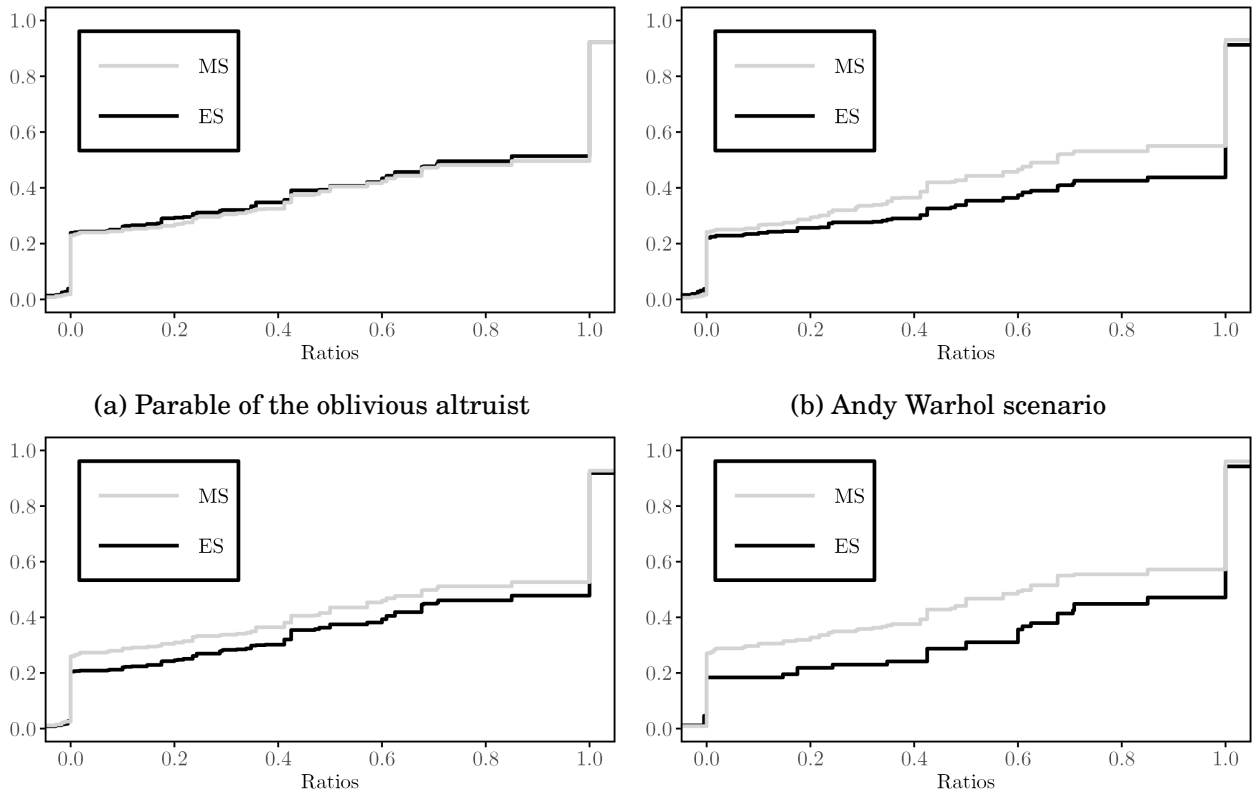
Panels (a) and (b) in Figure 3 display the cumulative distributions of the ratio of W_{NL} to W_L for these two questions, grouped by participants' answers. While responses to "The parable of the oblivious altruist" do not seem to correlate with our incentivized welfare assessments (a two-sided K-S test 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 dominate the values for those who gave the mental-statism answer (p-value of 0.001).[28] In particular, the share of pure external statists ($W_{NL}/W_L = 1$) is larger in the former group (47.51% vs. 38.06%; p-value of 0.001), and the share of pure mental statists ($W_{NL}/W_L = 0$) is larger in the latter group (18.29% vs. 22.42%; p-value of 0.078).

Result 5. *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."*

To assess the extent to which individuals hold different welfare notions for the self than for others, i.e., the extent to which *ideals projection* holds, we present participants with the Experience Machine thought experiment (Nozick, 1974) and ask "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. Panel (c) in Figure 3 displays the cumulative distributions of the ratio of W_{NL} to W_L, again 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 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 is larger in the latter group (17.80% vs. 23.51%; p-

[27]For 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.

[28]While "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 about how the policy might have "indirect" material consequences over Norman, for example, via equilibrium effects.



(a) Parable of the oblivious altruist

(b) Andy Warhol scenario

(c) Experience Machine

(d) All combined

Figure 3: 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%).

value of 0.015). Thus, participants giving the external-statism answer indeed put relatively more weight on external states.

Result 6. *There is a positive correlation between the incentivized choosing-for-others welfare assessment and the unincentivized choosing-for-self response.*

This correlation suggests participants' surrogate choices directly speak to their own preferences over the mental and external states—i.e., there is support for *ideals projection*—in line with Ambuehl et al. (2021). In the traditional revealed preference paradigm, this means the welfare notions their choices adhere to reflect their own welfare.

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 3 displays the cu-

mulative 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.

Taken together, this section's results suggest 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.

5. Application to normative economic analysis

How do our results inform welfare analysis? In this section, we provide a proof-of-concept example that shows how mental and external states can be *jointly* incorporated into welfare assessments. We illustrate how doing so can overturn conventional welfare judgments about correcting agents' incorrect beliefs.

In recent years, many studies have examined the consequences of incorrect beliefs.[29] The typical welfare analysis evaluates agents' choices by assessing their consumption using objective probabilities and then deriving implications for policy. For example, Conlon et al. (2018) show that biased beliefs about job-offer distributions distort acceptance decisions, lowering welfare relative to a complete-information benchmark; Spinnewijn (2015) demonstrates that workers save too little for unemployment due to overly optimistic beliefs, and derives the implications for optimal insurance design; and Balleer et al. (2025) show workers' optimistic bias about their future labor market transitions implies a substantial loss in individual welfare compared to a full-information benchmark. In all these examples, welfare is measured using pure external statism as a welfare criterion: Workers' welfare is

[29]Examples are biased expectations about inflation (Ortoleva and Snowberg, 2015; Malmendier and Nagel, 2016; Andre et al., 2024; Gennaioli et al., 2024), house prices (Cheng et al., 2014), unemployment (Spinnewijn, 2015; Andre et al., 2022; Jäger et al., 2024), financial market outcomes (Greenwood and Shleifer, 2014; Malmendier and Nagel, 2011; Shiller, 2014; Andre et al., 2023), social mobility and future earnings potential (Alesina et al., 2018), their health (Oster et al., 2013), among others (see also Bordalo et al., 2022 for a review of biased beliefs about financial and macroeconomic variables).

determined entirely by what their actual returns from employment and unemployment are and not at all by what they expect those returns to be—their mental states.[30]

At the same time, there is much evidence that anticipatory utility—the experience of pleasant or aversive emotions when thinking about future welfare (e.g., health or illness, wealth or bankruptcy)—can guide choices and affect welfare. That is, hope, fear, anxiety, and related belief-driven emotions are increasingly being considered important determinants of well-being, not only through their material consequences, but also as pure "mental consumptions" (Schelling, 1984; Loewenstein, 1987; Caplin and Leahy, 2001; Brunnermeier and Parker, 2005). In this way, this literature, at least implicitly, considers welfare to go beyond external states, to encompass the individuals' mental states.[31] The most explicit such case is Bernheim et al. (2024), who assume that people are pure mental statists who associate each choice option with an anticipated mental state bundle and then select their favorite bundle from the resulting menu. A policy, then, improves their welfare if and only if it provides them with a better mental state bundle according to those preferences.

We present a proof-of-concept demonstration of how mental and external states can be jointly incorporated into welfare assessments. In doing so, we use our experimental estimates to bridge traditional external-statist welfare analysis with emerging mental-statist approaches.

To conduct our exercise, we consider the quantitative life-cycle model of Balleer et al. (2025), henceforth, "BDFG," in which agents hold optimistic beliefs about their labor market outcomes. In BDFG, the welfare effects of such optimism are evaluated by comparing the equilibrium consumption in the biased economy, c_t, and the unbiased economy, $\bar{c}_t$. That is, their external statist welfare function compares $\sum_{t\geq 0}\beta^t E_0[u(c_t)]$ to $\sum_{t\geq 0}\beta^t E_0[u(\bar{c}_t)]$, where expectations E_0 are taken with respect to the *true* probabilities.

Specifically, BDFG assess the welfare costs of optimistic beliefs by computing the equivalent variation in expected lifetime consumption, ϕ_0, that makes an agent in the biased

[30] In the online appendix, Spinnewijn (2015) considers welfare concepts that place positive weight on the agent's perceived expected utility (i.e., their mental states). The distinction between a social planner who maximizes an agent's true versus subjective expected utility is conceptually explored by Salanié and Treich (2009), who introduce a framework to analyze the trade-offs that arise when evaluating the welfare of agents who misperceive risks. Both of these studies refer to a mental statist welfare notion as the agent's *perceived* welfare (e.g., which a populist policy maker would care about), rather than an alternative notion of *true* welfare for the social planner to consider.

[31] For example, Brunnermeier and Parker (2005) write: "Beliefs impact well-being directly through anticipation of future flow utility and indirectly through their effects on agent behavior. Optimal beliefs trade off the incentive to be optimistic in order to increase expected future utility against the costs of poor outcomes that result from decisions made based on optimistic beliefs."

economy as well off as in the counterfactual economy without bias:

$$\sum_{t\geq 0}\beta^t E_0[u((1+\phi_0)c_t)] = \sum_{t\geq 0}\beta^t E_0[u(\bar{c}_t)]. \tag{1}$$

Incorporating mental states. Consider now a social planner who uses both external *and mental* states, weighted by ω, to assess agents' welfare, e.g., via

$$W = \sum_{t\geq 0}\beta^t E_0[\omega MS_t + (1-\omega)ES_t]. \tag{2}$$

While it is clear that $ES_t = u(c_t)$, we must clarify what constitutes a mental state in this context. Suppose that at time t the agent's mental state is simply their current consumption, i.e., $MS_t = u(c_t)$—in particular, anticipatory utility is not part of the mental state. In this case, the welfare function in (2) coincides with the external statist function implicit in (1) and is therefore independent of ω. In this case, incorporating mental states has no effect on welfare assessments.

Now, suppose that at time t the agent's mental state is also driven by anticipatory utility; that is,

$$MS_t = \quad \alpha \quad \underbrace{\hat{E}_t\left[\sum_{\tau\geq t}\beta_{AU}^{\tau-t}u(c_{i\tau})\right]}_{\text{anticipatory utility } (AU_t)} + (1-\alpha) \quad \underbrace{u(c_t)}_{\text{consumption}}, \tag{3}$$

where $\hat{E}_t$ is the agent's subjective expectation at time t, and α captures how mental states weight anticipatory utility relative to current consumption.[32]

A social planner who incorporates mental states into welfare assessments would then use (2) and (3) to obtain the equivalent variation ϕ that would make an agent in the biased economy as well off as in the counterfactual economy without bias.

Quantitative exercise. We now conduct this exercise using the data available in BDFG and our own experimental estimate of ω, the relative importance of mental and external states. Compared to the welfare assessments conducted in BDFG, estimating (2) requires substantially more demanding data, since the planner needs to know the *subjective* expectation of lifetime consumption utility at every state and at every time t to construct the random stream of MS_t.

For simplicity, and given the data reported in BDFG, we illustrate the mental statists' welfare criterion by assuming that the anticipatory utility term at time t (henceforth, AU_t) equals AU_0 for all t (henceforth the "constant-AU" simplification; alternatively, we set $AU_t =$

[32]Note that the social planner judges the agent's welfare using their mental states, MS_t, even though they are formed with incorrect beliefs, but aggregates them according to correct probabilities, E_0.

0 for all $t > 0$, the "one-shot-AU" simplification). The constant-AU simplification allows us to make progress in our proof-of-concept exercise because BDFG report an equivalent variation ϕ_1 which, when further assuming $\beta_{AU} = \beta$, equates the anticipatory utility term at time 0 across the two economies:[33]

$$\sum_{t\geq 0} \beta^t \hat{E}_0\big[u\big((1+\phi_1)c_t\big)\big] = \sum_{t\geq 0} \beta^t E_0[u(\bar{c}_t)]. \tag{4}$$

Crucially, knowing the sequence of AU_t, ϕ_0, and ϕ_1 allows us to calculate equivalent variations between the biased and unbiased economies, for arbitrary α—the weight on anticipatory utility in the agent's mental state—and ω—the weight on mental states (see Appendix B for details). Henceforth, denote $\phi(\alpha,\omega)$ the equivalent variation given the values of α and ω.

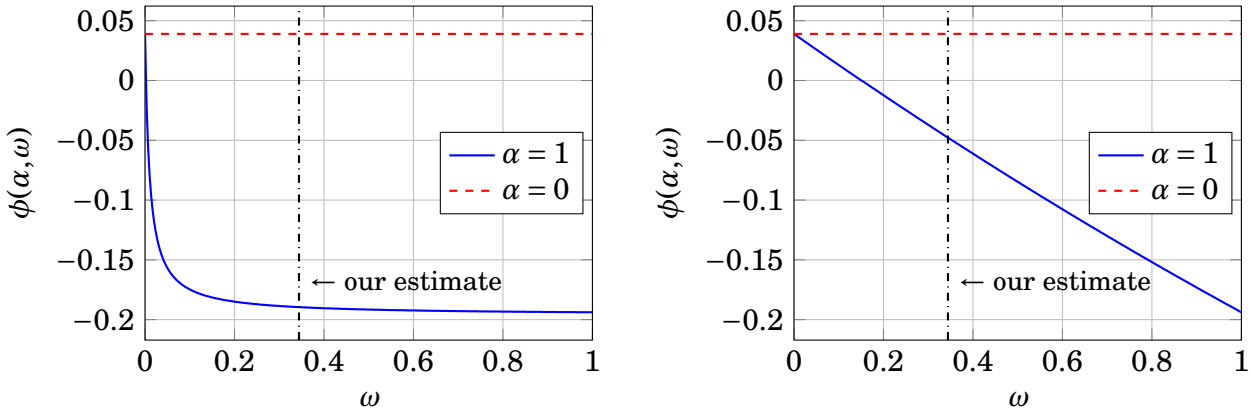
Figure 4 panel (a) reports the equivalent variations estimated for two extreme cases—$\alpha = 0$ and $\alpha = 1$—where we further let $\beta_{AU} = \beta = 0.9887$ as in BDFG. When mental states are only determined by contemporaneous consumption, i.e., $\alpha = 0$, then welfare is independent of ω, and so the equivalent variation ϕ equating welfare in the biased and unbiased economy is the same as that when the planner employs external statism in (1), i.e., $\phi(0,\cdot) = 0.039$ in BDFG. That is, agents in the biased economy need to be compensated with 3.9% extra consumption to be as well off as in the unbiased economy. This is because individuals make poor decisions due to their incorrect beliefs.

When anticipatory utility affects mental states, i.e., $\alpha > 0$, agents in the biased economy may attain *higher* welfare than those in the unbiased economy—provided the weight on mental states, ω, is sufficiently large—since their optimism generates welfare-relevant emotions. To estimate the associated equivalent variation, consider a social planner who uses our experimental results as a measure of ω. For participants in *Baseline* with $W_L > 0$ and $0 \leq W_{NL} \leq W_L$, we interpret W_{NL}/W_L as the weight on external states, i.e., $1-\omega$, which yields $\omega = 0.3441$ (that is, we are ignoring the heterogeneity in welfare notions for simplicity; Naik and Reck (2024) provide a framework to incorporate uncertainty directly into the social planner's objective function). We then obtain that agents in the biased economy are better off as long as $\alpha \geq 0.0059$.

Figure 4 panel (b) reports the same equivalent variations as panel (a) but under a different simplification, by which $AU_t = 0$ for all $t > 0$. In this case, under the same assumptions as before, the biased economy is better as long as $\alpha \geq 0.4378$. The difference in the α thresholds across the constant-AU and the one-shot-AU simplifications is expected since

[33] They report ϕ_1 because they are interested in the equivalent variation that makes an agent who maximizes their expected consumption indifferent between the economy with and without bias.

the discounted stream of AU_t is mechanically smaller under one-shot-AU, which increases the minimum weight on AU_t that makes welfare in the biased economy larger than that in the unbiased one.



(a) Constant-AU: $AU_t = AU_0$ for all t

(b) One-shot-AU: $AU_t = 0$ for all $t > 0$

Figure 4: Comparison of $\phi(\alpha,\omega)$ under constant-AU and one-shot-AU specifications, for $\alpha = 1$, $\alpha = 0$, $\beta_{AU} = \beta = 0.9887$.

Discussion. We provide a proof-of-concept illustration of how welfare assessments can jointly incorporate mental and external states, using our experimental estimates to connect traditional external-statist analysis with emerging mental-statist approaches.

The precise quantitative estimates will depend on a variety of factors. First, note that our simplifying assumptions regarding the stream of AU_t can lead to a misestimation of ϕ. For example, if AU_t actually decreases in t, the constant-AU assumption leads to an underestimation of ϕ. AU_t might decrease over time as the agent learns about the (negative) material consequences of their (biased) actions, or under the intuition that there is less utility to anticipate as one grows older. Second, the planner might want to estimate the agent's mental states discounting anticipated future consumption at a rate β_{AU} that need not equal the discount rate by which the planner aggregates mental states over time, β. In particular, if $\beta_{AU} < \beta$, our estimates of the difference between the welfare in the biased economy and that in the unbiased one will be too large. For example, Loewenstein (1987) assumes $\beta_{AU} \leq \beta$ to capture the intuition that far-away future consumption generates little anticipatory feelings. Third, while Loewenstein (1987), for instance, argues that $\alpha > 0$, there are no empirical estimates of its value. Fourth, note that, while we try to keep most of our social planner's choices grounded in the agent's choices (as per the revealed preferences welfare criterion), introducing anticipatory utility to assess the welfare of agents for whom anticipatory utility does not guide their choices constitutes a deviation from the revealed preferences criterion. Absent data limitations, we would want to assess the welfare of agents

who choose their stream of consumption considering anticipatory utility, i.e., who maximize (2) using subjective probabilities $\hat{E}_0$, and where MS_t is as in (3).

As our estimates illustrate, neglecting the joint role of external and mental states can substantially alter welfare assessments when agents hold incorrect beliefs. Pure external statist approaches, such as that by Balleer et al. (2025), overestimate the welfare cost of optimistic beliefs. Similarly, pure mental statist approaches, such as that hinted at by Oster et al. (2013) and embraced by Bernheim et al. (2024), may underestimate it. Future work that calibrates the key parameters for evaluating mental states would enable more precise quantitative welfare statements.

6. 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 and a person's beliefs separately matter 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 preference approach, has a wider domain of applications; in particular, it allows us to experimentally study how external states—i.e., "reality"—are perceived to affect welfare beyond the individual's mental state—i.e., their beliefs. These perceptions are highly relevant to a social planner not only because people's welfare views may reflect true welfare but also because democratic principles imply that policy ought to reflect citizens' preferences.

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, participants' welfare assessments reflect complementarities between external and mental states, suggesting a value of truth. 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. Ultimately, there could be specific aspects about our setting and design choices that affect the estimated magnitudes. For instance, the elicited welfare no-

tions may depend on the demographics or cultural background of the participants. Hence, we invite future work on the basis of our proof-of-concept that further tests people's views about these centuries-old welfare notions.

Our illustrative application to a quantitative life-cycle model shows that, when both external and mental states are welfare-relevant, policies that correct misperceptions need not always enhance welfare. Hence, incorporating empirically elicited heterogeneity in welfare notions into quantitative models can refine welfare assessments and guide policy prescriptions. Future work could contribute to this integration by estimating or calibrating the various parameters forming mental states for social planners to draw from in forming welfare assessments (as per the standard revealed preferences welfare criterion).

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) document 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 in specific settings.

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, agents 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). As our application in Section 5 shows, whether this individual should be informed about their incorrect beliefs crucially depends on whether the beliefs are welfare-relevant. Thus, novel and testable predictions between welfare notions, as elicited in our experiment, and the demand for self-deception emerge.

References

Alesina, A., Stantcheva, S., and Teso, E. (2018). Intergenerational mobility and preferences for redistribution. *American Economic Review*, 108(2):521–54.

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.

Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2024). Narratives about the macroeconomy. Technical report, SAFE Working Paper.

Andre, P., Pizzinelli, C., Roth, C., and Wohlfart, J. (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *The Review of Economic Studies*, 89(6):2958–2991.

Andre, P., Schirmer, P., and Wohlfart, J. (2023). Mental models of the stock market. *Univ. of Copenhagen Dept. of Economics Discussion Paper*, (07/23).

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.

Balleer, A., Duernecker, G., Forstner, S., and Goensch, J. (2025). The effects of biased labor market expectations on consumption, wealth inequality, and welfare. *American Economic Journal, Macroeconomics (Forthcoming)*.

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.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

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.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2022). Overreaction and diagnostic expectations in macroeconomics. *Journal of Economic Perspectives*, 36(3):223–244.

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.

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.

Cheng, I.-H., Raina, S., and Xiong, W. (2014). Wall street and the housing bubble. *American Economic Review*, 104(9):2797–2829.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–77.

Conlon, J. J., Pilossoph, L., Wiswall, M., and Zafar, B. (2018). Labor market search with imperfect information and learning. Technical report, National Bureau of Economic Research.

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 Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.

Eden, M. and Piacquadio, P. G. P. (2025). The ethical mirror. Discussion Paper 20624, CEPR.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Gennaioli, N., Leva, M., Schoenle, R., and Shleifer, A. (2024). How inflation expectations deanchor: The role of selective memory cues. Technical report, National Bureau of Economic Research.

Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746.

Hindriks, F. and Douven, I. (2018). Nozick's experience machine: An empirical study. *Philosophical Psychology*, 31(2):278–298.

Jäger, S., Roth, C., Roussille, N., and Schoefer, B. (2024). Worker beliefs about outside options*. *The Quarterly Journal of Economics*, 139(3):1505–1556.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kőszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *Economic Journal*, 97(387):666–684.

Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3):166–167.

Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking?*. *The Quarterly Journal of Economics*, 126(1):373–416.

Malmendier, U. and Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, 131(1):53–87.

Mill, J. (1879). *Utilitarianism*. Longmans, Green and Company.

Naik, C. and Reck, D. (2024). Intrapersonal utility comparisons as interpersonal utility comparisons: Welfare, ambiguity, and robustness in behavioral policy problems. Working Paper 32813, National Bureau of Economic Research.

Nozick, R. (1974). *Anarchy, state, and utopia*. John Wiley & Sons.

Nunn, N. and Sierra, R. S. d. l. (2017). Why being wrong can be right: Magical warfare technologies and the persistence of false beliefs. *American economic review*, 107(5):582–587.

Ortoleva, P. and Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, 105(2):504–535.

Oster, E., Shoulson, I., and Dorsey, E. R. (2013). Optimal expectations and limited medical testing: Evidence from huntington disease. *American Economic Review*, 103(2):804–830.

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.

Salanié, F. and Treich, N. (2009). Regulation in happyville. *The Economic Journal*, 119(537):665–679.

Schelling, T. C. (1984). *Choice and Consequence: Perspectives of an Errant Economist*. Harvard University Press, Cambridge, MA.

Sen, A. (1985). *Commodities and Capabilities*. North-Holland, Amsterdam. New Delhi: Oxford University Press, 1987; Italian translation: Giuffre Editore, 1988; Japanese translation: Iwanami, 1988.

Shiller, R. J. (2014). Speculative asset prices. *American Economic Review*, 104(6):1486–1517.

Smith, B. (2011). Can we test the experience machine? *Ethical Perspectives*, 18(1):29–51.

Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13(1):130–167.

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*, 2:13–34.

Weijers, D. (2013). Intuitive biases in judgements about thought experiments: The experience machine revisited. *Philosophical Writings*, 41(1).

A. Additional tables and figures

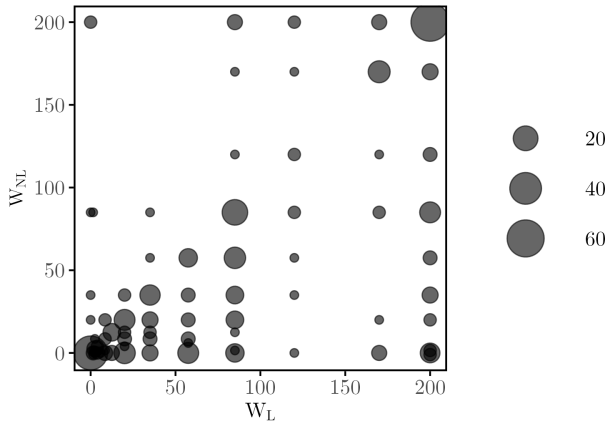


Figure 5: W_L, W_{NL} are the amounts of dollars added to the books with fake notes that make the participant indifferent between the Receiver getting the fake notes and the money, and the original notes, in the *Learns* and *NotLearns* case, respectively. The data is displayed for the *Baseline* treatment and restricted to the high-quality subsample.

	$W_L<0$			$W_L=0$			$W_L>0$				
	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}<W_L$	$W_L=W_{NL}$	$W_{NL}>W_L$
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. The data is given for the *Baseline*, *LowMS*, and *HighMS* treatments.

	$W_L<0$			$W_L=0$			$W_L>0$				
	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}<W_L$	$W_L=W_{NL}$	$W_{NL}>W_L$
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. The data is given for participants grouped by their answers to the unincentivized questions.

	$W_L<0$			$W_L=0$			$W_L>0$				
	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}>0$	$W_{NL}<0$	$W_{NL}=0$	$W_{NL}<W_L$	$W_L=W_{NL}$	$W_{NL}>W_L$
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. The data is given for participants grouped by their answers to the unincentivized questions.

B. Details for Section 5

We consider the constant-AU case. We seek ϕ such that

$$\underbrace{\sum_{t\geq 0}\beta^t E_0\left[\omega\left(\alpha\hat{E}_0\Big[\sum_{\tau\geq 0}\beta_{AU}^{\tau}u\big((1+\phi)c_{i\tau}\big)\Big]+(1-\alpha)\,u\big((1+\phi)c_t\big)\right)+(1-\omega)\,u\big((1+\phi)c_t\big)\right]}_{\text{welfare in the biased economy}}$$
$$=\underbrace{\sum_{t\geq 0}\beta^t E_0\left[\omega\left(\alpha E_0\Big[\sum_{\tau\geq 0}\beta_{AU}^{\tau}u(\bar{c}_{i\tau})\Big]+(1-\alpha)\,u(\bar{c}_t)\right)+(1-\omega)\,u(\bar{c}_t)\right]}_{\text{welfare in the unbiased economy}}. \tag{5}$$

From the paper, ϕ_0 and ϕ_1 are such that

$$E_0\left[\sum_{t\geq 0}\beta^t u\big((1+\phi_0)c_t\big)\right] = \hat{E}_0\left[\sum_{t\geq 0}\beta^t u\big((1+\phi_1)c_t\big)\right] = E_0\left[\sum_{t\geq 0}\beta^t u(\bar{c}_t)\right].$$

For $u = \ln$, we have

$$E_0\left[\sum_{t\geq 0}\beta^t u(c_t)\right] = E_0\left[\sum_{t\geq 0}\beta^t u(\bar{c}_t)\right] - \tfrac{1}{1-\beta}\ln(1+\phi_1),$$

$$\text{and}\quad \hat{E}_0\left[\sum_{t\geq 0}\beta^t u(c_t)\right] = E_0\left[\sum_{t\geq 0}\beta^t u(\bar{c}_t)\right] - \tfrac{1}{1-\beta}\ln(1+\phi_0).$$

Substituting into (5) and solving yields

$$\phi = \left(1+\phi_1\right)^{\frac{\omega\alpha}{\omega\alpha+(1-\omega\alpha)(1-\beta)}} \cdot \left(1+\phi_0\right)^{\frac{(1-\omega\alpha)(1-\beta)}{\omega\alpha+(1-\omega\alpha)(1-\beta)}} - 1. \tag{6}$$

For the "one-shot-AU" case,

$$\phi = \left(1+\phi_1\right)^{\omega\alpha} \cdot \left(1+\phi_0\right)^{1-\omega\alpha} - 1. \tag{7}$$

C. Experimental instructions

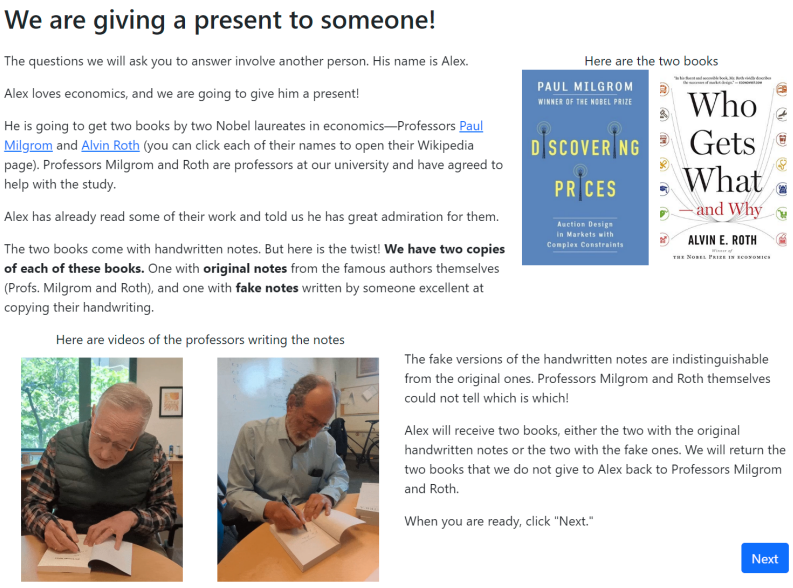
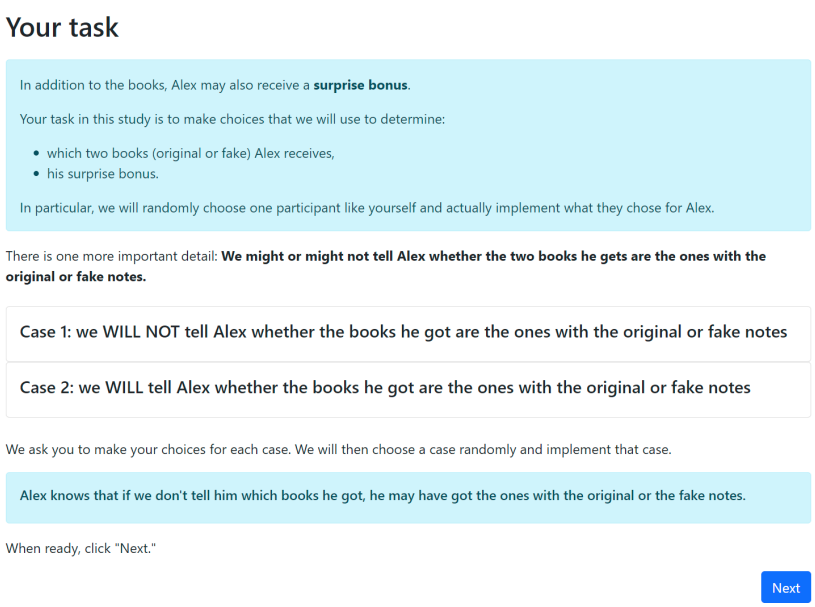
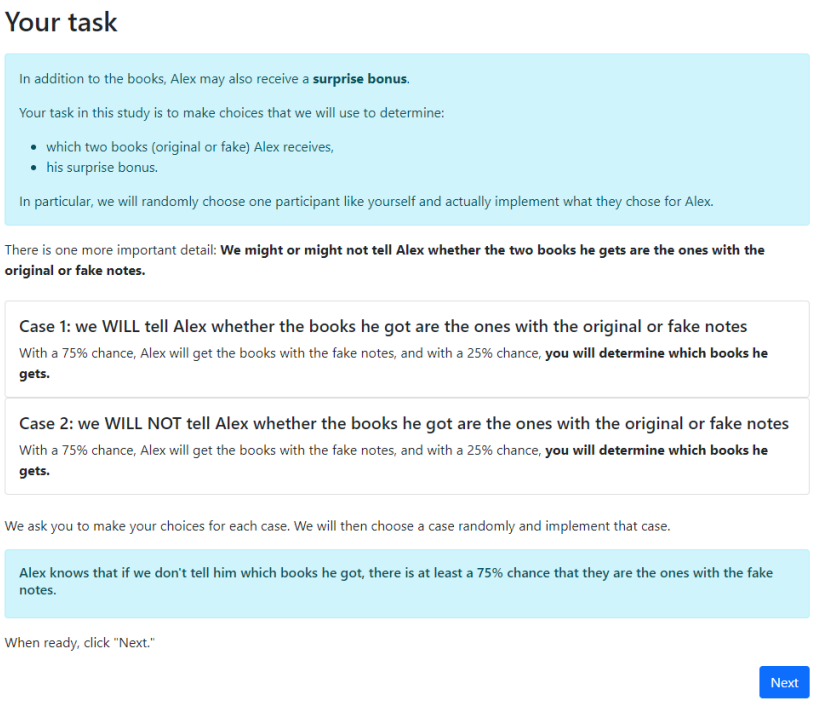


Figure 6

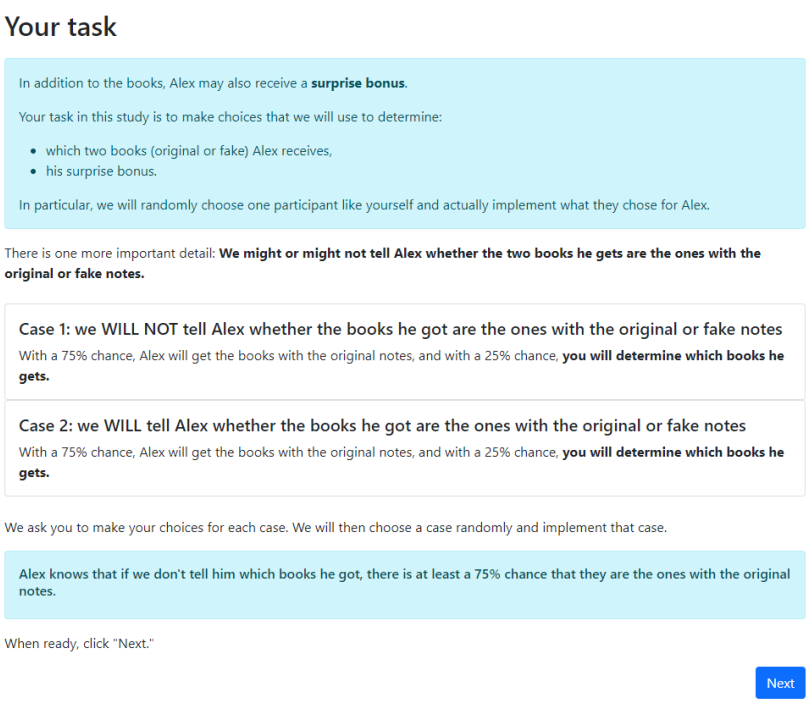


(a) *Baseline* treatment



(b) *LowMS* treatment

Figure 7



(c) *HighMS* treatment

Figure 7

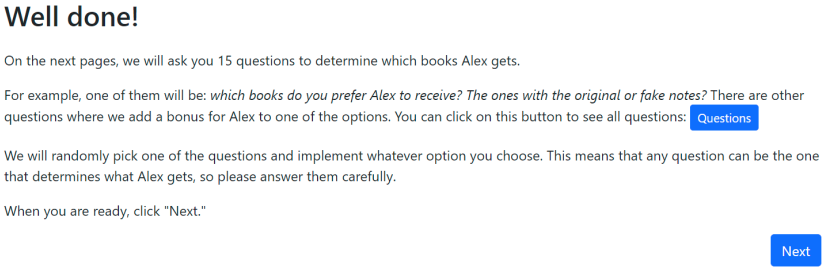
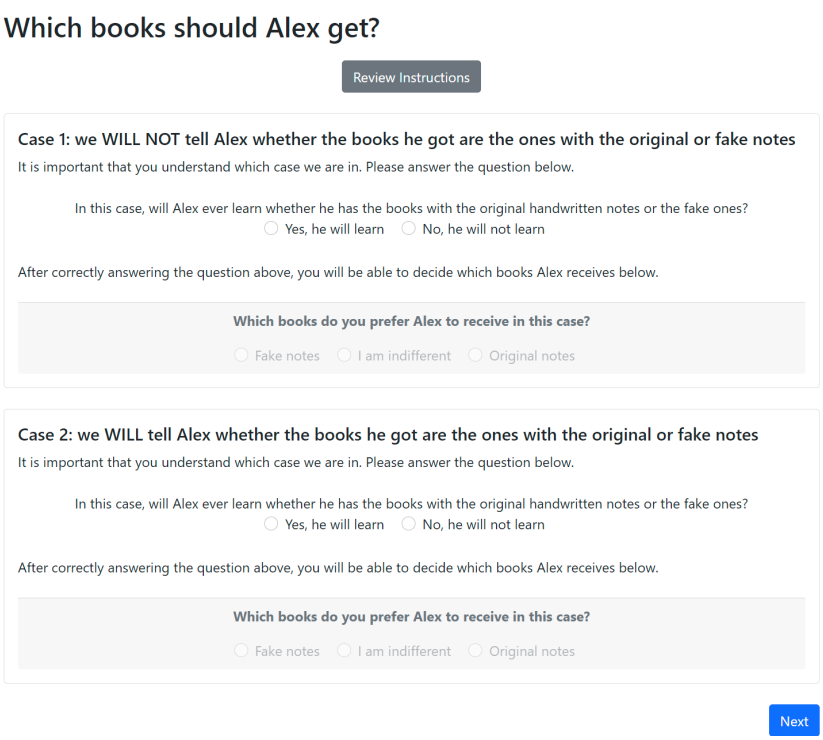
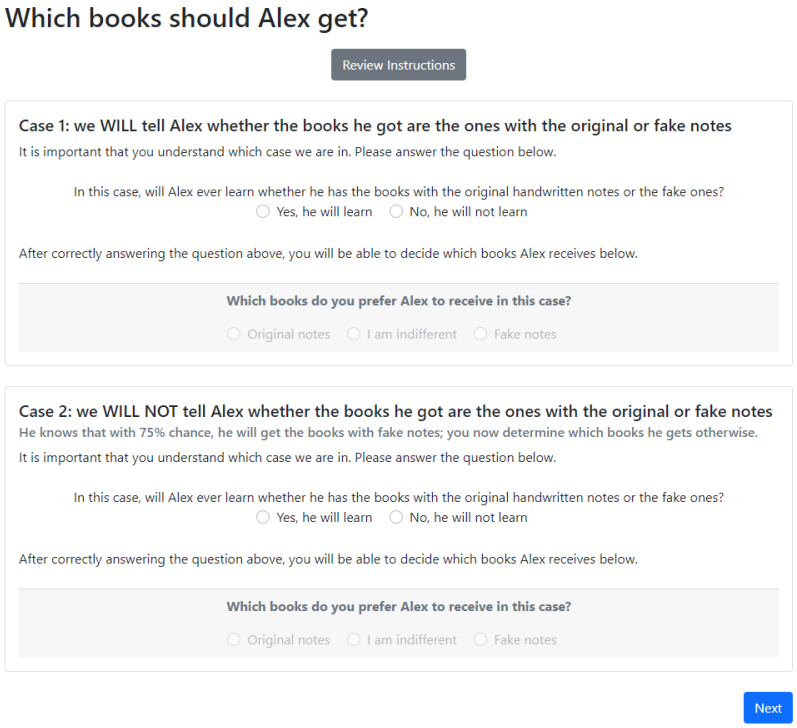


Figure 8

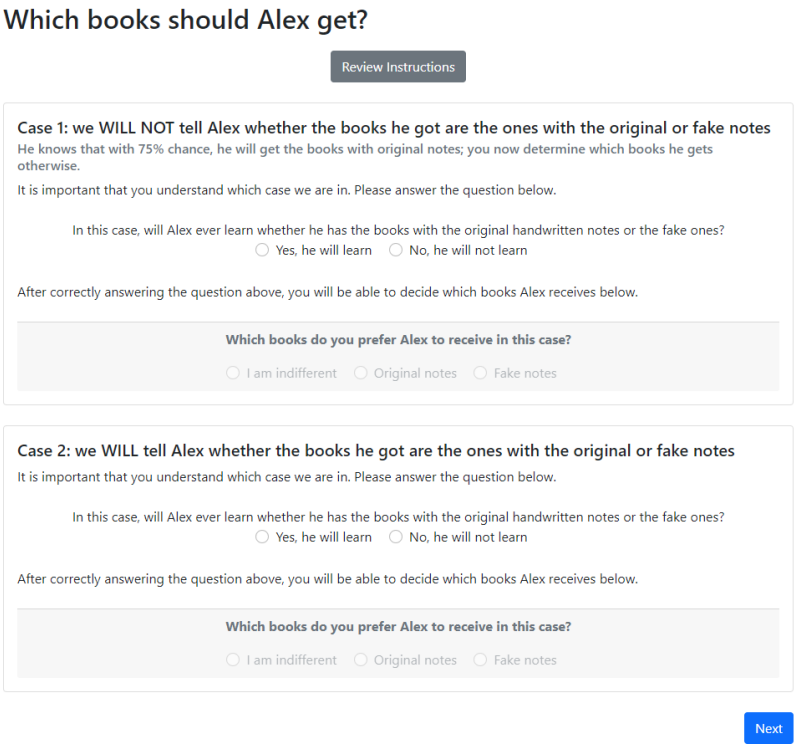


(a) *Baseline* treatment



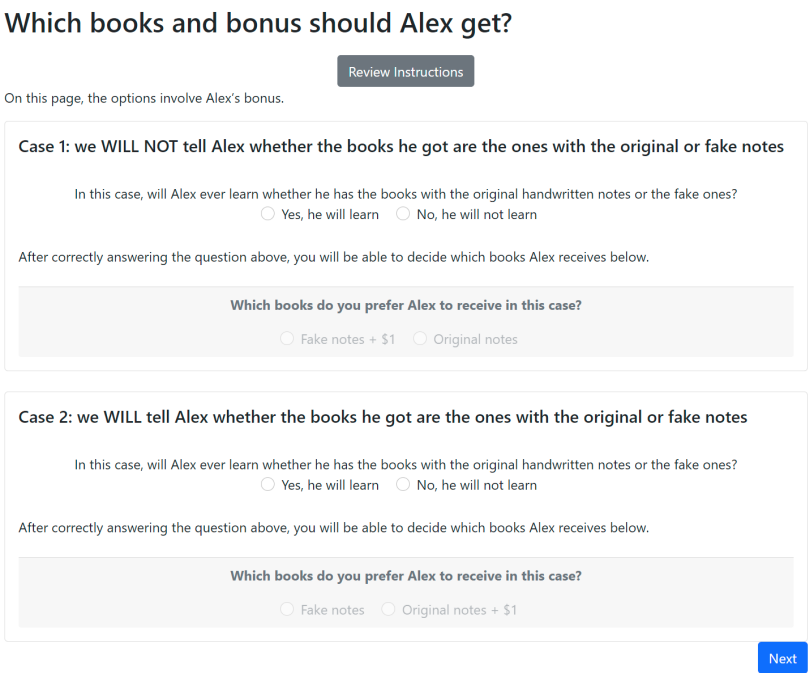
(b) *LowMS* treatment

Figure 9

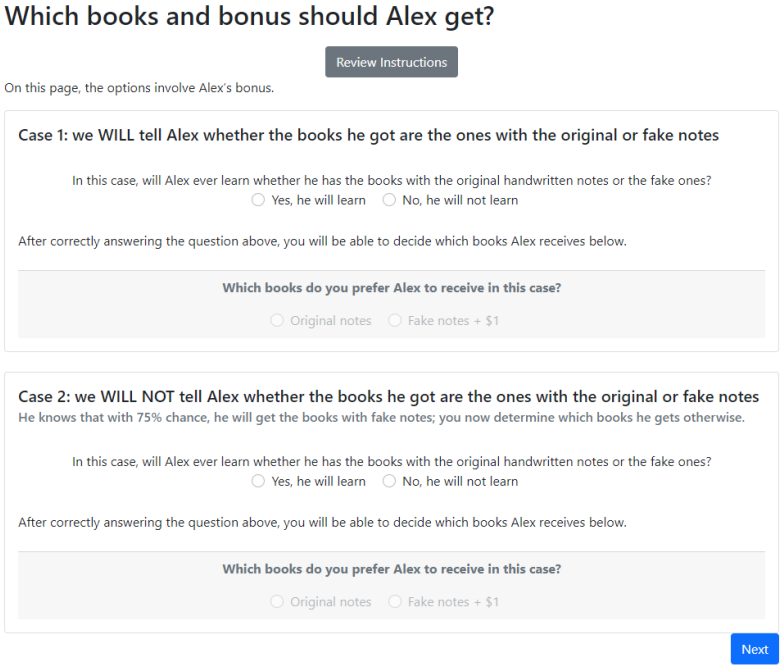


(c) *HighMS* treatment

Figure 9

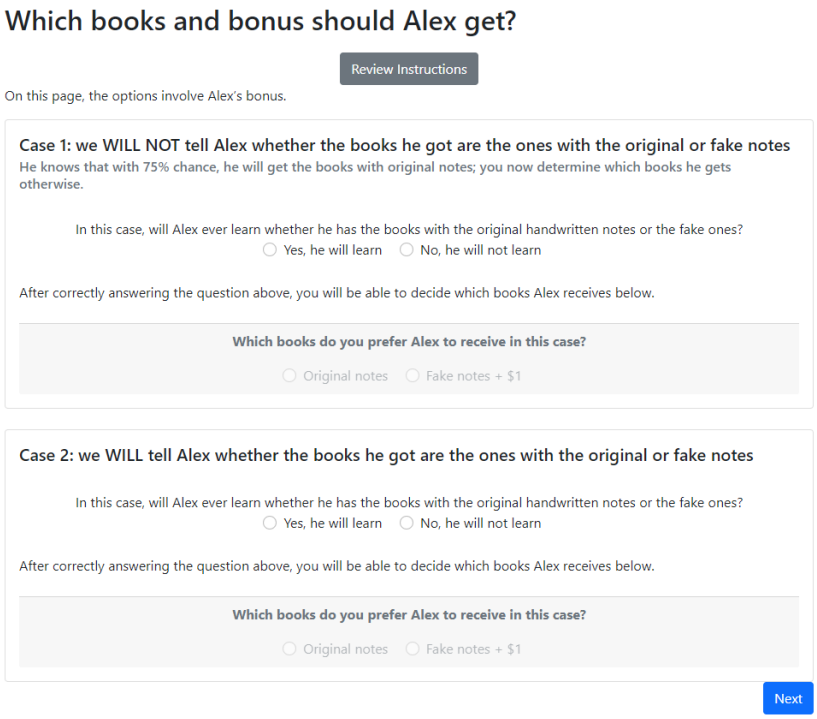


(a) *Baseline* treatment



(b) *LowMS* treatment

Figure 10



(c) *HighMS* treatment

Figure 10

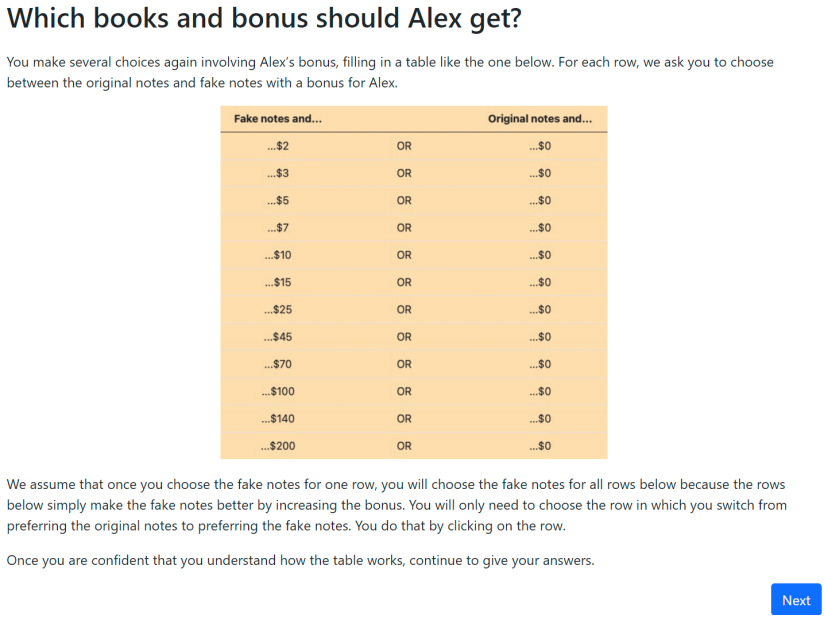
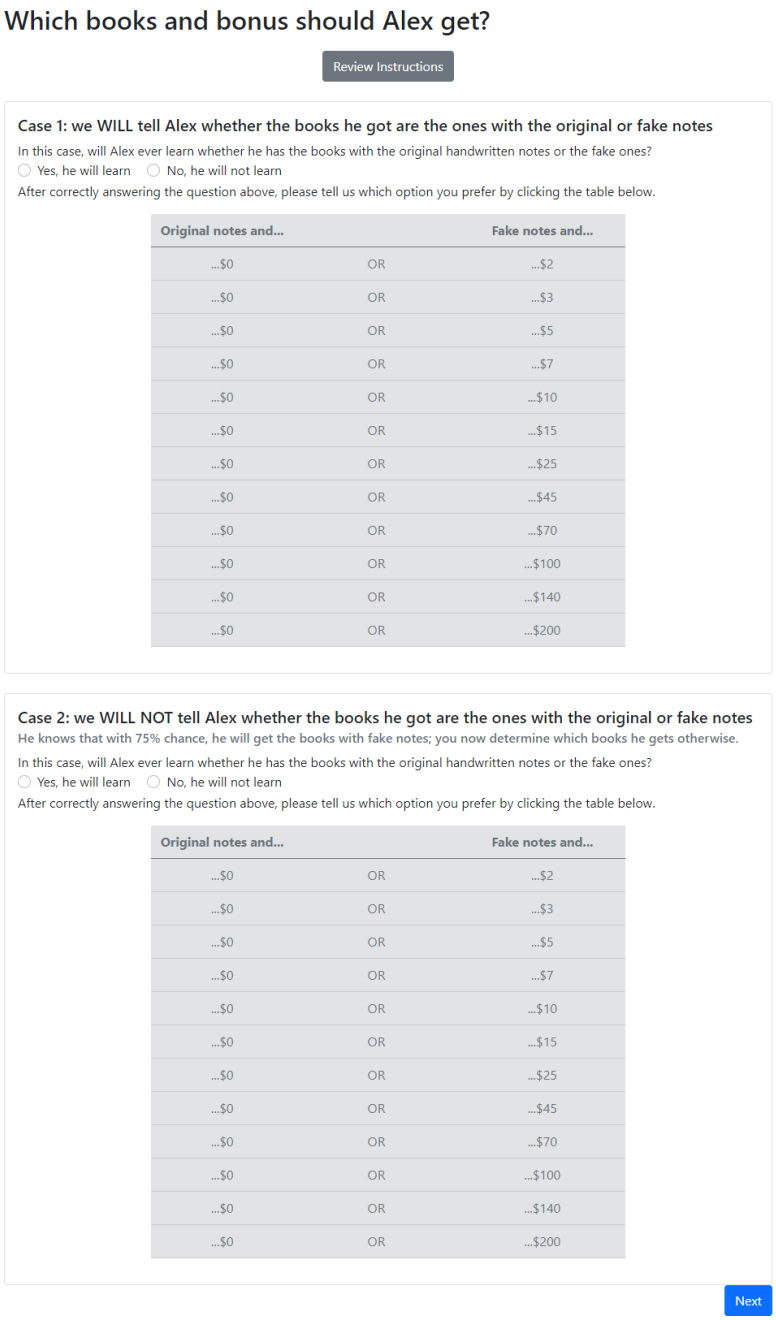


Figure 11

(a) *Baseline* treatment

Figure 12



(b) *LowMS* treatment

Figure 12

(c) *HighMS* treatment

Figure 12

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

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 14

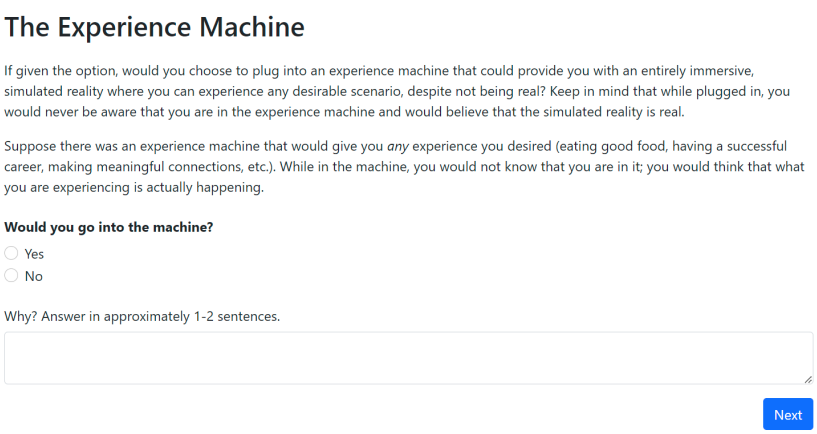


Figure 15

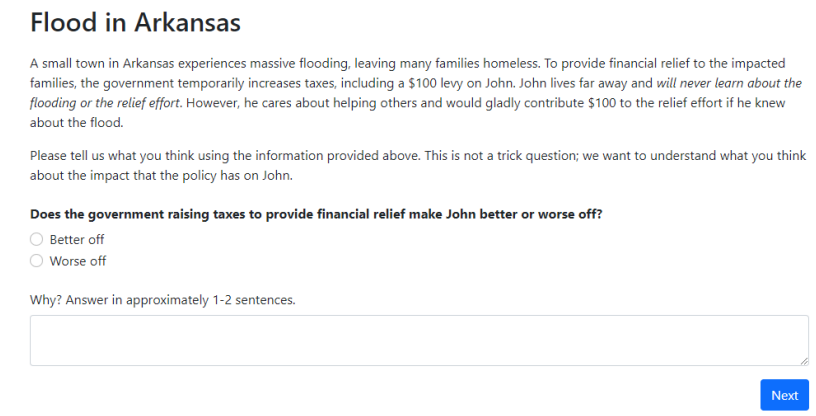


Figure 16

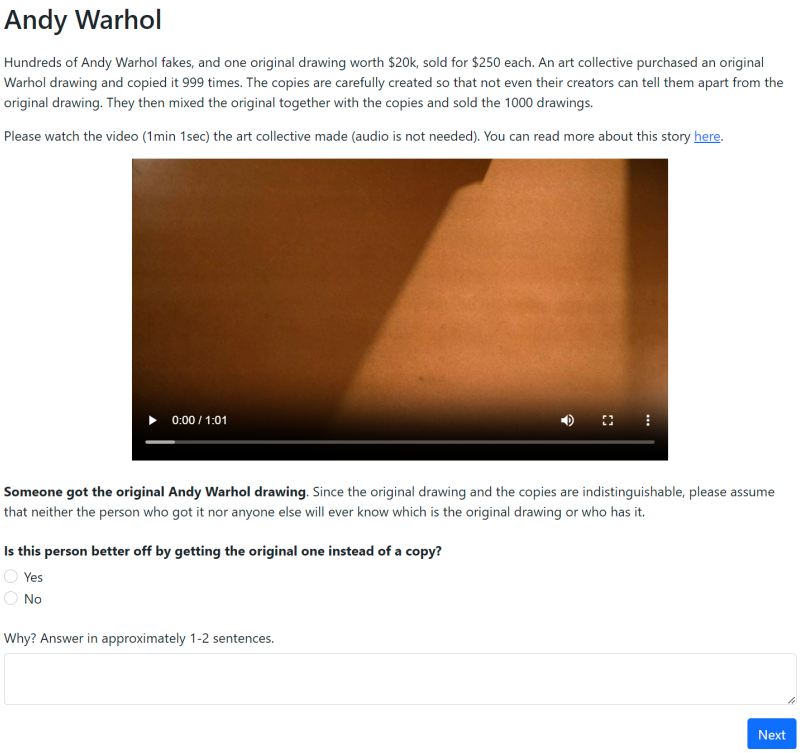


Figure 17: Click here to see the video: `https://moforgeries.org/images/showcase.mp4` (accessed 2023/08/23).