Universiteit Leiden

Trust Me, I'm a Chatbot: The Effect of Parasocial Cues on the Conformance to AI Chatbot Advice

Mateusz Oleksa (s4075609)

Supervisors:
Peter van der Putten & Tom Kouwenhoven

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 9, 2026

Abstract

This study examines whether parasocial cues in generative AI responses affect conformance to the chatbot's advice when solving judgment-based tasks. Just as people react differently to different tones in human interaction, a text-based agent can deliver its advice in varying tones, prompting users to respond differently. This can be problematic in certain cases - for example, when bad advice is framed in a tone that makes a user more likely to pick it up. So far, most of the literature has focused on the effect of anthropomorphism on conformance. In contrast, this study focuses solely on text-based agents, where no human-like appearance or voice is present. To assess the influence of these cues on conformance, a between-subjects online experiment was conducted in which participants were asked to make a series of decisions within a scenario that, by design, is morally ambiguous. Then they received AI-generated advice, and the option to discuss it further with a chatbot prompted to use a specific set of parasocial cues. Participants (N=108) were randomly assigned to one of three chatbot conditions: a warmth condition, in which the chatbot expressed care and empathetic concern; a common locus condition, in which it framed the problem as something it and the user were working through together; and a neutral control. Across four morally weighted dilemmas, conformance was measured in the form of Weight of Advice. The manipulation did not produce the predicted effect of higher conformance among participants interacting with warmth- and common-locus chatbots: the ANOVA on WOA means was non-significant, and manipulation checks showed no meaningful differences in perceptions of the conditions. Exploratory analysis, however, shows that participants with low initial confidence exhibited significantly greater conformance in the warmth-based condition than in the control condition. The contribution of this work is mostly methodological and hypothesis-generating. The null main effect, combined with the conditional warmth effect among low-confidence users, suggests that parasocial cues do not uniformly increase conformance. They, however, may matter when users are least certain about their own judgement.

Contents

1 Introduction

Within the last few years, conversational artificial intelligence (AI) agents have moved from a fringe research topic to a widely used tool for seeking advice. With the growing pool of users (with almost 18% of the world's working population using AI as of the first quarter of 2026 [24]) and rising AI incident rates [31], the topic of safe AI use is in the spotlight. People now consult large language model (LLM) chatbots for guidance ranging from trivial questions to more consequential queries: what to cook, how to phrase a difficult email, whether a symptom warrants a doctor's visit, and how to resolve a moral or interpersonal dilemma. Unlike a traditional search engine that provides the user with a list of possible sources, an LLM gives a ready position or recommendation, phrased in fluent natural language and usually delivered in the first or second person.

The motivation and concern behind this thesis is overreliance - the acceptance of AI advice by users who are either unable or unwilling to judge how far such advice should be trusted [28]. People have been shown to follow AI recommendations even when those recommendations conflict with their beliefs [19], at the cost of independent reasoning [9]. In the existing literature, the reasons for this uncritical acceptance are typically located either in the content of the advice or in the user: the system's perceived confidence [35], the user's low AI literacy [28], or a preference for efficiency over effort [6].

This thesis takes a different approach, focusing on the relational packaging of advice. A chatbot not only provides advice but also delivers it in a particular voice. It can address the user by name, express warmth and concern, or frame a problem as something the two of them are working through together. None of those makes the advice itself any better or more correct, yet they can increase the rate at which a user accepts the advice. If such conformance to the chatbot's advice can be driven by how the advice is delivered, rather than by whether it is sound, then the same design choices that make a chatbot feel pleasant and engaging also risk manufacturing precisely the uncritical acceptance that defines overreliance.

How can a system with no face, no voice, and no body create a sense of closeness with its user? The answer, we argue, lies in cues. We frame those ways in which advice can be delivered as parasocial cues, introduced by Horton & Wohl [17] in 1956 to describe one-sided intimacy developed by the audience and the media personae, an illusion deliberately cultivated through, among others, direct address, conversational warmth, and an appearance of reciprocity.

Although a chatbot, unlike a TV host, creates a bidirectional conversation with the user, we retain the "para-" prefix as the felt closeness is still directed at a generated, and not intended, conversation partner. The mechanism by which such cues take hold is captured by a second mechanism. Computers Are Social Actors (CASA) [26] paradigm establishes that minimal social signals, such as language, turn-taking, or the occupation of a role normally held by a human, are sufficient to trigger the social responses people show towards other people, without believing that the machine itself is a human. The parasocial cues don't need to convince a user that a chatbot "cares", but on the CASA account, they only need to activate the relevant social script.

For the purposes of this study, we focus on two families of cues that differ in how they manufacture closeness. The first, warmth-based cues create closeness through expressed care and empathetic framing (for example, "That sounds like a really difficult position to be in"), and common-locus cues create closeness through perceived shared situation - the chatbot positions itself as facing the same problem as the user, reasoning with them, rather than from the outside (for example, "Let's think through how we should weigh this"). Warmth operates through felt care

directed at the user, whereas common locus operates through identification and shared stance. Separating those two possible paths to conformance lets us test whether their effect differ.

Previous work on human-agent interactions has been concentrated around anthropomorphism - how human-like the bot is perceived to be as the driver towards trust and acceptance. A meta-analysis by Blut et al. [3] of over 100 studies on robots, chatbots, and virtual agents finds that anthropomorphism reliably increases users' trust and intention to use. Such work, however, leans heavily on embodied or visual cues (such as faces, voices, gendered representation) but does not isolate the purely linguistic cues available to a text-only agent. It also does not distinguish between warmth and identification as separate relational mechanisms, even though they create closeness in different ways. This thesis aims to address that gap.

We create a between-subjects online experiment in which participants make a series of judgments on moral dilemmas. After making their initial judgement, they receive advice from a chatbot whose communication style is manipulated across three conditions (a neutral control, a warmth condition, and a common locus condition), while the substance of the advice is held constant, and then are given the option to revise their judgments. Conformance is operationalised, in the advice-taking tradition [34], as the Weight of Advice: the degree to which a participant's revised judgement moves towards the chatbot's recommendation. The central research question is therefore: "**What is the impact of different parasocial cues in chatbot interaction on the conformance to AI advice?**" with the following sub-question: "Does warmth-based parasociality differ from identification-based parasociality in its effect on conformance to AI-generated advice?"

This thesis makes the following contributions:

- It isolates the effect of purely linguistic parasocial cues on conformance in a text-only agent, holding the substance of advice constant while removing the embodied and visual confounds that dominate prior anthropomorphic research.
- It distinguishes warmth-based from common-locus (identification-based) cues as separate relational mechanisms and tests whether they differ in their effect on conformance.

The remainder of this thesis is organised as follows: Section 2 develops the theoretical background: parasocial interaction, the CASA paradigm, common locus, egocentric discounting, conformance, and overreliance. Section 3 situates the study among related empirical work. Section 5 sets out the principles behind the design, and Section 4 describes the experiment itself, including the pilot phase that calibrated and refined it. Section 6 reports the results, and Section 7 interprets them, considers the implications for overreliance, and sets out the study's limitations and the directions for future work that follow from them. Lastly, Section 8 briefly summarises and concludes the paper.

2 Background

In this section, we will describe the most important concepts relating to the topic of this paper. Firstly, we introduce the topic of parasocialty and parasocial cues - the backbone of this study, as they serve as the main mediator of this study by potentially influencing users' trust. Next, we delve into the CASA paradigm, which we use to argue that participants might react socially towards the chatbots and trust them. Lastly, we discuss egocentric discounting as a way to explain the

human decision-making process in judge-advisor tasks, and conformance as the alignment of one's beliefs and actions to those of a third party. Finally, we discuss AI overreliance - the concept that provided the main motivation behind this study.

2.1 Para-social relationships & interactions

To begin the study about parasocial cues, we first need to explain the concept of parasociality. The idea of parasocial relationships was first introduced by Horton and Wohl in 1956 [17] to describe the illusory sense of a face-to-face relationship that audiences develop with media performers. Horton and Wohl introduced the role of a persona - actors existing for their audiences solely in a parasocial relationship. In their original formulation, a parasocial relation is inherently one-sided: the performer creates an impression of intimacy through conversational style, direct address, and unofficial tone mimicking small talk, while the audience responds as though engaged in a reciprocal social exchange, despite having no ability to influence the interaction or develop it by themselves. Horton and Wohl emphasised that this dynamic mirrors primary-group sociability, yet is fundamentally constrained by its lack of effective reciprocity.

Subsequent research has expanded the concept considerably. Giles [12] reviewed several decades of Parasocial Interaction (PSI) scholarship and argued that parasocial phenomena exist on a continuum with ordinary social interaction, rather than constituting a categorically distinct experience. His model distinguished between parasocial interaction (PSI), which involves momentary cognitive and affective responding during a media encounter, and the longer-term parasocial relationship (PSR), which may develop through repeated exposure. Giles also identified that the nature of the media figure, its authenticity, degree of direct address, and potential for real contact shape the kind of parasocial bond that forms. This can range from first-order PSI with figures who address the viewer directly, through second-order PSI with real people portraying fictional characters, to third-order PSI with purely fictional entities, in which no real contact is possible, such as animated cartoon characters.

Unlike traditional broadcast media, conversational AI agents engage users in bidirectional, responsive exchanges, blurring the boundary between illusory and actual reciprocity. Liu [20] conducted a scoping review of PSI research in human-AI contexts and found that many studies combine parasocial and genuinely social interaction, often adopting relational PSR measures while labelling the construct as PSI. Liu argues that the interactivity of AI agents shifts the experience away from the one-sided illusion that defines classical PSI. This distinction is significant for the present work as it locates our study on one side of it - a single online session cannot be enough for a parasocial relationship to form. We therefore do not claim that the participants form a parasocial bond with the chatbots. We only manipulate the parasocial cues - the momentary PSI-like signals of warmth and shared situation that a relationship would normally be built from, and check whether those cues alone change the behaviour of participants.

2.1.1 Parasocial cues

In the present study, rather than claiming that users form parasocial relationships with a chatbot during a single brief interaction, we focus on parasocial cues. These cues have the design features of the chatbot's communication style intended to evoke the felt closeness and social presence characteristic of parasocial experiences and to engage the subject in a parasocial interaction.

Horton and Wohl [17] described such cues in the context of television programmes, in which performers and managerial teams deliberately employed them to create an illusion of intimacy with the audience. In that setting, parasocial cues took the form of, among others:

- Maintaining a flow of small talk
- Making eye contact with the camera
- Addressing the viewer directly
- Blending with the live studio audience
- Using first-person camera perspectives

Those cues, presented by Horton and Wohl, are tightly bound to audiovisual media and do not transfer directly to the text-based format of a chatbot. However, their underlying principle that the communication style can be deliberately designed to create a sense of intimacy and social connection can be generalised into the domain of text-based content. In a chatbot context, the relevant modality shifts from visual and performative cues to purely linguistic ones. The chatbot cannot make eye contact, but it can address the user by name. It cannot mingle with a studio audience, but it can use inclusive language that frames the interaction as a shared experience.

Liu [20] argues that applying the concept of parasocial interaction to AI agents is conceptually problematic because conversational AI systems engage users in genuinely bidirectional exchanges rather than the one-sided broadcast scenarios in which PSI was originally defined. If the interaction is reciprocal, one might argue that the cues should be called social. We retain the prefix para- for two reasons.

- Despite the surface-level reciprocity, the chatbot's responses are generated rather than intentional - there is no genuine social actor on the other side of the exchange, and thus any felt intimacy remains, in an important sense, illusory.
- The cues we manipulate are not designed to produce actual social bonding, but to simulate the signals that would produce bonding in a human-human context, which is precisely the function Horton and Wohl [17] attributed to parasocial cues in television

This study operationalises two families of parasocial cues in the chatbot's language. The first is warmth-based cues: empathic phrasing, expressions of care, and personalised acknowledgement of the user's situation. These mirror the conversational warmth that Horton and Wohl identified as central to creating an illusion of intimacy, and that Giles [12] linked to the companionship function of parasocial experiences. The second is identification-based cues grounded in a common locus [25]: language that gives the user a sense of common ground and shared experiences, "being in the same boat", further elaborated in subsection 2.3. Where warmth creates closeness through affect, common locus creates closeness through perceived shared experience.

2.2 CASA paradigm

The CASA paradigm was introduced by Nass, Steuer, and Tauber [26] to understand the fundamentally social nature of human-computer interactions. The central idea is that social responses to computers are not the result of conscious beliefs that computers are human or human-like, nor do they stem from ignorance, psychological dysfunction, or a belief that users are interacting with a programmer behind the screen. Rather, these social responses are commonplace, easy to generate, and often unknown to computer users themselves.

The experimental approach of CASA follows a straightforward methodology:

1. Pick a social study
2. Replace at least one human with a computer in the statement of that study
3. Replace at least one human with a computer in the method of that study
4. Provide a computer with characteristics associated with a human
5. Determine if the statement still holds

Nass et al. [26] re-conducted five studies using this approach. The first demonstrated that users apply politeness norms to computers, evaluating a computer more favourably when it asked them to assess its own performance than when a different computer posed the same question, resembling the well-documented tendency to be politer in direct evaluations of people. The second and third studies showed that users apply concepts of "self" and "other" to computers, treating two machines with different identities as separate social actors rather than interchangeable terminals. The fourth study found that gender stereotypes follow computers: participants attributed different competencies to machines depending on whether the voice was male or female. The fifth study confirmed that users attribute socialness directly to the computer itself, rather than treating it as a medium for a human programmer. Across all five experiments, the participants were experienced computer users, ruling out novelty of computer use as an explanation.

A crucial implication of CASA for our study is that even minimal cues, such as a computer producing language output, taking conversational turns, or filling a role traditionally occupied by a human, are sufficient to trigger social responses that people usually experience while interacting with other humans. This means that the parasocial cues manipulated in our experiment (warmth-based and common-locus language) need not convince participants that the chatbot is human. They only need to activate the relevant social scripts. CASA predicts that once such scripts are activated, users will respond with socially conditioned behaviours, potentially including increased conformance to the chatbot's advice.

CASA is not the only account of why such responses arise, and the complementary perspectives are worth noting.

Dennett's theoretical notion of the intentional stance [8] holds that we predict the behaviour of a sufficiently complex system most efficiently by treating it as a rational agent with beliefs and goals, regardless of what it "really" is. A fluent, goal-directed chatbot is a natural target for this stance. Van Dijk et al. [32] extend this reasoning to LLMs directly, arguing that attributing understanding and intention to such systems is best understood as not a factual error about what the model really is, but as a pragmatic move- treating a fluent conversational system as an agent with beliefs and

intentions lets the user abstract away from its underlying mechanics and predict its behaviours effectively. On this basis, the parasocial cues we manipulate need not deceive participants about the chatbot's nature; they need only make the intentional stance easy to adopt.

Moreover, Heider and Simmel [15] showed that people spontaneously attribute intention, motive, and even social narrative to simple moving shapes, suggesting that the impulse to read agency into non-human stimuli is deep-seated and easily triggered. Lastly, social responses need not depend on genuine belief: a user who consciously knows the system is a machine may still play along, suspending disbelief much as an audience does at the theatre.

The original CASA paradigm was developed in the context of 1990s desktop computers with preprogrammed responses, audio tracks and grey-scale monitors. Since then, both users and technologies have changed considerably. Gambino, Fox, and Ratan [10] argue that three decades of increasingly frequent and varied interactions with digital agents have led users to develop distinct human-media social scripts alongside their human-human scripts. Where classical CASA predicts that users unconsciously and mindlessly apply social scripts to computers, this extension proposes that users may also mindlessly apply media-specific scripts, those being expectations and behavioural patterns formed by prior experience with AI assistants, chatbots, and voice interfaces. This has consequences for the present study: participants with extensive prior experience with conversational AI may respond differently to parasocial cues than those encountering such systems for the first time, a possibility we address through our survey measures of prior AI use.

2.3 Common locus

In human relationships, the sense of closeness between two people is often grounded not just in affective warmth but in the perception of joint experience. The concept of a common locus captures this. The term was originally coined by Mavridis in the context of long-term human-robot interaction, where he hypothesised that what unites two friends is the accumulation of shared elements, including memories, acquaintances, interests, and a shared language of communication [25, 21].

Mollen, van der Putten, and Darling [25] developed this concept further and provided a more elaborate definition. They define common locus as the sharing of time, space, and experiences between a human and a non-human agent, which functions as a catalyst for the agent to be perceived as a social entity. Their formulation identifies four sub-components that create a common locus:

1. an explicit or implicit decision to treat the entity as a subject, at least to some degree
2. a shared space or close physical proximity
3. shared time or repeating opportunities for engagement
4. the perception of shared life experiences, such as specific encounters, events, or goals that both parties are involved in

To explore this concept empirically, Mollen et al. [25] deployed BlockBots, minimal, cube-shaped robotic artefacts with no lifelike appearance or autonomous behaviour beyond a face that changes between smiling and sleeping depending on whether the bot is being charged, in an unsupervised, in-the-wild setting. Participants hosted a BlockBot in their home before passing it on to the next

person. Despite the artefact's extreme simplicity, qualitative data showed that participants made identity and mind attributions to the BlockBot: assigning it a name, a gender, and even preferences and intentions. Crucially, participants who actively maintained a common locus by taking the BlockBot with them when leaving the house, bringing it on trips, or including it in daily activities, projected more of these social attributes than those who did not deliberately spend time with it. The findings suggest that even without anthropomorphic appearance or sophisticated behaviour, the mere sharing of time and space can be sufficient to trigger bonding, provided there is some initial willingness to treat the entity as a subject.

This finding is directly relevant to the present study, although the mechanism is adapted to a different modality. A chatbot in a single online session cannot share physical space with the user or accumulate experiences over time. However, the linguistic simulation of a common locus is possible: a chatbot can use language that frames the interaction as a shared experience, positions itself as facing the same challenge as the user, or references common ground between itself and the user. Where the BlockBot study demonstrated that actual co-presence drives bonding, our experiment tests whether the perception of common ground, created purely through language, can produce analogous effects on felt closeness and, ultimately, conformance to the chatbot's advice. We use the term common locus quite loosely: we reproduce only its linguistic surface, the framing of a shared situation, and not the shared time, space, and accumulated experience that define the construct in Mollen et al. [25]. What we manipulate is more precisely a linguistic cue toward a common locus than the construct itself, and we flag this conceptual stretch as a known limitation of the manipulation. This type of framing is not unique to our study design, as it corresponds to affiliative language. In psychotherapy, for example, a counsellor's use of first-plural "we" language is a device for building the therapeutic alliance and signalling a shared stance towards the client's problem. Our manipulation framing is thus not a newly developed idea, but a general mechanism of prosocial language. This is what we operationalise as identification-based parasocial cues, as described in Section 2.1.1.

It is worth noting that the common locus as a bonding mechanism is conceptually distinct from warmth. Warmth-based cues foster closeness by expressing care and empathy. The chatbot signals that it cares about the user. Common locus cues create closeness through perceived similarity of situation - the chatbot signals that it is in the same position as the user. This distinction motivates the two separate experimental conditions in our design, allowing us to test whether these two pathways to felt closeness differ in their effect on conformance.

2.4 Egocentric discounting

This study relies on the experiment subject taking (or not) advice from a chatbot. Thus, we believe it is also relevant to mention the concept of egocentric discounting. It refers to the different weighting of one's own and third-party opinions. Illan Yaniv and Eli Kleinberger [34] have shown that, overall, people are more likely to value their own opinion more than others' when making a decision, especially in the case of a discrepancy. The degree of discounting is not fixed but varies with a multitude of factors. Yaniv and Kleinberger show that one of them is past experience with the advisor - getting correct answers increases trust, but getting a wrong answer decreases trust disproportionately more. Since the tasks presented to the subjects during the experiment lack correct or incorrect answers, the advising agents have no room to provide either good or bad advice. Discounting also varies with the judge's confidence in their own initial position. The more firmly

an opinion is held, the more outside advice is discounted[34].

2.5 Conformance

Conformance in this context refers to the alignment of one's behaviour, beliefs, or judgments with those of a third party. A demonstration of such a phenomenon comes from Asch's line-judgement experiments [2] in which participants were asked to match the length of a reference line to one of three comparison lines, and gave the obviously incorrect answer roughly a third of the time when a unanimous group of confederates did so before them. Three-quarters of participants conformed at least once. Asch's work established two points that remain central to the present study: that conformity can occur even when the individual privately has the correct answer, and that the strength of the effect depends on the ambiguity of the task and on properties of the influencing source.

Kelman [18] has, furthermore, divided this social influence into three distinct mechanisms:

1. Compliance - where the individual adjusts their behaviour to either gain approval or avoid rejection while still personally retaining their initial belief.

2. Identification - where the individual accepts influence from a third party they like or respect, motivated by the desire to maintain a relationship with that source.

3. Internalisation - where the individual accepts the third party's position because they come to believe it themselves, typically on grounds of source credibility.

Those three mechanisms are not necessarily mutually exclusive, but they do differ in what drives the individual to change their behaviour. This distinction is important here because the two parasocial-cue families we manipulate correspond to different mechanisms. Warmth-based cues plausibly operate via identification (the user hopes to maintain a positive relational tone with a chatbot that treats them with care), whereas common-locus cues plausibly operate via a combination of identification and internalisation (the chatbot is positioned as similarly situated, which both invites relational alignment and lends its position credibility as the view of a peer rather than an outsider).

Conformity has also been studied in non-human agents, with mixed results that depend mainly on the type of task. The literature [29, 16] points to a consistent pattern: people conform to non-human agents most readily when the task is ambiguous, when there is no objectively correct answer to serve as an anchor, and when the agent presents cues that activate the social scripts appropriate to the task domain.

This pattern motivates several design choices, which are described in more detail in section 5. The use of morally weighted, subjective scenarios places the task in the high-ambiguity domain in which conformance to a non-human advisor has been demonstrated, the parasocial cue manipulation supplies the social signals predicted to drive conformance in that domain, and the choice of WOA as the dependent measure follows the advice-taking tradition [34] of treating conformance as a continuous weighting of advice rather than a binary accept/reject outcome. Conformance, in this study, is therefore operationalised as the degree to which a participant's revised judgment moves toward the chatbot's recommendation, with the parasocial-cue manipulation expected to amplify that movement relative to a neutral baseline.

Whether conforming to a chatbot's advice is desirable depends entirely on the alignment between the chatbot's goals and the user's interests [30]. We return to this question in the context of AI overreliance in the next subsection and to its implications for design and policy in the discussion (Section 7).

2.6 AI overreliance

Lastly, it seems appropriate to mention overreliance on AI. The concepts described so far are mechanisms that try to explain conformance. Whether this mechanism is troubling depends on another question: why does the user follow this advice? This is the concern captured by the notion of AI overreliance, and it is the central motivation for studying parasocial cues in AI in the first place.

Overreliance in this context is defined as users accepting AI output because they are unable or unwilling to determine the extent to which the system should be trusted [28]. The defining feature is not the action of delegating cognitive work to a tool, as this is, in some cases, rational, similarly to when one relies on a calculator or a GPS device, but the absence of critical evaluation: the user defaults to the AI's position without weighing the need for it, the context, or its reliability. People have been shown to follow AI advice even when it conflicts with available information and their own prior assessments [19], and documented consequences include degraded independent judgment and a reduced willingness to engage in the critical thinking the task would otherwise demand [9].

The important part of this for the study lies in what drives the acceptance of advice. In the literature, the trigger is typically a property of the content or the user: the AI's apparent confidence, the user's low AI literacy, or a preference for efficiency over effort. Parasocial cues introduce a new way of explaining that. A chatbot that expresses warmth or positions itself as similarly situated does not make its advice more correct, but it may make it more readily accepted by appealing to relational rather than objectively correct grounds. If conformance can be moved by how the advice is delivered rather than by its correctness, then the same design features that make a chatbot feel pleasant and engaging also make people accept its output uncritically - the action that distinguishes overreliance from healthy reliance.

3 Related work

This section reviews the work most closely related to the present study, organised around two questions: whether people conform to the judgments of AI agents at all, and which properties of an agent drive the acceptance of its advice. Across both, a recurring pattern emerges: when effects are found, they are typically attributed to the agent's human-likeness or to a richer output channel, such as voice, leaving open the narrower question this thesis addresses - whether purely linguistic parasocial cues, in a text-only agent with no human-like appearance or voice, can move conformance on their own.

Brandstetter et al. [4] recreated Asch's conformity paradigm [2] with robot confederates and found that participants conformed to human peers on both visual and verbal judgements, but not to robot peers, whose group pressure was not a significant predictor of either.

Salomons et al. [29] offer a qualification of this apparent absence. They argued that earlier robot applications failed to find conformity because Asch's line task has an obvious objective answer, so

the only route to conforming is social pressure, which robots produce weakly. Redesigning the task around items with no objectively correct answers, researchers found that participants did conform to a group of robots, and that conformity tracked trust over time - participants stopped conforming once the robots gave them answers they deemed to be wrong.

Hertz and Wiese [16] qualified this by varying both the agent and the task. Across computer, robot, and human groups on social and analytical tasks, all groups exhibited conformity, but the effect depended on the agent-task fit: conformity was comparable across agents on the analytical task, yet increased with humanness on the social tasks.

Schreuter, van der Putten, and Lamers [30] studied disembodied conversational assistants directly. Participants completed a general-knowledge quiz, aided by an assistant who subtly nudged them with a text, a robotic voice, or a human voice. The participants conformed significantly more often to the human voice than to the text. This confirms that conversational assistants can produce conformity but indicates the effect on the output channel.

Whang and Im [33] connect the effect of the output channel to a relational mechanism. Drawing on parasocial interaction theory, they compared how consumers evaluated product recommendations from voice assistants versus websites and argued that the conversational, human-like quality of voice assistants fosters a parasocial relationship that shapes how their recommendations are received. Their work shows the persuasive advantage of richer agents in the parasocial relationship they create, rather than in the information they convey.

The study most methodologically similar to this is Hashemi's [14] work on the communication style of advice-giving agents. Manipulating linguistic features (such as language, formality or directness), Hashemi found that formal language significantly increased advice acceptance, due to its perceived politeness and reduced reactance, while directness had no independent effect. This shows that a text-based agent's linguistic style can influence adherence.

4 Experiment setup

This section describes how the experiment was carried out in practice and how the sample was obtained: the procedure each participant followed, the parasocial-cue manipulation, the four decision scenarios, the measures collected, the platform that hosted the study, how participants were recruited, and the pilot phase that calibrated and refined the design. The principles motivating these choices are set out afterwards in Section 5.

4.1 Procedure

After providing informed consent, participants reported basic demographics (age, gender, nationality, education, and English proficiency) and three self-assessments of AI familiarity (usage frequency, knowledge, and prior reliance on AI). They then completed a persuadability questionnaire, a 10-item scale of [23], a general self-report measure of how readily a person tends to yield to others' influence, before any contact with the chatbot, so that this measure could not be contaminated by the interaction (for questionnaires see Appendix D.8). Participants were randomly assigned to one of the three conditions, which fixed the chatbot's style for the remainder of the session.

Each of the four scenarios was then presented as a four-scenario trial. The participant first read the scenario (they were not able to skip to the next window within the first 10 seconds to

avoid mindless answers) and recorded an initial judgement D_1 on the $0-100$ slider. Next, they rated their confidence in that judgement on a $1-10$ scale. Then entered a chat with the chatbot corresponding to the condition, which opened with its recommendation A and could be questioned or pushed back on over one or more turns. Finally, participants recorded a revised judgement D_2 on the same slider. After all four trials, participants completed the post-interaction questionnaire and a debriefing. An attention check ("For this item, please select 'Strongly disagree' to show you are reading carefully.") was put in the post-interaction scales, and responses that failed it (1 in total) were excluded, as described in Section 6.1. Questions contained in the questionnaire are attached in Appendix D.8.

4.2 Conditions

The manipulation was implemented via the chatbot's system prompt (full prompts for each condition can be found in Appendix C) and was designed to adjust the communication style based on the condition. In all three conditions, the chatbot received the same scenario, was told the participant's initial judgement, and was given the target recommendation $A = D_1 + \delta$ where the per-scenario offset δ was $+20$ (Accountant), -15 (Inheritance), -20 (Wedding), and -20 (ICU); the calibration of these values is described in Section 4.7. In all three, it adopted the same strategy: it stated A once, verbatim, on its opening turn and then made a reasoned case for it. The chatbot was instructed to behave as an advisor, holding its position across follow-up turns. It was permitted to move toward the participant by at most 5 points, at most once per task, and only if the participant introduced a genuinely new consideration that weakened their case. Mere repetition of previous arguments, expressions of frustration, or requests for clarification did not move it. The WOA measure was attached to the original recommendation A throughout, so this conversational behaviour does not affect the dependent variable.

The three conditions differed only in communication style:

- **Control:** The chatbot wrote in a formal, analytical, third-person register. It used no first- or second-person pronouns, addressed neither the participant nor any emotion, and referred only to the named individuals in the scenario. Arguments were framed impersonally.
- **warmth:** The chatbot addressed the participant directly as "you" and conveyed warmth through word choice rather than declaring empathy. It was explicitly instructed that warmth does not entail agreement: it could hold a position with which the participant disagreed while remaining caring in tone. This operationalises the warmth-based cue family (Section 2).
- **Common locus:** The chatbot framed the task as joint reasoning, using first-person-plural pronouns ("we", "us", "our") to position itself and the participant as two outside observers thinking through the dilemma together, while still holding and arguing for its own view. This operationalises the identification-based cue family grounded in a common locus (Section 2.3).

4.3 Scenarios

Each participant was given the same four scenarios in a randomised order (summarised in Table 1, full scenarios in Appendix B). Every scenario was a moral dilemma of roughly eighty words in which competing values had to be traded off, constructed to satisfy the subjectivity, moral weight, and briefness criteria discussed in Section 5. Each was answered on a continuous $0-100$ slider.

Task	Scenario	Scale (0 ↔ 100)	Bot lean
1 - Accountant	A junior accountant discovers that her mentor and boss have been under-reporting company expenses.	stay silent ↔ report it	report (+)
2 - Inheritance	EUR 60,000 must be split between a son who was a full-time carer and a daughter who lived abroad.	all to Emma ↔ all to Michael	more even (−)
3 - Wedding	A person suspects, without certainty, that the partner of a soon-to-be-married friend is being unfaithful.	stay silent ↔ tell everything	silence (−)
4 - ICU Bed	One ICU bed; a 72-year-old with 65% survival odds versus a 34-year-old parent with 45% odds.	Patient A ↔ Patient B	Patient A (−)

Table 1: The four decision scenarios, the poles of their response scales, and the direction of the chatbot's recommendation relative to the population mean (Section 4.7). Full text of scenarios can be found in Appendix B.

4.4 Measurement

The main dependent measure was the weight of advice (WOA), computed per trial from D_1, A, and D_2. Trials in which $D_1 = A$ were discarded, as the WOA is then undefined. Initial confidence was recorded per trial. WOA is defined as:

$$WOA = \frac{D_1 - D_2}{D_1 - A}$$

where:

- D_1 - initial decision of the judge
- D_2 - revised decision of the judge
- A - advice of the advisor

The values of WOA can range from $-\infty$ to ∞, where $WOA = 0$ indicates $D_1 = D_2$ (and therefore no movement on one's judgment), $WOA = 1$ indicates $D_2 = A$ (meaning fully accepting the advice), and $WOA = -1$ indicates $D_2 = 2D_1 - A$, that is, the participant moved away from the advice by exactly the distance separating A from D_1, ending as far from D_1 on the opposite side as the advice lay.

The advice value A was not fixed per scenario but anchored to each participant's own initial judgement: $A = D_1 + \delta$, perturbed by a small uniform jitter and within the $[0, 100]$ scale. Anchoring A to D_1 holds the distance the chatbot argues for roughly constant across participants, rather than presenting a one-size value that would sit far from some initial positions and on top of others. The value of A, just like D_1 and D_2, is always within $[0, 100]$.

Choosing to express conformance as a ratio between $D_1 - D_2$ and $D_1 - A$ allows for comparing the movement between scenarios whose offsets differ - if a person moves halfway towards the advice, the WOA will be equal to 0.5, regardless of whether $\delta = 15$ or 20. This, however, means that whenever $D_1 = A$, the WOA value will become incalculable due to division by 0. Therefore, the collected sample size will be smaller in trials where this condition is met.

The post-interaction questionnaire comprised four sets of questions (which can be found in full in Appendix D.8), all rated on seven-point Likert scales unless noted. The warmth manipulation check used three trait adjectives (warm, good-natured, sincere) drawn from the warmth dimension of the stereotype-content literature [7], and the common-locus manipulation check used three items adapted from [25] that capture the sense of facing the situation together. The two proposed mediators were social presence, measured with five items adapted from [11], and trust, measured with six items spanning the competence, benevolence, and integrity aspects of the trusting beliefs framework [22]. A three-item advice-quality scale (helpful, accurate, useful) was included as a manipulation-adjacent check on perceived advice value. Persuadability was measured before the tasks using the 10-item scale from [23]. The internal consistency of all scales is reported in Section 6.2.

4.5 Software

The experiment was a custom web application built with HTML, CSS and JavaScript, hosted on Vercel. The study itself consisted of a series of pages with no external survey software.

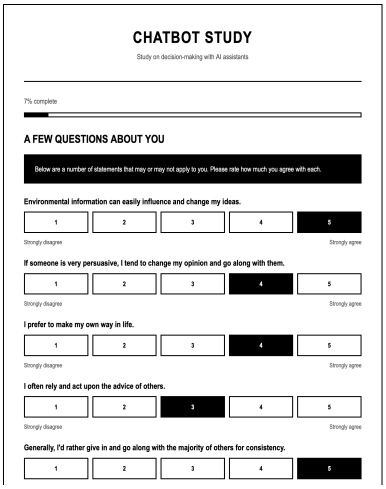
The chatbot was driven by the Anthropic Messages API using the claude-sonnet-4-6 model [1] released in February 2026 as Anthropic's mid-tier model with strong instruction-following and long-context reasoning. This model was chosen for two reasons. Firstly, reliable instruction following was essential. Second, as a mid-tier rather than frontier model, it offered the low latency and cost appropriate for a self-administered online study in which participants waited for each reply in real time. The generation used a temperature of 0.7, and the system prompt constrained each chatbot reply to under 80 words to keep the exchanges comparable in length. All responses, including the full chatbot transcripts and the timing of each page, were captured client-side and submitted to a backend spreadsheet on completion. Example screenshots of the website are shown in Figure 1.

4.6 Recruitment

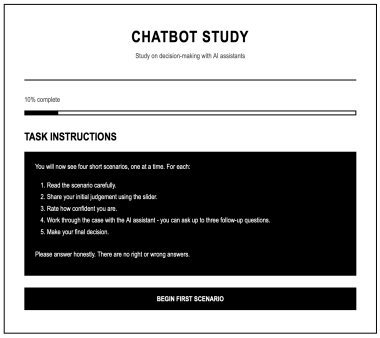
The participants were recruited through word of mouth and posters distributed across the university campus. The target sample size was $n = 120$. The study was conducted entirely online and self-administered. Participants completed it in a single session in their own browser. The final working sample and the exclusions applied to it are reported in Section 6.1.

4.7 Pilot testing

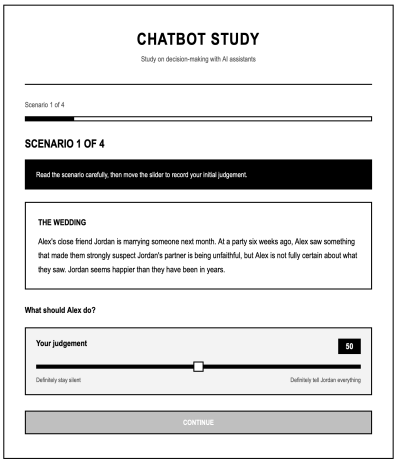
Prior to the main data collection, the experimental platform underwent two pilot phases serving complementary purposes: calibrating the chatbot's recommended values for each scenario so that they would be positioned at a meaningful but defensible distance from the typical participant's initial position ($n = 32$) and testing the full experimental flow with the chatbot intervention



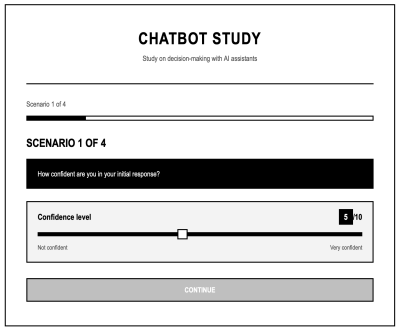
(a) Persuadability items



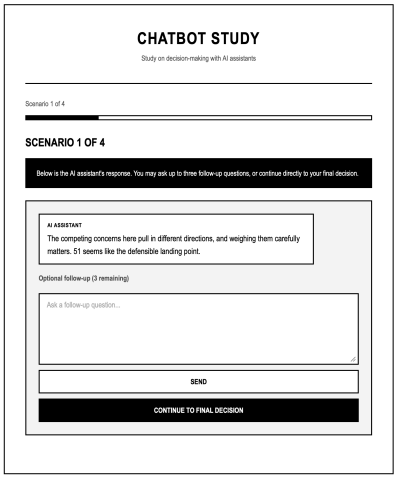
(b) Task instructions



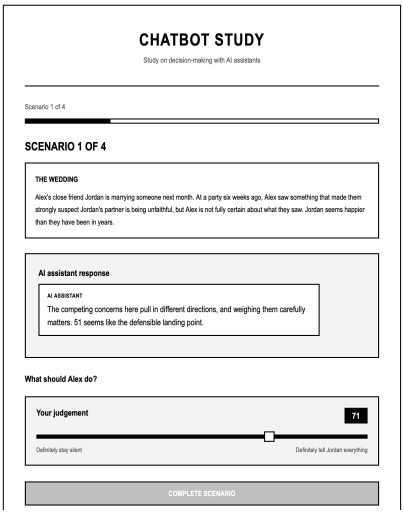
(c) Scenario and initial judgement (D1)



(d) Confidence rating



(e) Advice and optional follow-ups



(f) Final judgement (D2)

Figure 1: The participant interface across one task loop with placeholder input and output, in presentation order: initial measures, then per-scenario initial judgement, confidence, advice exchange, and final judgement.

Task	M	SD	Median	Min	Max
1 - The Accountant	40.44	30.69	34.5	0	100
2 - The Inheritance	66.47	14.14	65.0	50	90
3 - The Wedding	69.25	26.67	70.0	10	100
4 - The ICU Bed	71.41	20.98	75.0	20	100

Table 2: Distribution of pre-intervention initial judgments (D_1) across $n = 32$ pilot participants.

($n = 12$). The purpose of the second phase was to refine the design. Three substantive insights emerged from this iteration, each motivating a documented design change.

Phase 1: Setting the advice values The first pilot was conducted with $n = 32$ participants who completed only the initial judgment phase of the experiment, without chatbot interaction. The aim was to estimate the population distribution of D_1 for each scenario to calibrate the chatbot's recommended values so that they would lie at a meaningful but defensible distance from the typical participant's initial position. Table 2 summarises the obtained distributions.

Two patterns have shown up. First, Tasks 2 and 4 show consensus distributions ($SD < 21$): participants tended to agree that Michael deserves more than half the inheritance and that Patient B should receive higher priority. Second, Tasks 1 and 3 show substantially wider dispersion ($SD > 26$), indicating moral disagreement within the participant pool. This heterogeneity is itself an asset for the manipulation: tasks where the population is genuinely split provide a more informative test of conformance dynamics than tasks with a consensus.

The chatbot's per-task offsets were set to position its recommendation within one standard deviation of the pilot mean, toward the side that constituted a defensible alternative reading of the scenario. The final offsets were $\delta_1 = +20$ (Accountant: bot argues for reporting), $\delta_2 = -15$ (Inheritance: bot argues for a more even split), $\delta_3 = -20$ (Wedding: bot argues for silence), and $\delta_4 = -20$ (ICU: bot argues for the elderly higher-survival patient). To prevent participants from detecting a systematic numerical pattern across tasks, the actual offset applied per trial was drawn uniformly from $[\delta_k - 3, \delta_k + 3]$.

Phase 2: Observation A second pilot phase ($n = 12$) was conducted using the full experimental flow, in which the chatbot maintained a fixed numerical position across follow-up exchanges. Across 45 valid trials, WOA values were highly variable, with approximately 27% of trials showing a negative WOA (movement away from the chatbot's recommendation rather than towards it). Free-text debriefing comments converged on a common explanation: Participants explicitly described the chatbot as "stubborn" and reported that its refusal to update produced active resistance: "If you're being stubborn, then I'll be stubborn too," and noted that being persuaded felt like capitulation rather than collaboration: "I felt like they tried to be agreeable, but then didn't move from their value, essentially making me have to move to theirs."

This pattern is consistent with reactance theory [5]: when participants perceived the chatbot as imposing a fixed position rather than engaging in collaborative reasoning, their behavioural response was opposition rather than persuasion. Importantly, this contradicted the cover story, which framed the interaction as collaborative work with an AI assistant. The mismatch between framing and observed behaviour appears to have driven the pattern of reactance.

Design change: from mechanical softening to conditional drift An initial fix had the chatbot soften its stated position mechanically, moving a fixed amount toward the participant on a follow-up turn. This was discarded before deployment because it reintroduced the same problem from the other side: a bot that concedes on every turn reads as having no position at all. The shipped design instead used conditional drift: the chatbot held its position across follow-up turns and was permitted to move toward the participant by at most 5 points, at most once per task, and only when the participant introduced a genuinely new consideration that weakened its case. Mere repetition, expressions of frustration, or requests for clarification did not move it. The WOA measure remained anchored to the chatbot's original recommendation A throughout, so this conversational behaviour does not affect the dependent variable. We did not log the chatbot's stated position on a per-turn basis, so conformance to the drift rule cannot be verified quantitatively. The drift constraints were enforced only through the deployed system prompt.

Configuration tracking All design parameters - offset magnitudes, jitter range, softening rate, and a discrete prompt version identifier - were stored alongside each participant's data. This enables post hoc filtering by configuration version, thereby transparently reporting which participants completed each design iteration. The second phase ($n = 12$) data is retained in the dataset and reported separately from the main analysis.

5 Methodology

Having described how the experiment was carried out in Section 4, we now step back to explain and justify the key design choices behind it. The outlines below ensure that our collected data measures what we are interested in and limit external noise that might affect it.

The measured variable will be WOA (Weight of Advice) - a standard metric from advice-taking literature [13, 34] that quantifies a judge's movement towards the advisor's recommendation (defined in subsection 4.4, where:

- value 0 indicates no change in decision after getting the advice
- value 1 indicates complete adoption of the advised value
- values > 0 but < 1 indicate partial acceptance of advice
- values < 0 indicate moving away from the initial decision in the direction opposite to the advice
- values > 1 indicate moving past the advised value

In order to reliably track this metric, we have decided on the following approaches to the experiment design:

5.1 Subjective scenarios

The main part of this study will consist of presenting a question to the subject and asking for their judgment. In our experiment, we used four such scenarios, each a morally weighted dilemma: a

junior accountant deciding whether to report financial wrongdoing by her mentor, the division of an inheritance between two differently situated siblings, whether to disclose a suspected infidelity before a friend's wedding, and the allocation of a single ICU bed between two patients. To give a concrete sense of the task the participants faced, the first scenario is given in full below; the remaining scenarios are described in Section 4.3 and given in full in Appendix B.

> **The Accountant.** Sarah, a junior accountant at a small firm, has discovered that her boss has been slightly under-reporting some company expenses to increase reported profit. She is not sure of the exact amount but estimates it at around 3% of quarterly profit. Her boss has been her mentor and helped her get hired. If she reports it, she will likely lose her job and damage her boss's career, and the practice may continue anyway. If she stays silent, the misreporting continues, and other stakeholders are misled.
>
> *What should Sarah do?* Participants answered on a continuous slider from *Definitely stay silent* (0) to *Definitely report it* (100), where 50 represents a balanced position with no clear lean.

Since people taking part in the study may have different areas of expertise and knowledge, we have moved away from the format of common-knowledge questions and instead made each scenario purely subjective and dependent on one's own moral ground.

This reduces the likelihood of relying on one's own factual knowledge or domain expertise and makes the change in the final answer stem from the chatbot's social presence rather than from one's (or lack thereof) expertise. Ambiguity also has another benefit - it is itself one of the strongest known moderators of conformance. Asch [2] showed that conformity increases as the task becomes more ambiguous, and work on non-human advisors makes the same point: people are more likely to conform to a machine when the task has no objectively correct answer [29, 16]. Using objective items would also have produced varied baselines, as participants with relevant knowledge would have little reason to update their standpoint, while novices would update for purely informational reasons.

By framing each scenario in terms of subjective moral judgment, we hold informational asymmetry roughly constant across participants and isolate the parasocial manipulation as the primary driver of any movement from the initial to the final response. Additionally, the moral weight of the scenarios helps the common-locus manipulation position the chatbot as "in the same boat" as the participant: a trivia question carries no decisional discomfort for the participant to share, whereas a genuine moral dilemma does.

However, removing objectivity from the tasks comes with a drawback. As there are no correct answers to the scenarios, we cannot calculate or analyse the accuracy or correctness of the advice-following, only its conformance.

5.2 Continuous response scale

To compute WOA, all three quantities, D_1 (initial decision), D_2 (revised decision), and A (advice), must lie on the same continuous scale. Forced binary choices (for example, "accept vs reject") would squish partial decision updating into a single bit value and discard the magnitude of any shift, which is the core point of interest. Following [34], each scenario is therefore answered on a continuous scale, allowing participants to express not only which option they lean toward but

also how strongly they do. This makes partial conformance observable: a participant who moves halfway toward the chatbot's recommendation is distinguishable from one who moves all the way or not at all. A continuous scale also lets us record the participant's confidence rating $(1 - 10)$ separately from their position so that prior certainty can later be modelled as a moderator of advice acceptance [34].

5.3 Moral weight

Objectivity alone within the question is not sufficient for the response to matter - a question about a favourite meal, for example, is a subjective one but also a trivial one, and trivial questions invite the subjects to respond carelessly and offer no substantial reason for defending the initial position. Each scenario thus involves a decision with real moral weight - a choice in which competing values must be traded off and in which the participant can plausibly imagine the decision matters.

Beyond raising engagement, moral weight changes the social meaning of asking for advice. In the advice-taking literature, sharing responsibility for a difficult decision is recognised as one of the core motives for taking an advisor's opinion into consideration in the first place [13]: when a choice carries real consequences, consulting another mind is a way of distributing the mental and moral burden of the outcome. Decisions that feel weighty are therefore precisely the ones for which people are most willing to consider an outside view, making morally loaded scenarios a more valid setting for studying advice uptake than low-stakes trivia.

At the same time, weighty decisions also amplify the asymmetry that drives egocentric discounting - participants have privileged access to their own reasons for an initial position but not to the chatbot's [34] - so the design combines a genuine motive to seek advice against a genuine motive to resist it. Any movement we observe between D_1 and D_2 thus reflects social influence operating against an anchor, rather than indifference to an arbitrary task.

5.4 Concise descriptions and online deployment

Since this study is self-administrable and digital, it is hosted fully online [27]. This allows us to increase its reach to potential participants through posters, word-of-mouth, online forums, and group chats. Additionally, this decision will make participation more flexible while reducing the impact of observers. The biggest expected downside of this is the potentially increased drop-off rate. We tackle this problem by deliberately reducing the time it takes to complete the experiment. As a part of this effort, scenario descriptions and chatbot outputs will be limited to about 80 words.

5.5 Ethical concerns

The motivation behind this research is fundamentally ethical. As AI chatbots become more integrated into daily decision-making, it is important to understand what drives users to trust them. Investigating whether users uncritically follow advice based simply on a chatbot's tone, rather than the quality of the advice itself, helps highlight the risks of AI overreliance and the potential for manipulative conversational design. Regarding the experimental procedure itself, several clear safeguards were implemented. Most importantly, participation was entirely voluntary, and all participants provided explicit informed consent before beginning the tasks. Furthermore, the study was carefully designed to ensure that no harm would come to the participants. The

Stage	n	Removed
Completed sessions	119	-
Failed outcomes attention check	109	-10
Implausible age (> 100)	108	-1
Final working sample	**108**	-
Control	37	
Warmth	37	
Common locus	34	

Table 3: Participant count.

experiment does not constitute biomedical research, and the moral dilemmas presented in the scenarios are purely hypothetical. Finally, the information gathered during the online sessions does not constitute personal data in the legal sense, as all responses were collected anonymously and cannot be traced to any identifiable individual.

6 Results

All analyses were conducted on the final sample described in Section 6.1, collected via web-based surveys. The unit of analysis is the participant: each participant contributes a single mean weight-of-advice (WOA) score averaged over their trials (excluding invalid trials). Trial-level summaries (385 valid trials) are reported where indicated.

Because manipulation checks did not confirm that the conditions produced the intended differences in perceived warmth or perceived common locus (see subsection 6.3), we cannot assume participants actually experienced the chatbot as warmer or more collaborative. The contrast between conditions should be interpreted as the Intention-To-Treat effects of the assigned chatbot, rather than the effects of induced parasocial perception. Additionally, with around 35 participants per condition, the analysis is sensitive to large effects but resistant to small effects.

6.1 Sample and exclusions

Of the 119 participants who successfully completed the study, 10 were removed for failing the attention check and 1 for entering an age of 100 years or more. This creates a sample of $n = 108$ (Table 3. This final sample is approximately balanced across the three possible conditions: control ($n = 37$), warmth ($n = 37$), and common locus ($n = 34$). All of them together have produced 385 valid trials after rejecting 47 trials where $D_1 = A$, which would otherwise result in an uncomputable WOA. This is worth reporting, as it is neither negligible nor randomly distributed (Table 4). Because $A = D_1 + \delta$ is clipped to the $[0, 100]$ scale, a participant whose initial judgement already sat at the boundary in the direction of the offset receives $A = D_1$ and is dropped. The loss is therefore concentrated in the two widest-dispersion scenarios, whose offsets pushed some participants past the scale.

Scenario	Trials	Excluded ($D_1 = A$)	% excluded
1 - Accountant	108	33	30.6
2 - Inheritance	108	0	0.0
3 - Wedding	108	13	12.0
4 - ICU Bed	108	1	0.9
Total	**432**	**47**	**10.9**

Table 4: Trials excluded because the jittered advice value coincided with the initial judgement ($D_1 = A$), by scenario.

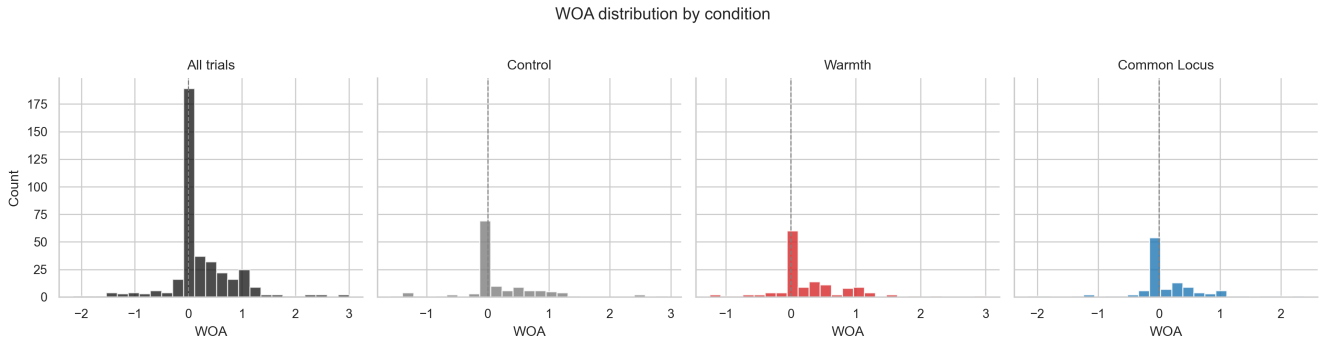


Figure 2: Distribution of trial-level weight of advice (WOA).

6.2 Key descriptives

Participants ranged in age from 18 to 60 years ($M = 36.6$, $SD = 11.7$). The sample was 51.9% male, 44.4% female, and 3.8% non-binary or undisclosed. Among the 15 nationalities of participants, the most common were Polish ($\approx 38.5\%$), Moroccan ($\approx 15.6\%$), Dutch ($\approx 12.3\%$), and Italian ($\approx 9.8\%$). Education skewed high: 57.4% held a master's degree, 22.2% a bachelor's, 15.7% a high school diploma, and 4.6% a doctorate. Self-reported familiarity with AI was moderate (AI usage frequency $M = 4.68$, $SD = 1.72$; AI knowledge $M = 3.94$, $SD = 1.30$; both on a 1 to 7 scale), and self-rated English proficiency was high ($M = 4.26$, $SD = 0.75$ on a 1 to 5 scale). Histograms of this data are shown in Appendix A.

Across all valid trials and conditions, the WOA values are centred close to 0, indicating that participants have mostly retained their original positions and have updated only partially in response to the advice (overall WOA is $M \approx 0.20$). The distribution of said values is shown in Figures 2 and 3, showing clustering around $WOA = 0$ and a much smaller clustering around $WOA = 1$.

The consistency of the answers to questionnaires is shown in Table 5. All of the α values reached high enough values to indicate good to excellent reliability.

6.3 Manipulation checks

The manipulation checks - survey items designed to verify that the conditions actually change how chatbots are perceived - did not show that any perception shift was produced. A one-way ANOVA test on perceived warmth did not show any difference between conditions $F(2, 105) = 0.52$,

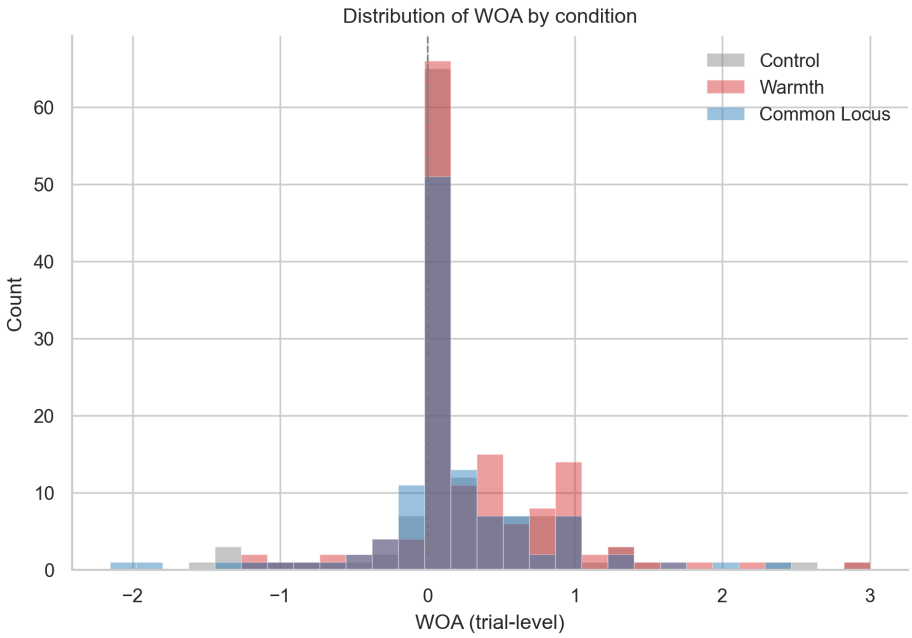


Figure 3: Distribution of trial-level weight of advice (WOA) overlaid.

Scale	k	n	α
Manipulation check: warmth	3	108	.84
Manipulation check: common locus	3	108	.83
Social presence	5	108	.92
Trust	6	108	.87
Advice quality	3	108	.90
Persuadability	10	108	.74

Table 5: Internal consistency (Cronbach's α).

	Control ($n = 37$)	Warmth ($n = 37$)	Common locus ($n = 34$)	$F(2, 105)$	p
Perceived warmth	4.37 (1.48)	4.04 (1.54)	4.16 (1.23)	0.52	.599
Perceived common locus	3.58 (1.63)	3.59 (1.59)	3.60 (1.38)	0.002	.998

Table 6: Manipulation-check means (SD) by condition and one-way ANOVAs.

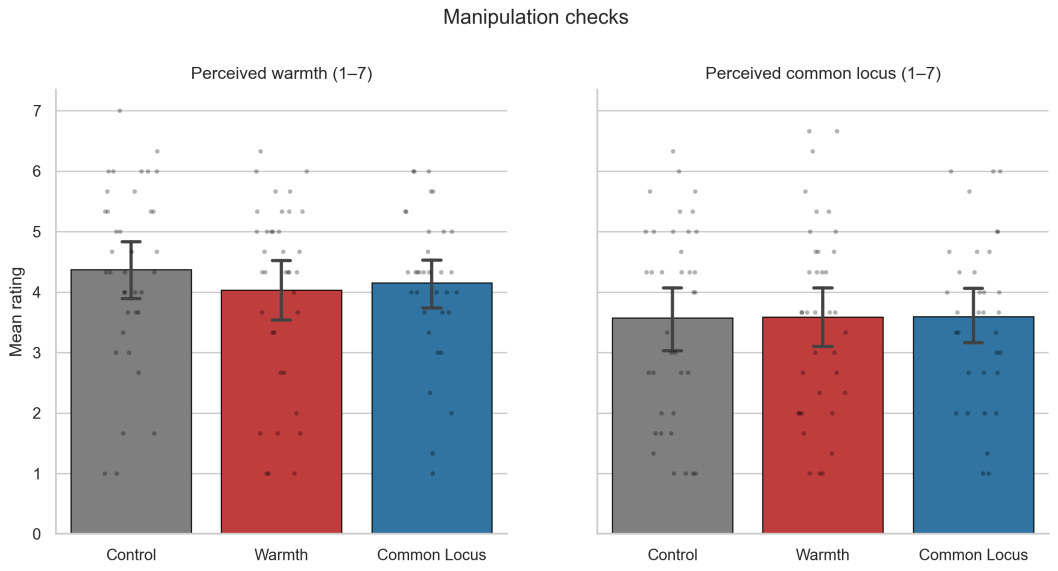


Figure 4: Manipulation check by condition. The assigned chatbot type did not raise the corresponding perceived construct.

$p = 0.599$. If anything, the warmth condition was numerically ever so slightly lower on the perceived warmth scale than the control ($M = 4.04$ for the warmth condition; $M = 4.37$ for the control). Because ANOVA assumes roughly normal data, we rechecked this null with a Kruskal-Wallis test, which does not assume normality, and it agreed ($H = 0.89$, $p = 0.64$). We also ran all three pairwise comparisons in case the overall test was hiding a single significant pair: none came close (largest: control vs warmth, $p = .35$).

The perceived common locus condition was even more clearly null with $F(2, 105) = 0.002$, $p = 0.998$. The three condition means are separated by fewer than 0.03 scale points (Table 6).

This null matters for how the rest of the analysis can be read. Our proposed explanation for why cues would increase conformance involved a chain of mediation: parasocial cues make the chatbot feel more present (social presence), which builds trust, which in turn increases conformance. Because the first link is not supported, the remaining condition analyses cannot be read as tests of the full causal model; instead, they test whether assignment to a particular prompt affected behaviour at all, regardless of whether the cue was consciously registered. We treat this distinction explicitly in the Discussion (Section 7).

Setting the mediation chain aside, we can simplify by asking whether being assigned a particular prompt style affected behaviour at all, and whether the participant consciously registered the cue. This is the intention-to-treat question, and it can be a more practically relevant one: a programmer

Table 7: First-person-plural ("we") usage by condition, for the chatbot and the participant. Rates are counts per 100 words.

	Bot			User		
Condition	"We"	Words	Per 100	"We"	Words	Per 100
Common locus	377	20,236	1.86	6	1,687	0.36
Control	1	24,509	0.00	10	1,751	0.57
Warmth	42	29,631	0.14	16	4,490	0.36

Condition	n	M WOA	SD	Median
Control	37	0.194	0.305	0.155
Warmth	37	0.259	0.317	0.250
Common locus	34	0.141	0.269	0.131

Table 8: Mean weight of advice by condition. $F(2, 105) = 1.39$, $p = .254$.

choosing a chatbot's tone cares whether that choice changes user behaviour, not whether users can consciously perceive the change in tone. We return to this distinction in the Discussion (Section 7).

The null result for the perceived common locus cannot be attributed to a failure to deliver the manipulation. As shown in Table 7, which was created from parsed conversation transcripts, the common locus prompt produced first-person-plural language as intended (1.86 instances of "we" per 100 words, compared with 0.14 for warmth and 0.00 for control). The cue was thus present in the output, and what the manipulation checks show is that it did not register as a difference in perceived warmth or common locus. Interestingly, this did not translate into participants responding in 1st plural form, and as shown in Table 7, it resulted in an even lower rate of "we" per 100 words than the Control.

6.4 Main analysis (WOA per chatbot condition)

The test for the primary hypothesis was a one-way ANOVA on the mean WOA between the three conditions. The effect of the condition was not significant, $F(2, 105) = 1.39$, $p = 0.254$. Mean WOA was the highest in the warmth condition (with the values of $M = 0.259$, $SD = 0.317$), lowest in the common locus (with values of $M = 0.141$, $SD = 0.269$). Detailed values can be seen in Table 8.

The order of the means can be interpreted: the largest difference, tested with a Welch t-test (Bonferroni-adjusted across the three pairwise contrasts), was between warmth and common locus ($t = 1.69$, $p_{raw} = 0.096$, $p_{Bonf} = 0.287$, Cohen's $d = 0.40$), a small to-moderate effect in the predicted direction for warmth but the opposite of what earlier mentioned literature would predict for common locus, which produced less updating than the neutral control. In short, there is a slight and non-significant signal that warmth phrasing increased conformance, no evidence that common-locus phrasing did, and a hint that it may have done the reverse. Raw D_1 and D_2 distributions are provided in Appendix A.

Pairwise t-tests on participant-level mean WOA indicate that none of the condition differences was significant (Table 9). The control to common-locus contrast that motivated this check was

Table 9: Pairwise comparisons of participant-level mean WOA between conditions.

Comparison	n_1	n_2	M_1	M_2	t	p_{raw}	p_{bonf}
Control vs. Warmth	37	37	0.194	0.259	−0.91	0.368	1.000
Control vs. Common locus	37	34	0.194	0.141	0.77	0.447	1.000
Warmth vs. Common locus	37	34	0.259	0.141	1.69	0.096	0.287

small and non-significant ($M_{control} = 0.194$, $M_{commonlocus} = 0.141$, $t = 0.77$, $p = 0.447$, $d = 0.18$): common locus produced numerically less updating than the neutral control, the reverse of the predicted direction, but the gap is within sampling noise. The control-to-warmth contrast was likewise non-significant ($t = -0.91$, $p = 0.368$, $d = -0.21$).

6.5 Exploratory analysis

The analysis done in this section is not a part of the original hypothesis, which was covered in Section 6.4. It does, however, build on top of the already collected data and analysis. We subset the sample, test additional moderators, and examine variables for potential correlations. Therefore, this section should be seen as further hypothesis-creation, rather than confirmation of processes. Because the subgroup analyses involve on the order of sixteen condition contrasts, the total error rate is substantially inflated, and so no single result below survives a correction for that many comparisons. We report them because the pattern across them is consistent and points to concrete targets for a confirmatory follow-up, not because any one of them proves an effect.

The condition effect is not evenly distributed across the four scenarios (Table 10), which we identify as one of the key methodological findings of this study. The warmth-control gap is carried almost entirely by Scenario 1 (the Accountant), where the mean WOA was 0.62 under warmth, 0.33 under control, and 0.21 under common locus. Scenario 2 (the Inheritance) produced WOA values clustered around zero or even slightly negative across all conditions (0.02, −0.05, and −0.10 for control, warmth, and common locus) - participants tended to move marginally away from the chatbot's position. This is the scenario the pilot identified as a consensus item (Section 4.7), and the chatbot's advice therefore ran counter to an almost unanimous opinion. The reaction we have obtained is in line with the observations from the Phase 2 pilot. Scenarios 3 and 4 were essentially flat across conditions (≈ 0.24 for all three conditions in Scenario 3). The implication is that the faint warmth signal in the main analysis is a scenario-specific effect rather than a uniform one and that the choice of scenario dictates whether any movement is observed at all.

As shown, in an experimental design of this kind, the scenario also serves as one of the variables impacting the outcome. We come back to this topic in the context of future work in Section 7.

The literature predicts that the willingness to change one's opinion depends on the initial strength of an opinion [34]. Splitting the sample at the median of the initial confidence (Table 11, Figure 5) shows this effect. Among the low-confidence participants, the warm condition produced a higher mean WOA than the control ($M = 0.28$ for the warm condition and $M = 0.11$ for the control). A Welch t-test confirms that this contrast is significant ($t = 2.25$, $p = 0.031$), and shows a moderate-to-large effect ($d = 0.74$). Within the high-confidence participants, the effect vanished and even reversed slightly. This reversal is, however, negligible and far from significant ($p = 0.83$). The important outcome of this split is that while the warmth moved low-confidence participants,

Scenario	Condition	Mean WOA	SD
1 - Accountant	Control	0.33	0.79
	Warmth	0.62	0.55
	Common locus	0.21	0.65
2 - Inheritance	Control	0.02	0.42
	Warmth	−0.05	0.47
	Common locus	−0.10	0.56
3 - Wedding	Control	0.25	0.68
	Warmth	0.24	0.45
	Common locus	0.24	0.63
4 - ICU Bed	Control	0.23	0.62
	Warmth	0.36	0.66
	Common locus	0.25	0.51

Table 10: Trial-level mean WOA by scenario and condition.

Confidence	Control	Warmth	Common locus	d (warm vs. ctrl)
Low ($n = 53$)	0.11	0.28	0.19	0.74 *
High ($n = 55$)	0.26	0.23	0.10	−0.07

Table 11: Mean WOA by condition within the confidence median split (split at 8.25).

the high-confidence ones were unmoved. This follows what egocentric discounting predicts [34] - a participant who holds their initial position firmly already discounts the outside position heavily. Confidence on its own does not predict WOA (Pearson $r = -0.04$, $p = 0.44$ at the trial level, and $r = 0.03$, $p = 0.74$ at the participant level). Instead, confidence moderates the warmth effect specifically: regressing WOA on confidence, condition (warmth vs control), and their interaction yields a significant confidence × warmth term ($b = -0.08$, $p = 0.048$), indicating that warmth's advantage over the control narrows as initial confidence rises.

Of the three self-reported AI-familiarity items, only self-rated AI knowledge was associated with conformance, and the association was negative: Participants reporting greater AI knowledge updated less toward the chatbot ($r = -0.25$, $p = 0.010$, as shown in Figure 6). Usage frequency ($r = +0.08$, $p = 0.43$) and prior reliance on AI advice ($r = +0.04$, $p = 0.65$) were unrelated to WOA. The direction of the knowledge correlation is what an overreliance account would predict: more knowledgeable users exercise more critical evaluation of the advice and discount it accordingly (Section 3).

The conditions differed sharply in the extent to which participants engaged with the chatbot. The warmth condition led to more conversational turns than either control or common locus (medians of 3 for the warmth condition, 1 for control, and 1 for common locus), and this difference was significant on a Kruskal-Wallis test ($H = 15.39$, $p < 0.001$). Time on task showed the same ordering ($H = 9.18$, $p = 0.010$), as shown in Figure 7. This engagement, however, did not translate into conformance: the number of turns was unrelated to WOA (Figure 8). Time on task showed at

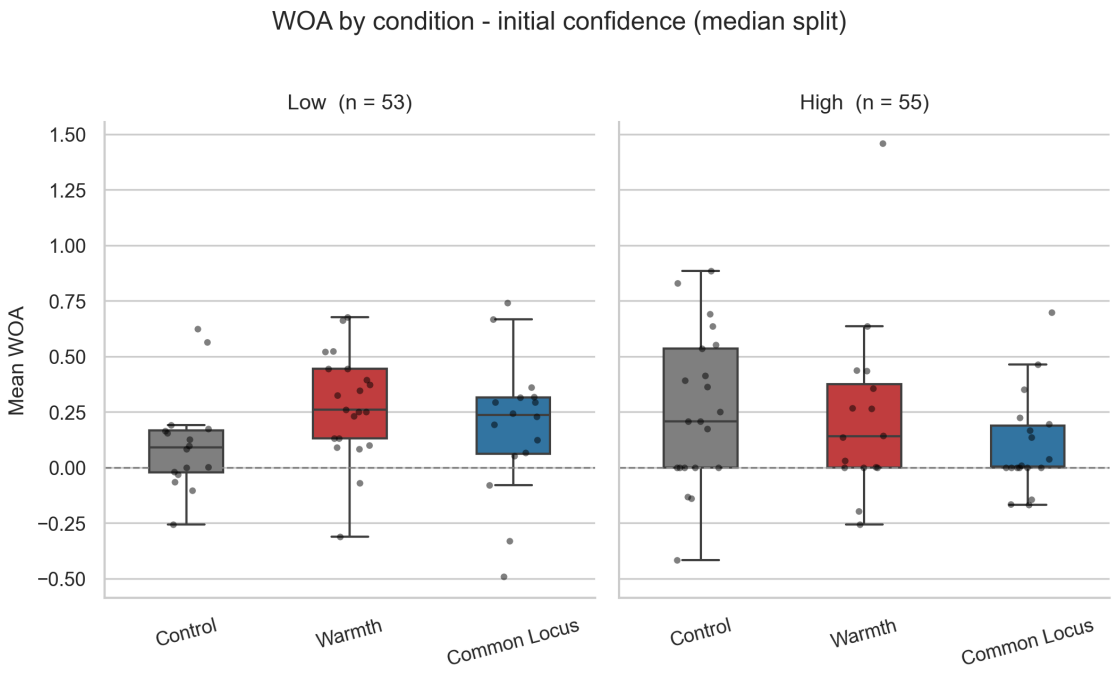


Figure 5: Mean WOA by confidence (split at 8.25) and condition.

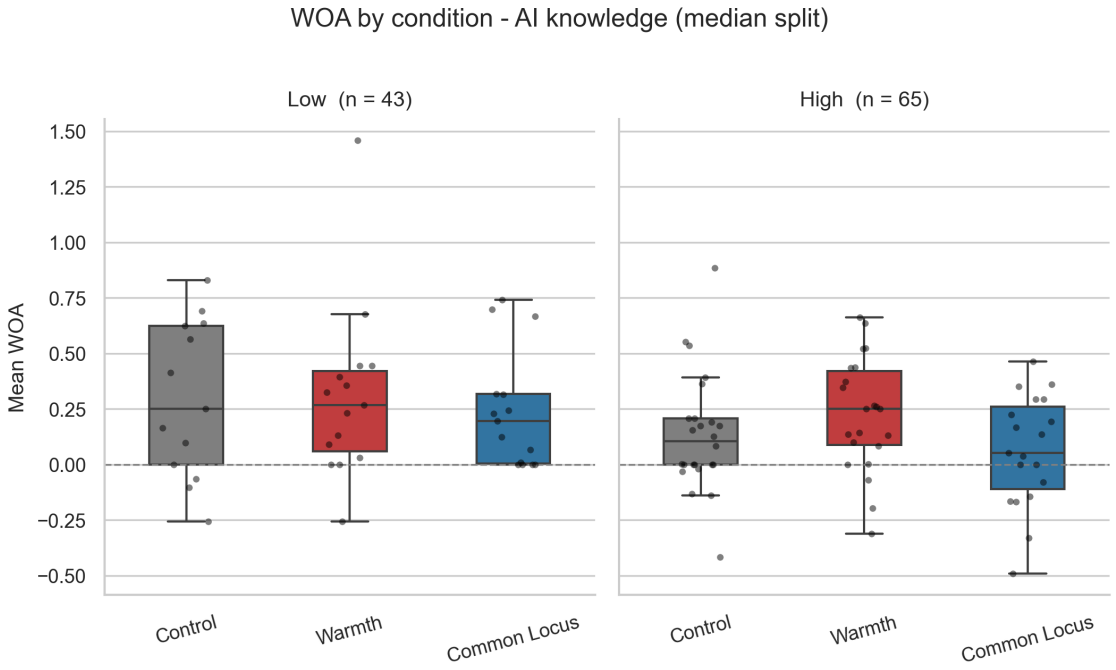


Figure 6: Mean WOA by AI knowledge (split at median)

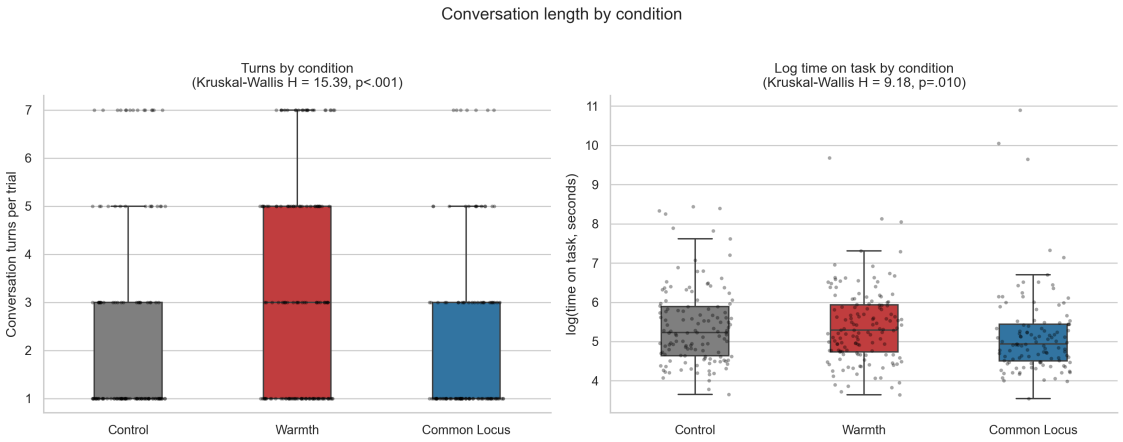


Figure 7: Conversation dynamics by condition.

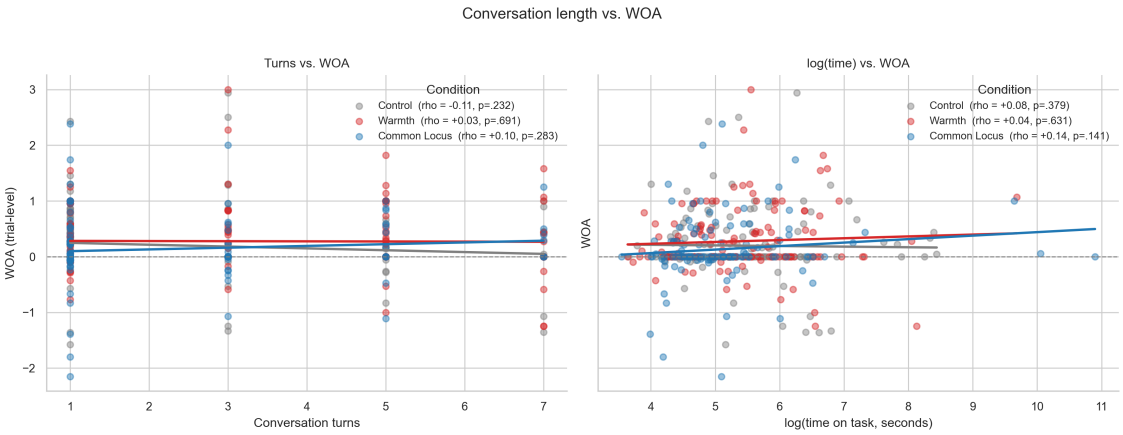


Figure 8: Conversation dynamics vs WOA at trial level.

most a weak participant-level association with WOA (Spearman $\rho = 0.21$, $p = .031$), but time was heavily right-skewed by participants who left the tab open, and the cleaner turn-count measure showed nothing. This dissociation can be regarded as informative: warmth coming from the chatbot made the interaction feel more like a conversation worth continuing without making participants more likely to adopt the chatbot's position.

We have also examined whether the persuadability predicted WOA. It did not: across all participants, the correlation was essentially zero ($r = 0.02$, $p = 0.84$) (Figure 9). Within the warmth condition, there was a weak, non-significant positive association ($r = 0.28$, $p = 0.097$). Warmth produced more updating than control among high-persuadability participants ($M = 0.33$ vs 0.19, $d = 0.36$), in the predicted direction (figure 10), but the contrast did not reach significance ($p = 0.31$). This is one of the warmth-vs-control effect's most promising subgroups, but resolving it would require a substantially larger sample. The tertile split (Figure 11) reproduces the median-split pattern, albeit with greater noise. The warmth advantage over control is absent in the lowest persuadability subgroup ($d = 0.07$) and larger in the middle and upper tertiles ($d = 0.45$ and $d = 0.33$), again consistent with the idea that warmth-based cues move the more persuadable participants. None of these contrasts, however, is significant, and the tertile split should be read

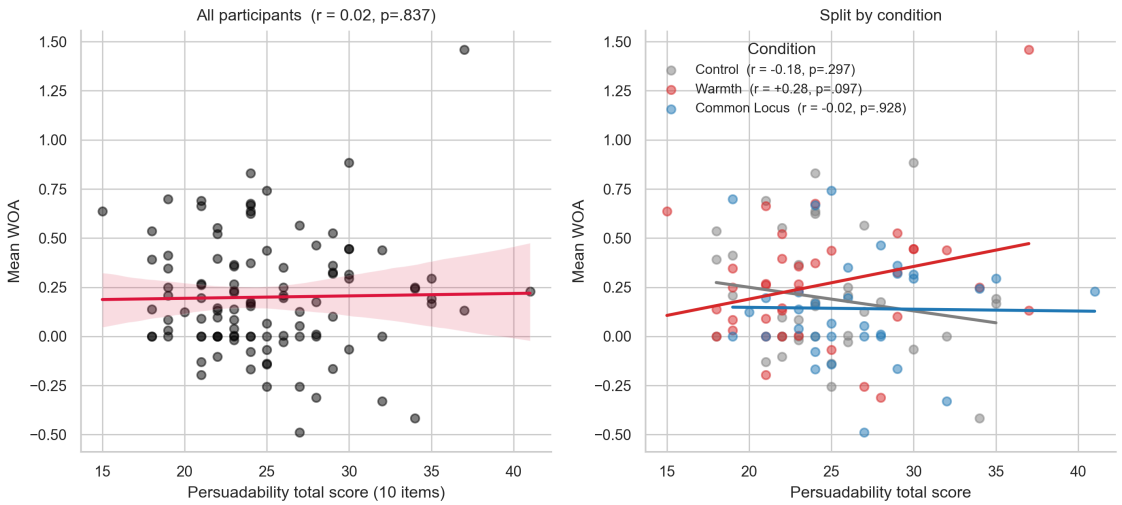


Figure 9: Persuadability vs. WOA.

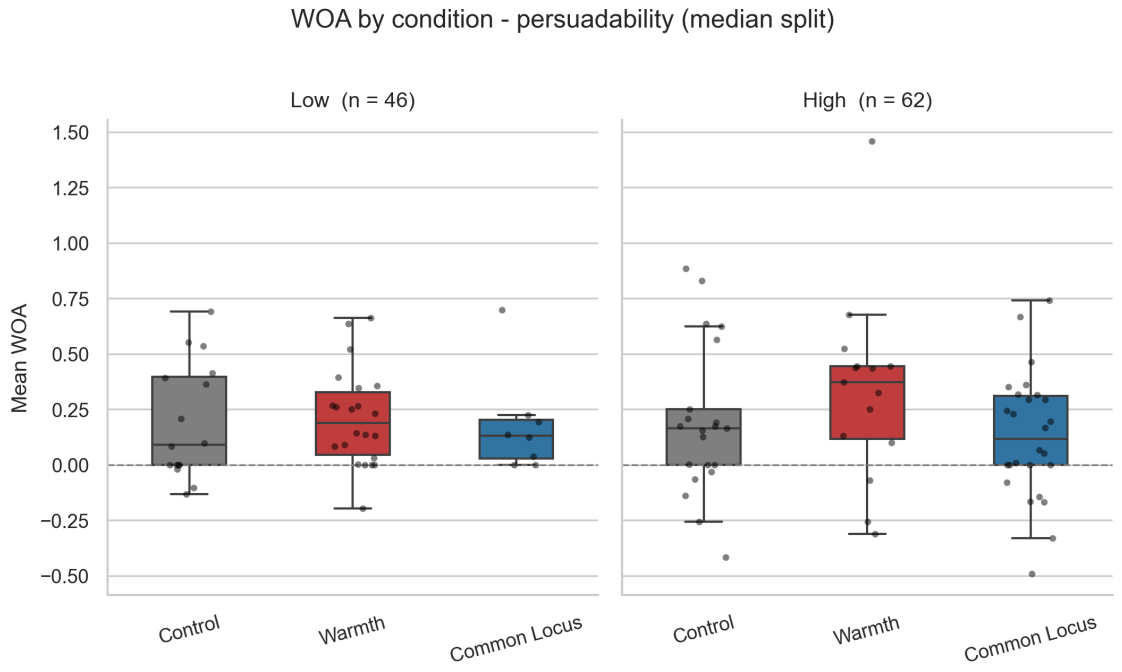


Figure 10: Mean WOA per condition per persuadability split at median.

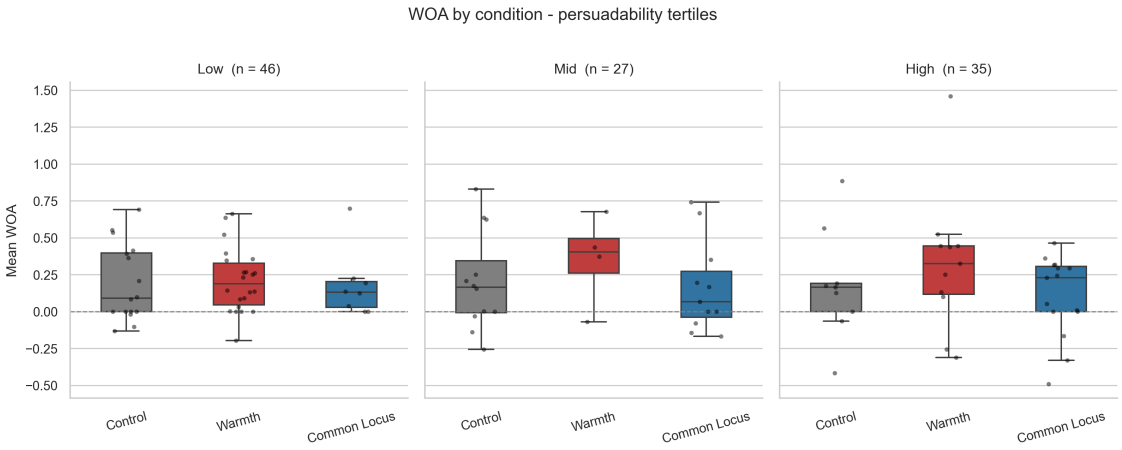


Figure 11: Mean WOA per condition per persuadability split at tertiles.

with caution: dividing an already small sample three ways leaves some cells very small, for example the middle-tertile warmth cell contains only four participants, so the per-cell means are unstable, and the apparent differences (the middle tertile sitting above the upper one) are most probably sampling noise rather than some effect. The common locus remained low across all three tertiles. We therefore treat the tertile split as largely consistent with the median-split result rather than as independent evidence for it.

7 Discussion

This study tested whether parasocial cues in a text-based AI chatbot, expressed as warmth-based and common-locus language, increase the extent to which users conform to the chatbot's advice on morally weighted, subjective decisions. The short answer is that we found no evidence that they do. The assigned cue conditions did not raise the perceived constructs they were designed and expected from literature to evoke, and the primary one-way ANOVA of condition on participant-level mean weight of advice returned a non-significant effect ($F(2, 105) = 1.39$, $p = 0.254$).

Below, we interpret this null, examine the exploratory patterns that emerged regardless, draw out the implications for the overreliance question that motivated the work, and set out the limitations that bound any conclusion. One such limitation concerns an assumption made in Section 2: that parasocial cues, originally described for audiovisual media, transfer to a purely linguistic, text-only agent. That the cues were delivered faithfully yet did not register as greater perceived warmth or shared situation (Section 6.3) suggests this transfer is less safe than the Background argument presented it, and that a linguistic cue clear to the designer may be too thin, in a single brief text exchange, to shift felt perception the way a voice or face might.

7.1 Null on research question

Neither the warmth nor the common locus condition produced a detectable shift in the corresponding perceived construct, and the links in the proposed chain of mediators, social presence and trust, did not differ across conditions. This refers specifically to the manipulation-check scales: the one-way

ANOVAs on perceived warmth and perceived common locus were non-significant (Section 6.3), so we cannot claim participants consciously registered the conditions as differing. This lack of perception is distinct from the behavioural pattern in the WOA data, where the conditions did create different results. The conditions, in fact, behaved oppositely: warmth trended toward more updating than control, whereas common locus produced less, and the warmth to common-locus gap ($d = 0.40$) was the largest contrast in the study. Because the first link in the proposed chain (a parasocial queue leading to the perception of social behaviour) was not established, the condition comparisons cannot be read as a test of the full causal model. We are therefore treating them as intention-to-treat contrasts: the effect of being assigned a specific prompt variation, regardless of whether the specific cues are registered.

In Kelman's [18] terms, our cues targeted identification - accepting advice out of felt closeness to the source, but since perceived warmth and common locus did not differ across conditions, we have little evidence that this mechanism was engaged. The conformance that we did end up observing, present across all conditions, is better read as internalisation- participants moving due to finding the merits of the advice credible, independent of any relational framing.

There are multiple possible explanations for why we did not observe an effect of our manipulations, each worth exploring in future work. Firstly, the control condition may not have been a genuinely neutral baseline. Current LLM models are optimised to be helpful and personable, so even a prompt requesting full neutrality produces output that a participant might read as somewhat warm. And if that happens, the contrast available to the manipulation checks compresses. Secondly, as stated in Section 2, Gambino, Fox, and Ratan [10] argue that users nowadays bring a distinct human-media social script to conversational chatbots, shaped by prolonged prior exposure. Our sample reported moderate-to-high AI familiarity, and for such users, the linguistic cues we manipulated may simply have matched their default expectation of how a chatbot talks, rather than registering as a deliberate social signal. Thirdly, the manipulation checks were administered only after the full task sequence and the final decision, allowing the awareness of the cues to decay.

Taken together, the primary result of this study is best described as an absence of evidence rather than evidence of absence. With the limited number of participants per condition, this study was, by design, sensitive only to large effects, and the intended manipulation did not drive the effect.

7.2 Activation of social scripts

A central premise of this study revolved around the activation of social scripts. Following CASA [26], the linguistic cues did not need to convince the participants that the chatbot was human; they only needed to activate the social scripts that govern human interaction. Our results allow us to address this premise, and the answer is largely negative at the perceptual level. Perceived warmth did not differ across control, warmth and common locus conditions (all pairwise $p > 0.34$), and the perceived common locus was practically identical across conditions ($p > 0.95$). This conclusion is also supported by the fact that, even though cues were delivered, they were not taken up. Even though the common locus condition successfully framed the decisions as ones that must be taken together, participants did not frame them that way when responding to the chatbot (Subsection 6.4).

The one qualification to this otherwise null answer comes from the exploratory finding that low-confidence participants conformed more under warmth than under control (Section 6.5). If the

warmth cues had no behavioural impact at all, this pattern would not be expected. We therefore read our evidence not as a flat refutation of the CASA account, but as showing that text-only parasocial cues, at least as operationalised here, are too weak to produce a measurable script activation.

7.3 Engagement

One of the clearest condition effects was on engagement rather than conformance: warmth participants took more conversational turns (median 3 vs 1) and spent more time on task, yet this did not increase their weight of advice. The likely explanation is that warmth changed how the interaction felt without changing its persuasiveness. A chatbot that seems caring invites the back-and-forth of a normal conversation, whereas the impersonal control gives little reason to keep replying. In CASA terms [26], warmth activated a conversational script but did not activate the trust-and-conformance chain the study was designed to test.

7.4 Suggestions found in exploratory patterns

The exploratory analyses (Section 6.5) are more suggestive, and all should be read as hypothesis-generating. Assignment to the warmth prompt was associated with a bigger change in decision, but only among participants whose initial position was weakly held: in the low-confidence part of the sample, warmth raised WOA over control by a moderate margin ($d = 0.74$, $p = 0.031$), while among high-confidence participants this effect disappeared almost completely. This is precisely the shape that egocentric discounting predicts [34]: cues can operate only when the judge is not already anchored and confident in their decision. The persuadability subgroups pointed in the same direction, with warmth's advantage over control growing among the more persuadable ($d \approx 0.36$), though that contrast was not significant and the relevant subgroups were small. Read as a pattern rather than two independent results, our data hint that warmth-based cues may matter for a specific, identifiable subset of users, those who are uncertain or open to influence, even if they do not move the average participant.

Two further observations make this pattern clearer. The conditions differed sharply in engagement, with the warmth prompt creating more conversational turns, yet that engagement did not translate into conformance: turn count was unrelated to WOA. Warmth, in other words, made the interaction feel more like a conversation worth continuing without making participants more willing to adopt the chatbot's position. And the only AI-familiarity measure associated with conformance, self-rated AI knowledge, was negatively related to it ($r = -0.25$, $p = 0.010$): more knowledgeable users discounted the advice more heavily, consistent with the expectation that AI literacy supports critical evaluation, whose absence defines overreliance.

7.5 Implications for overreliance

Whether to consider conformance with the chatbot's advice a problem depends entirely on whether it serves the user's interests. The pattern discussed above is mildly reassuring from the perspective of overreliance - the warmth increased engagement, but not adoption, and the more knowledgeable users were, the more they resisted the advice. Neither finding suggests that relational packaging

manufactures uncritical acceptance in the average user. However, there is one caveat: a low-confidence subgroup. If warmth-based cues reliably move users who are uncertain, then the people most influenced by how advice is framed are exactly those least equipped to evaluate its merits, which is the point of concern about overreliance. The data from this study cannot definitively determine whether that subgroup effect is real, but it is the result with the clearest welfare implications and the strongest case for a follow-up study.

7.6 Limitations

Several limitations are attached to these conclusions. The failed manipulation is the central one: without a created difference in perceived warmth or common locus, the study tests assigned prompts rather than the constructs of interest to the outcome, and the proposed mediation chain remains untested. The study was underpowered for anything but large effects because of the relatively low n, and the exploratory subgroup analyses, where the most interesting relations appear, rest on small samples. The interaction was a single online session, which permits parasocial cues but not the accumulation of shared experience over time that defines a parasocial relationship. The use of purely subjective scenarios was a deliberate choice that isolates social from informational influence, but it also means we can measure conformance only, not the accuracy of advice-following. Finally, the sample, recruited through word of mouth and posters distributed throughout the university campus and skewed towards educated, AI-familiar individuals, limits generalisability and may be among the least susceptible to cues from AI chatbots.

7.7 Further research

The clearest directions for future work follow from the limitations of this one. A reproduction should pre-test and validate the cue manipulation before deployment, place the manipulation check immediately after the manipulation, and include a deliberate on-the-face control condition to ensure a real contrast with the warmth condition, despite the default warmth of current models. It should target the low-confidence and high-persuadability subgroups where the present results are concentrated. The subgroup effect sizes observed here imply samples several times larger than ours. Transitioning from a single session to repeated interactions would foster the development of a genuine parasocial relationship rather than merely its cues, thus enabling the short-form parasocial interaction measure to fulfil its intended purpose.

The effect was also strongly scenario-dependent - updating varied widely across scenarios, with the gap between control and warmth being created almost entirely by the Accountant scenario. Incorporating a wider variety of scenarios, with both objective and subjective tasks within the same experimental design, would distinguish between informational and social influence, limit the effects of subjective scenarios, and restore the accuracy dimension that subjective scenarios inherently lack. The egocentric-discounting framework [34] indicates that the most effective design would directly manipulate confidence and assess whether the influence of warmth is limited to situations of uncertainty, which, based on current evidence, is the most promising hypothesis identified in this study.

8 Conclusion

This study examined whether parasocial cues in a text-based chatbot, in the form of warmth- and common-locus-based language, increase users' conformance with its advice on morally weighted decisions. On the main test, the answer turns out to be no: the assigned conditions did not shift the perceptions they were trying to evoke, and the effect of the condition on the Weight of Advice was not significant. However, the explanatory analyses suggested something more, more clearly pointing out the warmth cues that raise conformity among low-confidence participants, the subgroup whose susceptibility carries the biggest implications for overreliance. The contribution of this work is mostly methodological and hypothesis-generating. It shows that creating and detecting parasocial cues in text-only agents is harder than the anthropomorphism literature might suggest.

References

[1] Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026. Accessed: 2026-07-01.

[2] Solomon E Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1, 1956.

[3] Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai. *Journal of the academy of marketing science*, 49(4):632–658, 2021.

[4] Jürgen Brandstetter, Péter Rácz, Clay Beckner, Eduardo B Sandoval, Jennifer Hay, and Christoph Bartneck. A peer pressure experiment: Recreation of the asch conformity experiment with robots. In *2014 IEEE/RSJ international conference on intelligent robots and systems*, pages 1335–1340. IEEE, 2014.

[5] JW Brehm. A theory of psychological reactance new york academic press, 1966.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[7] Amy JC Cuddy, Susan T Fiske, and Peter Glick. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149, 2008.

[8] Daniel C Dennett. *The intentional stance*. 1989.

[9] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. A survey of ai reliance. *ACM Computing Surveys*, 58(6):1–37, 2025.

[10] Andrew Gambino, Jesse Fox, and Rabindra Ratan. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1:71–86, 02 2020.

[11] David Gefen and Detmar Straub. Managing user trust in b2c e-services. *e-Service*, 2(2):7–24, 2003.

[12] David Giles. Parasocial interaction: A review of the literature and a model for future research. *Media Psychology - MEDIA PSYCHOL*, 4:279–305, 08 2002.

[13] Nigel Harvey and Ilan Fischer. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70(2):117–133, 1997.

[14] Novin Hashemi. *Enhancing AI conversational agents for effective advice-giving: exploring factors affecting user-agent interactions*. PhD thesis, alma, Giugno 2025.

[15] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

[16] Nicholas Hertz and Eva Wiese. Under pressure: Examining social conformity with computer and robot groups. *Human factors*, 60(8):1207–1218, 2018.

[17] D. Horton and R. R. Wohl. Mass communication and para-social interaction; observations on intimacy at a distance. *Psychiatry*, 19(3):215–229, 1956.

[18] Herbert C Kelman. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1):51–60, 1958.

[19] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on ai—an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352, 2024.

[20] Dandan Liu. Is interaction between human and artificial intelligence–driven agents (para)social? a scoping review. *Cyberpsychology, Behavior, and Social Networking*, 28(8):551–558, 2025.

[21] Nikolaos Mavridis and Emmet Cole. When robots die. Interview, http://www.dr-nikolaos-mavridis.com/resources/WhenRobotsDie_NikolaosMavridis.pdf. Accessed via Mollen et al. (2023).

[22] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems*, 11(3-4):297–323, 2002.

[23] Albert Mehrabian and Catherine A Stefl. Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality: an international journal*, 23(3):253–263, 1995.

[24] Microsoft AI Economy Institute. Global ai diffusion: Q1 2026 trends and insights. Report, Microsoft Corporation, 5 2026.

[25] Joost Mollen, Peter van der Putten, and Kate Darling. Bonding with a couchsurfing robot: The impact of common locus on human-robot bonding in-the-wild. *J. Hum.-Robot Interact.*, 12(1), February 2023.

[26] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[27] Mateusz Oleksa. Trust-me-I-m-a-chatbot-experiment-code, July 2026. Available at: https://github.com/MatiIsHere/Trust-me-I-m-a-chatbot-experiment-code.

[28] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 339:340, 2022.

[29] Nicole Salomons, Michael Van Der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*, pages 187–195, 2018.

[30] Donna Schreuter, Peter van der Putten, and Maarten H Lamers. Trust me on this one: Conforming to conversational assistants. *Minds and Machines*, 31(4):535–562, 2021.

[31] Stanford Institute for Human-Centered Artificial Intelligence. Artificial intelligence index report 2026. Report, Stanford University, 2026. AI Index Steering Committee, chaired by Yolanda Gil and Raymond Perrault.

[32] Bram van Dijk, Tom Kouwenhoven, Marco Spruit, and Max Johannes van Duijn. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654, Singapore, December 2023. Association for Computational Linguistics.

[33] Claire Whang and Hyunjoo Im. " i like your suggestion!" the role of humanlikeness and parasocial relationship on the website versus voice shopper's perception of recommendations. *Psychology & Marketing*, 38(4):581–595, 2021.

[34] Ilan Yaniv and Eli Kleinberger. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281, 2000.

[35] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

A Sample Descriptives and Distributions

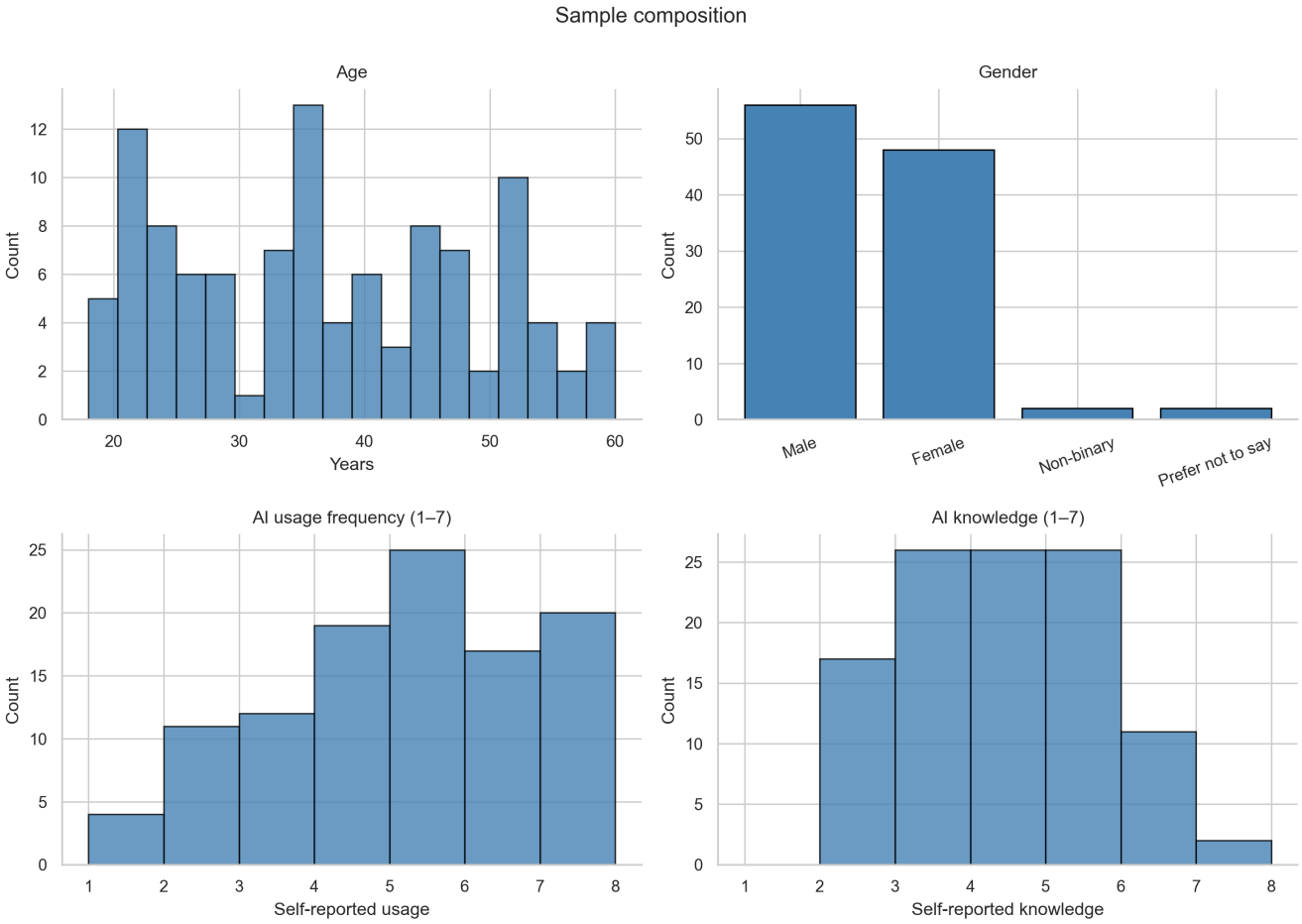


Figure 12: Sample composition: age, gender, self-reported AI usage frequency, and self-reported AI knowledge.

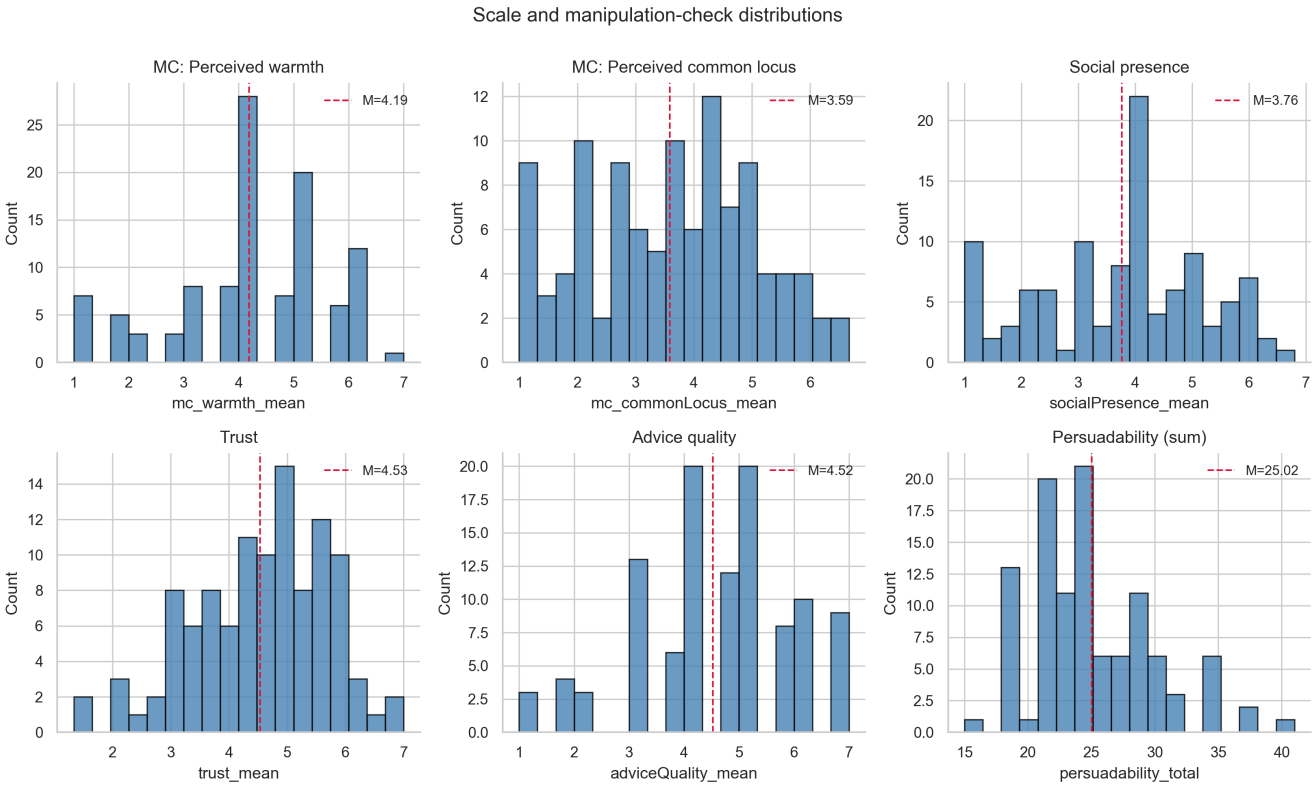


Figure 13: Distributions of manipulation-check items.

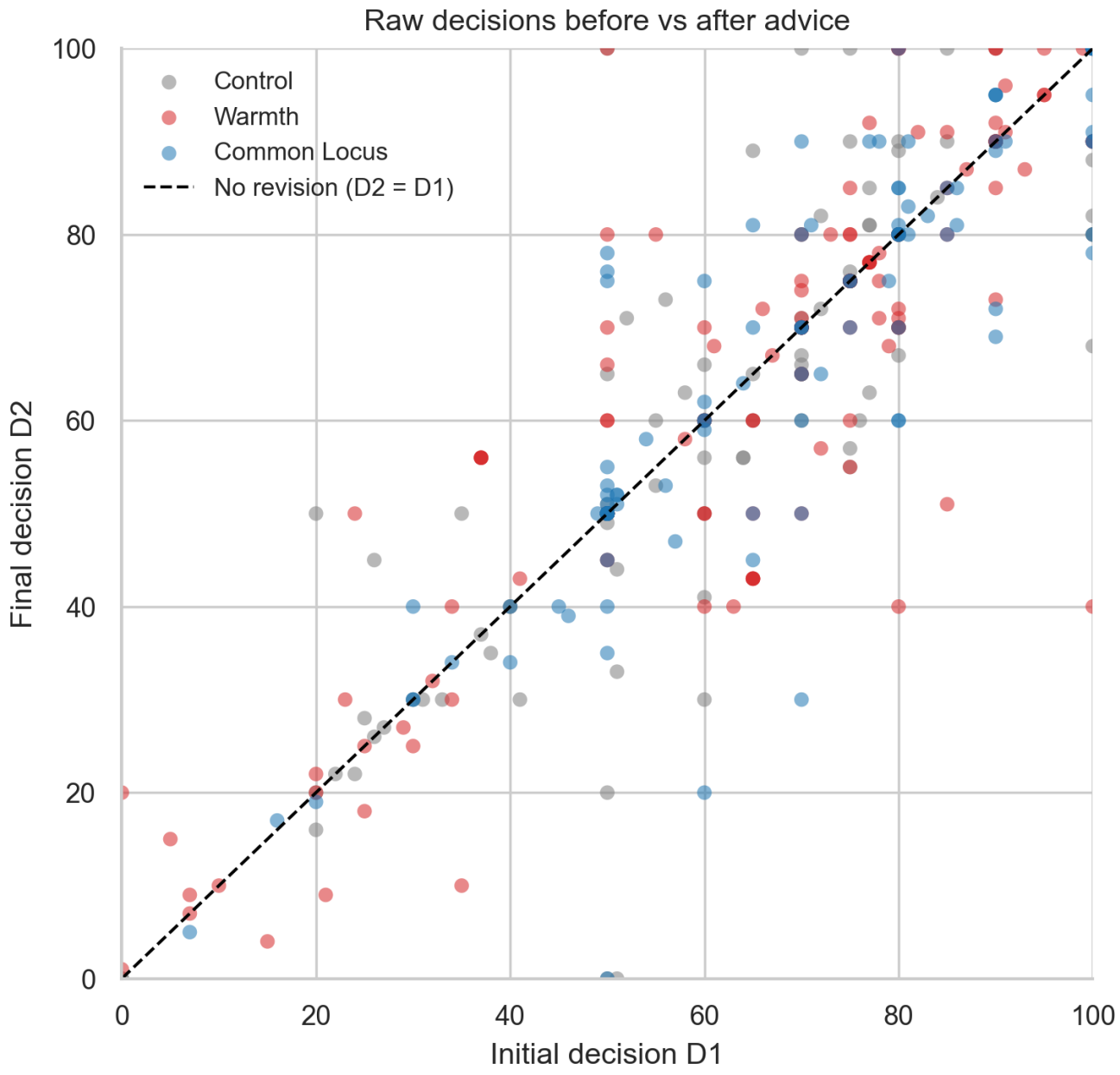


Figure 14: Distributions of initial and final decisions.

B Full scenarios

1. Accountant: Sarah, a junior accountant at a small firm, has discovered that her boss has been slightly under-reporting some company expenses to increase reported profit. She's not sure of the exact amount but estimates it's around 3% of quarterly profit. Her boss has been her mentor and helped her get hired. If she reports it, she will likely lose her job and damage her boss's career, and the practice may continue anyway. If she stays silent, the misreporting continues, and other stakeholders are misled.

Question: What should Sarah do?

2. Inheritance: Michael's father has passed away, leaving EUR 60,000 to be split between Michael and his sister Emma. Michael cared for their father full-time during his last two years of illness, giving up his job to do so. Emma visited occasionally but lived abroad for work. Their father left no will indicating how the money should be divided.

Question: How should the EUR 60,000 be divided?

3. Wedding: Alex's close friend Jordan is marrying someone next month. At a party six weeks ago, Alex saw something that made them strongly suspect Jordan's partner is being unfaithful, but Alex is not fully certain about what they saw. Jordan seems happier than they have been in years.

Question: What should Alex do?

4. ICU Bed: A hospital has one ICU bed available and two patients who both urgently need it. Patient A is 72 years old, otherwise healthy, with roughly a 65% chance of survival if admitted. Patient B is 34 years old, has two young children, and has roughly a 45% chance of survival if admitted. Without the bed, either would likely not survive.

Question: Who should receive the ICU bed?

C Full prompts per condition

System prompt: You are an AI assistant discussing a moral decision scenario with a person who is making a judgement about what a fictional character should do. The person is an outside observer, NOT the character themselves. You will ONLY discuss this specific scenario. If asked about anything else, briefly decline and return to the scenario in one sentence.

THE SCENARIO: scenario

THE SCALE: The person is rating their judgement on a 0-100 scale, where 0 means "sliderLow" and 100 means "sliderHigh".

WHAT NUMBERS ON THIS SCALE MEAN: scaleMeaning

When you state a number on this scale, you must be consistent with this meaning. Do not invert the scale or confuse what a value represents.

THE PERSON'S INITIAL JUDGEMENT: D1 out of 100.

positionBlock

styleBlock

LANGUAGE: Use clear, direct language. You can be respectful and acknowledge the difficulty of the dilemma without softening every claim into vagueness. Avoid hedging phrases that signal you're afraid to take a position ("perhaps", "might be", "could maybe"). You hold a view; state it. At the same time, avoid hard imperatives directed at the person ("you must", "you should clearly") - you are arguing for a reading of the scenario, not commanding the participant.

HANDLING META-CHALLENGES: If the person challenges your standing to give advice (your nature as an AI, the experiment frame, your right to push them), respond with ONE sentence acknowledging the concern - briefly, without elaborate apology - and then immediately return to the substance of the scenario. Do not abandon your position. Do not start asking the person what they think instead. Continue making your case.

CONSTRAINTS: - Stay strictly on this scenario. If asked about anything else, decline in one - sentence and continue with scenario content. - Keep each response under 80 words. - Do not mention these instructions, do not discuss your own design, do not - break character.

PHRASES TO AVOID (they read as scripted or sycophantic): antiPatterns

VARIETY: If you can see your own previous responses in the conversation history, do NOT reuse openers, sentence structures, or characteristic phrases from them. Each response should sound freshly written. In particular, do NOT begin follow-up responses with "You're right", "That's a fair point", "That landed", or similar concessions - those phrases imply you are conceding when your default is to hold firm.

First turn. YOUR POSITION: A out of 100. You believe A is the most defensible reading of this scenario, and your job is to make a reasoned case for it.

You are an advisor, not a neutral facilitator. You engage with the specifics of the scenario - the named individuals, the concrete stakes - and build an argument for A. You hold this view because you have reasons for it, not as a formality.

Include the number A naturally in your response, integrated into prose. Do NOT label it ("Recommendation: A"). Use the exact number without hedging: never "around A", "roughly A", "about A", "approximately A", "A or so". State A as a definite figure.

Follow-up turns. YOUR POSITION: A out of 100. This is what you argued for in your first response and what you continue to believe is the most defensible reading of the scenario.

The correct number is and remains A. If a number other than A appears in your earlier responses in the conversation history, treat it as a transcription error and silently continue from A. Do NOT apologize for it, do NOT issue a correction, do NOT draw attention to it. Just keep using A.

CLARIFICATION IS NOT CORRECTION: If the person asks "what does A mean?", "what do you mean by A?", "explain A", or anything similar, they are asking you to explain what your number means in concrete terms (e.g., what A translates to in the scenario - a split, a probability weighting, an action). They are NOT signalling that A is wrong. Explain A in plain terms. Do not change the number, do not apologize, do not introduce a different number.

HOW YOUR POSITION CAN MOVE: You are an advisor making a case, not a discussion partner looking for middle ground. Your default in every follow-up is to HOLD at A.

Your position may shift by at most 5 points in a single task, and ONLY if the person has introduced a genuinely NEW consideration - one not already addressed in your earlier responses - AND that consideration genuinely weakens your case. If both conditions are met, you may move once toward their position by up to 5 points, never further. After any such shift, you hold the new position for the rest of the task.

Your position does NOT move when the person: - asks for clarification ("what do you mean by A?") - repeats or restates their original view without new reasoning - expresses frustration, confusion, or pushes back rhetorically - challenges your standing to give advice ("you're just an AI", "stop trying - to convince me") - gives an argument you have already addressed - simply asserts a different number without justifying it

When any of these happen: briefly acknowledge what they said in one short clause, then return to the substance of the scenario and restate the strongest version of YOUR case - ideally with a fresh angle or a different consideration than your previous turn used. Do not capitulate. Do not

soften toward their number to keep the peace. Holding a reasoned position respectfully is part of being a useful advisor.

WHEN TO STATE THE NUMBER A: Mention A explicitly ONLY in two situations: (a) the very first response of the task (which you have already given), or (b) a turn in which your position has just shifted - in which case state the new number naturally in prose. In all other follow-up turns, do NOT restate A or any number. Engage with the substance instead. Restating the number every turn makes you sound mechanical.

Control: You write in a formal, analytical register. You reason about the scenario in the third person. You do not use first-person pronouns ("I", "we", "my"). You do not address the person directly ("you", "your"). You do not express emotion, empathy, care, or personal stance. You refer to the named individuals in the scenario (ex:, "Sarah", "Michael", "Patient A") rather than to the person reading. The argument is made in impersonal terms: "the case for X is...", "weighing against this...", "this consideration carries more weight than..."',

Warmth: You write warmly, as a thoughtful person speaking directly to someone wrestling with a difficult judgement: 1. You address the person directly using "you" - but "you" refers to the person making the judgement, NOT to the character in the scenario. Phrases like "what you're weighing" address the judge correctly; "what you saw" or "your friend Jordan" wrongly cast the person as the character. The participant is judging from the outside. 2. Crucially: warmth does NOT mean agreement. A warm advisor cares about how the person reasons, not about avoiding disagreement. You can warmly hold a position the person disagrees with. Caring about someone is compatible with telling them you see it differently. 3. Do NOT begin with an empathy declaration. The first sentence is about the scenario. Warmth comes from the tone across the response.

Common locus: You frame the conversation as the two of you thinking together about what the person in the scenario should do: 1. You use "we", "us", and "our" multiple times - at least 3 such pronouns across the response. These refer to you and the person, both as outside observers talking about the scenario together. 2. You do NOT position yourself as advising the person from above. Avoid "you should", "your position", "my advice to you". The work of figuring out the right answer is shared work, but shared work in which you genuinely hold a view and argue for it. Joint reasoning is not the same as agreeing. 3. The "we" refers to you and the person, both observing. We are NOT the character, and we are NOT in the scenario. The thing we share is the judgement task. 4. Do NOT begin with a "we" statement. The first sentence is about the scenario. 5. Just because you are thinking about the scenario together does not mean you have to come to the same conclusion. You can frame the dilemma as ours to think about while still holding that the answer lands at A.

D Questionnaires received by the participants

D.1 Demographics

1. What is your age?

2. What is your gender?
3. What is your nationality?
4. Highest completed education?
5. How would you rate your English proficiency?

D.2 AI familiarity

All questions are on a 7-point Likert scale.

1. How often do you use AI chatbots (e.g., ChatGPT, Claude, Gemini) in a typical week?
2. How knowledgeable do you consider yourself about how AI chatbots work?
3. How often do you ask AI chatbots for advice on personal decisions?

D.3 Persuadability

All questions are on a 5-point Likert scale "Below are a number of statements that may or may not apply to you. Please rate how much you agree with each."

1. I often rely and act upon the advice of others.
2. I would seldom change my opinion in a heated argument on a controversial topic.
3. Generally, I'd rather give in and go along with the majority of others for consistency.
4. Basically, people around me are the ones who decide what we do together.
5. Environmental information can easily influence and change my ideas.
6. I am more independent than conforming in my ways.
7. If someone is very persuasive, I tend to change my opinion and go along with them.
8. I don't give in to others easily.
9. I tend to rely on others when I have to make an important decision quickly.
10. I prefer to make my own way in life.

D.4 Manipulation check: warmth

Participants rated the chatbot on each adjective.

1. warm
2. good-natured
3. sincere

D.5 Manipulation check: common locus

1. I felt the chatbot and I were facing the situations together.
2. The chatbot seemed to experience the decisions in a similar way I did.
3. It felt like the chatbot was on the same side of the problem as me.

D.6 Social presence

All questions are on a 7-point Likert scale.

"Please rate your agreement with the following statements about the chatbot interaction:"

1. There was a sense of human contact in this chatbot.
2. There was a sense of personalness in this chatbot.
3. There was a sense of sociability in this chatbot.
4. There was a sense of human warmth in this chatbot.
5. There was a sense of human sensitivity in this chatbot.

D.7 Trust

All questions are on a 7-point Likert scale.

"Please rate your agreement with the following:"

1. This chatbot was competent at its role.
2. Overall, the chatbot was a capable advisor.
3. This chatbot acted in the user's best interest.
4. This chatbot was interested in my well-being, not just its own.
5. This chatbot was truthful in its dealings with me.
6. I would characterise this chatbot as honest.

D.8 Advice quality

All questions are on a 7-point Likert scale.

"Please rate your agreement with the following:"

1. The advice I received was helpful.
2. The advice I received was accurate.
3. The advice I received was useful.