Classification exercise for: "Which notion of welfare do people adhere to when choosing for others?"

Gonzalo Arrieta Lukas Bolte

September 5, 2023

Hello!

Thank you for agreeing to help us with our research project.

We ran an experiment, and we would like you to help us judge participants' answers.

To be able to help us, it is important that you understand the idea of the project and the experiment. We appreciate it if you could read the description below carefully.

Project description

Suppose you get a book by your favorite author with a personalized note.

We can probably all agree that getting this book is good for your "welfare" (a difficult-to-define term that describes how good your life is). But why?

When receiving the book with the note, two things happen:

1) You now believe you have a book with a note by the author.

2) You actually have a book with a note by the author.

Philosophers disagree on which one of the two consequences actually matters for your welfare (or if both do). There are two schools of thought in both extremes of the disagreement—and a spectrum in between. One school of thought, called mental statism, says that your welfare is a function of your mental state only (as the name suggests), and so what you believe to be true is all that matters, and what is actually true does *not* matter. By this school of thought, only point 1) above matters for your welfare, regardless of what point 2) is. Another school of thought takes a so-called preference theoretic approach and argues that welfare has to do with having your preferences satisfied—whether or not you know about it. By this school of thought, only point 2) above matters for your welfare, regardless of what point 1) is.

In this project, we are empirically testing which school of thought participants recruited in an online experiment adhere to when choosing for others.

Going back to the example of the books with the notes, suppose that the notes might be fake (but indistinguishable). If you will never know whether the note was actually written by the author or not, does whether it is or is not affect your welfare? Or does it not matter, given that you will never know? What we are asking here is, if you fix what 1) is (i.e., fix the beliefs), does changing 2) affect your welfare?

Mental statism suggests that *it doesn't matter* whether the book has the fake or the original note since you never learn about it (i.e., it does not matter what 2) is; only 1) matters). The preference-theoretic welfare notion suggests you are better off if the book actually has the original note by the author since you prefer the original note, and hence satisfying those preferences increases welfare, *regardless of whether you know it or not* (i.e., it does not matter what 1) is; only 2) matters).

In our experiment, we then ask someone who cares about you (well, not you specifically, but another participant called "Alex" that we recruited to get books!) whether (and by how much) they would decrease a random bonus payment you will receive for you to have the original note (i.e., how much are they willing to change 2) given that 1) won't change?). If they do, then this might suggest that they lean toward a preference-theoretic welfare notion since they want to change what is actually true even if you never learn about it. If they don't, then they might agree more with mental statism, saying that it really doesn't matter whether you have the original not or not if what you believe doesn't change. Of course, people can be anywhere in between the two extreme schools of thought.

Experimental design

To best understand the experiment, we think it would be easiest if you just went through it. Please go through one of our treatments using this link: `https://decision-making-study-2009-e671db18ce5c.herokuapp.com/demo`
Please:

1. Click on the link above.
2. Click on `welfare` when choosing between `welfare` and `welfare_dev`
3. Click on the "Session-wide demo link" that is generated to start the study.

You can do this as many times as you want. So feel free to go back to the study in case you want to check something.

You just went through one of the three treatments; there are two more. The idea is that whether participants want Alex to have the original notes when Alex won't learn might depend on what Alex expects to get. To test this, our treatments vary what Alex expects to get. In the `high` treatment, we tell participants that if Alex doesn't learn, he knows that he will get the original notes with a probability of at least 75%. In the `low` treatment, we tell participants that if Alex doesn't learn, he knows that he will get the original notes with a probability of at most 25%.

The treatments are called `low, middle, high`, referring to the induced likelihood of Alex receiving the original notes if he won't learn (the `middle` treatment is the one you can go through in the link above).

Figure 1 below shows how we induce the treatment variation in the experiment. In particular, it shows how we induce the `low` belief. The other treatment variations are similar.

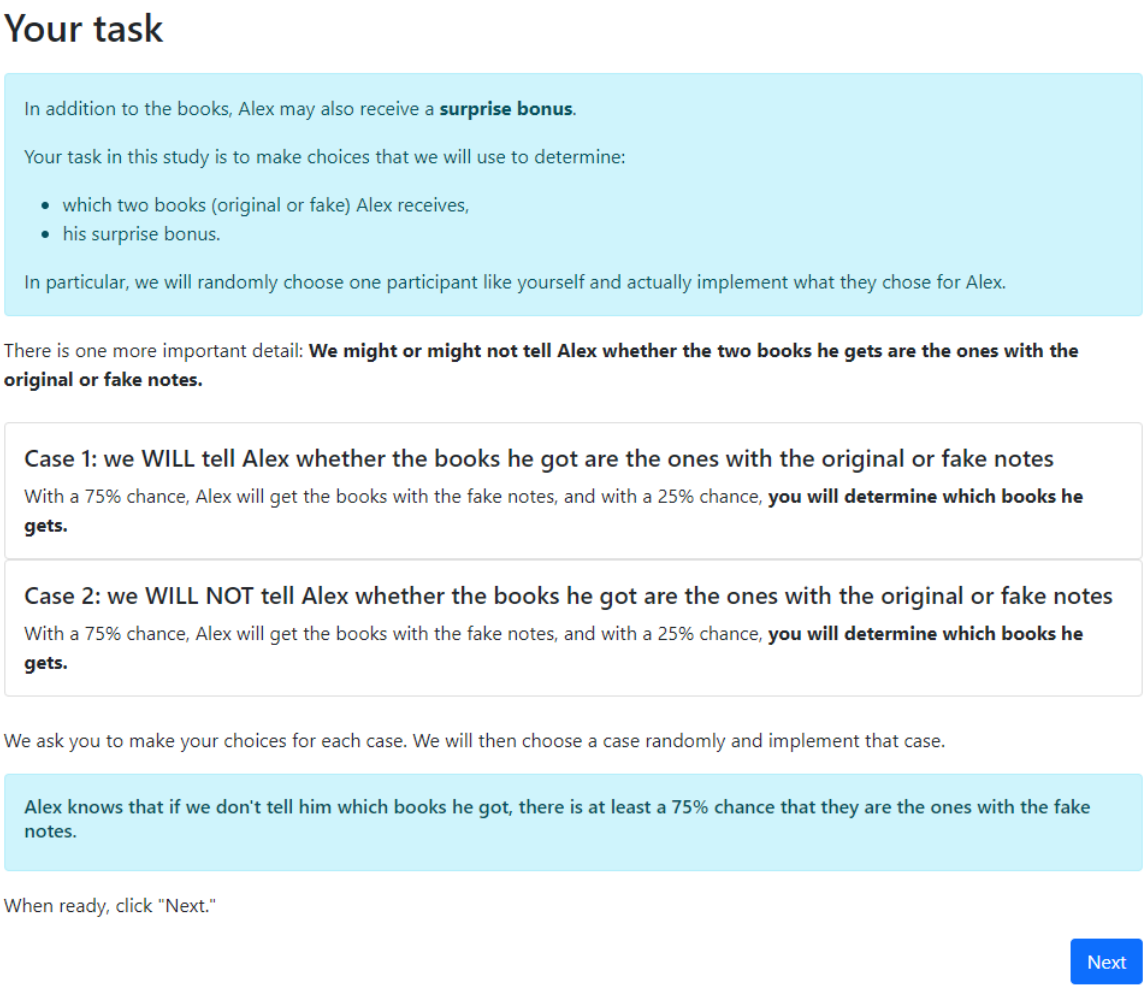

Figure 1: Screenshot from the experiment

Outcome variables

We have a few outcome variables. The most important ones, and the ones that we need your help with, are these:

- `TradWTP`: How much do participants value Alex getting the original notes when Alex will learn? This is a range. For example, [70, 100] means they prefer to give Alex the books with fake notes over the ones with the original notes if the bonus that comes with the fake notes exceeds some value between 70 and 100 (we don't know the exact value, and that's okay). We know that for a bonus that exceeds $100, they definitely prefer to give Alex the books with fake notes and the bonus, and for bonuses below $70, they definitely prefer to give Alex the ones with original notes and no bonus. If you're familiar with the "willingness-to-pay" language, this range represents the participants' willingness-to-pay to give Alex the book with original notes over the one with fake ones, in the case when Alex will know what he gets.
- `ESWTP`: How much do participants value Alex getting the original notes when Alex will NOT learn? This is a range, too, and works in the same way. This range represents the participants'

willingness-to-pay to give Alex the book with original notes over the one with fake ones, in the case when Alex will not know what he gets.

Each participant answers both `TradWTP` and `MPLWhy`. In the experiment, we refer to them as Case 1 and Case 2, with which case corresponds to what is being randomized.

- `MPLWhy`: Participants are prompted to explain their answers. When they gave the same answers for both cases, we asked:
  - You gave the same 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)

  And when they gave different answers, we asked:
  - 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)

  Remember, which case corresponds to `TradWTP` and which to `MPLWhy` is randomized. So in participants' answers for `MPLWhy`, some may refer to Case 1 meaning the case corresponding to `TradWTP` while others refer to Case 1 meaning the case corresponding to `ESWTP`.

Classification exercise

We want you to help us assess the quality of the data we have collected. To do so, we need to, for each response, answer the following four questions:

1. Is the text inconsistent with the responses? [Yes, No—accepting near-consistency]
2. Does the text exhibit a misunderstanding of a part of the instructions? [Yes, No]
3. Does the text exhibit an understanding of the difference between the two cases, either explicitly or implicitly? [Yes, No]
4. Is the text likely to be generated by ChatGPT or similar? [Yes, No]

Here are examples of how to answer each question:

1. Is the text inconsistent with the responses? [Yes, No—accepting near-consistency]

Example 1:
`TradWTP` = (25, 45); `ESWTP` = (25, 45); `Treatment` = low.

`MPLWhy` = *"Because it doesn't matter whether he knows or not. The outcome is essentially the same to me."*

This participant is not inconsistent because they say it doesn't matter whether Alex knows or not, and indeed, they give the same range in both cases.

Example 2:
`TradWTP` = (200, inf); `ESWTP` = (-inf, -1); `Treatment` = middle.

`MPLWhy` = *"In case 1 since we won't tell him what he got, fake notes would work fine. In case 2 we tell him so it needs to be originals."*

This participant is also not inconsistent for our purposes. However, note that they say "fake notes would work fine" but then seem to strictly prefer fake notes over original ones because they choose the fake ones over the original ones and a dollar.

Example 3:
`TradWTP` = (15, 25); `ESWTP` = (140, 200); `Treatment` = middle.

`MPLWhy` = *"Tring to maximize profits for Alex..he will not know if they are fake or not."*

This participant is inconsistent because they say they want to maximize profits for Alex and that "he will not know if they are fake or not," but then pick the second highest ESWTP range of (140, 200), which is inconsistent with both, trying to maximize Alex's profits, and saying it doesn't matter if he won't know if they are fake or not. The suspicion is that this participant probably didn't understand how the multiple price list works.

2. Does the text exhibit a misunderstanding of a part of the instructions? [Yes, No]

Example 4:
`TradWTP` = (45, 70); `ESWTP` = (-inf, -1); `Treatment` = high.

`MPLWhy` = *"I don't want Alex to be told about the books all the time."*

This participant exhibits a misunderstanding of a part of the instructions. While it may be the participant's preference to not have Alex be told about the books all the time, this cannot be a reason for *why* they made their choices since the choices do not affect whether Alex is told or not. This strongly suggests that this participant thought they could influence whether Alex is told or not, i.e., they misunderstood the instructions.

3. Does the text exhibit an understanding of the difference between the two cases, either explicitly or implicitly? [Yes, No]

Example 1 from above is an example of a participant who exhibits an understanding of the difference between the two cases because when asked "why did you pick differently for each case?" their response points directly to the difference between the cases: "Because it doesn't matter whether he knows or not. The outcome is essentially the same to me."

Example 5:
`MPLWhy` = *"I think he would like the original notes based on his interest in economics."*

4. Is the text likely to be generated by ChatGPT or similar? [Yes, No]

This participant does not exhibit an understanding of the difference between the two cases. Note that there is nothing wrong with their response, it just doesn't show that they understand the difference, so the answer to this question should be "no."

Example 6:

`MPLWhy` = *"In Case 1, where Alex is not informed about whether the books have original or fake notes, I focused on providing advice that would allow Alex to make an informed decision without any preconceived bias. This approach was aimed at helping Alex navigate the situation independently. In Case 2, where Alex is informed about the authenticity of the notes, I tailored my responses to align with the assumed scenario that Alex is aware of the notes' authenticity. This allowed me to provide more specific guidance based on the information available to Alex."*

This participant has likely used a program such as ChatGPT. The text just does not seem like something a human on Prolific would write.

We need two research assistants to classify all answers ($N = 1,478$). Our best guess is that this will take about 15 hours.

The first research assistant adds their classifications to this list.

List 1:

- `https://docs.google.com/spreadsheets/d/1FdBM6eQ9vIBOGYqkzRbUxTnyiCTjIyEFUF78RaylJqs/edit?usp=sharing`

The second adds their classifications to this list.

List 2:

- `https://docs.google.com/spreadsheets/d/1NhwweBpLo2kEq_7mRBx_S7odnpCfi5Dp7SSg_Bok0ao/edit?usp=sharing`

Both, please start at the top and work your way down. It is also important that you do these classifications independently, i.e., that you don't discuss them with each other.

Questions, comments, concerns

We are two principal investigators on this project: Gonzalo Arrieta, who is a Ph.D. at Stanford University, and Lukas Bolte, who is a postdoctoral fellow in the Department of Social and Decision Sciences at CMU. You can contact either one of us.

Gonzalo's email is garrieta@stanford.edu and Lukas' email is lbolte@andrew.cmu.edu.

Please feel free to ask questions, give comments, or raise concerns.