### **Study Information**

1. **Title:** Expectations of artificial moral advisors
2. **Description**

In this experimental study, we look at how people judge (artificial) moral advisors who make different patterns of moral judgments in dilemmas that contrast utilitarian and non-utilitarian principles in both instrumental harm and impartial beneficence. By asking both about trust and expectations of which advisor is most likely to be AI, we will assess whether people expect artificial moral advisors to be more consistently utilitarian in their moral decision-making. In the study, participants will be presented with moral advice on different variants of a moral dilemma about either instrumental harm or impartial beneficence. We manipulate the moral advice to be either consistently deontological, consistently utilitarian, normatively sensitive, or non-normatively sensitive. Participants will report how trustworthy, competent, and empathic they think the advisor is and whether they think the advisor is more likely to be an AI or a human.

1. **Hypotheses**

Our primary hypotheses draw on previous work and focus on instrumental harm dilemmas (e.g., Everett et al. 2016; Myers & Everett, 2025).

**H1:** At baseline, participants will trust the consistently utilitarian advisor less than other advisors for instrumental harm dilemmas

**H2:** Participants will rate the consistently utilitarian advisor as more likely to be AI than other advisors for instrumental harm dilemmas

While previous work has mostly focused on instrumental harm, there have been mixed findings with impartial beneficence. Some work suggests that (artificial) advisors who make utilitarian decisions about impartial beneficence might also be distrusted (Myers & Everett, 2025), but other work finds that humans who make such decisions are trusted more (Everett et al. 2021). Because of these mixed findings, we do not pre-register explicit hypotheses about impartial beneficence here, though we will explore these patterns in the paper.

### **Design Plan**

1. **Study type**

Online experiment

1. **Blinding**

There is no blinding in this study

1. **Study design**

The study follows a 2x4 between-subjects experimental design. Participants will be randomly presented with a moral dilemma about either:

1. Instrumental harm; or
2. Impartial beneficence

Participants will also be presented with one of four advisor types:

1. Consistently deontological
2. Consistently utilitarian
3. Normatively sensitive
4. Non-normatively sensitive
5. **Randomization**

As well as randomizing participants to experimental conditions, we will also randomly choose the particular moral dilemma from a wider set of dilemmas. Participants in the instrumental harm group will either see the Bomb dilemma, the Enemy Spy dilemma, or the Hostage dilemma. Participants in the impartial beneficence group will either see the Donation dilemma, the Marathon dilemma, or the Volunteering dilemma. Further randomization includes counterbalancing of the dilemma variants (morally irrelevant first vs. morally relevant first) and the order of the advisor types for the direct comparison after the between-subjects manipulation.

### **Sampling Plan**

1. **Existing data**

Registration prior to creation of data: As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.

1. **Data collection procedures**

Participants will be recruited through Prolific Academic. Data will be collected online via Qualtrics survey forms.

1. **Sample size**

We will recruit 900 participants.

1. **Sample size rationale**

We determined our target sample size using a power analysis simulation based on pilot data. Our simulation suggested that a sample size of N=100 per condition would be required to detect a difference in perceived trust (at baseline) between the consistently utilitarian advisor and the other advisors with at least 80% power. This results in an overall target sample size of N=800. We aim to sample an additional 100 participants to account for exclusions.

1. **Manipulated & measured variables**

Manipulated (between-subjects)

* **Dilemma type:** instrumental harm vs. impartial beneficence
* **Advisor type:** consistently utilitarian, consistently deontological, normatively sensitive, or non-normatively sensitive
* **Counterbalancing:** morally relevant change first vs. morally irrelevant change first
* **Order of advisors for direct comparison:** [advisors randomly shuffled]

Outcome measures (1-7 Likert scales)

Several outcome measures are asked at baseline and then again after seeing the advisor respond to the morally relevant and morally irrelevant dilemma variants:

* **Trust:** “How trustworthy do you think this advisor is?”
* **Trust on other issues:** “Based on their advice, how willing would you be to trust this advisor on other issues?”
* **Empathy:** “How empathic do you think this advisor is?”
* **Competence:** “How competent do you think this advisor is?”

Other outcome measures are asked only after presenting the dilemma variants:

* **Likely human:** “How likely is this advisor to be AI or human?”
* **Surprised AI:** “How surprised would you be if you found out this advisor is AI?”

The final outcome variables are measured when the participant is asked to directly compare all four advisor types after the between-subjects manipulation:

* **Compare trust:** “How trustworthy do you think each of these advisors are?”
* **Compare likely human:** “How likely are each of these advisors to be AI or human?”

Moral judgements

At the end of the study, we will ask participants to report their own judgement of the moral dilemma:

* **Judgement baseline:** “Do you think that [person] should [utilitarian option]?”
* **Judgement morally-relevant:** “Do you think that [person] should [utilitarian option]?”
* **Judgement morally-irrelevant:** “Do you think that [person] should [utilitarian option]?”

Demographic measures

* **Age:** self-reported age in years
* **Gender:** self-reported gender identity

### **Analysis Plan**

1. **Analysis plan**

We will fit Bayesian multilevel ordinal models to the data using the brms R package (see [here](https://paulbuerkner.com/brms/)). We will assess support for our hypotheses by determining whether 95% credible intervals for key comparisons exclude zero. All models will use weakly regularising priors to impose conservatism on parameter estimates.

To test H1, we will fit the following model:

|  |
| --- |
| fit1 <- brm(formula = trust ~ 1 + advisor\_type \* time \* dilemma\_type + (1 | id) +  (1 + advisor\_type \* time | dilemma),  data = d, family = cumulative, ...) |

We will compare trust at baseline between the four advisor types.

To test H2, we will fit the following model:

|  |
| --- |
| fit2 <- brm(formula = likely\_human ~ 1 + advisor\_type \* dilemma\_type + (1 | id) +  (1 + advisor\_type | dilemma),  data = d, family = cumulative, ...) |

We will compare human-likelihood ratings between the four advisor types.

1. **Data exclusion**

Participants will be excluded if any of the following apply:

* they close their browser window without completing the experiment
* they have a low Captcha score (<0.5)
* they fail the attention check question at the beginning of the study
* they write a correct answer to the briefly-presented open-ended question “Prove that the square root of 2 is an irrational number” at the beginning of the study (indicating that they are a bot or are using AI to answer the survey)

1. **Missing data**

Since we are requesting responses on Qualtrics rather than forcing responses, there is a possibility that participants may not respond to some of the questions. In practice, this isn’t likely to be a frequent occurrence. But if this happens, we will exclude those cases from the analysis.

**References**

Everett, J. A., Colombatto, C., Awad, E., Boggio, P., Bos, B., Brady, W. J., ... & Crockett, M. J. (2021). Moral dilemmas and trust in leaders during a global health crisis. *Nature Human Behaviour*, *5*(8), 1074-1088.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

Myers, S., & Everett, J. A. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition*, *256*, 106028.