# Trust in Artificial Moral Advisors across Cultures

Scott Claessens1, Konrad Bocian2, Paulo Sérgio Boggio3,4, Grégory Fiorio5, Léo Fitouchi6, Zeynep Genç1, Ivar R. Hannikainen7, Lea C. Kamitz8, Tamino Konur1, Ethan Landes1, Peng Liu9,10, Katarzyna Miazek2, Waldir M. Sampaio3,4, Jorge Suárez7, Pierce Veitch1, Onurcan Yilmaz11, Hongbo Yu12, and Jim A.C. Everett1

1School of Psychology, University of Kent

2Department of Psychology in Sopot, SWPS University

3Social and Cognitive Neuroscience Laboratory, Mackenzie Presbyterian University

4National Institute of Science and Technology on Social and Affective Neuroscience (INCT-SANI)

5Institut Jean Nicod, Département d’Études Cognitives, École Normale Supérieure, Université PSL

6Institute for Advanced Study in Toulouse, Department of Social and Behavioral Sciences, Toulouse School of Economics

7Department of Philosophy, University of Granada

8IPPPRI, Anglia Ruskin University

9Center for Psychological Sciences, Zhejiang University

10Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore

11Department of Psychology, Kadir Has University

12Department of Psychological and Brain Sciences, University of California Santa Barbara

# Author Note

Scott Claessens  <https://orcid.org/0000-0002-3562-6981>

Konrad Bocian  <https://orcid.org/0000-0002-8652-0167>

Paulo Sérgio Boggio  <https://orcid.org/0000-0002-6109-0447>

Grégory Fiorio  <https://orcid.org/0009-0008-3731-5678>

Léo Fitouchi  <https://orcid.org/0000-0002-1724-0129>

Zeynep Genç  <https://orcid.org/0009-0004-8586-5771>

Ivar R. Hannikainen  <https://orcid.org/0000-0003-0623-357X>

Lea C. Kamitz  <https://orcid.org/0000-0003-1290-5088>

Tamino Konur  <https://orcid.org/0009-0005-4706-1200>

Ethan Landes  <https://orcid.org/0000-0002-1186-1717>

Peng Liu  <https://orcid.org/0000-0003-4929-0531>

Katarzyna Miazek  <https://orcid.org/0000-0001-9072-2688>

Waldir M. Sampaio  <https://orcid.org/0000-0002-6066-4314>

Jorge Suárez  <https://orcid.org/0000-0002-6215-2064>

Pierce Veitch  <https://orcid.org/0009-0005-3364-7470>

Onurcan Yilmaz  <https://orcid.org/0000-0002-6094-7162>

Hongbo Yu  <https://orcid.org/0000-0002-3384-7772>

Jim A.C. Everett  <https://orcid.org/0000-0003-2801-5426>

This pre-print is currently not yet peer-reviewed and may differ from the final version. Word count (excl. abstract, methods, figure legends, and references) = 4,908 words.

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: *Scott Claessens***:** conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, and writing – review & editing. *Konrad Bocian***:** methodology and writing – review & editing. *Paulo Sérgio Boggio***:** methodology and writing – review & editing. *Grégory Fiorio***:** methodology and writing – review & editing. *Léo Fitouchi***:** methodology and writing – review & editing. *Zeynep Genç***:** methodology and writing – review & editing. *Ivar R. Hannikainen***:** methodology and writing – review & editing. *Lea C. Kamitz***:** methodology and writing – review & editing. *Tamino Konur***:** methodology and writing – review & editing. *Ethan Landes***:** methodology and writing – review & editing. *Peng Liu***:** methodology and writing – review & editing. *Katarzyna Miazek***:** methodology and writing – review & editing. *Waldir M. Sampaio***:** methodology and writing – review & editing. *Jorge Suárez***:** methodology and writing – review & editing. *Pierce Veitch***:** methodology and writing – review & editing. *Onurcan Yilmaz***:** methodology and writing – review & editing. *Hongbo Yu***:** methodology and writing – review & editing. *Jim A.C. Everett***:** conceptualization, funding acquisition, methodology, supervision, writing – original draft, and writing – review & editing

Correspondence concerning this article should be addressed to Jim A.C. Everett, School of Psychology, University of Kent, Keynes College, Canterbury CT2 7NP, United Kingdom, Email: [j.a.c.everett@kent.ac.uk](mailto:j.a.c.everett@kent.ac.uk)

# Abstract

Developers of artificial intelligence are already building prototypes for artificial moral advisors: autonomous systems designed to provide humans with recommendations on ethical issues. Yet it remains unclear whether people will actually trust and adopt these technologies. Here, we investigate perceptions of human and artificial moral advisors in a large-scale cross-cultural experiment (12 countries; *N* = 6,896). We show that people trust artificial advisors less than human ones, especially when their advice aligns with utilitarian rather than deontological principles. Nonetheless, people update their own moral judgments in the direction of the advice – regardless of whether it came from a human or AI advisor – and do so more when the advice is deontological. These findings reveal a psychological paradox: people trust artificial moral advisors less, yet still defer to them in practice. This dissociation raises serious questions about the potential influence of machines on human morality, with implications for AI governance.

*Keywords*: artificial intelligence, morality, trust, moral influence

# Trust in Artificial Moral Advisors across Cultures

Artificial intelligence (AI) – any system that uses an algorithm or a statistical model to perform tasks that usually require human intelligence – is reshaping the world around us. By leveraging the computational power of AI agents, we are able to build on our own skills and therefore perform tasks more easily, more quickly, and perhaps even better than we could alone. Students use large language models (LLMs) to help them write assignments, banks rely on AI to help with credit checks, and hospitals use AI tools to provide diagnoses for diseases. But could artificial intelligence also be used to help us make better moral decisions?

Artificial moral advisors (AMAs) refer to AI systems designed to assist humans in making ethical decisions, leveraging artificial intelligence to analyse moral dilemmas, engage in Socratic dialogue about moral matters[1](#ref-Lara2020), and/or provide recommendations based on established ethical theories, principles, and guidelines[2](#ref-Giubilini2018). The idea, at root, is simple: Artificial intelligence provides a tool that can help enhance human activity in many different domains, so why not use AI to help people make better moral decisions? Given that human moral judgments are influenced by a variety of factors of questionable normative relevance[3](#ref-Haidt2001), AMAs could instead serve a function akin to the “ideal observer”[4](#ref-Firth1952) by offering dispassionate and consistent judgments free from human biases[2](#ref-Giubilini2018),[5](#ref-Sinnott2021). Over the last decade, there has been growing philosophical attention to the possible future role of AMAs[1](#ref-Lara2020),[2](#ref-Giubilini2018),[5](#ref-Sinnott2021)–[7](#ref-Liu2022), and now with the release of openly available LLMs like ChatGPT, Gemini, and Claude, this interest has rapidly moved from hypothetical to practical.

While contemporary LLMs like ChatGPT are not designed specifically to assist in morality, their flexibility and accessibility means that they do already give advice on moral dilemmas[8](#ref-Krugel2023). Perhaps concerningly, ethical advice surreptitiously written by ChatGPT is rated as more moral, thoughtful, and correct than advice written by a human moral expert[9](#ref-Dillion2025). Beyond general-purpose LLMs, companies are already working on prototypes for AI-powered systems designed specifically to model morality. For example, the Allen Institute’s Delphi[10](#ref-Jiang2025) is trained on a large corpus of moral judgments, and it is now even possible to converse with an AI chatbot trained on the writings of the renowned ethicist Peter Singer (<https://www.petersinger.ai/>).

Questions about how we should create AMAs, whether we should at all, and the long-term consequences of these technologies remain very much open[1](#ref-Lara2020),[7](#ref-Liu2022),[11](#ref-Landes_preprint). However, even if we set aside the crucial question of whether people *should* trust AMAs, it still remains unclear whether people will actually trust and adopt AMAs in practice. We can broadly define trust as “accept[ing] vulnerability based upon positive expectations of the intentions or behaviour of another”[12](#ref-Rousseau1998), but how this will apply to people’s interactions with AMAs is far from clear. Even if people report that they trust AMAs, there is no guarantee that they would actually listen to the moral advice by changing their judgment, and indeed theoretical models of trust in AI often distinguish between attitudinal trust and its behavioral consequences, such as willingness to use the technology[13](#ref-Lalot2024). Given the increasing role that AI plays in life, and in particular claims that AI-powered moral advisors could help us make better moral judgments, we need to understand whether people trust AMAs, whether this trust depends on the kinds of moral decisions that the AMAs make, and whether people will actually defer to AMAs on moral issues. Our paper addresses all of these questions.

Will people trust AMAs? A large body of research, mostly conducted prior to the release of contemporary LLMs, has documented the phenomenon of algorithm aversion: the tendency for individuals to distrust AI relative to humans, even when the AI boasts identical or even superior performance[14](#ref-Dawes1979)–[17](#ref-Meehl1957). This algorithm aversion leads to a specific distrust in AI making moral decisions, driven by the perception that AI lacks internal experience[18](#ref-Bigman2018). These results suggest that people will distrust AI models that give advice on moral issues. However, these findings may no longer generalize in a post-ChatGPT era in which people are interacting with LLMs on a regular basis. Moreover, what limited research there has been on trust in AI in the moral domain has tended to focus narrowly on Western populations[19](#ref-Henrich2010), while moral norms, prevalence of AI, and morally relevant cultural values vary dramatically across societies[20](#ref-Barnes2024)–[22](#ref-Jackson2024). It therefore remains unclear whether people around the world will trust AMAs and whether levels of trust will vary across cultures.

Will people trust AMAs differently depending on the moral decisions that they make? For AMAs to be trustworthy, they must be aligned with some normative standard, and there is unsurprisingly a great deal of interest in how AI can and should be aligned with our moral values – and which ones specifically. A fundamental challenge that developers of AMAs will face is determining which kind of ethical framework to benchmark against, especially in moral dilemmas where different ethical frameworks endorse mutually exclusive actions. For example, is it morally acceptable to break normal prohibitions against murder in order to prevent harm to a greater number? Consequentialist theories such as utilitarianism focus on the ‘greatest good for the greatest number’, positing that only consequences matter when making moral decisions[23](#ref-Bentham1983)–[25](#ref-Singer1993). In contrast, non-utilitarian deontological theories claim that we also have to consider rights, duties, and obligations when making moral judgments: for example, even if murder might bring about good consequences, it should still be judged as wrong since we have a moral duty not to harm others[26](#ref-Fried1978)–[28](#ref-Ross1930). AMAs could theoretically provide advice in line with either of these philosophical frameworks, and given the ubiquity of moral dilemmas in everyday life, an artificial moral advisor will be expected to advise on the appropriate action in these fault lines of our morality.

Deontological and utilitarian theories are both ethical theories, and so competing judgments in moral dilemmas can therefore be described as “moral” in the sense that they align with normatively sensitive standards of moral behavior. And yet, we know that people do not judge other humans who endorse deontological and utilitarian decisions as equally moral and trustworthy. In particular, a growing body of work suggests that people who give characteristically deontological responses to sacrificial moral dilemmas (by rejecting instrumental harm) are seen as more trustworthy than those who give characteristically utilitarian responses (by endorsing instrumental harm)[29](#ref-Brown2019)–[35](#ref-Sacco2017). A potential explanation for this asymmetry is that people who make deontological decisions are perceived as warmer, more predictable, and more committed to their social partners[31](#ref-Everett2016),[32](#ref-Everett2018),[34](#ref-Rom2017),[36](#ref-Turpin2021), whereas people who make utilitarian decisions are perceived as colder, less moral, and less empathic[34](#ref-Rom2017),[37](#ref-Kreps2014),[38](#ref-Uhlmann2013).

Based on this previous work, we might assume that people will trust AMAs that give deontological moral advice more than AMAs that give utilitarian advice, just as they do for humans[31](#ref-Everett2016)–[33](#ref-Everett2021). However, there is also evidence that preferences for deontological decision-makers can be sensitive to the type of role one is judging. For example, people show an increased preference for deontological decision-making in close personal relationships, but show more acceptability for utilitarian decision-making from political leaders in impartial roles[32](#ref-Everett2018). Moreover, people are more likely to make utilitarian judgments when they want to signal their competence, rather than their warmth[39](#ref-Rom2018). Given that AMAs are proposed to be able to provide more impartial and competent advice in line with the idea of an ideal observer, it is therefore possible that people might actually prefer more utilitarian advice from AMAs. This would be in line with work showing that people trust autonomous vehicles more when they make utilitarian decisions, rather than deontological decisions, in switch-style sacrificial dilemmas[40](#ref-Young2019).

Of most relevance to this paper, Myers and Everett[41](#ref-Myers2025) presented the results of four pre-registered studies looking at how people trust AMAs compared to humans. In a sample of British participants, they found that participants trusted AMAs less than humans who gave the same moral advice; that participants trusted AMAs less if they endorsed rather than rejected instrumental harm (in line with utilitarian philosophy); that even when participants agreed with the specific decision that the AMA gave, there remained a tendency to expect that they would disagree with decisions made by the AMA in future; and that participants not only distrusted utilitarian AMAs, but expected AI to give utilitarian advice.

While these results are important in helping us understand how AMAs are differently trusted based on the kinds of decisions they make, there is still much we do not know. First, Myers and Everett[41](#ref-Myers2025) only looked at participants in the UK, limiting cross-cultural generalizability. While cross-cultural research has shown that countries differ in their cooperation with AI[42](#ref-Karpus2025), fears of it occupying various positions[43](#ref-Dong2024), and legal concerns[44](#ref-Ikkatai2024), it remains unclear how people across different countries that vary across different cultural indices perceive AMAs. Second, Myers and Everett[41](#ref-Myers2025) examined ratings of trustworthiness, willingness to trust the AI on other issues, and expected agreement in the future; however, we do not know if people will actually follow the advice from AMAs.

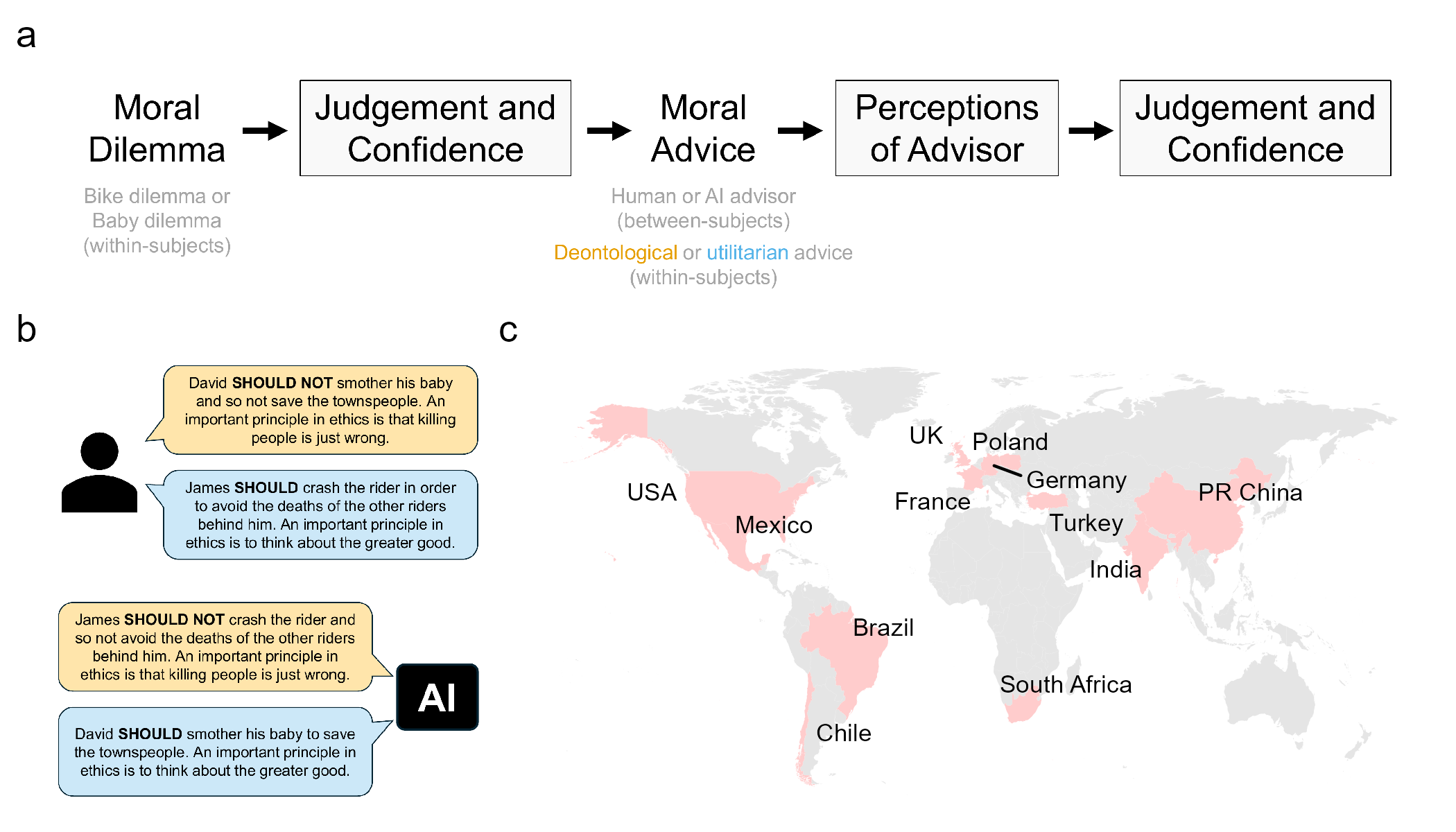
Will people listen to the advice provided by AMAs? Even if people report that they trust AMAs, this stated trust may not lead people to actually defer to the moral advice from AMAs in practice: perceived trustworthiness does not always entail trusting behavior. However, there is increasing evidence that people can be persuaded by contemporary consumer-grade LLMs in small but significant ways[45](#ref-Schoenegger2025). When directed to argue for a point in the non-moral domain, LLMs are capable of changing political beliefs[46](#ref-Durmus2024)–[48](#ref-Hackenburg2025), such as views on smoking bans and assault weapons bans[49](#ref-Bai2023), and can even reduce endorsement of conspiracy theories[50](#ref-Costello2024). There is less work on AI persuasion in the moral domain, but some research suggests that American participants listen to pre-generated LLM advice in sacrificial moral dilemmas by giving judgments in line with the advice provided[8](#ref-Krugel2023) and other work suggests that British participants update their moral judgments in everyday moral dilemmas when given advice from LLMs[51](#ref-Landes_under_review). This raises an important question: will people listen to AMAs less than humans, and might this vary depending on whether the AMAs give deontological or utilitarian advice?

## The Present Research

In this research, we explore whether people trust AMAs compared to human moral experts, whether this trust differs based on the kind of moral advice that the AMA gives in a sacrificial moral dilemma, and whether people update their own moral judgments when provided with advice from AMAs. Using a large cross-cultural sample ([Figure 1](#fig-study-overview)), we explore the generalizability of these effects across individuals and populations.

Figure 1

Overview of the study



*Note*. The flow of the experiment. We presented participants with a sacrificial moral dilemma (either the Bike dilemma or the Baby dilemma; within-subjects) and asked participants for their own moral judgment and their confidence in their judgment. We then presented participants with deontological or utilitarian advice (within-subjects) from either a human or an AI advisor (between-subjects). The advice was followed by several questions about the perceptions of the advisor, such as trust, and questions about participants’ revised moral judgment and confidence. Participants then repeated this flow for the other moral dilemma and the other direction of advice. (b) Examples of the deontological (orange) and utilitarian (blue) advice given by human and AI advisors in the study. (c) Countries included in the sample.

The flow of the experiment is shown in [Figure 1](#fig-study-overview). We presented participants with one of two sacrificial dilemmas reflecting a tension between utilitarian and deontological theories about the acceptability of harm to one to save a greater number of others (the “Bike” dilemma and the “Baby” dilemma; see Methods for more details). After asking participants what they thought should be done in the moral dilemma, we then presented participants with moral advice from an advisor that was described as either a human moral expert or an artificial moral advisor (between-subjects). The advisor either gave characteristically utilitarian advice to endorse the harm, justifying this in terms of maximizing the greater good, or gave characteristically deontological advice to reject the harm, appealing to certain moral actions being wrong even if they have good consequences. We asked participants to rate various perceptions of the advisor, such as trust. Participants were then presented with the dilemma again, and asked if, having seen this advice, they wanted to update their initial judgment. Finally, participants repeated this flow for the other sacrificial dilemma and the other direction of advice: for example, if participants initially saw an advisor give deontological advice for the “Bike” dilemma, they then saw a different advisor give utilitarian advice for the “Baby” dilemma. We conducted the experiment in 12 countries (Brazil, Chile, China, France, Germany, India, Mexico, Poland, South Africa, Turkey, UK, USA; overall *N* = 6,896) chosen to maximize variation in geographic spread, cultural backgrounds, and levels of AI progress according to global indices. In our pre-registration, we hypothesized that participants would trust AMAs less than human advisors, and utilitarian advisors less than deontological advisors (<https://osf.io/qa7gn/>).

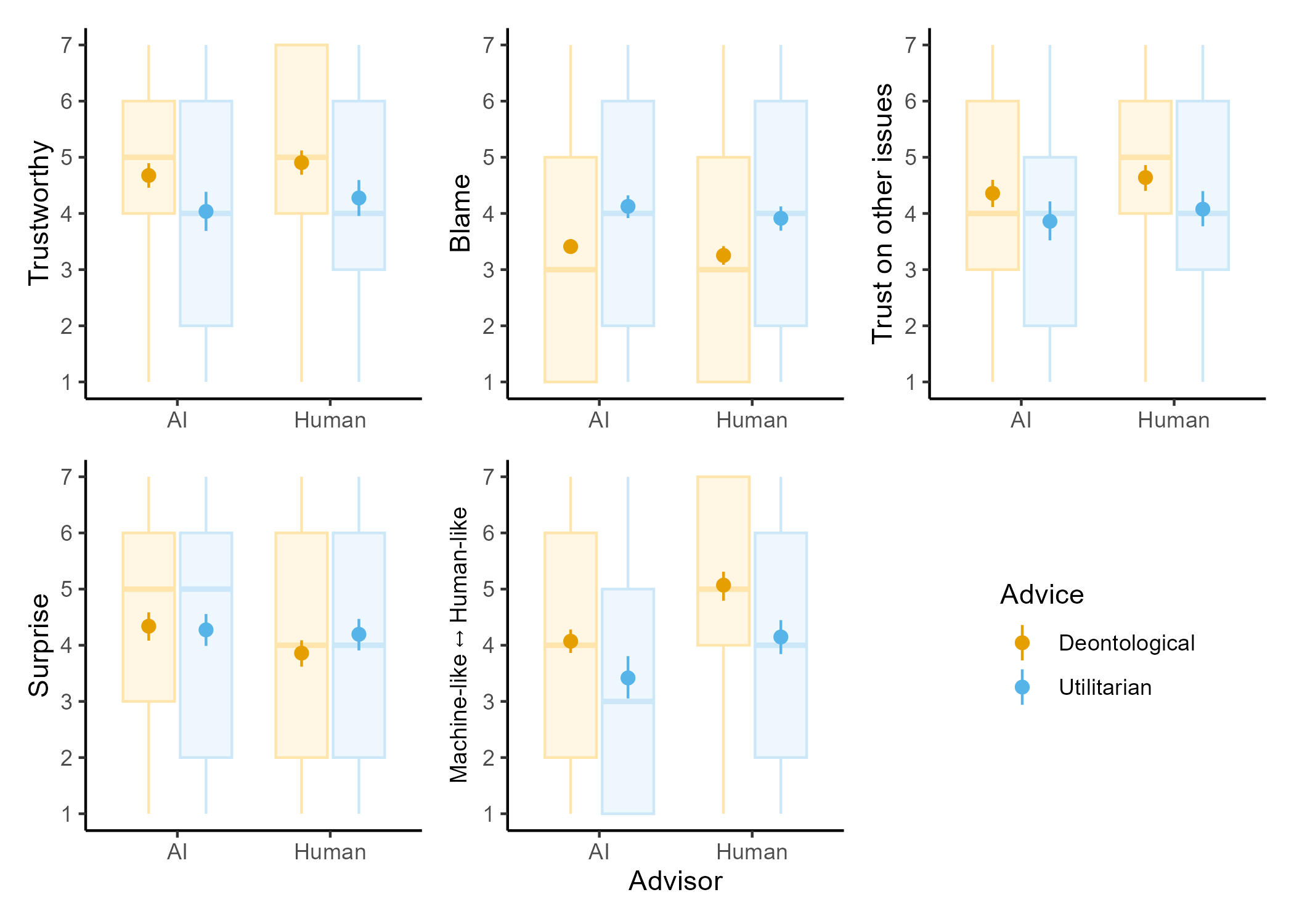
# Results

## Perceptions of the Moral Advisors

We first looked at participants’ overall perceptions of the moral advisors, pooling across dilemmas and countries ([Figure 2](#fig-means); [Table 1](#tbl-contrasts)). We found that the type of advice given (deontological vs. utilitarian) influenced participants’ perceptions of the advisors. In line with our pre-registered predictions, advisors who gave utilitarian advice in the sacrificial dilemmas were trusted less than advisors who gave deontological advice. Participants reported that they would blame someone else more for following utilitarian advice. In contrast to previous work[41](#ref-Myers2025), we did not find that people were more surprised when the AI gave deontological advice rather than utilitarian advice. However, we found that advisors were perceived as more machine-like when they gave utilitarian advice, suggesting at least some belief about the association of machines with utilitarian decision making.

Figure 2

Perceptions of AI and human moral advisors giving deontological (orange) and utilitarian advice (blue), pooling across dilemmas and countries



*Note*. Box and whisker plots show the distributions of the data, with horizontal lines representing sample medians, boxes representing interquartile ranges, and whiskers representing ranges. Point ranges show model-estimated marginal means, with points representing posterior medians and line ranges representing 95% credible intervals.

Table 1

Pairwise contrasts for perceptions of the moral advisors, pooling across dilemmas and countries

|  | Response | | | | |
| --- | --- | --- | --- | --- | --- |
| Contrast | Trustworthy | Blame | Trust other issues | Surprise | Human-like |
| **Effect of advice type** | | | | | |
| AI Utilitarian vs. AI Deontological | -0.59 [-0.92, -0.26] | 0.62 [0.42, 0.81] | -0.50 [-0.80, -0.17] | -0.05 [-0.21, 0.11] | -0.61 [-0.89, -0.32] |
| Human Utilitarian vs. Human Deontological | -0.60 [-0.92, -0.27] | 0.58 [0.37, 0.78] | -0.58 [-0.88, -0.24] | 0.28 [0.08, 0.47] | -0.90 [-1.18, -0.61] |
| **Effect of advisor type** | | | | | |
| Human Deontological vs. AI Deontological | 0.22 [0.10, 0.35] | -0.14 [-0.26, -0.01] | 0.29 [0.14, 0.44] | -0.40 [-0.53, -0.28] | 0.98 [0.69, 1.24] |
| Human Utilitarian vs. AI Utilitarian | 0.22 [0.09, 0.35] | -0.18 [-0.31, -0.05] | 0.21 [0.06, 0.37] | -0.06 [-0.23, 0.09] | 0.69 [0.40, 0.95] |
| **Interaction effect** | | | | | |
| Interaction effect | 0.00 [-0.14, 0.14] | -0.04 [-0.19, 0.10] | -0.07 [-0.21, 0.07] | 0.34 [0.17, 0.52] | -0.29 [-0.43, -0.15] |

*Note*. Numbers reflect differences in marginal means on the log-odds scale. For reference, on the log-odds scale, 0.23 is considered a small effect size, 0.54 a medium effect size, and 0.83 a large effect size[52](#ref-Chen2010). The bottom row represents the interaction between advisor type and advice type (i.e., the difference between the differences in the rows above). Main numbers are posterior medians, numbers in square brackets are 95% credible intervals.

In addition, we found that the type of advisor (human vs. AI) influenced participants’ perceptions. In line with our pre-registered predictions and with previous work on algorithm aversion in the moral domain[18](#ref-Bigman2018), artificial moral advisors were perceived as less trustworthy than human advisors. Participants reported that they would blame someone else more for following moral advice from an AI. Unsurprisingly, human advisors were seen as more human-like than AI advisors, especially when they gave deontological advice. Participants were also more surprised when deontological advice came from an AI, rather than from a human.

In addition to our main pre-registered analyses, we assessed the robustness of these effects in several ways. First, we looked at the results split across the two sacrificial dilemmas ([Supplementary Figure 1](#suppfig-means-by-dilemma); [Supplementary Table 1](#supptbl-contrasts-by-dilemma)). We found the same broad pattern of results for both dilemmas, though the type of advice (deontological vs. utilitarian) had a stronger influence on perceptions in the Baby dilemma compared to the Bike dilemma. Second, we looked at the results split by the order of the two blocks, finding little evidence of order effects ([Supplementary Figure 2](#suppfig-means-by-order); [Supplementary Table 2](#supptbl-contrasts-by-order)). Third, we determined whether the effects held in an “intention-to-treat” analysis, finding that our results did not differ when we included participants that we had excluded due to comprehension failures ([Supplementary Figure 3](#suppfig-means-itt); [Supplementary Table 3](#supptbl-contrasts-itt)).

## Variation across Individuals and Countries

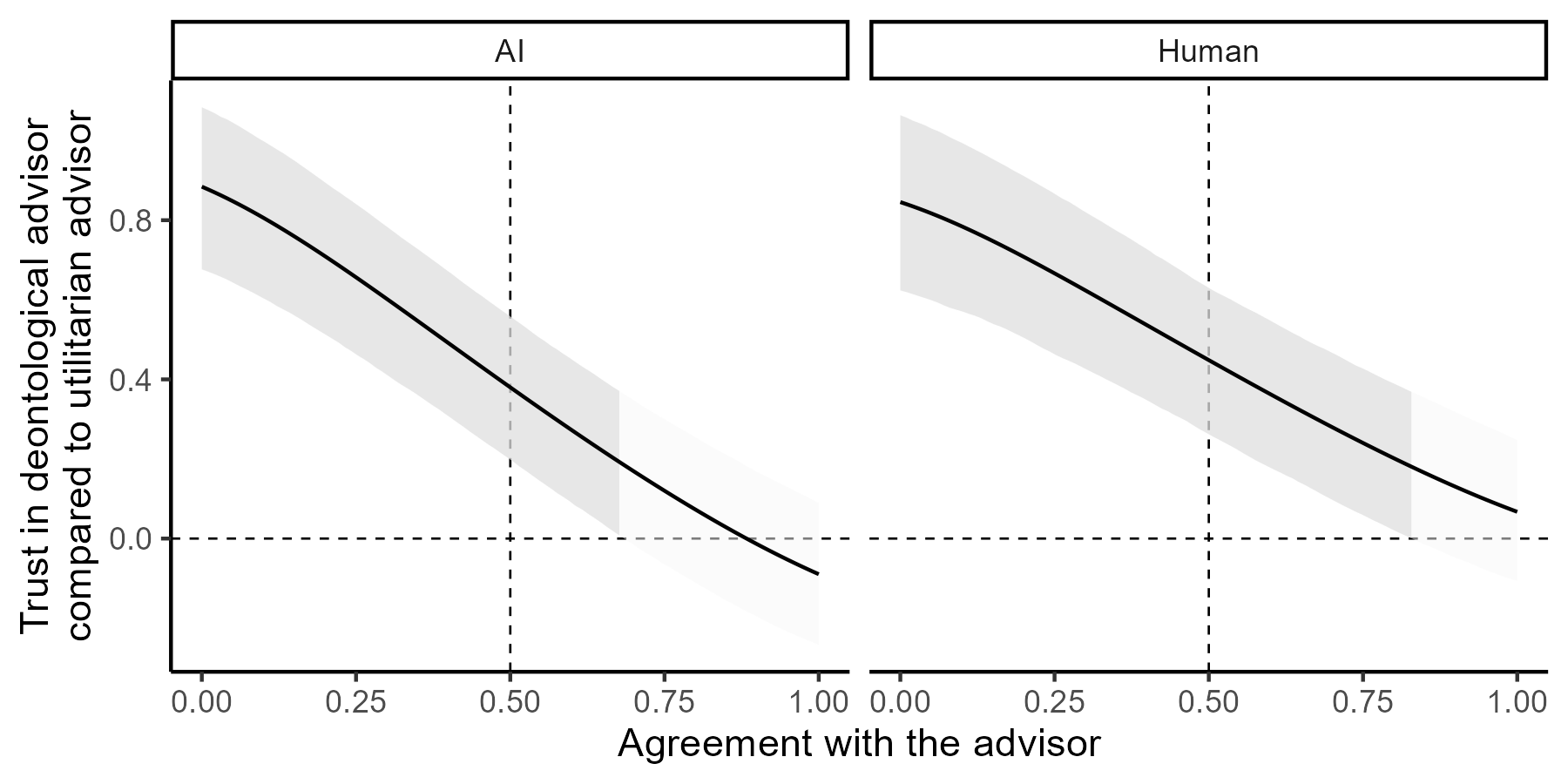
While overall effects capture broad trends, it is possible that our experimental manipulations worked differently for individuals with different demographics and moral preferences, and for countries with different cultural backgrounds. To investigate this potential heterogeneity, we ran several exploratory models including individual-level and cultural-level variables as moderators of the experimental effects, focusing on trust in the moral advisors as the outcome.

We first explored whether the experimental effects varied by demographics and other individual-level characteristics. In particular, we allowed participants’ religiosity and political ideology to moderate the effect of advice type (deontological vs. utilitarian) and allowed participants’ familiarity with AI and frequency of AI use to moderate the effect of advisor type (human vs. AI). We found tentative evidence that the experimental effects varied across these individual-level variables ([Supplementary Figure 4](#suppfig-ind-diffs)). In particular, we found that people trusted the deontological advisor more than the utilitarian advisor if they were more religious (interaction effect = 0.03, 95% credible interval [-0.01 0.07]) and more politically left-leaning (interaction effect = -0.05, 95% CI [-0.12 0.02]). We also found that people trusted the human advisor more than the AI advisor if they had less familiarity with AI (interaction effect = -0.02, 95% CI [-0.06 0.01]) and if they used AI less frequently (interaction effect = -0.03, 95% CI [-0.08 0.03]). However, these exploratory results should be interpreted cautiously as the credible intervals include zero.

Following previous work[41](#ref-Myers2025), we also explored whether the experimental effects varied depending on the participant’s own agreement with the moral advice. We were interested in participants’ agreement for two key reasons. First, the artificial moral advisors of the future are unlikely to know a specific user’s own preferences in a given moral dilemma in advance; and second, even if they could, the key theoretical appeal of artificial moral advisors is that they are thought to be able to provide impartial, disinterested advice that draws on normative principles, not merely repeat back what users already think. To explore the moderating role of agreement, we constructed an “agreement” variable from participants’ initial moral judgments, with zero indicating maximal disagreement with the moral advice and one indicating maximal agreement with the advice. In line with previous work[41](#ref-Myers2025), when we included this agreement variable as a moderator we found that participants trusted the deontological advisor more than the utilitarian advisor when they disagreed with the advisor (agreement = 0) and when they were neutral (agreement = 0.5; [Figure 3](#fig-agreement)). However, when they agreed with the advisor (agreement = 1), the 95% credible interval for the effect included zero. In other words, while participants trust advisors equally if their advice aligns with their own judgments, a relative distrust of utilitarian advisors over deontological advisors emerges when participants disagree with the advisor.

Figure 3

The estimated difference in trustworthiness between deontological and utilitarian advisors (on a 1-7 Likert scale) across different levels of agreement with the advisor

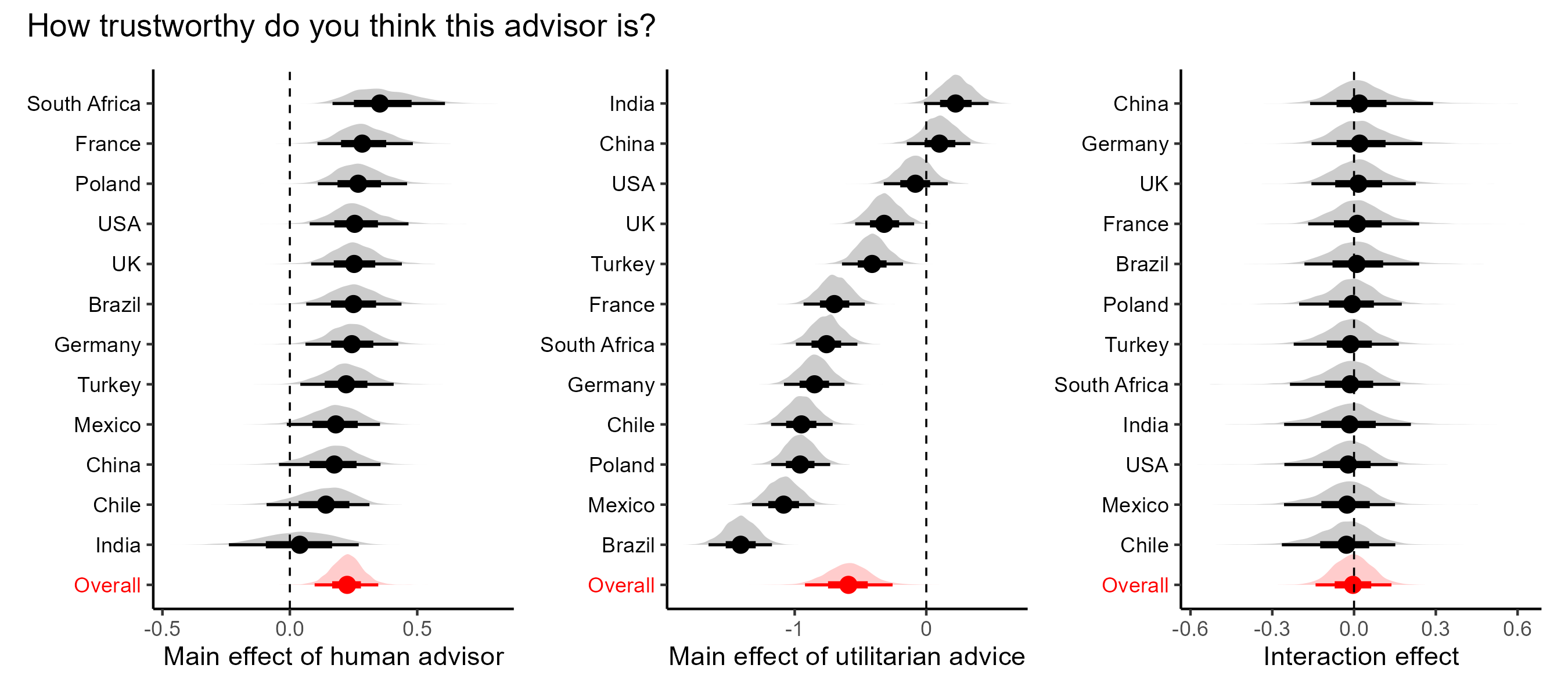


*Note*. The dashed horizontal line indicates no difference in trust between deontological and utilitarian advisors. The dashed vertical line indicates that the participant is neutral with regards to agreement with the advisor. Lines and shaded areas represent posterior medians and 95% credible intervals, respectively. Shaded areas become more transparent when the 95% credible intervals begin to include zero.

Finally, we explored whether the experimental effects varied across countries with different cultural backgrounds. Extracting the varying slopes from our multilevel model, we found that the experimental effects of advisor type (human vs. AI) and advice type (deontological vs. utilitarian) were relatively robust across countries ([Figure 4](#fig-country-slopes-trust); see Supplementary Figures [5](#suppfig-country-slopes-blame) - [8](#suppfig-country-slopes-humanlike) for other perceptions variables). However, there were some notable exceptions – the 95% CIs for the effect of advisor type included zero for India, Chile, China, and Mexico, and the 95% CIs for the effect of advice type included zero for India, China, and the USA. To explain the variation across countries, we explored whether the experimental effects varied by cultural factors highlighted by previous research[53](#ref-Awad2020),[54](#ref-Graham2016). Accounting for the geographic and cultural non-independence of countries[55](#ref-Claessens2023), we included global indices of AI progress and government readiness as moderators of the effect of advisor type (human vs. AI) and relational mobility, cultural tightness-looseness, and individualism as moderators of the effect of advice type (deontological vs. utilitarian). Of these cultural factors, we only found evidence that the experimental effects varied by tightness-looseness. Looser countries, like Brazil, Mexico, and Chile, trusted the deontological advisor relatively more than tighter countries like India and China (interaction effect = -0.37, 95% CI [-0.75 0.02]). Other cultural factors did not moderate the experimental effects ([Supplementary Figure 9](#suppfig-cross-cultural)).

Figure 4

Variation in model parameters across countries



*Note*. Estimates are the varying log-odds slopes from our multilevel model predicting trust, alongside overall fixed effects (red). Points and line ranges reflect posterior medians and 66% and 95% credible intervals, respectively. Grey densities reflect samples from the posterior distribution.

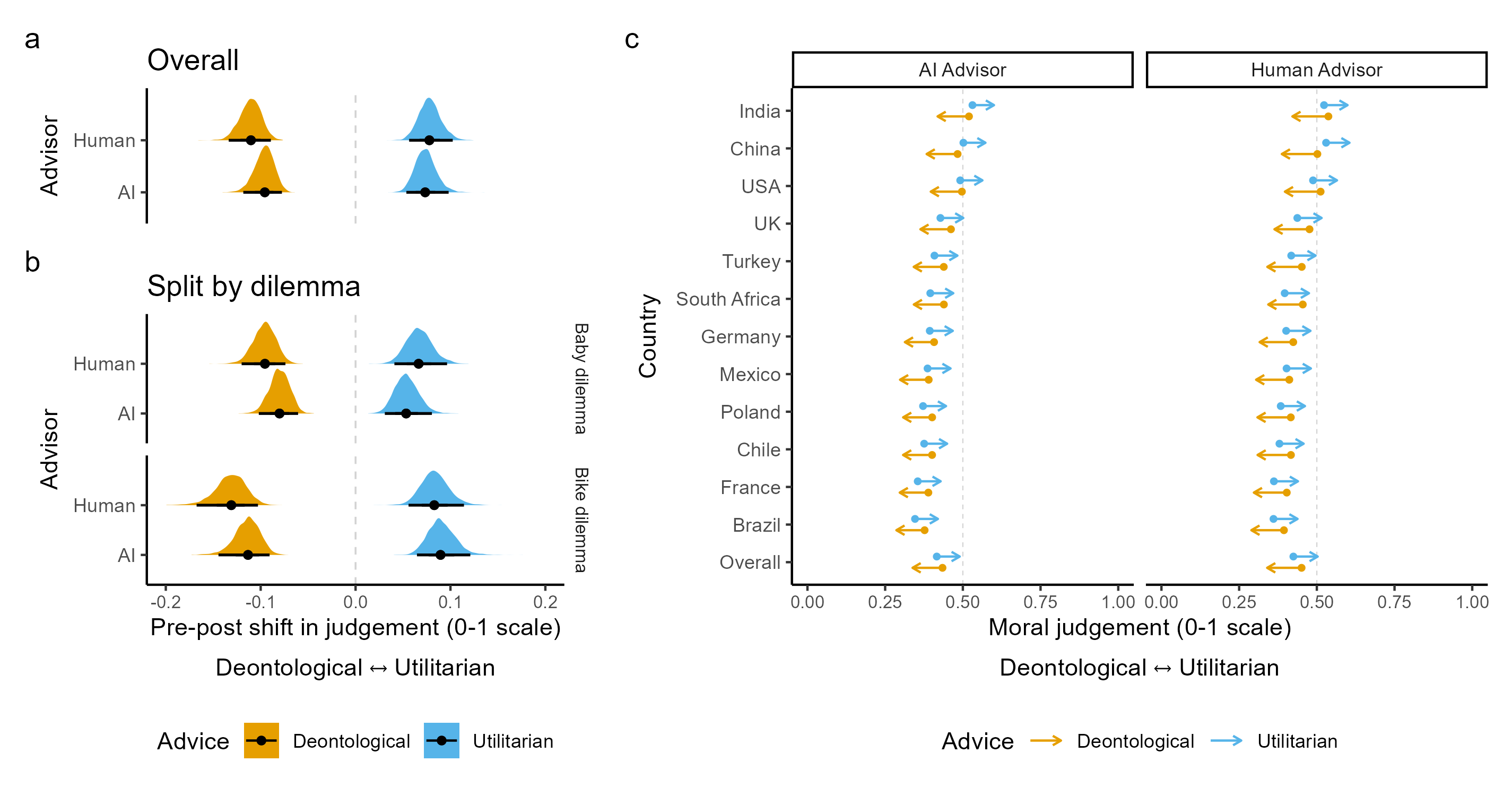
## Effects of Advice on Moral Judgment Updating

Given that participants generally trust artificial moral advisors less than human advisors, we might predict that they should also be less likely to follow AI advice. To test this, we looked at participants’ judgments of the moral dilemmas before and after seeing the advice. Do participants update their moral judgments in line with human and AI advice?

Yes. We found that participants listened to the moral advice, shifting their judgments in the direction of the deontological or utilitarian advice that they received ([Figure 5](#fig-judgment-shift)a). This was true for both artificial moral advisors and for human advisors, revealing a dissociation between the difference in rated trust and advice-updating for human vs. AI. In line with the finding that participants perceived advisors who gave utilitarian judgments to be less trustworthy, the absolute size of the pre-post shifts revealed that participants shifted their judgments more after receiving deontological advice compared to utilitarian advice, both when the advisor was human (difference in absolute shift = 0.03, 95% CI [0.01 0.06]) and when the advisor was AI (difference in absolute shift = 0.02, 95% CI [0.00 0.05]). By contrast, the absolute size of the pre-post shifts did not differ between human and AI advisors, either when they gave deontological advice (difference in absolute shift = 0.01, 95% CI [-0.01 0.03]) or utilitarian advice (difference in absolute shift = 0.00, 95% CI [-0.02 0.03]). This suggests that while people listened to advice, they particularly listened to deontological advice – but it did not matter whether this came from a human or an AI. This pattern of results was robust across both sacrificial dilemmas, though the advice tended to be more persuasive in the Bike dilemma compared to the Baby dilemma ([Figure 5](#fig-judgment-shift)b). The pattern also held across all the countries in our sample ([Figure 5](#fig-judgment-shift)c) and when we restricted our analysis to judgments in the first experimental block only, before participants knew they would be given the opportunity to revise their judgments ([Supplementary Figure 10](#suppfig-judgment-shift-block1)).

Figure 5

Effects of moral advice on participant judgments



*Note*. Estimated pre-post shifts in judgments on a 0-1 sliding scale, for human and AI advisors giving deontological and utilitarian advice. The grey dashed line indicates no pre-post shift in judgment. Points, line ranges, and densities reflect posterior medians, 95% credible intervals, and samples from the posterior distribution, respectively. (b) Estimated pre-post shifts in judgments, split by moral dilemma. (c) Estimated pre-post shifts in judgments, split by country. The grey dashed line indicates neutral judgments. Points and arrows reflect posterior median estimates before and after seeing the advice, respectively.

As well as shifting their judgments, participants’ confidence in their judgment increased after receiving the moral advice ([Supplementary Figure 11](#suppfig-confidence)). This pre-post increase in confidence was again slightly greater after seeing deontological advice compared to utilitarian advice, but did not differ between human and AI advisors. This pattern was robust across dilemmas and countries (Supplementary Figures [12](#suppfig-confidence-baby) and [13](#suppfig-confidence-bike)).

# Discussion

Can AI tools help us in the moral domain? Some have proposed that they can – and should[1](#ref-Lara2020),[2](#ref-Giubilini2018),[5](#ref-Sinnott2021). But it has remained unclear whether and how people trust AI advisors in the moral domain, and how the kind of normative ethical advice that the AI provides might influence this trust. In a large-scale, pre-registered experiment across 12 countries, we studied people’s trust in, and deference to, artificial moral advisors who give advice aligned with different ethical theories. We found that participants trusted artificial moral advisors less than humans, particularly when they gave advice based on utilitarian principles. Despite this relative distrust, however, people still listened to the moral advice from AI by updating their own moral judgments, especially when the advice was in line with deontological principles. Importantly, this judgment updating was observed regardless of whether the advice came from a human or an AI advisor. Taken together, these findings suggest that artificial moral advisors may yet play an influential role in human morality, for better or worse.

Our first key finding, that participants trusted artificial moral advisors less than humans, replicates previous work on algorithm aversion in general[14](#ref-Dawes1979)–[17](#ref-Meehl1957) and emerging findings in the moral domain[18](#ref-Bigman2018),[41](#ref-Myers2025). The relative distrust of AMAs was more pronounced for those who had less familiarity with AI and used AI less frequently, suggesting that as AI use becomes more widespread, negativity towards AMAs might reduce. We found that this pattern was observed for most of the countries in our sample and was not moderated by global indices of AI progress and government AI readiness.

Our second key finding is that the direction of advice affected participants’ perceptions of the advisors. Participants trusted AI (and human) advisors who gave characteristically deontological advice more than advisors who gave utilitarian advice. These results were observed overall and found when accounting for participants’ own moral judgments. We also found that looser countries like Brazil, Mexico, and Chile trusted deontological advisors relatively more than tighter countries like India and China, which may reflect a greater focus on preventing negative outcomes in tighter countries[56](#ref-Gelfand2006). Together, these findings suggest that for AI, like for humans, not all “moral” decisions equally signal trustworthiness[31](#ref-Everett2016)–[33](#ref-Everett2021). AMAs that are built to align with impartial utilitarian ethics may be particularly distrusted. By contrast, artificial moral advice that aligns with human norms of integrity and respect for persons may be better received by a public uneasy with AI’s perceived cold rationality, raising challenges to those who believe that utilitarianism is the correct normative standpoint by which AMAs should be programmed.

Our third key finding is that participants listened to the moral advice by updating their own moral judgments, regardless of whether the advisor was human or AI – but did so especially when the advice was deontological, in line with the effects on perceived trustworthiness. This finding extends the growing literature on the persuasive potential of AI[8](#ref-Krugel2023),[45](#ref-Schoenegger2025)–[50](#ref-Costello2024) into the moral domain and reveals a psychological paradox: people may claim to distrust AI relative to humans, yet still update their moral views based on AI advice. This is good news for those who argue for the possibility of AI to enhance moral judgments[2](#ref-Giubilini2018),[5](#ref-Sinnott2021), but is likely to ring alarm bells for those who have expressed concerns about outsourcing moral judgments to AI, including concerns about the lack of epistemic agency involved, the risk of moral deskilling, and potential issues with offloading moral responsibility onto machines[6](#ref-Landes2025),[7](#ref-Liu2022),[11](#ref-Landes_preprint),[57](#ref-Vallor2024). These concerns could be compounded if people follow advice from AMAs that they claim not to trust. For AI developers and policy-makers, our results suggest that further safeguards should be put in place for emerging AMA technologies, especially those operating in high-stakes domains like healthcare, military ethics, and criminal justice. When testing these systems, developers should be cautious not to conflate expressed trust with actual behavioral impact.

Our fourth key finding is that people seem to implicitly associate utilitarianism with being more machine-like. Utilitarianism, with its impartial focus on calculating predicted welfare and assessing consequences, has often been criticized for the way it seems to diminish humanity through its focus on algorithms – that it is a “a civilization of ‘things’ and not of ‘persons’”[58](#ref-JohnPaulII1995) and that when utilitarians make moral decisions by rationally calculating outcomes they are having “one thought too many”[59](#ref-Williams1981) rather than acting instinctively based on the pull of human ties. Unsurprisingly then, with its focus on computations of overall welfare and consequences, the utilitarian decision-making process is sometimes seen as akin to what a machine, not a human, would do. Indeed, there is evidence that people expect AI to be more utilitarian[41](#ref-Myers2025),[60](#ref-Malle2015) and here we show that people were more surprised when deontological advice came from an AI rather than a human; that humans who made utilitarian judgments were seen to be more machine-like than those who made deontological judgments; and that AI systems which made deontological judgments were seen as more human-like. This association between utilitarianism and machine-likeness highlights the broader challenges AI systems face in appearing not just intelligent, but *humane*. If AI developers want artificial moral advisors to be accepted, and this is by no means a given, then they should build systems that can do more than merely compute outcomes – the systems must also communicate care, character, and moral understanding in ways that align with human moral psychology.

Future work should build on these findings in several ways. First, as discussed above, artificial moral advisors need to be aligned with a normative standard, but there are many different ethical frameworks that AMAs could be aligned to provide advice from. While our results cohere with previous work showing that utilitarian advisors are trusted less[32](#ref-Everett2018),[33](#ref-Everett2021),[41](#ref-Myers2025), it will be important to explore how AMAs will be judged when their advice aligns with other philosophical theories, such as virtue ethics, in line with more general calls to move beyond sacrificial dilemmas as the workhorse of moral psychology[61](#ref-Kahane2023). Second, future work should assess the generalizability of these findings beyond abstract sacrificial dilemmas to everyday moral situations. While much discussion about AMAs has focused on moral dilemmas where normative theories conflict[62](#ref-Tasioulas2019) or in complex multi-chain cases where AMAs may be particularly helpful (e.g., kidney donation[5](#ref-Sinnott2021)), research should explore how people trust AI advisors in the kinds of everyday moral situations that LLMs are increasingly being used to used to provide advice for, such as parenting, workplace fairness, and relationships. These real-world contexts may trigger different patterns of trust and influence than abstract dilemmas. Third, future work should explore how and why people listen to AMAs and what the boundary conditions of these effects are. Philosophers have expressed concern about the possibility of people deferring entirely to AI advisors instead of critically evaluating the reasons that they give[1](#ref-Lara2020),[6](#ref-Landes2025). Our finding that people listened more to deontological than utilitarian advice suggests both that our results are not merely due to demand characteristics and that people are actively attending to the content of the advice. Still, it will be important for future work to look at the precise mechanism of belief change, the persistence of these effects over time, and potential influences of AMA advice on behavior in the real world when there is greater perceived risk – and reward – from receiving AI advice.

Overall, our findings reveal a paradox: people distrust artificial moral advisors relative to humans, and particularly distrust utilitarian advisors, yet still change their judgments in response to AI advice. This tension underscores both the promise and peril of using AI in the moral domain. As artificial moral advisors become more embedded in everyday life, it becomes ever more imperative that they do more than calculate, but align with the moral intuitions, emotions, and values that make us human.

# Methods

## Ethical Approval

Ethical approval was granted for the study by the University of Kent’s Psychology Research Ethics Panel (protocol ID: 202517411704229833). All participants provided informed consent at the beginning of the survey and were debriefed at the end of the survey.

## Participants and Sampling

To determine our sample size, we conducted a power simulation informed by the effect sizes from a previous study[41](#ref-Myers2025). This previous study found a small main effect of advice type (deontological vs. utilitarian) and a small main effect of advisor type (human vs. AI) on trust. Our simulation suggested that we would require 500 participants to detect both of these effects with 90% power in our proposed design. We aimed to recruit at least this number of participants in each country (overall minimum N = 6000 for 12 countries) to ensure that we were adequately powered within each country. After excluding and replacing participants who failed our exclusion criteria (see Exclusions), our final sample size was *N* = 6,896 (see [Supplementary Figure 14](#suppfig-sample) for overall sample characteristics and [Supplementary Table 4](#supptbl-sample-by-country) for a breakdown of the sample by country). These data were collected over a 23-day period from 8 April to 1 May 2025.

We recruited participants from 12 countries – Brazil, Chile, China, France, Germany, India, Mexico, Poland, South Africa, Turkey, the United Kingdom, and the United States – through the online panel aggregator Prime Panels (<https://www.cloudresearch.com/products/prime-panels/>; [Figure 1](#fig-study-overview)c). We chose this set of countries to maximize variation in geographic spread, cultural backgrounds, and levels of AI progress according to global indices ([Supplementary Figure 9](#suppfig-cross-cultural)). We used quotas to target nationally representative samples with respect to age and gender, relaxing these quotas when data collection slowed in the final 15% of participants in each country. Despite relaxing the quotas, we found that the differences between observed and targeted proportions in each quota category were less than 10% for all countries except China, which was skewed slightly towards younger participants ([Supplementary Figure 15](#suppfig-sample-representativeness)).

## Design

The general flow of the experiment is presented in [Figure 1](#fig-study-overview)a. The experiment followed a mixed design. We randomized participants to see advice from either human advisors or AI advisors throughout the whole experiment (between-subjects). We then presented participants with two experimental blocks in a within-subjects design. In the first block, participants saw either deontological or utilitarian advice (randomly counterbalanced) for either the Baby dilemma or the Bike dilemma (randomly counterbalanced; see Procedure for more details about the dilemmas). In the second block, participants then saw the other direction of advice for the other moral dilemma.

## Procedure

Participants completed the survey experiment in Qualtrics (<https://www.qualtrics.com/>). At the beginning of the survey, participants chose their survey language, completed a Captcha verification question, provided informed consent, reported whether they currently lived in the country of intended recruitment, provided their age and gender, and answered two attention check questions (see Supplementary Materials for attention check wording). Participants were unable to proceed with the survey if they met any of the following criteria: (1) they had a Captcha score of less than 0.5, (2) they did not provide informed consent, (3) they stated that they lived in a different country to that of intended recruitment, (4) they were in an age/gender quota group that had already been filled, or (5) they failed either of our two attention check questions.

If participants proceeded with the survey, we presented them with the two experimental blocks. At the start of each block, we presented participants with the sacrificial dilemma (see Supplementary Materials for full dilemma wordings). The Baby dilemma described a situation where townspeople are hiding from enemy soldiers, and someone must choose between smothering their crying baby to save everyone or letting the baby cry and risking everyone’s death. The Bike dilemma described a situation where a motorcycle rider is about to crash on the road, and someone must choose between causing the rider’s likely death by forcing them off the road or allowing a crash that could result in the deaths of many other riders behind them. These dilemmas were replicated from Myers and Everett’s[41](#ref-Myers2025) Study 2. On the same page as the dilemma, we asked participants for their own judgment on what they thought should be done in the dilemma, rated on a 0-1 sliding scale from deontological judgment (not smothering the baby, not crashing the rider) to utilitarian judgment (smothering the baby, crashing the rider). We also asked participants how confident they were in their judgment on a 1-7 Likert scale.

On the following page, we presented participants with advice about the dilemma from a moral advisor. In the human condition, we told participants that Dr.  Johnson or Dr. Smith (names randomly counterbalanced across blocks) drew on their extensive ethical and philosophical training. In the AI condition, we told participants that ETHIC-AI or VIRTUE-BOT (names counterbalanced) drew on advancements in machine learning about moral cases. We then presented the moral advice ([Figure 1](#fig-study-overview)b). When giving deontological advice, the advisor stated that “an important principle in ethics is that killing people is just wrong, and this duty to not kill should apply even if killing has good consequences in a specific case.” When giving utilitarian advice, the advisor stated that “an important principle in ethics is to think about the greater good, and in this specific case killing the one person would bring about better consequences overall.” On the same page as the advice, we then asked participants the following questions about their perceptions of the advisor on 1-7 Likert scales:

* How trustworthy do you think [the advisor] is?
* How much would you blame someone if they followed [the advisor’s] recommendation?
* Based on their advice, how willing would you be to trust [the advisor] on other issues?
* How surprised were you at [the advisor’s] recommendation?
* How well do the following words describe [the advisor]? Machine-like ↔ human-like

On the next page, we reminded participants about the moral advice and then asked them again to judge what they thought should be done (0-1 sliding scale) and report how confident they were in their judgment (1-7 Likert). Then on the final page of the block, we asked participants a binary-choice comprehension question (“What did the advisor recommend in this dilemma?”) to check that they could correctly recall the deontological or utilitarian advice. After seeing the first experimental block, participants repeated this process in a second experimental block with the other direction of advice and the other moral dilemma.

After the two experimental blocks, we asked participants some general questions about “AI tools like ChatGPT”, including how familiar they were with AI tools, how frequently they used AI tools, and how trustworthy they thought AI tools were. We then asked participants several demographic questions, including their highest education level, their subjective social status (measured with the MacArthur ladder), their political ideology (from left-wing to right-wing; not asked in China), and their religiosity. We then debriefed participants about the purpose of the study and gave them an open-ended form to leave any comments or remarks about the study.

## Translations

In non-English-speaking countries, we translated the survey using a forward- and back-translation procedure. First, a native speaker translated the survey from English into the target language. Then, another native speaker (who had not yet seen the survey) translated the survey back into English. Any discrepancies were discussed between forward- and back-translators and any issues were resolved. All translators then checked the final Qualtrics surveys to ensure that the translations had been implemented correctly.

## Pre-registration

We pre-registered our hypotheses, study design, sampling plan, exclusion criteria, and analysis plan on the Open Science Framework on 8 April 2025 (<https://osf.io/qa7gn/>).

## Exclusions

We excluded and replaced participants who completed the survey but met any of the following criteria: (1) they had taken the survey more than once as indicated by duplicate IP addresses or assignment IDs, (2) they straight-lined on both of the advisor-perceptions survey pages by providing the same responses to all items, (3) they sped through the survey, either by completing the survey in less than 100 seconds (~2 seconds per question/text) or completing the survey at least two median absolute deviations faster than the median response time, (4) they took more than 24 hours to complete the survey, (5) they failed to answer more than 50% of the questions in the survey, or (6) they wrote nonsensical responses in our open-ended question, as independently judged by translators who were blind to experimental condition and responses to our primary measures.

We also excluded responses to moral dilemmas if the participant failed the comprehension question for that moral dilemma. This exclusion strategy resulted in our final sample size of 6,896 participants. However, in additional “intention-to-treat” analyses, we retained these comprehension failures, resulting in a larger sample size of 7,096 participants.

## Statistical Analysis

To analyse the perceptions of the advisors, we fitted Bayesian cumulative-link ordinal models to the Likert scale data, modelling the interaction between advice type (deontological vs. utilitarian) and advisor type (human vs. AI). To split by moral dilemma, we included dilemma (Baby vs. Bike) as an additional moderator. In exploratory models, we interacted one or both of the experimental effects with individual-level characteristics (agreement with the advisor, religiosity, political ideology, familiarity with AI, frequency of AI use) or cultural-level variables (relational mobility[63](#ref-Thomson2018), tightness-looseness[64](#ref-Eriksson2021), individualism[65](#ref-Hofstede2010), the global AI index[66](#ref-TortoiseMedia), AI readiness[67](#ref-OxfordInsights)).

To analyse participants’ judgments and confidence, we fitted Bayesian zero-one-inflated beta models to the 0-1 slider data and Bayesian cumulative-link ordinal models to the Likert scale data, modelling the interaction between time (pre- vs. post- advice), advice type (deontological vs.  utilitarian), and advisor type (human vs. AI). To split by moral dilemma, we included dilemma (Baby vs. Bike) as an additional moderator.

For all models, we included varying intercepts for participants and varying intercepts and slopes for countries. When analyzing the moderating effects of cultural variables, we controlled for the non-independence of countries by allowing the country-level varying effects to covary according to geographic and linguistic proximity matrices constructed in previous research[55](#ref-Claessens2023). We used regularizing priors for all models to impose conservatism on parameter estimates. We used 95% credible intervals as the inference criteria. All models converged normally ( ≤ 1.01).

## Reproducibility

We conducted all analyses in R version 4.4.2[68](#ref-RCoreTeam). We fitted Bayesian regression models using the *brms*[69](#ref-Burkner2017) package. We produced visualizations using the *ggdist*[70](#ref-Kay2024), *ggplot2*[71](#ref-Wickham2016), and *patchwork*[72](#ref-Pedersen2025) packages. We reproducibly generated the manuscript using the *targets*[73](#ref-Landau2021) package and *quarto*[74](#ref-Allaire2024). All data, materials, and code to reproduce the analyses and figures in this paper can be found here: <https://github.com/ScottClaessens/trustAMAs>

# Acknowledgements

This work was generously supported by funding from the Economic and Social Research Council (ES/V015176/1), a Philip Leverhulme Prize (PLP-2021-095), and the Horizon Europe UK Guarantee via the UKRI (EP/Y00440X/1) awarded to JACE.

# Data and Code Availability

All data and original code can be found here: <https://github.com/ScottClaessens/trustAMAs>

# Statement of Interests

The authors have no conflicts of interest to disclose.

# References

1. Lara, F. & Deckers, J. [Artificial intelligence as a Socratic assistant for moral enhancement](https://doi.org/10.1007/s12152-019-09401-y). *Neuroethics* **13**, 275–287 (2020).

2. Giubilini, A. & Savulescu, J. [The artificial moral advisor. The "ideal observer" meets artificial intelligence](https://doi.org/10.1007/s13347-017-0285-z). *Philosophy & Technology* **31**, 169–188 (2018).

3. Haidt, J. [The emotional dog and its rational tail: A social intuitionist approach to moral judgment](https://doi.org/10.1037/0033-295X.108.4.814). *Psychological review* **108**, 814 (2001).

4. Firth, R. [Ethical absolutism and the ideal observer](https://doi.org/10.2307/2103988). *Philosophy and Phenomenological Research* **12**, 317–345 (1952).

5. Sinnott-Armstrong, W. & Skorburg, J. A. [How AI can aid bioethics](https://doi.org/10.3998/jpe.1175). *Journal of Practical Ethics* **9**, (2021).

6. Landes, E., Voinea, C. & Uszkai, R. [Rage against the authority machines: How to design artificial moral advisors for moral enhancement](https://doi.org/10.1007/s00146-024-02135-3). *AI & Society* **40**, 2237–2248 (2025).

7. Liu, Y., Moore, A., Webb, J. & Vallor, S. Artificial moral advisors: A new perspective from moral psychology. in *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* 436–445 (2022). doi:[10.1145/3514094.3534139](https://doi.org/10.1145/3514094.3534139).

8. Krügel, S., Ostermaier, A. & Uhl, M. [ChatGPT’s inconsistent moral advice influences users’ judgment](https://doi.org/10.1038/s41598-023-31341-0). *Scientific Reports* **13**, 4569 (2023).

9. Dillion, D., Mondal, D., Tandon, N. & Gray, K. [AI language model rivals expert ethicist in perceived moral expertise](https://doi.org/10.1038/s41598-025-86510-0). *Scientific Reports* **15**, 4084 (2025).

10. Jiang, L. *et al.* [Investigating machine moral judgement through the Delphi experiment](https://doi.org/10.1038/s42256-024-00969-6). *Nature Machine Intelligence* **7**, 1–16 (2025).

11. Landes, E. & Everett, J. A. C. AI should develop human empathy, not replace it. (2025) doi:[10.31234/osf.io/y3qzu\_v1](https://doi.org/10.31234/osf.io/y3qzu_v1).

12. Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. [Not so different after all: A cross-discipline view of trust](https://doi.org/10.5465/amr.1998.926617). *Academy of Management Review* **23**, 393–404 (1998).

13. Lalot, F. & Bertram, A.-M. [When the bot walks the talk: Investigating the foundations of trust in an artificial intelligence (AI) chatbot](https://doi.org/10.1037/xge0001696). *Journal of Experimental Psychology: General* **154**, 533–551 (2024).

14. Dawes, R. M. [The robust beauty of improper linear models in decision making](https://doi.org/10.1037/0003-066X.34.7.571). *American Psychologist* **34**, 571–582 (1979).

15. Dietvorst, B. J., Simmons, J. P. & Massey, C. [Algorithm aversion: People erroneously avoid algorithms after seeing them err](https://doi.org/10.1037/xge0000033). *Journal of Experimental Psychology: General* **144**, 114–126 (2015).

16. Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. (University of Minnesota Press, 1954).

17. Meehl, P. E. [When shall we use our heads instead of the formula?](https://doi.org/10.1037/h0047554) *Journal of Counseling Psychology* **4**, 268–273 (1957).

18. Bigman, Y. E. & Gray, K. [People are averse to machines making moral decisions](https://doi.org/10.1016/j.cognition.2018.08.003). *Cognition* **181**, 21–34 (2018).

19. Henrich, J., Heine, S. J. & Norenzayan, A. [The weirdest people in the world?](https://doi.org/10.1017/S0140525X0999152X) *Behavioral and Brain Sciences* **33**, 61–83 (2010).

20. Barnes, A. J., Zhang, Y. & Valenzuela, A. AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology* 101838 (2024) doi:[10.1016/j.copsyc.2024.101838](https://doi.org/10.1016/j.copsyc.2024.101838).

21. Inglehart, R. & Baker, W. E. [Modernization, cultural change, and the persistence of traditional values](https://doi.org/10.1177/000312240006500103). *American Sociological Review* **65**, 19–51 (2000).

22. Jackson, J. C. & Medvedev, D. [Worldwide divergence of values](https://doi.org/10.1038/s41467-024-46581-5). *Nature Communications* **15**, 2650 (2024).

23. Bentham, J. *The collected works of Jeremy Bentham: Deontology, together with a table of the springs of action and article on utilitarianism*. (Clarendon Press, 1983).

24. Mill, J. S. *Utilitarianism*. (Parker, Son,; Bourn, 1863).

25. Singer, P. *Practical ethics*. (Cambridge University Press, 1993).

26. Fried, C. *Right and wrong*. (Harvard University Press, 1978).

27. Kant, I. *Groundwork for the metaphysics of morals*. (Oxford University Press, 2002).

28. Ross, W. D. *The right and the good*. (Clarendon Press, 1930).

29. Brown, M. & Sacco, D. F. [Is pulling the lever sexy? Deontology as a downstream cue to long-term mate quality](https://doi.org/10.1177/0265407517749331). *Journal of Social and Personal Relationships* **36**, 957–976 (2019).

30. Crockett, M. J., Everett, J. A. C., Gill, M. & Siegel, J. Z. [The relational logic of moral inference](https://doi.org/10.1016/bs.aesp.2021.04.001). in *Advances in experimental social psychology* vol. 64 1–64 (Elsevier, 2021).

31. Everett, J. A. C., Pizarro, D. A. & Crockett, M. J. [Inference of trustworthiness from intuitive moral judgments](https://doi.org/10.1037/xge0000165). *Journal of Experimental Psychology: General* **145**, 772–787 (2016).

32. Everett, J. A. C., Faber, N. S., Savulescu, J. & Crockett, M. J. [The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence](https://doi.org/10.1016/j.jesp.2018.07.004). *Journal of Experimental Social Psychology* **79**, 200–216 (2018).

33. Everett, J. A. C. *et al.* [Moral dilemmas and trust in leaders during a global health crisis](https://doi.org/10.1038/s41562-021-01156-y). *Nature Human Behaviour* **5**, 1074–1088 (2021).

34. Rom, S. C., Weiss, A. & Conway, P. [Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others’ moral dilemma responses](https://doi.org/10.1016/j.jesp.2016.09.007). *Journal of Experimental Social Psychology* **69**, 44–58 (2017).

35. Sacco, D. F., Brown, M., Lustgraaf, C. J. & Hugenberg, K. [The adaptive utility of deontology: Deontological moral decision-making fosters perceptions of trust and likeability](https://doi.org/10.1007/s40806-016-0080-6). *Evolutionary Psychological Science* **3**, 125–132 (2017).

36. Turpin, M. H. *et al.* [The search for predictable moral partners: Predictability and moral (character) preferences](https://doi.org/10.1016/j.jesp.2021.104196). *Journal of Experimental Social Psychology* **97**, 104196 (2021).

37. Kreps, T. A. & Monin, B. [Core values versus common sense: Consequentialist views appear less rooted in morality](https://doi.org/10.1177/0146167214551154). *Personality and Social Psychology Bulletin* **40**, 1529–1542 (2014).

38. Uhlmann, E. L., Zhu, L. L. & Tannenbaum, D. [When it takes a bad person to do the right thing](https://doi.org/10.1016/j.cognition.2012.10.005). *Cognition* **126**, 326–334 (2013).

39. Rom, S. C. & Conway, P. [The strategic moral self: Self-presentation shapes moral dilemma judgments](https://doi.org/10.1016/j.jesp.2017.08.003). *Journal of Experimental Social Psychology* **74**, 24–37 (2018).

40. Young, A. D. & Monroe, A. E. [Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas](https://doi.org/10.1016/j.jesp.2019.103870). *Journal of Experimental Social Psychology* **85**, 103870 (2019).

41. Myers, S. & Everett, J. A. C. [People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors](https://doi.org/10.1016/j.cognition.2024.106028). *Cognition* **256**, 106028 (2025).

42. Karpus, J. *et al.* [Human cooperation with artificial agents varies across countries](https://doi.org/10.1038/s41598-025-92977-8). *Scientific Reports* **15**, 10000 (2025).

43. Dong, M., Conway, J. R., Bonnefon, J.-F., Shariff, A. & Rahwan, I. Fears about artificial intelligence across 20 countries and six domains of application. *American Psychologist* (2024) doi:[10.1037/amp0001454](https://doi.org/10.1037/amp0001454).

44. Ikkatai, Y. *et al.* [The relationship between the attitudes of the use of AI and diversity awareness: Comparisons between Japan, the US, Germany, and South Korea](https://doi.org/10.1007/s00146-024-01982-4). *AI & SOCIETY* **40**, 2369–2383 (2024).

45. Schoenegger, P. *et al.* Large language models are more persuasive than incentivized human persuaders. (2025) doi:[10.48550/arXiv.2505.09662](https://doi.org/10.48550/arXiv.2505.09662).

46. Durmus, E. *et al.* Measuring the persuasiveness of language models. <https://www.anthropic.com/news/measuring-model-persuasiveness> (2024).

47. Hackenburg, K. & Margetts, H. [Evaluating the persuasive influence of political microtargeting with large language models](https://doi.org/10.1073/pnas.2403116121). *Proceedings of the National Academy of Sciences* **121**, e2403116121 (2024).

48. Hackenburg, K. *et al.* [Scaling language model size yields diminishing returns for single-message political persuasion](https://doi.org/10.1073/pnas.2413443122). *Proceedings of the National Academy of Sciences* **122**, e2413443122 (2025).

49. Bai, H., Voelkel, J., Eichstaedt, J. & Willer, R. Artificial intelligence can persuade humans on political issues. *Research Square* (2023) doi:[10.21203/rs.3.rs-3238396/v1](https://doi.org/10.21203/rs.3.rs-3238396/v1).

50. Costello, T. H., Pennycook, G. & Rand, D. G. [Durably reducing conspiracy beliefs through dialogues with AI](https://doi.org/10.1126/science.adq1814). *Science* **385**, eadq1814 (2024).

51. Landes, E., Francis, K. & Everett, J. A. C. People attend to moral reasons when responding to AI advice. (under review).

52. Chen, H., Cohen, P. & Chen, S. [How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies](https://doi.org/10.1080/03610911003650383). *Communications in Statistics - Simulation and Computation* **39**, 860–864 (2010).

53. Awad, E., Dsouza, S., Shariff, A., Rahwan, I. & Bonnefon, J.-F. [Universals and variations in moral decisions made in 42 countries by 70,000 participants](https://doi.org/10.1073/pnas.1911517117). *Proceedings of the National Academy of Sciences* **117**, 2332–2337 (2020).

54. Graham, J., Meindl, P., Beall, E., Johnson, K. M. & Zhang, L. [Cultural differences in moral judgment and behavior, across and within societies](https://doi.org/10.1016/j.copsyc.2015.09.007). *Current Opinion in Psychology* **8**, 125–130 (2016).

55. Claessens, S., Kyritsis, T. & Atkinson, Q. D. [Cross-national analyses require additional controls to account for the non-independence of nations](https://doi.org/10.1038/s41467-023-41486-1). *Nature Communications* **14**, 5776 (2023).

56. Gelfand, M. J., Nishii, L. H. & Raver, J. L. [On the nature and importance of cultural tightness-looseness](https://doi.org/10.1037/0021-9010.91.6.1225). *Journal of Applied Psychology* **91**, 1225–1244 (2006).

57. Vallor, S. *The AI mirror: How to reclaim our humanity in an age of machine thinking*. (Oxford University Press, 2024).

58. John Paul II. [Evangelium vitae](http://w2.vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_25031995_evangelium-vitae.html). (1995).

59. Williams, B. Persons, character, and morality. in *Moral luck: Philosophical papers 1973-1980* (ed. Williams, B.) 1–19 (Cambridge University Press, 1981). doi:[10.1017/CBO9781139165860.002](https://doi.org/10.1017/CBO9781139165860.002).

60. Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* 117–124 (Association for Computing Machinery, 2015). doi:[10.1145/2696454.2696458](https://doi.org/10.1145/2696454.2696458).

61. Kahane, G. & Everett, J. A. C. Trolley dilemmas from the philosopher’s armchair to the psychologist’s lab. in *The trolley problem* (ed. Lillehammer, H.) 134–157 (Cambridge University Press, 2023). doi:[10.1017/9781009255615.009](https://doi.org/10.1017/9781009255615.009).

62. Tasioulas, J. First steps towards an ethics of robots and artificial intelligence. *Journal of Practical Ethics* **7**, 49–83 (2019).

63. Thomson, R. *et al.* [Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat](https://doi.org/10.1073/pnas.1713191115). *Proceedings of the National Academy of Sciences* **115**, 7521–7526 (2018).

64. Eriksson, K. *et al.* [Perceptions of the appropriate response to norm violation in 57 societies](https://doi.org/10.1038/s41467-021-21602-9). *Nature Communications* **12**, 1481 (2021).

65. Hofstede, G., Hofstede, G. J. & Minkov, M. *Cultures and organizations: Software of the mind*. (McGraw Hill, 2010).

66. Tortoise Media. [The global AI index](https://www.tortoisemedia.com/data/global-ai). (2024).

67. Oxford Insights. [Government AI readiness index](https://oxfordinsights.com/ai-readiness/ai-readiness-index/). (2024).

68. R Core Team. [*R: A language and environment for statistical computing*](https://www.R-project.org/). (R Foundation for Statistical Computing, 2022).

69. Bürkner, P.-C. [brms: An R package for Bayesian multilevel models using Stan](https://doi.org/10.18637/jss.v080.i01). *Journal of Statistical Software* **80**, 1–28 (2017).

70. Kay, M. [ggdist: Visualizations of distributions and uncertainty in the grammar of graphics](https://doi.org/10.1109/TVCG.2023.3327195). *IEEE Transactions on Visualization and Computer Graphics* **30**, 414–424 (2024).

71. Wickham, H. [*ggplot2: Elegant graphics for data analysis*](https://ggplot2.tidyverse.org). (Springer-Verlag New York, 2016).

72. Pedersen, T. L. [*patchwork: The composer of plots*](https://patchwork.data-imaginist.com). (2025).

73. Landau, W. M. [The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing](https://doi.org/10.21105/joss.02959). *Journal of Open Source Software* **6**, 2959 (2021).

74. Allaire, J. J. *et al.* Quarto. (2024) doi:[10.5281/zenodo.5960048](https://doi.org/10.5281/zenodo.5960048).

# Supplementary Materials

# Moral Dilemma Wordings

The wording for the “Baby” dilemma was as follows:

Enemy soldiers have taken over David’s village. They have orders to kill all remaining civilians over the age of two. David and some of his townspeople have sought refuge in two rooms of the cellar of a large house. Outside, he hears the voices of soldiers who have come to search the house for valuables. David’s baby, who is with him in the room, begins to cry loudly. David puts his hand over the baby’s mouth to block the sound. If David removes his hand from the baby’s mouth, its crying will summon the attention of the soldiers who will spare his baby’s life, but will kill David and the others hiding in both rooms. To save himself and the others, David must keep his hand on the baby’s mouth and smother the baby to death.

The wording for the “Bike” dilemma was as follows:

James is an expert motorcycle rider who has gone on vacation in order to participate in Bike Week. Thousands of other motorcycle riders from across the country have come to ride in this event. As he is riding down the road in front of a large group of other riders, he sees that someone up ahead is losing control of their bike. As he speeds up to pull alongside the unstable rider, he realizes that this person is going to crash at any second. This would certainly result in a large pile-up and several deaths as the riders behind James run over each other trying to avoid the crashed rider. James realizes that he could physically run this rider off the road and into some trees. This would cause the other rider to crash and, at his current speed, almost certainly die, but it would prevent a crash in the middle of the street and the large pile-up of riders.

# Supplementary Figures

|  |
| --- |
| Supplementary Figure 1: Perceptions of AI and human moral advisors giving deontological (orange) and utilitarian advice (blue), pooling across countries and split by dilemma. Box and whisker plots show the distributions of the data, with horizontal lines representing sample medians, boxes representing interquartile ranges, and whiskers representing ranges. Point ranges show model-estimated marginal means, with points representing posterior medians and line ranges representing 95% credible intervals. |

|  |
| --- |
| Supplementary Figure 2: Perceptions of AI and human moral advisors giving deontological (orange) and utilitarian advice (blue), pooling across dilemmas and countries and split by the first and second experimental blocks. Box and whisker plots show the distributions of the data, with horizontal lines representing sample medians, boxes representing interquartile ranges, and whiskers representing ranges. Point ranges show model-estimated marginal means, with points representing posterior medians and line ranges representing 95% credible intervals. |

|  |
| --- |
| Supplementary Figure 3: Perceptions of AI and human moral advisors giving deontological (orange) and utilitarian advice (blue), pooling across dilemmas and countries and including participants who failed the comprehension checks (“intention to treat” analysis). Box and whisker plots show the distributions of the data, with horizontal lines representing sample medians, boxes representing interquartile ranges, and whiskers representing ranges. Point ranges show model-estimated marginal means, with points representing posterior medians and line ranges representing 95% credible intervals. |

|  |
| --- |
| Supplementary Figure 4: Variation in the experimental effects across individuals. The top row shows how the estimated difference in trust (on a 1-7 Likert scale) between deontological and utilitarian advisors changes with participants’ religiosity and political conservatism. The bottom row shows how the estimated difference in trust (on a 1-7 Likert scale) between human and AI advisors changes with participants’ familiarity with AI and frequency of AI use. Lines and shaded areas reflect posterior medians and 50% and 95% credible intervals. Dashed horizontal lines indicate no difference. |

|  |
| --- |
| Supplementary Figure 5: Variation in model parameters across countries. Estimates are the varying log-odds slopes from our multilevel model predicting blame, alongside overall fixed effects (red). Points and line ranges reflect posterior medians and 66% and 95% credible intervals, respectively. Densities reflect samples from the posterior distribution. |

|  |
| --- |
| Supplementary Figure 6: Variation in model parameters across countries. Estimates are the varying log-odds slopes from our multilevel model predicting trust on other issues, alongside overall fixed effects (red). Points and line ranges reflect posterior medians and 66% and 95% credible intervals, respectively. Densities reflect samples from the posterior distribution. |

|  |
| --- |
| Supplementary Figure 7: Variation in model parameters across countries. Estimates are the varying log-odds slopes from our multilevel model predicting surprise, alongside overall fixed effects (red). Points and line ranges reflect posterior medians and 66% and 95% credible intervals, respectively. Densities reflect samples from the posterior distribution. |

|  |
| --- |
| Supplementary Figure 8: Variation in model parameters across countries. Estimates are the varying log-odds slopes from our multilevel model predicting human-like perceptions, alongside overall fixed effects (red). Points and line ranges reflect posterior medians and 66% and 95% credible intervals, respectively. Densities reflect samples from the posterior distribution. |

|  |
| --- |
| Supplementary Figure 9: Variation in the experimental effects across different cultural variables. The top row shows how the estimated difference in trust (on a 1-7 Likert scale) between deontological and utilitarian advisors varies by relational mobility, tightness-looseness, and individualism. The bottom row shows how the estimated difference in trust (on a 1-7 Likert scale) between human and AI advisors varies by government AI readiness and the global AI index. Lines and shaded areas reflect posterior medians and 50% and 95% credible intervals. Letters are ISO-2 country codes. All models control for spatial and cultural non-independence between countries. |

|  |
| --- |
| Supplementary Figure 10: Effects of moral advice on participant judgments in the first block only. Estimated pre-post shifts in judgments on a 0-1 sliding scale, for human and AI advisors giving deontological and utilitarian advice. The grey dashed line indicates no pre-post shift in judgment. Points, line ranges, and densities reflect posterior medians, 95% credible intervals, and samples from the posterior distribution, respectively. |

|  |
| --- |
| Supplementary Figure 11: Pre-post increase in participants’ confidence in their judgment after seeing the moral advice. Points and line ranges represent posterior medians and 50% and 95% credible intervals, respectively. |

|  |
| --- |
| Supplementary Figure 12: Pre-post increase in participants’ confidence in their judgment after seeing the moral advice, for the Baby dilemma only. Points and line ranges represent posterior medians and 50% and 95% credible intervals, respectively. |

|  |
| --- |
| Supplementary Figure 13: Pre-post increase in participants’ confidence in their judgment after seeing the moral advice, for the Bike dilemma only. Points and line ranges represent posterior medians and 50% and 95% credible intervals, respectively. |

|  |
| --- |
| Supplementary Figure 14: Overall characteristics of the sample |

|  |
| --- |
| Supplementary Figure 15: Representativeness of the final sample. Bars show the observed proportions in each age-gender category, lines show the targeted proportions in our nationally representative quotas. Differences between observed and targeted proportions in each quota category were less than 10% for all countries except China, which was skewed slightly towards younger participants. |

# Supplementary Tables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | Response | | | | | | --- | --- | --- | --- | --- | --- | | Contrast | Trustworthy | Blame | Trust other issues | Surprise | Human-like | | **Baby dilemma** | | | | | | | AI Utilitarian vs. AI Deontological | -0.89 [-1.26, -0.51] | 0.84 [0.62, 1.05] | -0.80 [-1.16, -0.45] | 0.13 [-0.06, 0.31] | -0.90 [-1.21, -0.56] | | Human Utilitarian vs. Human Deontological | -0.98 [-1.35, -0.59] | 0.82 [0.58, 1.06] | -1.07 [-1.45, -0.71] | 0.68 [0.46, 0.89] | -1.39 [-1.71, -1.05] | | Human Deontological vs. AI Deontological | 0.26 [0.11, 0.41] | -0.08 [-0.25, 0.08] | 0.37 [0.20, 0.54] | -0.53 [-0.69, -0.37] | 1.07 [0.77, 1.35] | | Human Utilitarian vs. AI Utilitarian | 0.17 [0.01, 0.33] | -0.10 [-0.27, 0.08] | 0.10 [-0.09, 0.28] | 0.02 [-0.18, 0.21] | 0.58 [0.27, 0.87] | | Interaction effect | -0.09 [-0.28, 0.09] | -0.02 [-0.21, 0.18] | -0.27 [-0.46, -0.07] | 0.55 [0.32, 0.78] | -0.49 [-0.70, -0.28] | | **Bike dilemma** | | | | | | | AI Utilitarian vs. AI Deontological | -0.29 [-0.61, 0.04] | 0.41 [0.18, 0.64] | -0.23 [-0.55, 0.09] | -0.22 [-0.44, 0.01] | -0.33 [-0.63, 0.00] | | Human Utilitarian vs. Human Deontological | -0.24 [-0.55, 0.09] | 0.38 [0.13, 0.62] | -0.12 [-0.45, 0.20] | -0.08 [-0.33, 0.17] | -0.44 [-0.74, -0.11] | | Human Deontological vs. AI Deontological | 0.19 [0.02, 0.36] | -0.21 [-0.38, -0.02] | 0.20 [0.00, 0.39] | -0.27 [-0.43, -0.11] | 0.88 [0.57, 1.17] | | Human Utilitarian vs. AI Utilitarian | 0.25 [0.05, 0.43] | -0.24 [-0.42, -0.05] | 0.30 [0.09, 0.50] | -0.13 [-0.34, 0.08] | 0.78 [0.44, 1.09] | | Interaction effect | 0.06 [-0.15, 0.26] | -0.03 [-0.23, 0.17] | 0.10 [-0.10, 0.31] | 0.14 [-0.10, 0.38] | -0.11 [-0.33, 0.11] | |

Supplementary Table 1: Pairwise contrasts for perceptions of the moral advisors, pooling across countries and split by dilemma. Numbers reflect differences in marginal means on the log-odds scale. The interaction effects represent the interaction between advisor type and advice type (i.e., the difference between the differences in the rows above). Main numbers are posterior medians, numbers in square brackets are 95% credible intervals.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | Response | | | | | | --- | --- | --- | --- | --- | --- | | Contrast | Trustworthy | Blame | Trust other issues | Surprise | Human-like | | **Block 1** | | | | | | | AI Utilitarian vs. AI Deontological | -0.59 [-0.88, -0.29] | 0.52 [0.34, 0.71] | -0.48 [-0.74, -0.22] | -0.35 [-0.55, -0.15] | -0.66 [-0.91, -0.38] | | Human Utilitarian vs. Human Deontological | -0.50 [-0.80, -0.19] | 0.48 [0.27, 0.68] | -0.44 [-0.71, -0.18] | 0.05 [-0.17, 0.26] | -0.79 [-1.05, -0.51] | | Human Deontological vs. AI Deontological | 0.20 [0.06, 0.33] | -0.17 [-0.34, -0.01] | 0.23 [0.07, 0.38] | -0.40 [-0.56, -0.25] | 0.90 [0.62, 1.18] | | Human Utilitarian vs. AI Utilitarian | 0.29 [0.14, 0.44] | -0.22 [-0.38, -0.05] | 0.26 [0.10, 0.42] | 0.00 [-0.20, 0.19] | 0.76 [0.47, 1.05] | | Interaction effect | 0.10 [-0.08, 0.28] | -0.05 [-0.24, 0.14] | 0.03 [-0.15, 0.22] | 0.40 [0.18, 0.61] | -0.13 [-0.33, 0.07] | | **Block 2** | | | | | | | AI Utilitarian vs. AI Deontological | -0.59 [-0.93, -0.22] | 0.73 [0.49, 0.95] | -0.56 [-0.92, -0.19] | 0.23 [0.01, 0.43] | -0.57 [-0.89, -0.25] | | Human Utilitarian vs. Human Deontological | -0.70 [-1.07, -0.32] | 0.70 [0.44, 0.95] | -0.75 [-1.12, -0.37] | 0.52 [0.25, 0.79] | -1.02 [-1.35, -0.66] | | Human Deontological vs. AI Deontological | 0.25 [0.08, 0.42] | -0.11 [-0.28, 0.05] | 0.34 [0.13, 0.54] | -0.43 [-0.60, -0.27] | 1.07 [0.78, 1.36] | | Human Utilitarian vs. AI Utilitarian | 0.14 [-0.05, 0.32] | -0.14 [-0.33, 0.05] | 0.15 [-0.07, 0.37] | -0.13 [-0.37, 0.10] | 0.63 [0.32, 0.91] | | Interaction effect | -0.12 [-0.32, 0.10] | -0.03 [-0.24, 0.18] | -0.19 [-0.40, 0.03] | 0.29 [0.03, 0.56] | -0.45 [-0.66, -0.23] | |

Supplementary Table 2: Pairwise contrasts for perceptions of the moral advisors, pooling across dilemmas and countries and split by the first and second experimental blocks. Numbers reflect differences in marginal means on the log-odds scale. The interaction effects represent the interaction between advisor type and advice type (i.e., the difference between the differences in the rows above). Main numbers are posterior medians, numbers in square brackets are 95% credible intervals.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | Response | | | | | | --- | --- | --- | --- | --- | --- | | Contrast | Trustworthy | Blame | Trust other issues | Surprise | Human-like | | **Effect of advice type** | | | | | | | AI Utilitarian vs. AI Deontological | -0.69 [-1.00, -0.37] | 0.58 [0.37, 0.78] | -0.57 [-0.86, -0.30] | -0.06 [-0.21, 0.09] | -0.63 [-0.87, -0.40] | | Human Utilitarian vs. Human Deontological | -0.67 [-0.99, -0.35] | 0.54 [0.33, 0.77] | -0.67 [-0.95, -0.39] | 0.26 [0.06, 0.46] | -0.94 [-1.18, -0.70] | | **Effect of advisor type** | | | | | | | Human Deontological vs. AI Deontological | 0.19 [0.07, 0.30] | -0.14 [-0.27, 0.01] | 0.28 [0.14, 0.41] | -0.36 [-0.50, -0.22] | 0.98 [0.70, 1.25] | | Human Utilitarian vs. AI Utilitarian | 0.21 [0.08, 0.33] | -0.17 [-0.31, -0.03] | 0.18 [0.04, 0.32] | -0.04 [-0.21, 0.14] | 0.68 [0.40, 0.94] | | **Interaction effect** | | | | | | | Interaction effect | 0.02 [-0.10, 0.15] | -0.03 [-0.16, 0.10] | -0.10 [-0.22, 0.03] | 0.32 [0.14, 0.51] | -0.30 [-0.43, -0.17] | |

Supplementary Table 3: Pairwise contrasts for perceptions of the moral advisors, pooling across dilemmas and countries and including participants who failed the comprehension checks (“intention to treat” analysis). Numbers reflect differences in marginal means on the log-odds scale. The bottom row represents the interaction between advisor type and advice type (i.e., the difference between the differences in the rows above). Main numbers are posterior medians, numbers in square brackets are 95% credible intervals.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Country | Overall N | Female | Male | Age | Education (1-8) | SES (1-10) | Religiosity (1-7) | Political ideology (1-7) | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Brazil | 678 | 388 (57%) | 284 (42%) | 39.2 (13.8) | 5.7 (1.6) | 5.9 (1.8) | 4.7 (1.8) | 4.3 (1.9) | | Chile | 563 | 307 (55%) | 255 (45%) | 40.9 (14.9) | 5.4 (1.0) | 5.5 (1.4) | 3.7 (2.0) | 4.2 (1.6) | | China | 501 | 215 (43%) | 282 (56%) | 35.4 (10.8) | 5.9 (1.0) | 5.4 (1.8) | 2.3 (1.6) | - | | France | 585 | 325 (56%) | 256 (44%) | 49.1 (16.9) | 5.0 (1.4) | 5.2 (1.6) | 2.6 (1.8) | 4.4 (1.7) | | Germany | 532 | 284 (53%) | 245 (46%) | 47.6 (17.2) | 4.6 (1.4) | 5.3 (1.7) | 2.8 (1.9) | 3.8 (1.3) | | India | 558 | 259 (46%) | 293 (53%) | 36.8 (13.2) | 6.2 (1.1) | 6.8 (1.7) | 5.2 (1.6) | 4.8 (1.4) | | Mexico | 576 | 321 (56%) | 253 (44%) | 38.8 (13.5) | 5.4 (1.1) | 5.8 (1.6) | 4.1 (1.7) | 3.8 (1.5) | | Poland | 610 | 334 (55%) | 275 (45%) | 45.0 (16.5) | 5.3 (1.8) | 5.4 (1.5) | 3.7 (1.9) | 4.1 (1.6) | | South Africa | 593 | 306 (52%) | 276 (47%) | 39.2 (14.4) | 5.0 (1.1) | 5.3 (1.8) | 5.0 (1.8) | 4.1 (1.2) | | Turkey | 600 | 282 (47%) | 313 (52%) | 38.1 (13.0) | 5.5 (1.2) | 5.9 (1.7) | 4.2 (1.9) | 3.6 (1.8) | | UK | 524 | 277 (53%) | 242 (46%) | 48.3 (17.8) | 5.2 (1.4) | 5.1 (1.8) | 3.0 (2.0) | 4.2 (1.4) | | USA | 576 | 305 (53%) | 269 (47%) | 49.4 (16.2) | 5.3 (1.6) | 5.5 (2.3) | 4.4 (2.2) | 4.6 (1.7) | |

Supplementary Table 4: Breakdown of the sample by country