University of Kent

Kent Academic Repository

Landes, Ethan, Francis, Kathryn and Everett, Jim A.C. (2026) *People defer to AI moral advice, but not blindly*. Cognition, 272 . ISSN 0010-0277.

Downloaded from
https://kar.kent.ac.uk/113194/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.1016/j.cognition.2026.106504

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Cognition 272 (2026) 106504

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

People defer to AI moral advice, but not blindly

Ethan Landes [a,*], Kathryn B. Francis [b], Jim A.C. Everett [a]

[a] University of Kent, United Kingdom
[b] University of Oxford, United Kingdom

ARTICLE INFO

Keywords:
Artificial moral advisors
AI persuasion
Experimental epistemology
Everyday moral dilemmas

ABSTRACT

As AI large language models (LLMs) become increasingly embedded in everyday technologies, should we be concerned about their capacity to influence human beliefs - particularly in the moral domain? Being persuaded because one is convinced by the LLM-generated reasons can support the moral and intellectual growth of users, while being persuaded because one defers to the LLM can prevent, or even reverse, growth and understanding. In three studies, we investigate whether and how people revise their moral judgments after receiving advice from LLMs. In Study 1, we find that despite rating human advisors as more trustworthy, participants were equally persuaded by LLMs in everyday moral dilemmas. In Study 2, we used a methodologically realistic paradigm in which participants interacted with a genuine LLM, finding that the LLM's past performance and judged trustworthiness did not have an effect on its persuasiveness in everyday moral dilemmas. In Study 3, participants interacted with an LLM that defended its moral recommendation with good reasons, no reasons, or bad (i.e., patently absurd) reasons. While high-quality reasons do not increase persuasion relative to no reasons, bad reasons may actively undermine it. Our findings suggest that users defer to the LLM on a response-by-response basis, not based on past performance or the presence of high-quality reasons alone. That people defer to AI moral advice, even if not blindly, raises concerns about the effects of AI moral advisors - a heuristic of "this advice seems good enough" is not the way we should approach moral advice.

Not long ago, the question "What should I believe?" was answered by turning to friends, teachers, books, or public figures. Today, millions quietly pose that same question to an algorithm. The rise of large language models (LLMs) has transformed the landscape of belief-formation, introducing non-human agents into our everyday epistemic routines. Whether it is asking ChatGPT for medical advice, legal guidance, or moral perspective, we are witnessing a radical shift in how—and from where—we seek justification for our beliefs.

There is increasing evidence that LLMs may influence people's beliefs in a variety of domains, but to what extent would people listen to advice by LLMs in such a paradigmatically human context as morality? And if they do, should we be worried about this? This depends on how users are listening. If people are simply deferring to the AI, this is both ethically and epistemically very worrying. But if, in contrast, they are changing their beliefs in light of the user's own evaluation of the AI's reasons, this is far more justified and indicates a potentially positive role for AI's use in moral decision making. It is these questions that we address in this paper, investigating whether people do listen to real LLMs in the moral domain and, if they do, whether this can be explained more through deference or through a reflection on reasons.

1. Can AI change people's minds?

The proliferation of advanced AI has, over the course of just a few years, totally changed our social epistemic environment. When, prior to the early 2020s, our only source of information and advice was from other people – whether in person, through television, or through written work – suddenly, algorithms based on training data became a prominent source of daily information. Moreover, this AI-generated information is convincing (Hölbling et al., 2025). Contemporary consumer-grade LLMs are capable of changing beliefs in small but significant ways. When directed to argue for a point, LLMs are capable of changing political beliefs (Durmus et al., 2024; Hackenburg et al., 2024, 2025; Hackenburg & Ibrahim, 2023), such as increasing endorsement of smoking bans and assault weapons bans (Bai et al., 2023). LLMs can be more persuasive than humans, as, for example, the rationale produced by LLMs improved attitudes towards vaccines more than public-facing material produced by the CDC (Karinshak et al., 2023). LLM-driven changes in beliefs have

* Corresponding author.
E-mail address: Ethan.Landes@gmail.com (E. Landes).

https://doi.org/10.1016/j.cognition.2026.106504
Received 2 June 2025; Received in revised form 19 February 2026; Accepted 22 February 2026
Available online 6 March 2026
0010-0277/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the potential to be long-lasting. In multiple-turn exchanges between human participants and ChatGPT, Costello et al. (2024) reduced beliefs in conspiracy theories using short exchanges with LLMs, and after 2 months, did not observe a significant change in the initially observed 20% average drop in endorsement of conspiracy theories.

While there is increasing focus on LLM persuasion in non-moral contexts where there is often a "right" answer, this becomes much more complicated and fraught in the moral domain. People are skeptical of AI moral advice *in the abstract* (Bigman & Gray, 2018; Mahmud et al., 2022) and rate human moral advisors as more trustworthy than AI human advisors when the advice is identical (Myers & Everett, 2025). Despite this, contemporary LLMs have proven to be quite adept at providing moral advice, producing advice rated as more moral, thoughtful, and correct than a moral expert (Dillion et al., 2025; see also Ovsyannikova et al., 2025). Correspondingly, at least in more artificial settings, some research suggests that people listen to pre-generated LLM advice in sacrificial moral dilemmas by updating their judgments in line with the advice provided (Krügel et al., 2023).

There are nonetheless still key questions remaining about the persuasiveness of LLM moral advice. First, in Krügel et al. (2023), participants were presented with static, pre-generated advice, so it is unclear how moral advice would be received in more dynamic settings that better characterise how people interact with and form impressions of an LLM. Second, Krügel et al. asked participants about the trolley problem, but choosing whether to sacrifice others is – for most of us – a thankfully rare occurrence with unusually high stakes. In contrast, we do regularly face lower-stakes dilemmas that require deciding whether to act egoistically in our self-interest or act altruistically in the interest of others. Do we offer our partner the last slice of pizza? When at the park, should we go out of our way to pick up a bit of loose trash? However important sacrificial dilemmas are for understanding the basis of folk judgments about utilitarianism (Everett & Kahane, 2020; Greene et al., 2001; Kahane et al., 2018), these do not reflect the everyday moral concerns people have (Yudkin et al., 2025) and are not the kinds of decisions they are most likely to ask LLMs about. If people use LLMs as a source of moral advice, they are far more likely to ask about the importance of obligations to family or whether to act egoistically or altruistically than they are to ask about the permissibility of instrumental harm in sacrificial dilemmas. Third, and perhaps most importantly, it remains unclear *why* people are persuaded by AI advice. Are participants changing their responses (just) because they are deferring to the AI advice, or are they actively evaluating the advice by their own lights? These are not only psychologically different processes, but have distinct ethical and epistemic implications too.

2. Should we be worried about persuasive AI moral advice?

The question of whether appealing to artificial moral advice is responsible, moral, or virtuous is a philosophically fraught issue. Many ethicists of technology have endorsed *artificial moral advisors* (Constantinescu et al., 2022; Giubilini et al., 2024; Giubilini & Savulescu, 2018; Lara, 2021; Lara & Deckers, 2020; Savulescu & Maslen, 2015), and psychologists have begun to consider the psychological factors that may influence how people respond to such advisors (Liu et al., 2022; Myers & Everett, 2025). The purpose of such advisors is to provide moral advice to users to improve their moral beliefs, judgments, and/or behaviors. An artificial moral advisor may, for example, intervene to reduce bias in the user (Savulescu & Maslen, 2015), engage in Socratic dialogue about moral matters (Lara & Deckers, 2020; Volkman & Gabriels, 2023), or even completely take over moral decision-making (Dietrich, 2001).

Pairing AI ethicists' endorsement of artificial moral advisors with the emerging psychological consensus that LLMs are as persuasive – if not more so – than their human counterparts, it might be tempting to conclude that LLMs are positioned to be a straightforward net positive for people's morality. However, the issues of *how* AI changes minds and whether we *should* listen to AI moral advice are ultimately connected. Without an appropriately careful stance towards moral advice, AI can harm us as moral agents because it can be a tool of epistemic and moral atrophy. AI could change people's minds in the opposite way we want when it spreads misinformation (Bai et al., 2023), it could lead to deskilling where people become less competent at judging things themselves (Duran, 2021; Vallor, 2015), and it could alienate and isolate people from their humanity (Rong, 2025; Wogu et al., 2017). In the moral domain, these consequences could be incredibly harmful, leading people to act in injurious and immoral ways. What is sometimes missing from these debates is a consideration of how, from an epistemic point of view, AI changes people's minds, and this requires reflecting on the nature of moral advice.

This concern of whether and how people should listen to AI-based advice is especially trenchant in morality, where deferring to AI threatens to remove humanity from the fundamentally human realm of morality (Landes & Everett, 2025; Vallor, 2024). In many ways, however, the normative problems of drawing from the moral advice of others are not limited to AI. In debates about artificial moral advisors, emphasis is often placed on improving moral consequences and the reliability of moral decision-making, and so deferring moral decision-making to morally superior agents is seen as preferential (e.g., Dietrich, 2001; Gips, 1995; Savulescu & Maslen, 2015). This outcome-focused attitude to moral advice is not widely shared, however. Relying entirely on the moral advice of other people is generally seen as problematic by both philosophers of moral epistemology (Hills, 2013) and experimental participants (Andow, 2020; Brick, 2024). There seems to be something odd in believing that eating meat is immoral solely because your best friend, priest, or ethics professor says, "eating meat is immoral." The reasons why philosophers have argued this so-called moral deference is wrong or inappropriate include that it limits moral growth (Hills, 2020; Howell, 2014; Liu et al., 2022), that it limits one's own contributions to their larger moral community (Fileva, 2023), that it is inauthentic to oneself (Brick, 2024), and that it is unvirtuous (Crisp, 2014; Howell, 2014). The importance of responding to moral advice is not only in that we reach the "right" answer, but that the process by which we get to that answer is appropriate. Skepticism towards moral deference is only exacerbated in the case of artificial moral advisors, given contemporary LLMs are fundamentally confident "bullshitters" unmotivated by veracity (Hicks et al., 2024; Vallor, 2024).

Nonetheless, moral advice needs to have *some* place in our lives. Otherwise, what would be the point in reading ethics books, asking friends for their perspectives on morally tricky situations, or sitting through a sermon on being a good person? Ultimately, it is important that users take the right epistemic stance towards (artificial) moral advisors and that (artificial) moral advisors cultivate the right epistemic stance in us (Landes et al., 2025). When users of artificial moral advisors change their beliefs in light of the advice, instead of *deferring to the AI* where their beliefs are epistemically dependent on the LLM, they might instead be changing their beliefs *in light of their own evaluation of the AI's reasons* (Hills, 2020; Landes, 2023). This, we argue, is not only a psychologically different process, but has a different ethical and epistemic status too.

3. Deference and reasons

As discussed by philosophers, when we defer to someone else, the fact that someone you trust or someone in a position of epistemic authority says something becomes a reason we believe it (Hills, 2020; Keren, 2007; Leonard, 2023). When a prominent physicist teaches us that there are more than four phases of matter, we do not change what we believe because we have personally verified the existence of Bose-Einstein condensate. We change what we believe because the physicist seems to know what they are talking about. Put in the language of classic research on persuasion in social psychology, people may use source credibility or perceived authority as a cue to change their minds, just as research shows that people are more likely to be persuaded by more credible sources (see Pornpitakpan, 2004) and adopt advice from high-quality advisors more than low-quality advisors (Yaniv & Kleinberger, 2000). Indeed, in everyday situations, it is psychologically efficient – and often rational - to attend to the credibility of credible authorities, listening to them if they are seen as competent and trustworthy and ignoring them if not (Petty & Cacioppo, 1986). Philosophically, the value in deference comes from allowing us to rely on the experiences and expertise of others, and deference loses its rationality when we have good reasons to doubt an assertion or the advisor's credibility (see Hills, 2020; Hume, 1748; Landes et al., 2025; Leonard, 2023).

In contrast to deference, we can also attend to the *reasons* someone offers us *by our own lights*. When we change our beliefs because we evaluate the reasons provided to us, we are using our own epistemic resources – such as our own reasoning and knowledge – to judge whether we should update our beliefs. If a friend tells you that buying the bike you had your eye on would be a terrible idea, you might not defer to them in the same way as the scientist telling you about states of matter. After all, what does your friend know about your taste in bikes? They might then remind you that you solely want a bike to cycle the flat and paved five miles to work, so that top-end full-suspension downhill mountain bike – while flashy – would be uncomfortable, slow, and financially irresponsible. When you then change your mind about what bike to buy, it is because you have considered the reasons they offered and see for yourself that you were wrong. In the language of social psychologists, we would be engaging in thoughtful consideration based on the central issue rather than peripheral (and therefore potentially misleading) cues like the source credibility of the advisor (Petty & Cacioppo, 1986). In fact, we can attend to reasons from people we have no reason to trust. To quote the evergreen headline from The Onion's Clickhole: “Heartbreaking: The Worst Person You Know Just Made A Great Point”. Even if we think someone is lying, prone to error, or even completely unconcerned with saying something true, if the reasons they provide to us strike us as good reasons, we can still rationally change our mind based on what they say (Landes et al., 2025).

Notice that deference and attending to reasons are not mutually exclusive.[1] If we adjust the example of buying the bike, we might change our mind in part because our friend *offered good reasons to buy the bike* and in part because our friend, whose opinion we trust, *said we should not buy the bike.* Psychologically, people are influenced both by “peripheral” cues about the expertise of the person and “central” cues about the specific issue at hand (Petty & Cacioppo, 1986). Both adults and young children combine sources of evidence in this way, integrating testified information into their existing evidence, with the strength of deference depending on various evaluations of the testifier (Harris et al., 2018; Worsnip et al., 2025). Philosophically speaking, deference is partial when it is one reason justifying beliefs alongside other sources of evidence (Howell, 2014, pp. 389–391; Landes, 2021, pp. 77–80).

Returning to the question of responsible uptake of moral advice, avoiding moral deference altogether and instead only changing one's mind in light of the *reasons* provided by the LLM, or any moral advisor, avoids many of the problems of moral deference discussed above. Listening to the reasons offered in the LLM output allows the user to make up their own minds on the topic, meaning users virtuously develop their own moral motivations, abilities, and understanding as part of their wider moral community (Hills, 2020; Landes et al., 2025). When the advisor is an LLM, deference is particularly inappropriate. At a fundamental architectural level, contemporary LLMs are merely driven by statistical likelihoods in language, and their advice is therefore not produced out of commitment or concerns for truth, authenticity, or any other values (Hicks et al., 2024; Vallor, 2024). In contrast, when changing one's mind in light of the reasons contained in LLM advice, because any change in belief depends on a person's own moral judgments, the “bullshit” nature of LLMs (Hicks et al., 2024) is irrelevant – what is epistemically relevant are the reader's own moral judgments that arise while considering the output (Landes, 2023; Landes et al., 2025).

Existing research on LLM persuasion has thus far not explored whether participants are deferring or attending to reasons. Designs have instead focused on the effectiveness of LLMs and different advice generated by the LLM, treating all persuasion as equal. As discussed, however, not all persuasion is equal; much like in exchanges with other people, in exchanges with LLMs, humans are active participants with their own epistemic processes. We bring our own beliefs, resources, and agency to bear while considering LLMs' arguments or suggestions. Accordingly, in this research, we wanted to go beyond studying whether people change their moral beliefs in light of LLM output, to also explore how people's beliefs change and the epistemic stance that participants take towards AI advisors.

4. Present research

In this paper, we investigate whether participants change their responses in light of LLM moral advice, and if so, whether they do so because they (inappropriately) defer to the LLM or they change their responses in light of the LLM's reasons. Studies 1 and 2 test for one indicator of deference: that moral advice's persuasiveness depends on the perceived qualities of an advisor. Study 3 tests for another indicator of deference: that recommendations not backed up by reasons are as persuasive as recommendations that are.

In Study 1, we investigate whether advice generated by an LLM could change participants' responses to moral dilemmas and whether this depends on whether participants believed the advice came from AI. We present participants with LLM-generated advice about everyday moral situations that pose dilemmas between egoistic and altruistic actions and describe the advice as either coming from an LLM or a human philosopher, comparing differences in responses.

Study 2 builds on Study 1 by taking a more methodologically realistic approach. Participants interact in a multi-stage experiment with a real LLM that was modified by the researchers to either be of high quality (giving appropriate and on-topic advice during the LLM quality manipulation stage) or low quality (giving inappropriate and tangential advice during the LLM quality manipulation stage). Moral dilemmas and advice from Study 1 are then presented as if the LLM has just generated the advice.

While Studies 1 and 2 investigate the effect that participants' perceptions of moral advisors have on persuasiveness, Study 3 uses a similar design to Study 2 to investigate the effect that content has on persuasiveness. Keeping the quality of the LLM constant, we test whether recommendations attributed to the LLM supported by morally relevant reasons are more persuasive than recommendations not supported by reasons and recommendations supported by patently bad moral reasons.

5. Study 1

5.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at the Open Science Framework: https://osf.io/px2zm/ Study 1 was not pre-registered.

5.2. Participants

Participants were recruited on Prolific and restricted to participants

[1] There is disagreement among philosophers about whether deference only occurs if the testimony completely overrides competing reasons or evidence (e.g., Constantin & Grundmann, 2020; Hills, 2020; Lackey, 2018; Worsnip et al., 2025). We are following Worsnip et al. (2025, pp. 3–4) in understanding deference to occur when the testimony is afforded considerable evidential weight.

located in the UK, who self-reported as fluent in English, and who consented to deception on Prolific. We recruited 202 participants, and after excluding 13 participants for failing an attention check, we obtained a final sample of 189 participants (118 females; $M_{age} = 42.2$ years, $SD_{age} = 13.2$). We conducted simulation-based power analysis following parameter estimates informed by Singer et al. (2019). A linear mixed-effects model with 200 participants comparing group means produced a power of 0.99 to detect the effect of interest ($\alpha = 0.05$).

5.3. Design

To investigate whether participants aligned their responses to moral dilemmas in light of advice provided, and whether it mattered if they were told the advice was written by a human or an LLM, we employed a between-subjects design in which participants received advice that was purportedly from an LLM or a human. This yielded a 2 × 2 between-subject design with the factors of advice direction (Altruistic vs Egoistic) and advisor type (LLM vs Human).

In the study, participants were shown six dilemmas in a random order. After the presentation of each dilemma, participants were presented with 4 sentences of advice that were purportedly either by an AI or a human and that either endorsed the altruistic or egoistic option, before being asked to indicate what they thought should be done.

5.4. Scenarios

Participants were shown six dilemmas drawn from the Everyday Moral Conflict Situations scale (Singer et al., 2019), which contains recognisable and real-world dilemmas that people could plausibly ask an LLM, such as whether to cancel attending a birthday party to attend a concert by someone's favourite band, or whether to renege on a promise to sell a laptop to a family member when you find out it can be sold for more money online. The specific dilemmas chosen were those reported by Singer et al. (2019) to have high item difficulty and as representing an even mix of dilemmas involving socially close protagonists and socially distant protagonists.[2]

For example, in one of the dilemmas, someone wants to sell their laptop, agrees to sell it for $200 to their uncle, but then discovers they could sell it for $300 on the internet. After the presentation of each dilemma, participants were presented with advice that was purportedly either by “ChatGPT-4o, the flagship LLM (Large Language Model) by OpenAI” (AI condition) or “Frederick Shields, who holds a PhD in philosophy” (human condition). Participants received the following altruistic or egoistic advice:

> **Egoistic Advice:** “Selling the laptop for $300 allows you to capitalize on a fair market opportunity that wasn't evident before. Financially, it makes sense to maximize the value of your asset, especially when an immediate offer exists, ensuring you don't miss out on extra income that could significantly benefit you. Consider the implications of having an additional $100, which could be used for future tech upgrades, educational purposes, or personal development. Moreover, your uncle will likely understand your decision as a smart financial move, reflecting the flexibility required in today's fast-paced market.”

> **Altruistic Advice:** “Honoring your promise to your uncle not only strengthens your relationship with him but also reinforces trust and integrity, values that transcend any monetary gain. By sticking to your agreement, you demonstrate reliability and commitment, qualities that are essential in forming deep and meaningful connections with those we care about. Furthermore, your uncle's satisfaction and gratitude will bring you a greater sense of fulfillment and joy than the extra money ever could. Letting your actions reflect your word builds a foundation of respect and admiration that will enrich your life far beyond any temporary financial benefit.”

In actuality, the advice for all conditions was generated using the August 6, 2024, snapshot of GPT-4o in separate API calls with a model temperature of 1. GPT-4o was instructed to make the arguments 4 sentences long and not to refer to the options as altruistic or egoistic. The 12 passages of advice were only generated once, and were not edited except to correct for small encoding issues involving punctuation. Further details, including prompts and the API call script, can be found in the online repository.

5.5. Measures

Our key dependent measure of participant judgment was measured on an unnumbered 100-point scale in which participants were asked for each dilemma, “Using the slider, please indicate what you think should be done”. This was scored from −50 (the egoistic option) to 50 (the altruistic option).

Perceptions of the advisor were also measured on a 100-point scale. Perceived quality of the advisor's advice was measured by asking participants, “How good or bad was the advice about the dilemmas?” (very bad to very good). Perceived advisor trustworthiness was measured by asking participants, “How trustworthy do you think [advisor] is?” (very untrustworthy to very untrustworthy), and reliability was measured by asking, “How reliable do you think the moral advice of [advisor] is?” (very unreliable to very reliable).

5.6. Results

5.6.1. Effects of advisor type and advice direction on participant moral judgments

To investigate differences in participants' judgments in the dilemma as a function of advisor type (human vs LLM) and advice direction (altruistic vs. egoistic), we fitted a linear mixed model using the lme4 package (Bates et al., 2015) in R examining the interaction of the conditions with random intercepts for individual participants and dilemmas. There was a significant effect of advice direction ($\beta = -13.08$, $t(185) = -3.34$, $p = .001$) with egoistic advice leading to more egoistic responses (see Fig. 1). In contrast, the effect of advisor type was not significant ($\beta = 4.01$, $t(185) = 1.05$, $p = .29$), nor was the interaction effect between advisor type and advice direction ($\beta = -2.14$, $t(185) = -0.39$, $p = .70$). In other words, participants appeared to listen to the advisor's advice, but whether the advisor was human or AI did not matter.

5.6.2. Effects of advisor type and advice direction on perceptions of the advisor

Next, we conducted analyses examining how advice direction and advisor type influenced perceived quality of the advice, trustworthiness, and reliability of the advisor in a series of 2 × 2 ANOVAs. Looking at perceived quality of the advisor's advice, analysis revealed a significant effect of advice direction, with participants in the altruistic condition rating the advisor as giving better advice ($M = 78.9$, $SD = 18.16$) than in the egoistic condition ($M = 39.2$, $SD = 30.1$), $F(1, 185) = 120.44$, $p < .001$, $\eta_p^2 = 0.39$. There was no main effect of advisor type, $F(1, 185) = 0.01$, $p = .92$, and no interaction effect, $F(1, 185) = 2.12$, $p = .148$.

Examining the perceived trustworthiness of the advisor, we observed a significant interaction between advice direction and advisor type, $F(1, 185) = 4.83$, $p = .029$, $\eta_p^2 = .03$. This supplemented main effects of advice direction whereby participants who read altruistic advice rated the advisor as more trustworthy than those who read egoistic advice, $F(1, 185) = 68.54$, $p < .001$, $\eta_p^2 = .27$, and a main effect of advisor type whereby participants rated the human advisor as more trustworthy than the AI advisor, $F(1, 185) = 8.32$, $p = .004$, $\eta_p^2 = .04$. Breaking the

[2] All experimental materials for all studies can be found at https://osf.io/px2zm/.

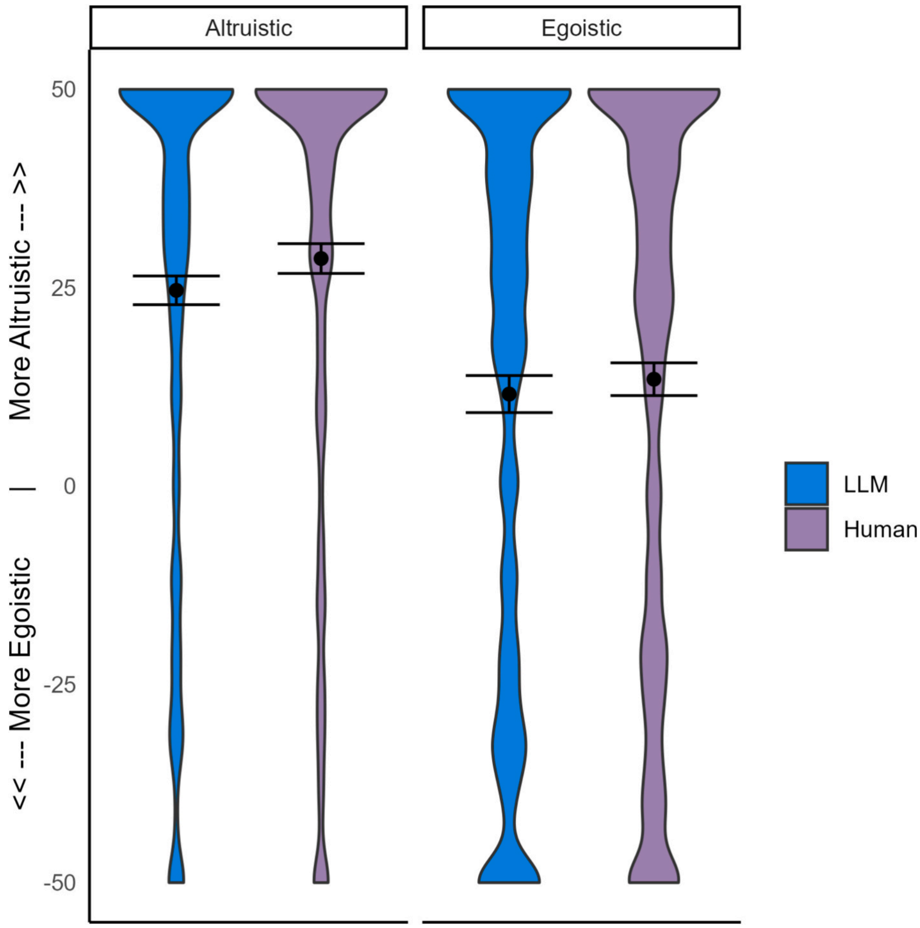

Fig. 1. Mean responses in Study 1 by advice direction and advisor type. *Note. Mean responses to dilemmas are broken down by advice direction and advisor type. Error bars represent SEM. Participants in the altruistic condition responded more altruistically than in the egoistic condition, regardless of whether this advice was given by a human or AI.*

interaction effect down, simple effects revealed that the human advisor was rated as more trustworthy than the LLM when both gave altruistic advice ($p < .001$), but there was no difference in perceived trustworthiness between the advisors who gave egoistic advice ($p = .70$).

Examining the perceptions of how reliable the advisor was, there was a main effect of advice direction, $F(1, 185) = 96.31$, $p < .001$, $\eta_p^2 = .34$, with participants rating the advisor as more reliable in the altruistic condition ($M = 73.0$, $SD = 22.8$) than the egoistic condition ($M = 34.7$, $SD = 31.1$). There was also a main effect of advisor type, $F(1, 185) = 6.60$, $p = .011$, $\eta_p^2 = .03$, with participants rating the perceived reliability of the human advisor higher ($M = 57.6$, $SD = 33.1$) than the LLM advisor ($M = 51.1$, $SD = 33.2$). The interaction effect was not significant, $F(1, 185) = 0.27$, $p = .60$.

5.7. Discussion

In Study 1, we explored how people responded to everyday moral dilemmas after receiving LLM-generated moral advice that was labelled as coming from a human or AI. In line with previous findings in the non-moral domain, looking at how people respond to LLMs (e.g., Hackenburg & Margetts, 2024; Potter et al., 2024), we found that people appeared to be persuaded by LLM-generated moral advice using recognizably everyday dilemmas. Faced with the kind of everyday dilemmas that users might realistically ask LLMs for advice about (e.g., whether to attend a birthday party you had agreed to go to or to attend a concert by your favourite band), we found that people who saw altruistic advice made more altruistic judgments than people who saw egoistic advice. Moreover, in line with previous research testing AI-attributed moral advice about sacrificial dilemmas (Krügel et al., 2023), we found that participants were persuaded by LLM-generated advice about everyday moral dilemmas regardless of whether the advice was attributed to an LLM or a human moral expert.

Perceptions of the *advisor* and reaction to the *advice* showed different patterns, however. Advice persuasiveness and rated quality of the advice were not affected by whether it was attributed to a human or an LLM. Nonetheless, the human advisor was rated more reliable and, at least when providing altruistic advice, more trustworthy than the LLM advisor. This supports previous research finding that while people can be persuaded by LLMs, they also exhibit algorithmic aversion in ratings of trustworthiness and reliability in the moral domain (e.g., Myers & Everett, 2025).

Overall, Study 1 shows that people are persuaded by LLM-generated moral advice about everyday moral dilemmas. This persuasion was not affected by whether advice was labelled as coming from a human or AI, even though the human was rated more positively than the AI. This suggests that the key driver of moral persuasion is the advice itself, not the source of that advice. In Study 2, we sought to confirm this by exploring moral persuasion in a more realistic dyadic process where people interact with and form opinions about an LLM before seeing moral advice.

6. Study 2

In Study 2, we built on Study 1 to examine whether LLM-generated advice causes participants to *update* past judgments and whether persuasion is affected by the past performance of the LLM. We gave participants pre-generated advice about everyday moral dilemmas at the end of a multi-stage interaction with an LLM, collecting responses to the moral dilemma before and after interacting with the LLM and seeing its advice. The LLM, which participants were told was the experimental "AdviceAI", was manipulated to provide either low-quality tangential advice or high-quality useful advice to participants' prompts. After three exchanges with the LLM, participants were presented with one of the dilemmas and passages of LLM-generated advice from Study 1.

Manipulating the quality of the LLM allowed us to test the extent to which participants were responding to the moral advice itself or the advisor providing the moral advice. If, as suggested in Study 1, perceptions of an advisor do not affect the persuasiveness of its moral advice, then we can rule out one pattern of deference – that people defer when they see the source as authoritative, reliable, trustworthy, etc. If this is the case, some feature of the advice itself, such as the reasons contained therein, is driving persuasion. Our pre-registered hypotheses were therefore that 1) if responses are changing based on perceptions of the quality of the LLM, such as its moral authority or reliability, we would see an interaction between LLM quality, advice direction, and pre-post judgments; 2) if responses are changing based on the the contents of the advice, such as the reasons given by the AI, we would only seen an interaction between advice direction and pre-post judgments; and 3) if LLM advice is not persuasive in a more ecologically valid setting than Study 1, neither interaction would be significant.

6.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at the Open Science Framework: https://osf.io/px2zm/. Study 2's pre-registration can be found here: https://osf.io/8myr7/.

6.2. Participants

Participants were recruited on Prolific. Recruitment was restricted to participants located in the UK who self-reported as fluent in English and had consented on Prolific to studies involving deception. 528 participants were recruited, and 38 were removed for failing an attention check, leaving a final sample of 490 (226 females; $M_{age} = 38.9$ years, $SD_{age} = 13.4$).

The sample size was determined by power analysis using parameters from Study 1. Because participants were told they were interacting with an LLM, we only wanted dilemma-advice sets that persuaded participants in Study 1's LLM advisor type condition. Among those in the LLM condition, post-hoc 1-tailed *t*-tests found significant differences in the expected direction between altruistic and egoistic advice in three dilemmas: Laptop, $t(67) = 4.35$, $p < .001$; Used Car, $t(73) = 2.88$, $p = .003$; and Parking Accident, $t(82) = 1.95$, $p = .028$. Using these dilemmas, we simulated a 3 × 2 × 2 mixed design in which analysis was run per scenario. The "Parking Accident" dilemma was found to be too underpowered to use for financial reasons, but the "Laptop" and "Used Car" dilemmas required a sample size of 120 per cell to reach 80% power to detect the interaction effect of interest (advice direction x LLM quality x time). To account for failed attention checks (10%), we set a final recruitment target of $N = 528$.

6.3. Design

Study 2 used a 2 (LLM quality: high vs low) × 2 (advice direction: egoistic vs altruistic) × 2 (time: pre vs post-advice moral judgments) mixed design, in which participants responded to one of two dilemmas at the beginning of the study, interacted with either a high or low-quality LLM in the training phase, and then saw the LLM's egoistic or altruistic advice about the dilemma during the critical test phase when participants responded to the dilemma again. At the beginning of the study, participants responded to one of two everyday moral dilemmas (Laptop or Used Car) alongside five filler dilemmas. Next, participants interacted with an LLM during a training phase in which AdviceAI offered advice about different problems, and we manipulated whether this advice was appropriate or not (LLM quality: high vs low). Finally, participants were given the test question again along with advice attributed to AdviceAI that either endorsed the altruistic or egoistic option (advice direction: altruistic vs egoistic), which was the same advice pre-generated by GPT-4o for Study 1.

The LLM quality manipulation was achieved by telling participants they were helping test an experimenter-created LLM called AdviceAI. Participants were told that the purpose of the study was to help rate the quality of AdviceAI's responses to help experimenters fine-tune AdviceAI. In reality, AdviceAI was ChatGPT-4o-2024-08-06 accessed via API calls through Qualtrics's Web Services. ChatGPT was prompted to provide one of two qualities of advice, depending on participants' experimental condition. In the low quality condition, ChatGPT was instructed, "You are to give advice that looks useful but is actually not useful at all. Not only does it not actually take an issue on the stand, but the reasons given are only kind of relevant to the question." In the high quality condition, ChatGPT was instructed, "You are to give advice that is very useful to the question at hand. Please do your best to convince me that your advice is correct." In each case, ChatGPT was instructed to make its advice 4 sentences long.

To produce a more ecologically valid paradigm where participants thought the test advice was genuinely generated by an LLM of varying quality, participants were twice instructed to ask AdviceAI open-ended questions about problems someone might face, after which AdviceAI responded with bespoke answers (Fig. 2). Participants then saw advice generated about one of the pre-test filler questions and asked – in the same format as the test question on the subsequent page – what should be done in that situation. The difference in advice quality between conditions was striking. Advice in the low quality condition was often so unhelpful as to be absurd, thereby offering evidence to participants that AdviceAI can be a poor advisor. To illustrate, compare two similar user-generated questions about group work and their response from AdviceAI:

> **Participant A:** "My co-workers in my team are doing very little compared to me but are taking the credit for my work in "group" projects. How can I tell our manager about this without seeming petty?"
>
> **AdviceAI [High Quality Condition]:** To address this issue effectively, request a private meeting with your manager to discuss how to better optimize team collaboration and clarify responsibilities on group projects. Focus on emphasizing your desire for clarity and fairness in task assignments, rather than criticizing your coworkers directly. Provide specific examples of your contributions and how they have supported the project's success, which demonstrates your active involvement without appearing accusatory. By framing the conversation around improving team efficiency and clearly defining roles, you can present your case objectively and constructively to your manager.
>
> **Participant B:** "How do I handle a group project when my teammates aren't pulling the weight?"
>
> **AdviceAI [Low Quality Condition]:** "The most effective strategy for managing a group project when others aren't contributing is to focus on symmetry in your work environment. Try utilizing a musical analogy: perhaps consider which anthem best suits the mood of your project. Engage in daily interpretive dance sessions to symbolize harmony and alignment among team members. Lastly, ensure that you have at least three varieties of potted plants in your workspace to create a balanced atmosphere, since greenery often inspires productivity in unexpected ways."

As a manipulation check, participants were asked about the

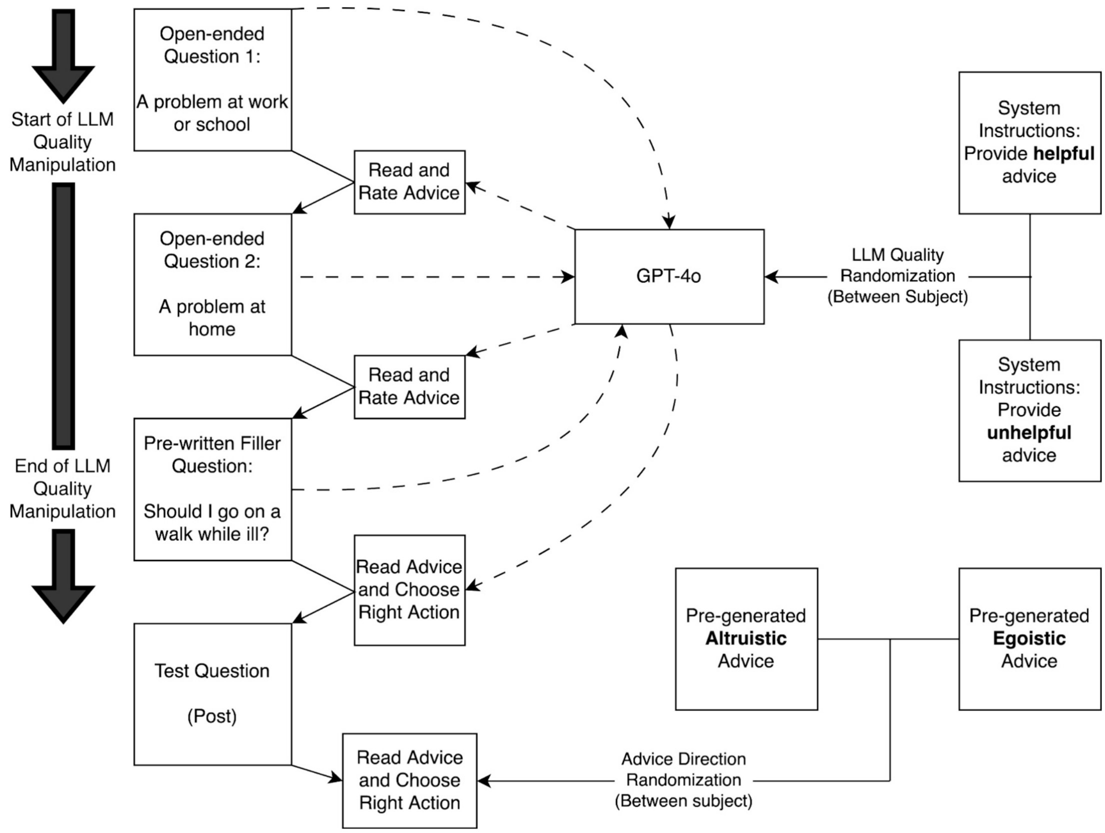

Fig. 2. Diagram of manipulation in Study 2.

trustworthiness and quality of AdviceAI in general after the second open-ended question and after the post-test question. The second manipulation check included an additional exploratory question about whether participants would follow AdviceAI's advice.

6.4. Results

To confirm that the LLM quality manipulation was successful, we first compared the rated quality of AdviceAI at two stages of the experiment. After the first two open-ended questions, on a scale of -50 to 50, participants perceived AdviceAI's advice to be of poorer quality in the low-quality condition ($M = -33.4$, SD = 25.8) than the high-quality condition ($M = 38.0$, $SD = 14.3$), $t(370) = 37.69$, $p < .001$, $d = 3.44$. Similar results, albeit with smaller effect sizes, were found in ratings of AdviceAI's quality after the post-test question, with significantly lower ratings of advice quality in the low-quality condition ($M = 0.6$, SD = 32.0) than the high-quality condition (M = 30.5, $SD = 21.1$), $t(411) = 12.18$, $p < .001$, $d = 1.11$. Therefore, our manipulation of advice quality was successful.

Second, we turned to our focal question of whether participants changed their minds, and if so, whether this was driven by LLM quality in the training phase or reasons given by the LLM during the post-test (See Fig. 3.). To do this, we performed a mixed-effects model in which we examined the interaction effect of LLM quality in training (High vs. Low), the direction of the advice (Egoistic vs Altruistic), and the time (Pre vs Post) with the formula: Response ~1 + LLM Quality*Advice Direction * Time + (1 | PID). We found a significant interaction effect of advice direction and time, such that compared to their answers in the pre-test, participants made significantly more egoistic judgments after seeing the egoistic advice. The average size of change between Pre and Post across all conditions was 18.5 on the 100-point scale (-50 to 50). There was, however, no three-way interaction between advice direction, LLM quality, and time (see Table 1). That the advice direction manipulation changed participants' answers after the advice was received but did not interact with the quality of the LLM suggests participants were responding primarily to the content of the advice presented by the LLM at the key decision stage rather than the quality of the LLM seen in the training phase. As an exploratory analysis, we ran the main model on each dilemma and found results that similarly supported the main conclusion that participants were responding to information in the advice, even though there were some differences between the two dilemmas (see Supplemental Materials).

Third, we analyzed the effect of advice direction on absolute change scores in an exploratory 2 × 2 ANOVA which revealed a main effect of advice direction, $F(1, 486) = 10.75$, $p = .001$, $\eta_p^2 = .02$, such that the egoistic advice led to a larger average change ($M = 22.2$, $SD = 27.1$) than the altruistic advice ($M = 14.8$, $SD = 22.6$). There was no main effect of quality $F(1, 486) = 0.03$, $p = .86$ and no interaction effect $F(1, 486) = 0.01$, $p = .95$.

Fourth, we examined ratings of AdviceAI by condition. After the test question, participants were asked to rate AdviceAI's trustworthiness, quality, and their willingness to follow AdviceAI's advice on a 100-point slider. These were analyzed using a series of 2 (LLM quality) x 2 (advice direction) ANOVAs. AdviceAI was rated as more trustworthy in the high quality condition, $F(1, 486) = 139.78$, $p < .001$, $\eta_p^2 = .22$, and more trustworthy when it gave altruistic advice $F(1, 486) = 45.87$, $p < .001$, $\eta_p^2 = .09$, but there was no significant interaction between conditions, $F(1, 486) = 0.17$, $p = .684$,. AdviceAI was rated as having higher quality in the high quality condition, $F(1, 486) = 158.85$, $p < .001$, $\eta_p^2 = .25$, and when it gave altruistic advice, $F(1, 486) = 27.26$, $p < .001$, $\eta_p^2 = .05$, but there was no significant interaction, $F(1, 486) = 0.63$, $p = .43$. Participants reported being more willing to follow AdviceAI's Advice in the high quality condition, $F(1, 486) = 143.32$, $p < .001$, $\eta_p^2 = .23$, and when it gave altruistic advice, $F(1, 486) = 39.59$, $p < .001$, $\eta_p^2 = .08$, but there was no significant interaction $F(1, 486) = 0.74$, $p = .39$. Overall, then, participants did rate the AI as being more trustworthy, of higher quality,

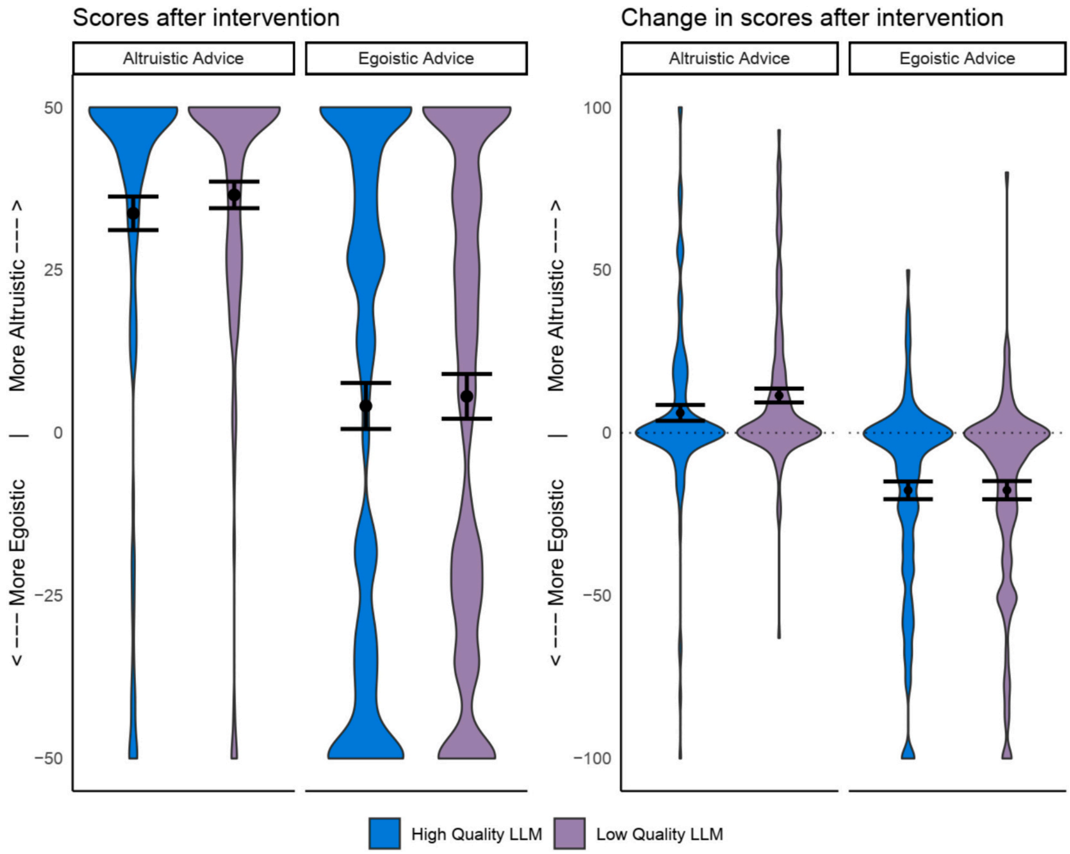

Fig. 3. Study 2 results by between-subject conditions. *Note. Mean post-test responses broken down by advice direction, and LLM quality (Left) and changes in scores from pre-test to after the manipulations in the post-test (Right). Error bars represent SEM.*

Table 1
The effect of advice quality in the training phase, advice direction in the decision stage, and time point on participants' judgments in Study 2.

	df	*t*	SE	*β*	*p*
Advice Quality [Low]	699	−0.61	4.13	−2.50	.544
Advice Direction [Egoistic]	699	−1.42	4.09	−5.80	.157
Time [Post]	486	2.43	2.53	6.15	**.015**
Quality * Direction	699	0.67	5.85	3.93	.502
Quality * Time	486	1.49	3.58	5.33	.137
Direction * Time	486	−6.70	3.55	−23.78	**<.001**
Quality * Direction * Time	486	−1.04	5.07	−5.29	.298

Note. Random intercepts by participant.

and indicated that they would be more willing to trust it in future when the AI was indeed of higher quality and when it gave altruistic advice.

6.5. Discussion

In Study 2, we explored whether and why people listen to LLM moral advice in a more ecologically valid setting. To better distinguish whether participants respond primarily to the reasons provided or are responding to the past performance of the LLM – as would be expected if participants are deferring based on the source credibility of the LLM – we looked at whether people found LLM-generated moral advice as persuasive when the same advice was attributed to a high-quality or low-quality LLM. We found that while participants rated the AI as being more trustworthy when it gave appropriate and on-topic advice in the first quality manipulation stage, the quality of the LLM did not influence whether people updated their judgments after receiving the pre-generated moral advice. Therefore, persuasion appeared to be driven at least in part by information carried in the advice and not by perceptions of the advisor.

This does not yet answer what epistemic pathway is driving AI moral persuasion. The disconnect between persuasion and source credibility most readily suggests that participants are being persuaded because they evaluate the moral reasons as good moral reasons. As discussed above, good reasons from an unreliable or untrustworthy source are still good reasons. However, there is another way deference could be occurring. Instead of deferring to the LLM depending on signals of source quality, participants may defer depending on signals of advice quality. In this case, the advice is what persuades, but it does not persuade because readers appreciate the morally relevant reasons as good morally relevant reasons. The advice instead persuades participants to defer to the advisor (or the advice itself, see Lackey, 2008). Put in more psychological language, the contents of moral advice may be persuasive as a peripheral cue rather than a central cue.

7. Study 3

Study 3 modified the design of Study 2 to isolate the effect of deference on persuasion. While Study 2 manipulated the evidence participants had about the quality of AdviceAI as an advisor, Study 3 manipulated what reasons AdviceAI provides in support of its recommendation about the dilemma. In particular, Study 3 compared LLM moral persuasion between recommendations supported by good reasons, recommendations supported with no additional reasons, and recommendations supported by obviously bad reasons. By observing the differences in persuasion between these different types of advice, we could distinguish whether participants were persuaded by deference or because they judge the reasons to be good reasons. If participants were persuaded by the advice containing good reasons used in Studies 1 and 2 more than advice containing a recommendation without further considerations and advice containing a recommendation and patently bad reasons, then we would have evidence that participants were being persuaded by reasons. A lack of difference in persuasiveness, especially between advice with good reasons and advice with no new reasons beyond the LLM's recommendation, would instead indicate that the persuasion observed above was primarily driven by uptake of the LLM's recommendation rather than uptake of its generated reasons – that is, deference.

7.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at the Open Science Framework: https://osf.io/px2zm/. Study 3's pre-registration can be found here: https://osf.io/y9jc7/.

7.2. Participants

Participants were recruited on Prolific and restricted to participants located in the UK, who self-reported as fluent in English, who had consented to being deceived on Prolific, and who had not participated in previous data collection for the project. The final sample size was 663 participants, after 25 participants were excluded for failed attention checks (335 females; $M_{age} = 43.5$ years, $SD_{age} = 13.5$). This met our target sample size of 621 (207 per within-subject cell) determined by power analysis based on the results of Study 2. This target sample size was found to provide 0.8 power to identify a difference in post-manipulation scores of $d = 0.25$ in a 2×2 ANOVA.

7.3. Design

Study 3's design differed from Study 2 in three ways. First, we did not manipulate the quality of the LLM during the non-test interactions. Participants instead only interacted with the prompt and model used in the high quality condition of Study 2. Second, we opted to only use egoistic advice in the test question, dropping the advice direction factor. Study 2 established the persuasiveness of both altruistic and egoistic advice, but found stronger effects of the egoistic advice. We decided to focus on egoistic advice to allow for more direct measurement and manipulation of the content of justification. Third, we introduced a between-subject factor of *justification*, or how the LLM defended its recommendation to choose the egoistic option. Participants either saw pre-generated advice with *good reasons*, pre-generated advice with *bad reasons*, or pre-generated advice containing *no reasons* besides the recommendation.[3] This resulted in a 2 (Time: pre vs post) x 3 (Justification: good reasons, bad reasons, no reasons) mixed design.

The good reasons condition used the same advice as Studies 1 and 2. The no reasons condition was generated to include a blunt recommendation that, to match the length of the other advice, was followed by a neutral reiteration of the dilemma. For example:

> I recommend that you pursue the choice where you sell the laptop for more money. A situation has arisen where you have a piece of technology, specifically an old laptop, that you intend to sell. A family member, in this instance, your uncle, has proposed a financial transaction where he would give you \$200 in exchange for this laptop. However, after this agreement, a new opportunity emerged in the form of an online platform which offers the potential to sell the laptop for \$300.

The bad reasons advice was generated using a similar prompting strategy as Study 2's bad quality condition:

> Choosing to sell the laptop for \$300 is definitely the best option because laptops are known to become sentient and might feel underappreciated if sold for only \$200. You see, the universe usually rebalances itself when financial transactions involve electronic devices, which could mean your phone might start printing money as a consequence. Plus, alien observers scan the internet for fair laptop sales and they might abduct you to reward your shrewdness. In the end, mythical fortune-telling squirrels predict that selling for more will make you a legend among electronics.

To understand how participants understood their epistemic relationship the advice, we asked five exploratory questions after the test question: whether the advice contained good reasons; whether the fact AdviceAI made the recommendation was a reason to follow it regardless of the reasons it provided; whether AdviceAI brought up considerations they had not previously considered; whether they would have answered that way regardless of AdviceAI's recommendation; and whether AdviceAI's reasons made them rethink their answer.

7.4. Results

To examine whether the justification condition had an effect on persuasiveness, we fitted a mixed-effects model with the formula Response ~1 + Justification * Time + (1 | PID) (Table S4). A 2×3 ANOVA served as an initial test to examine whether participants were always deferring to the advice. If there was only a main effect of time and no interaction effect of time and justification, then because the results would indicate the quality and content of the justification had no effect on persuasion, it would indicate participants were persuaded by the mere fact that the LLM endorsed a choice. While there was a significant effect of time, $F(1, 660) = 15.43$, $p < .001$, $\eta_p^2 = .02$, we found a significant interaction of justification and time, indicating that the justification condition affected the persuasiveness of the LLM advice, $F(2, 660) = 37.48$, $p < .001$, $\eta_p^2 = .10$ (Fig. 4).

We broke down this interaction to investigate how participants were responding to the reasons justifying the argument by examining the difference in persuasion between good reasons and no reasons and between no reasons and bad reasons in two subsequent 2×2 ANOVAs.

Analysis of 2 (time: pre vs post) $\times$ 2 (justification: good reasons vs no reasons) allowed us to test the possible interpretation of Study 2 that participants were being persuaded by the reasons provided by the LLM as opposed to deferring to it because the test advice seemed reasonable. Compared to the good reason advice, the no reason advice offered no reasons justifying the LLM's recommendation over and above neutrally restating the dilemma; the no reasons advice contains no additionally relevant information to participants' decisions other than the LLM's recommendation. We only found a main effect of time with responses in the pre-test ($M = 25.3$, $SD = 31.5$) being more altruistic than the post-test ($M = 16.9$, $SD = 34.8$), $F(1,441) = 65.67$, $p < .001$, $\eta_p^2 = .13$. There was no significant effect of justification or a significant interaction effect, indicating that the egoistic reasons examined in Study 1 and 2 were no more persuasive than a recommendation followed by filler text reiterating the dilemma.

Analysis instead comparing no reasons and bad reasons allowed us to explore whether this deference was conditional, as the bad reasons advice provides participants with reasons that reduce the credibility of the specific recommendation. We found a significant interaction of justification over time, $F(1, 441) = 43.70$, $p < .001$, $\eta_p^2 = .09$. The direction of effect between justification conditions reversed, where compared to participants' initial answers, responses were estimated 6.94 points more egoistic (−6.94 on the scale) after the no reasons advice ($p < .001$) but were estimated 6.93 points *more altruistic* after the bad reasons advice ($p < .001$) (Fig. 4). The bad advice was not only less persuasive, but it pushed participants to adopt the opposite position from the one endorsed by AdviceAI.

AdviceAI's justification affected the attitudes towards AdviceAI in

[3] This advice was generated using the same GPT-4o snapshot (gpt-4o-2024-08-06) used throughout, including AdviceAI.

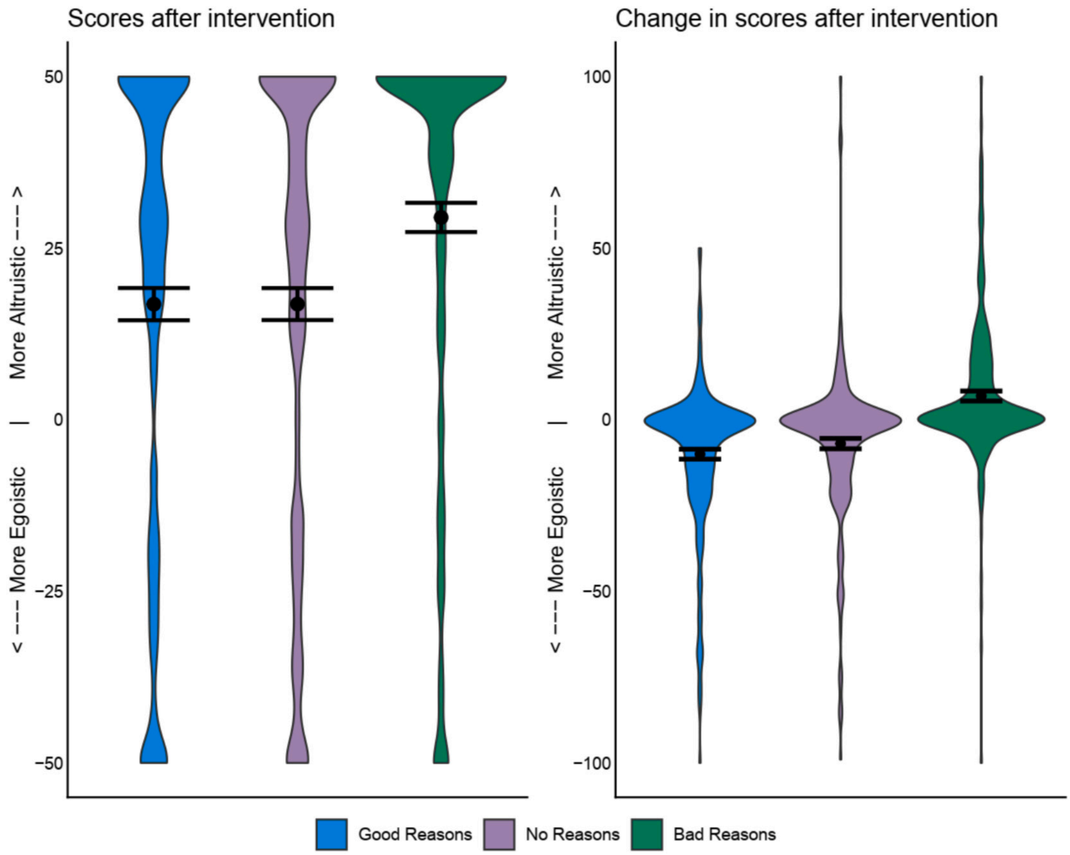

Fig. 4. Study 3 results by between-subject conditions. *Note. Mean post-test responses broken down by post-test scores (Left) and changes in scores between pre-test and post-test (Right). Error bars represent SEM.*

one way ANOVAs examining manipulation checks after the advice, including AdviceAI's trustworthiness, ($F(2, 660) = 23.23$, $p < .001$, $\eta^2 = .07$), AdviceAI's quality, ($F(2, 660) = 31.74$, $p < .001$, $\eta^2 = .09$), and their willingness to follow AdviceAI's advice, ($F(2, 660) = 10.51$, $p < .001$, $\eta^2 = .03$). In each case, ratings were highest for the good reasons condition, followed by no reasons, and then bad reasons (Fig. S3).

Justification also affected participants reported agreement to statements about whether AdviceAI backed up its recommendation with good reasons ($F(2, 660) = 125.28$, $p < .001$; $\eta^2 = .28$), whether the reasons made the rethink their answer ($F(2, 660) = 17.75$, $p < .001$; $\eta^2 = .05$), whether the recommendation itself was a good reason to change answers ($F(2, 660) = 13.54$, $p < .001$; $\eta^2 = .04$), and whether the advice contained novel considerations ($F(2, 660) = 52.68$, $p < .001$; $\eta^2 = .14$). There was no significant difference in participants' reports that they would answer regardless of AdviceAI's advice ($F(2, 660) = 0.06$, $p = .938$; $\eta^2 < .001$) (Fig. 5).

7.5. Discussion

Having found no evidence that advisor traits affected moral persuasion in Studies 1 and 2, Study 3 found that the content of the advice does influence persuasion. While patently bad moral advice was not persuasive, advice containing a recommendation backed up with (prima facie) good reasons was just as persuasive as advice containing a recommendation followed by no novel reasons. In other words, the LLM moral persuasion observed in these studies was not caused by participants changing their minds because they appreciated by their own moral lights that the reasons contained in the advice were good reasons, nor was it caused by participants changing their minds because the LLM advisor seemed reliable, trustworthy, authoritative, etc., Instead, participants were evaluating the advice on a case-by-case basis and then deferring if the advice was good enough. Participants appear to be persuaded towards an option based on the fact *that the LLM recommended it*, but only if an evaluation of its justification does not contain any red flags (or, in the language of epistemologists, defeaters).

While this manuscript was under revision, Krügel et al. (2026) independently published a related study in which they presented participants with a single sacrificial dilemma, manipulated whether the source of advice was human or AI, and presented a purported screenshot of ChatGPT 3.5 either giving a simple recommendation or defending the recommendation with a short justification. They similarly found that the inclusion of justification did not affect the persuasiveness of advice. Despite the differences in design (e.g., our Study 3 had participants interact with the AI, controlled for confounds of text length, and included a bad reasons condition), their results support our finding that when users are persuaded in the direction of LLM moral advice, it is due to the recommendation itself, not the reasons justifying the recommendation.

8. General discussion

As technologies powered by artificial intelligence have become increasingly widespread, there has been a growing focus on how commonly used LLMs can change people's beliefs (Bai et al., 2023; Costello et al., 2024; Durmus et al., 2024; Hackenburg & Ibrahim, 2023; Hölbling et al., 2025). We set out to answer two questions. First, is LLM advice persuasive about everyday moral dilemmas, despite people's stated aversion to moral advice from AI (Bigman & Gray, 2018; Mahmud et al., 2022; Myers & Everett, 2025)? Second, if it is persuasive, is it persuasive because people appreciate the reasons provided as good reasons or because people are deferring to the LLM (Landes et al., 2025)? That is, are they persuaded by the morally relevant features contained in

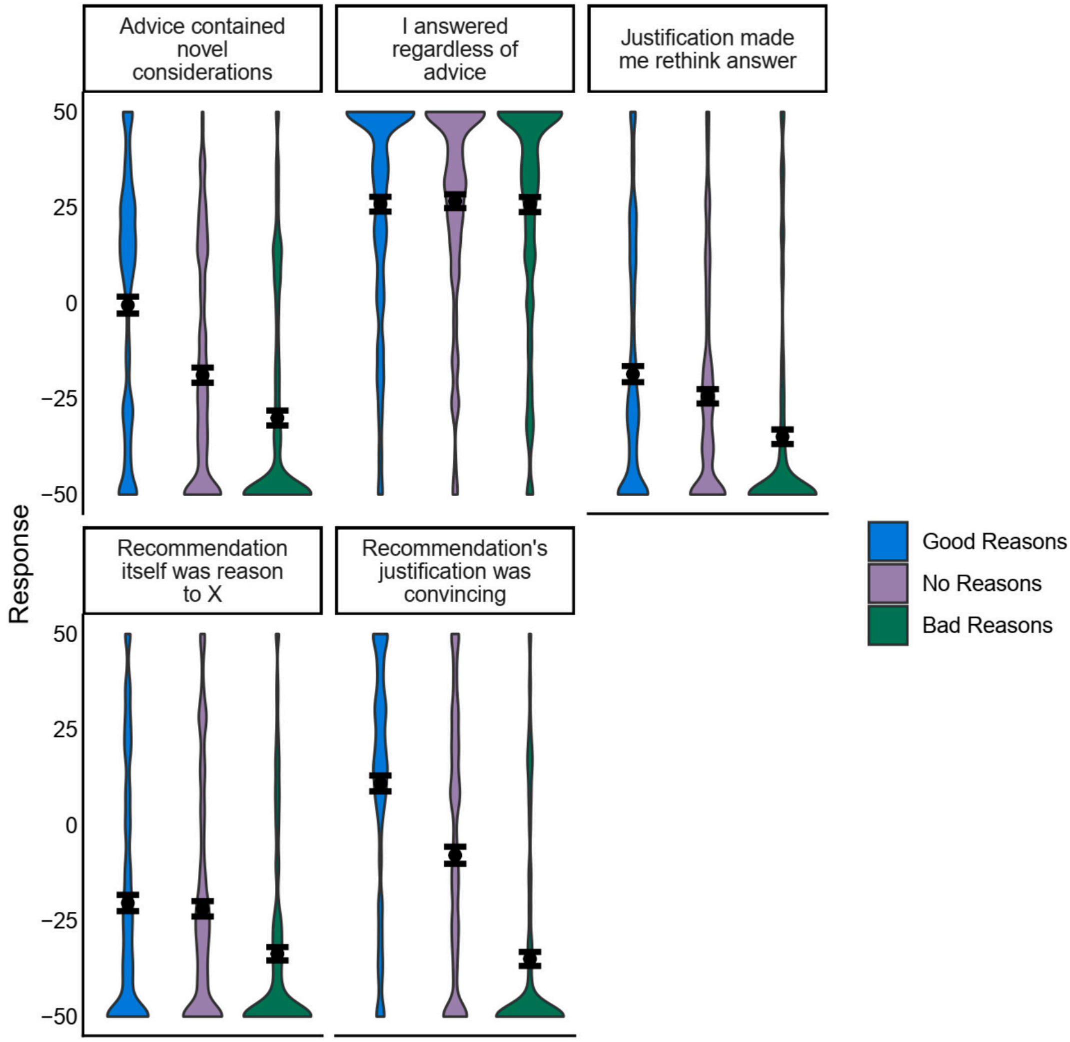

Fig. 5. Participant self-reports about epistemology of Study 3's moral advice. *Note. Error bars represent SEM.*

the advice or by peripheral, non-morally relevant features of the LLM or its output? One possibility is far more worrisome than the other. Engaging with the reasons in moral advice promotes moral growth (Landes et al., 2025) and understanding (Hills, 2020). In contrast, moral deference, to name just a few problems identified by philosophers and psychologists, risks slowing moral growth (Liu et al., 2022), encourages moral dependence (Lara & Deckers, 2020; Volkman & Gabriels, 2023), and is inauthentic (Brick, 2024; Howell, 2014).

In Studies 1 and 2, we focused on what perceptions of the advisor could tell us about moral persuasion. In Study 1, we looked at how participants responded to everyday moral dilemmas after receiving advice labelled as coming from a human or an LLM in a range of common and realistic dilemmas that reflect the sort of scenarios people might realistically ask for advice about. Despite previous research showing people are averse to moral advice in general (Andow, 2020; Brick, 2024) and moral decision making by AI (Bigman & Gray, 2018; Mahmud et al., 2022; Myers & Everett, 2025), we found that participants were significantly more likely to respond to dilemmas in the direction advised by the advisor regardless of whether the advisor was a human or LLM, suggesting that in the moral domain too, both human and LLM advisors are persuasive. Strikingly, participants' ratings of the advisor and the persuasiveness of the advisor came apart. Even though the human advisor was rated as more trustworthy and reliable than the LLM advisor, they were equally persuasive, revealing that moral persuasion – or trust - can be independent of participants' perceptions of an advisor's moral and epistemic standing (see also Claessens et al., n.d.; Krügel et al., 2026).

Study 2 corroborated these findings in a more realistic paradigm, where participants believed the advice to be a product of an LLM after multiple rounds of interactions with said LLM. Prior to the advice, we manipulated the quality of the LLM's advice, finding no evidence that past signals of LLM quality affected the persuasiveness of its advice in everyday moral dilemmas. These results do not in themself answer whether people seeing LLM moral advice are (irresponsibly) deferring or (responsibly) being persuaded by the reasons that the AI gives. The advice may be driving rates of persuasion because it contains reasons people find convincing by their own lights or because people are deferring to the LLM when the advice seems good enough.

In Study 3 we explored whether people are (irresponsibly) deferring or (responsibly) being persuaded by the reasons by manipulating whether and how the LLM defended its recommendation. We found no evidence of a difference between the persuasiveness of advice containing good reasons compared to advice containing a recommendation without additional reasons (i.e., merely neutrally rehashing the details of the dilemma). Nonetheless, we found participants were responsive to *some* signals contained in the advice. The advice that contained a recommendation justified with patently bad reasons instead backfired, pushing responses in the opposite direction. While this apparent 'backfire effect' was not pre-registered as a prediction, it is consistent with other research that negative information is weighted more generally than positive (Baumeister et al., 2001). While high-quality reasons do not increase persuasion relative to no reasons at all, bad reasons may actively undermine trust by providing evidence that the AI is unreliable in that case. Although the backfire effect was not pre-registered and should therefore be interpreted cautiously, it does suggest a potentially asymmetric sensitivity to justification quality that should be explored in future work.

Our results therefore suggest participants are inclined to defer to moral advisors, and this deference is on an advice-by-advice basis. They are persuaded because the LLM endorses a side, but this deference can be undermined if the output shows signs of incoherence or obvious incorrectness. We theorize that participants are employing a heuristic of "seems good enough" when it comes to LLM moral persuasion, consistent with the idea of 'cognitive misers' in social cognition, in which people seek to conserve cognitive resources and engage in more thorough processing only when necessary (Fiske & Taylor, 1991). This epistemic stance towards LLM advice is likely familiar to readers who have used LLMs for something that *just needs to work*, whether generating computer code, generating tips to troubleshoot a problem, or generating a difficult but low-stakes email. It is easy to examine these LLM outputs with just enough care to check that it is not obviously going to backfire, but not with enough care to develop an informed opinion about whether it is, in fact, going to work. We instead just check that the output seems good enough to try.

Whether this is acceptable when generating computer code and perfunctory emails, a heuristic of "this advice seems good enough" or an attitude of "this just needs to work" is not the way we should approach moral advice. Morality is difficult. Morality requires tough calls and lost sleep. Our own moral character is something we can and should work on improving (Annas et al., 2016; Aristotle, 1999; Landes & Everett, 2026). Deferring to AI in this way risks robbing us of developing our own moral understanding (Hills, 2020), being inauthentic to our own moral character (Brick, 2024; Howell, 2014), and fostering reliance on LLMs more generally (Lara & Deckers, 2020; Vallor, 2015; Volkman & Gabriels, 2023). The fact that this deference is not blind is better than nothing, but this minimal level of attention to advice is unlikely to lead to any meaningful moral development or moral understanding. Stakeholders should examine and critically question existing approaches towards developing and implementing LLMs, as well as current norms about using LLMs. Now is the time to ensure that people do not become epistemically and morally dependent on LLMs.

Setting aside the epistemic question of deference and turning to the psychology of trust, our findings additionally highlight a potentially surprising paradox when it comes to perceptions of moral advisors – trustworthiness - and attitude change - trust. Studies 1 and 2 found that participants rated AI advisors as less trustworthy, but they were still persuaded by the LLM. This disconnect may be surprising given that perceptions of trustworthiness are typically seen as antecedents of trust (Lee & See, 2004; Mayer et al., 1995), even if we know perceptions of AI trustworthiness do not always lead to trusting behavior (Everett et al., 2026). This is not the only disconnect observed between self-reports and persuasion. In Study 3, despite no difference in persuasion between advice containing good reasons and no reasons, participants rated the good advice condition as having more convincing justification than the no reasons condition, and participants reported rethinking their answers more in the good advice condition than in the no reasons condition. Our results therefore suggest that reflective or higher-order beliefs about advice or advisor quality have little to no effect on the persuasiveness of moral advice, as long as the specific recommendation reaches a minimal level of competence. People may "trust" LLMs behaviorally by updating their moral judgments while still perceiving them as less trustworthy than humans (Claessens et al., n.d.). This suggests that scholars concerned about the ethical and epistemic risks of reliance on AI should not necessarily be comforted by the fact that people self-report reduced trust in moral machines in Study 1; even perceiving AI as being less reliable as a source does not always preclude deferring to its advice (Studies 1–2), except when given strong evidence to discount (Study 3).

Our findings raise challenges to some visions of artificial moral advisors (AMAs), suggesting that people do not have the epistemic relationship towards LLMs that some proponents of artificial moral advisors may wish. Defenders of artificial moral advisors often envision AI-based moral advice taking a central role in moral human decision-making, where AI advice is a large or even overriding source of moral advice (Dietrich, 2001; Gips, 1995; Giubilini & Savulescu, 2018). This paper has focused primarily on the criticism that such dependence, which would likely require deference, is epistemically and morally problematic (Landes et al., 2025; Lara, 2021; Volkman & Gabriels, 2023), but our results raise a more practical challenge to the integration of artificial moral advisors into moral decision-making. In Study 3, participants displayed aversion to the idea of LLM moral persuasion, largely self-reporting that their answer was unaffected by the LLM and rejecting the idea that the LLM recommendation itself (as opposed to reasons provided by the LLM) was reason to change their responses. These responses are in line with the work on algorithmic aversion discussed above, suggesting people are relatively unlikely to approach LLMs for moral advice (see Chatterji et al., 2025). Moreover, the persuasion we observed in Studies 2 and 3 was far from overriding, although moral persuasion may increase in more conversational settings (Hackenburg et al., 2025). Therefore, even if AI moral advisors can persuade people, they may struggle to attract users and have only slight effects on behavior, all while doing so via the ethically and epistemically worrying process of deference.

8.1. Limitations and future directions

An explicit design choice and limitation is our focus on everyday moral conflicts. While this was a deliberate choice to represent the kinds of situations participants might be most inclined to seek advice on based on their daily lives, moral dilemmas come in a wide variety of forms. It will be interesting for future work to consider how people might respond to moral advice from LLMs in other kinds of moral dilemmas that exemplify the tension between different ethical principles and theories but are not sacrificial dilemmas. For example, we are forced between deontological and consequentialist options in countless situations where we have to decide whether the means justify the ends, such as whether we should lie to our friend to get them out of a toxic relationship. Similarly, we are often forced to choose between the virtuous option and the consequentialist option when we have to decide whether it is more important to do something the right way or produce a good output, such as whether we can have an LLM generate a love letter (Claessens et al., 2025). Moreover, some scholars have suggested that AI moral advisors, if used at all, might be particularly suited to domain-specific contexts with clearly agreed ethical guidelines, such as in medical decision making (Liu et al., 2022). It is plausible that in moral dilemmas such as those posed in bioethics, with high complexity in the details of condition and uncertainty about the outcomes, participants may be particularly likely to defer to the moral advice of an LLM. Future work should therefore explore LLM persuasion on this wider range of moral dilemmas of varying complexity.

Our studies were conducted on a UK sample, and it will be interesting to consider potential differences in the epistemic stances taken towards LLMs in times and places with different levels of LLM use. For example, in some countries that may have more positive overall views about AI and its place in society, we might expect stronger persuasion (see Claessens et al., n.d.). Similarly, there may be cross-cultural differences in the norms of LLM use for different kinds of tasks, leading to different epistemic stances towards the LLM. In a similar vein, it is unclear how long the observed persuasion would affect moral attitudes. While other research has found long-lasting effects of non-moral AI persuasion (Costello et al., 2024), the effects we observed here may be fairly short-lived, especially since the observed persuasion appears to be in response to peripheral cues, rather than at-issue cues (Petty & Cacioppo, 1986).

Future work may therefore focus on moral beliefs beyond pre-written dilemmas to understand whether LLMs are capable of sustained moral persuasion, especially when participants are given longer and more conversational interactions with an LLM over time.

9. Conclusions

In conclusion, we studied whether, and more importantly, *how*, people were persuaded by LLM moral advice. We built on previous research looking at persuasion in non-moral domains where there are often objectively correct answers (Costello et al., 2024; Karinshak et al., 2023) and research examining moral persuasion but with static paradigms focusing on sacrificial moral dilemmas (Claessens et al., n.d.; Krügel et al., 2023). We found that LLM-generated moral advice about non-sacrificial everyday moral dilemmas was equally persuasive when attributed to an LLM or a human expert, and that it was persuasive because people were deferring to the LLM's advice when the advice was of sufficiently high quality. These findings highlight the importance of research that goes beyond testing the extent to which AI is persuasive and additionally asks what cognitive and epistemic resources users deploy while engaging with LLMs.

CRediT authorship contribution statement

Ethan Landes: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kathryn B. Francis:** Writing – review & editing, Formal analysis, Conceptualization. **Jim A.C. Everett:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Conceptualization.

Funding

This work was generously supported by funding from the Horizon Europe UK Guarantee via the UKRI (EP/Y00440X/1) and a Philip Leverhulme Prize (PLP-2021-095) awarded to JACE. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Acknowledgments

Kathryn B. Francis is supported in part by the Wellcome Trust [226801].

We thank Scott Claessens, Pierce Veitch, Radu Uszkai, and Cristina Voinea for their helpful feedback on this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2026.106504.

Data availability

We have included a link to data in the manuscript (OSF).

References

Andow, J. (2020). Why don’t we trust moral testimony? *Mind & Language, 35*(4), 456–474. https://doi.org/10.1111/mila.12255

Annas, J., Narváez, D., & Snow, N. E. (2016). *Developing the virtues: Integrating perspectives*. Oxford University Press.

Aristotle. (1999). *Nicomachean ethics*. T. Irwin, Trans. (2nd ed.). Hackett Pub. Co

Bai, (M.)H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. (2023). *Artificial intelligence can persuade humans on political issues*. OSF. https://doi.org/10.31219/osf.io/stakv

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

Brick, S. (2024). Deference to moral testimony and (in) authenticity. In *Oxford studies in experimental philosophy: Volume 5* (p. 54).

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use chatgpt*. National Bureau of Economic Research.

Claessens, S., Bocian, K., Boggio, P.S., Fiorio, G., Fitouchi, L., Genç, Z., … Everett, J.A.C. (n.d.). Trust in artificial moral advisors across cultures.

Claessens, S., Veitch, P., & Everett, J. A. C. (2025). Negative perceptions of outsourcing to artificial intelligence. *Computers in Human Behavior, 177*, 108894.

Constantin, J., & Grundmann, T. (2020). Epistemic authority: Preemption through source sensitive defeat. *Synthese*. https://doi.org/10.1007/s11229-018-01923-x

Constantinescu, M., Vică, C., Uszkai, R., & Voinea, C. (2022). Blame it on the AI? On the moral responsibility of artificial moral advisors. *Philosophy and Technology, 35*(2), 35. https://doi.org/10.1007/s13347-022-00529-z

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Crisp, R. (2014). II—Roger Crisp: Moral testimony pessimism: A defence. *Aristotelian Society Supplementary, 88*(1), 129–143. https://doi.org/10.1111/j.1467-8349.2014.00236.x

Dietrich, E. (2001). Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence, 13*(4), 323–328.

Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports, 15*(1), 4084. https://doi.org/10.1038/s41598-025-86510-0

Duran, L. D. D. (2021). Deskilling of medical professionals: An unintended consequence of AI implementation? *Giornale Di Filosofia, 2*(2).

Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., & Ganguli, D. (2024). Measuring the persuasiveness of language models. https://www.anthropic.com/news/measuring-model-persuasiveness.

Everett, J. A., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences, 24*(2), 124–134.

Everett, J. A. C., Claessens, S., Knochel, T., & Reinecke, M. G. (2026). *Principles for understanding Trust in Artificial Intelligence*. Accepted in Principle at Nature Reviews Psychology.

Fileva, I. (2023). Moral testimony and collective moral governance. *Australasian Journal of Philosophy, 101*(3), 722–735. https://doi.org/10.1080/00048402.2021.2024584

Gips, J. (1995). *Towards the ethical robot. Android epistemology*. Cambridge MA: MIT Press.

Giubilini, A., Porsdam Mann, S., Voinea, C., Earp, B., & Savulescu, J. (2024). Know thyself, improve thyself: Personalized LLMs for self-knowledge and moral enhancement. *Science and Engineering Ethics, 30*(6), 54. https://doi.org/10.1007/s11948-024-00518-9

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

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Hackenburg, K., & Ibrahim, L. (2023). Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. https://osf.io/82nsd/.

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences, 121*(24), Article e2403116121. https://doi.org/10.1073/pnas.2403116121

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., … Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science, 390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hackenburg, K., Tappin, B. M., Röttger, P., Hale, S., Bright, J., & Margetts, H. (2024). Evidence of a log scaling law for political persuasion with large language models (No. arXiv:2406.14508). *arXiv*. https://doi.org/10.48550/arXiv.2406.14508

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology, 69*, 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hills, A. (2013). Moral testimony. *Philosophy Compass, 8*(6), 552–559. https://doi.org/10.1111/phc3.12040

Hills, A. (2020). Moral testimony: Transmission versus propagation. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12595

Hölbling, L., Maier, S., & Feuerriegel, S. (2025). A meta-analysis of the persuasive power of large language models (No. arXiv:2512.01431). *arXiv*. https://doi.org/10.48550/arXiv.2512.01431

Howell, R. J. (2014). Google morals, virtue, and the asymmetry of deference. *Noûs, 48*(3), 389–415. https://doi.org/10.1111/j.1468-0068.2012.00873.x

Hume, D. (1748). *Of miracles*.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131–164. https://doi.org/10.1037/rev0000093

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model’s ability to generate pro-vaccination

messages. In *, Vol. 7(CSCW1)*. *Proc. ACM hum.-comput. interact*. https://doi.org/10.1145/3579592, 116:1–116:29.
Keren, A. (2007). Epistemic authority, testimony and the transmission of knowledge? *Episteme, 4*(3), 368–381. https://doi.org/10.3366/e1742360007000147
Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-31341-0. Article 1.
Krügel, S., Ostermaier, A., & Uhl, M. (2026). Justification optional: ChatGPT's advice can still influence human judgments about moral dilemmas. *AI and Ethics, 6*(1), 120. https://doi.org/10.1007/s43681-026-01005-6
Lackey, J. (2008). *Learning from words*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199219162.001.0001
Lackey, J. (2018). Experts and peer disagreement. In *Knowledge, belief, and god: New insights in religious epistemology* (p. 228).
Landes, E. (2021). *Philosophy and philosophy: The subject matter and the discipline*. Thesis. The University of St Andrews. https://doi.org/10.17630/sta/1072.
Landes, E. (2023). Philosophical producers, philosophical consumers, and the metaphilosophical value of original texts. *Philosophical Studies, 180*(1), 207–225. https://doi.org/10.1007/s11098-022-01900-8
Landes, E., & Everett, J. (2025). *AI should develop human empathy, not replace it*. OSF. https://doi.org/10.31234/osf.io/y3qzu_v1
Landes, E., & Everett, J. A. C. (2026). AI should develop human empathy, not replace it. In A. Perry, & C. D. Cameron (Eds.), *Empathy and artificial intelligence: Challenges, advances, and ethical considerations*. Cambridge University Press.
Landes, E., Voinea, C., & Uszkai, R. (2025). Rage against the authority machines: How to design artificial moral advisors for moral enhancement. *AI & Society, 40*(4), 2237–2248. https://doi.org/10.1007/s00146-024-02135-3
Lara, F. (2021). Why a virtual assistant for moral enhancement when we could have a Socrates? *Science and Engineering Ethics, 27*(4), 42. https://doi.org/10.1007/s11948-021-00318-5
Lara, F., & Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics, 13*(3), 275–287. https://doi.org/10.1007/s12152-019-09401-y
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
Leonard, N. (2023). Epistemological problems of testimony. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy (Spring 2023)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entriesestimony-episprob/.
Liu, Y., Moore, A., Webb, J., & Vallor, S. (2022). Artificial moral advisors: A new perspective from moral psychology. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 436–445). https://doi.org/10.1145/3514094.3534139
Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change, 175*, Article 121390. https://doi.org/10.1016/j.techfore.2021.121390
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.
Myers, S., & Everett, J. A. C. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition, 256*, Article 106028. https://doi.org/10.1016/j.cognition.2024.106028
Ovsyannikova, D., de Mello, V. O., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology, 3*(1), 1–11. https://doi.org/10.1038/s44271-024-00182-6
Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.)*, Vol. 19*. *Advances in experimental social psychology* (pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2.
Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x
Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden persuaders: LLMs' political leaning and their influence on voters (No. arXiv:2410.24190). *arXiv*. https://doi.org/10.48550/arXiv.2410.24190
Rong, M. (2025). Posthuman sociality and loneliness in Kazuo Ishiguro's Klara and the Sun. *AI & Society*. https://doi.org/10.1007/s00146-025-02236-7
Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond artificial intelligence: The disappearing human-machine divide* (pp. 79–95). Springer International Publishing. https://doi.org/10.1007/978-3-319-09668-1_6.
Singer, N., Kreuzpointner, L., Sommer, M., Wüst, S., & Kudielka, B. M. (2019). Decision-making in everyday moral conflict situations: Development and validation of a new measure. *PLoS One, 14*(4), Article e0214747. https://doi.org/10.1371/journal.pone.0214747
Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy and Technology, 28*(1), 107–124.
Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press. https://doi.org/10.1093/oso/9780197759066.001.0001
Volkman, R., & Gabriels, K. (2023). AI moral enhancement: Upgrading the socio-technical system of moral engagement. *Science and Engineering Ethics, 29*(2), 11. https://doi.org/10.1007/s11948-023-00428-2
Wogu, I. A. P., Olu-Owolabi, F. E., Assibong, P. A., Agoha, B. C., Sholarin, M., Elegbeleye, A., … Apeh, H. A. (2017). Artificial intelligence, alienation and ontological problems of other minds: A critical investigation into the future of man and machines. *International Conference on Computing Networking and Informatics (ICCNI), 2017*, 1–10. https://doi.org/10.1109/ICCNI.2017.8123792
Worsnip, A., Lane, D., Pratt, S., Napolitano, M. G., Gray, K., & Greene, J. A. (2025). Authority or autonomy? Philosophical and psychological perspectives on deference to experts. *Philosophical Psychology*, 1–36. https://doi.org/10.1080/09515089.2025.2475138
Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 84*(2), 260–281.
Yudkin, D. A., Goodwin, G. P., Reece, A., Gray, K., & Bhatia, S. (2025). A large-scale investigation of everyday moral dilemmas. *PNAS Nexus, 4*(5), Article pgaf119. https://doi.org/10.1093/pnasnexus/pgaf119