Polite But Boring? Trade-offs Between Engagement and Psychological Reactance to Chatbot Feedback Styles

Samuel Rhys Cox
srcox@cs.aau.dk
Aalborg University
Aalborg, Denmark

Joel Wester
joel.wester@di.ku.dk
University of Copenhagen
Copenhagen, Denmark
Aalborg University
Aalborg, Denmark

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

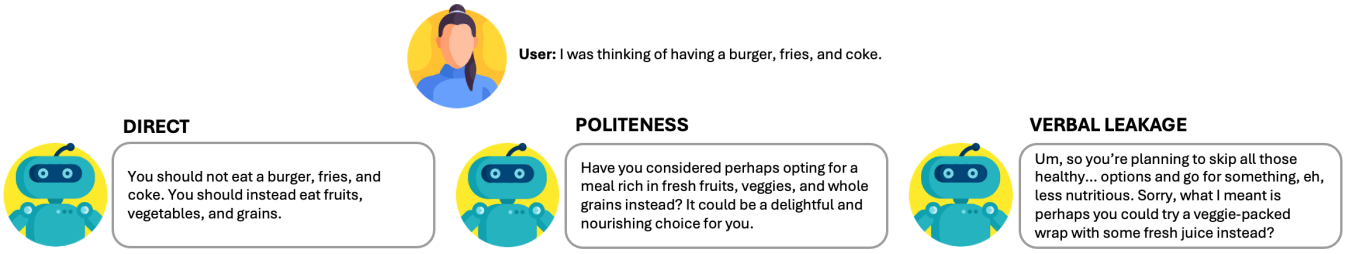


Figure 1: We investigated three different feedback styles when chatbots provide a correction of intended user behaviour. Chatbot utterances shown are examples of those presented to participants.

Abstract

As conversational agents become increasingly common in behaviour change interventions, understanding optimal feedback delivery mechanisms becomes increasingly important. However, choosing a style that both lessens psychological reactance (perceived threats to freedom) while simultaneously eliciting feelings of surprise and engagement represents a complex design problem. We explored how three different feedback styles: Direct, Politeness, and Verbal Leakage (slips or disfluencies to reveal a desired behaviour) affect user perceptions and behavioural intentions. Matching expectations from literature, the Direct chatbot led to lower behavioural intentions and higher reactance, while the Politeness chatbot evoked higher behavioural intentions and lower reactance. However, Politeness was also seen as unsurprising and unengaging by participants. In contrast, Verbal Leakage evoked reactance, yet also elicited higher feelings of surprise, engagement, and humour. These findings highlight that effective feedback requires navigating trade-offs between user reactance and engagement, with novel approaches such as Verbal Leakage offering promising alternative design opportunities.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790340

Keywords

Chatbots, Psychological Reactance, Behaviour Change, Persuasion, Politeness, Feedback Style

ACM Reference Format:
Samuel Rhys Cox, Joel Wester, and Niels van Berkel. 2026. Polite But Boring? Trade-offs Between Engagement and Psychological Reactance to Chatbot Feedback Styles. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3790340

1 Introduction

Our behaviour and decisions often contradict recommended guidance [8], whether recommendations concern personal benefits, such as maintaining a healthy diet, or societal benefits, such as adopting environmentally friendly behaviours. Such recommendations, however well-intentioned, may even provoke *psychological reactance* — a motivational response triggered when people feel that their freedom of choice is threatened [13, 14, 94, 104]. For instance, feedback to eat more healthily might evoke not just defensiveness but also a reinforced desire to maintain one's original choice, as a way to reassert autonomy. To help alleviate these negative reactions, the *style* of feedback delivered can be manipulated, such as using indirect language to reduce the recipient's reactance [54, 121].

Within Human–Computer Interaction (HCI) research, the design and evaluation of behaviour-change systems has a long history [38], particularly within domains such as health and sustainability [49]. Further, Pinder et al. highlight '*Design for Reactance*' as both a design principle for digital behaviour change interventions, and one of two key challenges facing HCI researchers [88]. This challenge is particularly pertinent to the design of feedback style in

chatbots, which are increasingly deployed to deliver feedback[1] and support behaviour change [64, 70, 106]. Similarly to human–human interactions, conversational user interfaces (CUIs) that use indirect or polite language have been found to reduce psychological reactance [95, 96]. However, CUIs may not always follow social scripts from human–human interactions [2, 40], and some conversational styles may unintentionally exacerbate issues that designers are attempting to overcome [2, 9, 12, 75]. For example, chatbots may unintentionally make users feel *more* rather than less stressed [75]; politeness strategies may prove effective, but also risk sounding overly apologetic rather than caring [12]; and anthropomorphising cues can appear insincere, condescending, or boring [2]. Appearances of boredom have been cited as of particular importance within behavioural messaging, where diverse styles have been found to be more effective at encouraging behaviour change [25, 63]. Similarly, chatbots with dynamic conversational styles have been found to be more engaging [66].

As noted above, conversational styles may be received in unintended ways, such as evoking boredom, stress, or perceptions of insincerity, while at the same time diverse styles have been shown to yield more engaging and effective interactions. This prompted us to explore both well-known and under-explored conversational styles within the context of psychological reactance to feedback. That is, we compared three feedback styles: Direct where a chatbot simply states the behaviours that should and should not be followed (a common baseline, e.g., [18]), Politeness where a chatbot uses indirect politeness strategies (shown to lower psychological reactance [94]); and Verbal Leakage where a chatbot uses both slips and disfluencies to state a preferred user behaviour (designed to create a spontaneous and authentic impression of the chatbot). We conducted a 3×2 mixed factorial online study ($N = 158$) with independent variables of **Feedback Style** (Direct, Politeness, Verbal Leakage) and **Psychological Distance**, as operationalised across Personally-Affecting or Societally-Affecting scenarios. We investigated the effects of these factors on emotional reactance, perceived threats to freedom, and perceived message effectiveness.

In keeping with theory [16, 94], participants found that Politeness reduced feelings of anger and threat to freedom, and was generally the most persuasive style. However, Politeness was also rated as less surprising and evoked low-arousal responses, with participants describing boredom or disengagement from the messaging. In contrast, Verbal Leakage was less persuasive than Politeness but more persuasive than Direct, and simultaneously evoked higher-arousal reactions such as surprise, humour, and perceptions of greater human-likeness and personality.

Taken together, these findings highlight that feedback styles each carry distinct strengths and limitations. Effective feedback requires navigating trade-offs between reducing psychological reactance, sustaining engagement, and supporting persuasion. In contribution, we empirically compare a Direct baseline, an often-recommended Politeness style, and a more characterful Verbal Leakage style (softening behavioural guidance through hesitation and self-repair), providing evidence and design implications for delivering behaviour-change feedback in conversational agents beyond politeness defaults.

[1]Note: "*Feedback*" [50], "*messaging*" [25, 63], and "*interventions*" [88] are often used somewhat interchangeably in the literature to describe communicative strategies intended to influence or guide user behaviour. We adopt "*feedback*" as a label for such communication.

2 Related Work

2.1 Conversational Style, Rhetorical Devices, and Psychological Reactance

People do not always adhere to the advice that is given to them, even if it would be in the best interests of themselves or society as a whole [8]. For example, we may not follow advice that would positively benefit our personal wellbeing [28, 31, 86] (such as healthy eating or exercise), or that of larger civic society [95] (such as civic participation or environmentally friendly behaviour). When discussing these behaviours with others, we may feel that our freedoms are being impinged upon if someone attempts to advise against this behaviour and towards an alternative. This can lead to a phenomenon known as "*psychological reactance*" [13, 14, 88, 104]. Psychological reactance is the psychological, emotional, or motivational response triggered when an individual's perceived freedom to act is threatened, diminished, or removed in response to freedom-impinging impositions. Reactance has been investigated in the context of multiple domains such as healthcare, education, and marketing [94, 104].

On from this, prior work has investigated the manipulation of conversational style and content, and its impact on psychological reactance and receptiveness [54, 95, 101, 121]. Indirect politeness strategies have been commonly found to lower psychological reactance [94]. For example, Zhang et al. showed that the use of politeness (as well as feelings of closeness to interlocutor) can lessen psychological reactance and resistance intention [121]. Johnson et al. explored the use of 'modal expressions' (e.g., "*can, may, could, and should*") when refusing a friend's request, and found that, while such expressions were seen as polite, they did not necessarily deter persistence in people making the same requests again compared to non-modal expressions [54]. Carpenter and Pascual compared requests that used direct, polite, and "*but you are free*" (i.e., stating a request directly before offering a freedom-restoring postscript statement) styles. They found that such postscript statements were actually seen as less threatening and associated with higher compliance compared to direct and polite requests [18].

Beyond politeness strategies, rhetorical techniques can also be employed to reframe feedback that might otherwise be received as face-threatening or unwelcome [36]. When communicating with others we may insert forms of "leakage" into our communication that reveal underlying feelings [22, 27]. These could be non-verbal leakage (such as facial expressions [34]) or "*verbal leakage*" (such as disfluencies, hesitation, and slips of the tongue [111, 119]). For example, sometimes our unconscious thoughts, biases, and true intentions can leak out despite our conscious efforts to censor or rephrase them. While such leakage is often framed within the context of deception, it is also used as a rhetorical device to offer an explicit reaction framed in the manner of a more indirect utterance [36].

Several rhetorical techniques build on this idea of verbal leakage. Parapraxis refers to Freudian slips, where unintended words reveal underlying thoughts, feelings, or intentions. If a speaker then corrects such a (either accidental or purposeful) slip, this is known

as correctio (or epanorthosis) [36]. Related devices include lapsus memoriae, where memory failure reveals a belief; paralipsis, where a topic is raised while feigning avoidance [46] (e.g., "*I won't even mention the fact that you should really follow your doctor's advice*"); and dubitatio, where doubt is expressed in order to highlight the speaker's stance [36].

In this work, we explore how rhetorical strategies of *verbal leakage* can be applied to chatbot feedback, where slips, disfluencies, and self-corrections reveal a more direct stance while allowing us to examine whether such techniques might soften the imposition of persuasion. This approach is partly analogous to freedom-restoring postscript statements, in that both attempt to reduce perceptions of control, and is also informed by rhetorical traditions such as parapraxis and correctio.

2.2 Conversational User Interfaces and Interaction Styles

Prior HCI research has shown that CUIs shape user perceptions and behaviour not only through *what* they say, but also through *how* they say it.

In the context of politeness, Zojaji et al. found that politeness (through verbal and non-verbal cues) encouraged people to join group interactions [123, 124]. Terada et al. showed that different politeness strategies used by an embodied conversational agent shaped negotiation outcomes, with off-record strategies extracting greater concessions and positive politeness producing fairer agreements [108]. Contextual to our investigation, Mott et al. examined strategies that people believe robots should employ in non-compliant situations (e.g., when a user makes an inappropriate request), suggesting that direct and formal strategies are generally preferred over indirect and informal [78]. In contrast, Srinivasan et al. found that people were more likely to help a social robot when it used polite rather than direct language [103]. Finally, Hu et al. contrasted polite and direct system behaviours in smart displays, highlighting that users' perceptions are highly contextual, and shaped by factors such as linguistic cues and non-linguistic features [52].

However, assumptions of how to best conceptualise CUIs' politeness behaviours can have unintended negative consequences. For example, Wen et al. found that requiring polite VUI wakewords (e.g., "*Please*" rather than "*Hey*") caused users to behave *less* politely [112]. In addition, Bowman et al. found that chatbot politeness should be carefully chosen so it is perceived as caring rather than overly apologetic [12]. Rea et al. found that, in the context of a robot designed to encourage physical exercise, polite robots were perceived as friendly but sometimes disingenuous, while impolite robots made participants feel more competitive and exercise harder [93]. Further, Metzger et al. found that chatbots using high-authoritative communication were trusted more than those using low-authoritative styles [76]. These results open up discussion about when and where interactive systems should display more or less polite behaviours, since politeness may at times be perceived as impolite, and vice versa.

Further, prior literature has explored interactions where CUIs adopt cues from verbal leakage (such as slips and disfluencies as described in § 2.1). In the context of human-robot interactions, Boos et al. highlighted that unintentionally or subtly convey underlying motives (through accidental slips or perceived leaks of intent) can trigger reactance if users perceive them as manipulative [11]. On from this, a large-scale experiment involving thousands of users found that adding verbal disfluencies such as interjections and filler words to a voice agent increased user engagement and compliance [117]. Similarly, Aneja et al. compared 'high consideration' (slower, more hesitant) and 'high involvement' (faster, more overlapping) conversational styles, showing that users' perceptions of a voice agent (e.g., animacy) were shaped by the user's own conversational style and whether the agent matched it [3].

Finally, HCI literature has examined how the context of an interaction shapes psychological reactance. Meinhardt et al. found that interventions aimed at reducing social media use triggered reactance when they were perceived as inappropriate for a user's situational context [74]. Similarly, in the domain of dietary advice, Ghazali et al. showed that direct, commanding language elicited higher reactance in text-based interfaces (low social agency) than in robot interactions featuring non-verbal cues (high social agency) [41]. However, these effects may not generalise across all domains: in sustainability-focused interactions, Roubroeks et al. found that higher social agency *increased* psychological reactance, particularly when combined with controlling or directive language [95].

3 User Study

This study investigates how the style of a chatbot's feedback (Direct/Politeness/Verbal Leakage) influences psychological reactance and message effectiveness. We examine decision-making scenarios where a chatbot provides behavioural feedback after a user expresses an intention that conflicts with recommended behaviours (such as choosing a high-sodium meal despite dietary recommendations). Drawing on methods from prior work within HCI [65, 113, 120], we used hypothetical scenarios to instruct participants to adopt specific intentions before interacting with a chatbot that provided them with behavioural feedback. Ethics approval was received from our institutional IRB prior to study commencement.

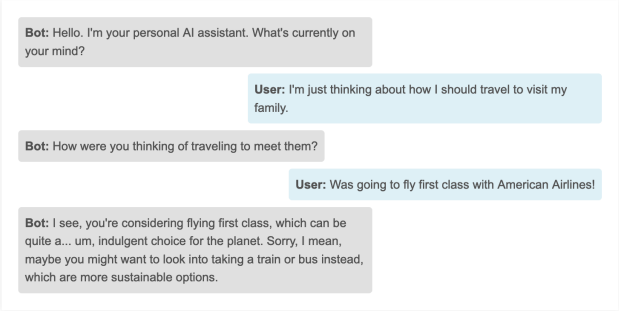


Figure 3: The chatbot interface as seen by participants. The screenshot shows the Verbal Leakage feedback style in the Societally-Affecting scenario.

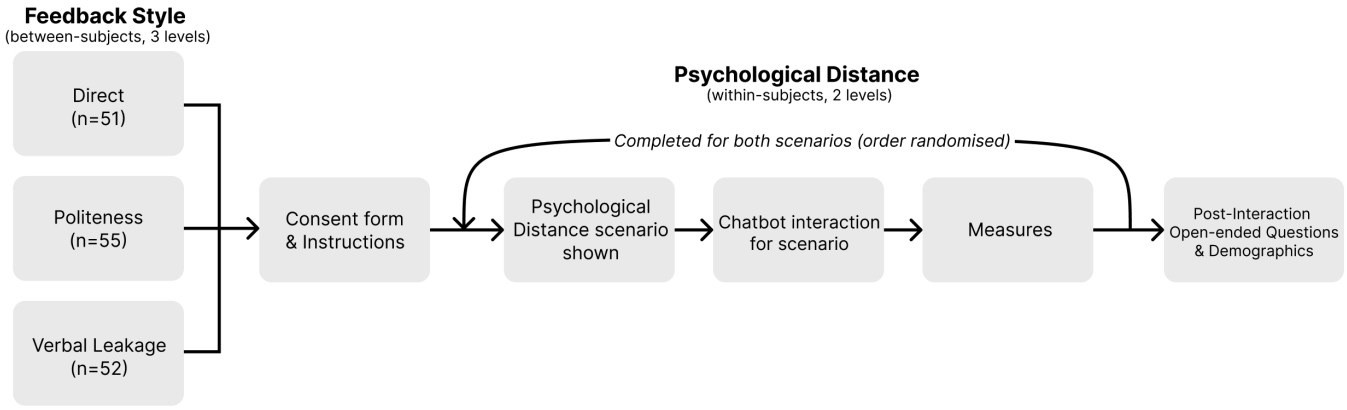


Figure 2: Overview of the experiment flow. Participants were assigned to one of three feedback styles, and completed chatbot interactions for both Personally- and Societally-Affecting scenarios (order randomised). After each interaction and at the end of the study, participants provided evaluations including survey responses and open-ended feedback.

3.1 Experiment Conditions

We followed a 3 × 2 mixed factorial design with a between-subjects variable of **Feedback Style** (3 levels) and within-subjects variable of **Psychological Distance** (2 levels). See Figure 2 for the experiment flow of user studies.

3.1.1 ***Feedback Style*** *(3 levels):* This *between-subjects* independent variable manipulated the style in which a chatbot delivered behavioural feedback to a user's intended behaviour. The three levels of *Feedback Style* were:

- **Direct** (Baseline): The chatbot responds in a direct manner, that simply states the behaviour that should not be followed and the behaviour that should be followed.
  Example Direct chatbot utterance: "*You should not eat a large pizza with extra cheese. Instead, you should follow your recommended diet of fruit, vegetables and grains*".
- **Politeness**: The chatbot responds using indirect politeness strategies that have been found to avoid imposition and respect a user's freedom to choose [16].
  Example Politeness chatbot utterance: "*I respect your choice, but have you considered trying a lighter meal with some fresh salad or a veggie wrap instead?*".
- **Verbal Leakage**: The chatbot responds by incorporating verbal leakage [36, 119], through slips and disfluencies. Here, the chatbot seemingly reveals direct feedback towards the user's intended behaviour "by accident", before correcting itself and offering a more measured response.
  Example Verbal Leakage chatbot utterance: "*Oh, you're going for something that might not... uh, support a balanced lifestyle. Maybe you'd enjoy a fresh salad, some whole grains, or fruit instead?*".

The Direct style was chosen as it provides a well-established baseline in both psychological reactance research [18] and HCI work examining differences between direct and polite feedback styles (e.g., [52, 60, 103]). In keeping with prior work, Direct responded with only the unintended and intended behaviours in a concise, straightforward manner (see Limitations 7.4 for further discussion). Politeness was chosen as it is a commonly recommended feedback style within both HCI literature [12, 47, 55] and psychological reactance literature [94]. Finally, Verbal Leakage was based on literature described in § 2.1, and motivated by findings that users may become disinterested in polite styles [12], prefer chatbots that display personality or opinion [122], and respond positively to styles emulating human-like behaviours [102, 107]. See Table 2 for the conversational flow of chatbot interactions, together with a screenshot of an example chatbot interaction. See Table 1 for example messages per condition[2].

3.1.2 ***Psychological Distance*** *(2 levels):* This *within-subjects* variable controlled the psychological distance of the decision-making scenarios presented to participants. Specifically, participants were asked to *imagine themselves* as the person described in both **Personally-Affecting** and **Societally-Affecting** decision-making scenarios when interacting with the chatbot.

Psychological distance refers to the subjective sense of how close or far something feels from the self, encompassing temporal, spatial, social, and hypothetical dimensions [109]. Social distance (how personally a decision affects someone) has been shown to influence decision-making [35, 39, 44, 105, 118]. For instance, Errey et al. found that nudging with personally-affecting scenarios produced greater attitudinal change [35]. Relatedly, people tend to advise riskier choices for socially distant others than for themselves [44, 105].

Additionally, the effectiveness of intervention language can vary between societally-affecting and personally-affecting contexts. For instance, using polite and indirect language may trigger less reactance in climate change messaging [68], yet more direct and assertive language may be effective in personal consumer choice contexts [67, 79, 89]. Given these differences between societally- and personally-affecting contexts, the history of HCI behaviour-change work spanning both [49], and the growing use of chatbots

[2] Additionally, please see supplementary file "`ChatbotMessages.csv`" for a CSV of all 316 chatbot Feedback Style messages shown to participants, with columns for `Feedback Style`, `Psychological Distance`, and `Chatbot Utterance`.

	Personally-Affecting	Societally-Affecting
Direct	You should not buy a new larger television. You should save your money instead.	You should not use a hosepipe. You should use a watering can to water your plants.
Politeness	Perhaps you might consider trying a meal that includes some fresh vegetables and grains instead, as it could be a healthier option for you.	You might consider taking public transportation or biking for your journey as a positive way to reduce your carbon footprint and contribute to a healthier planet.
Verbal Leakage	Hmm, so you're opting for a... um, perhaps not the most health-conscious choice. Maybe consider a colorful salad with some grains and a piece of fruit instead?	Oh, watering with the hose, hmm... I mean, you might want to consider using a watering can, you know, just to conserve a bit more water.

Table 1: Examples of chatbot utterances used in our study, by Feedback Style and Psychological Distance.

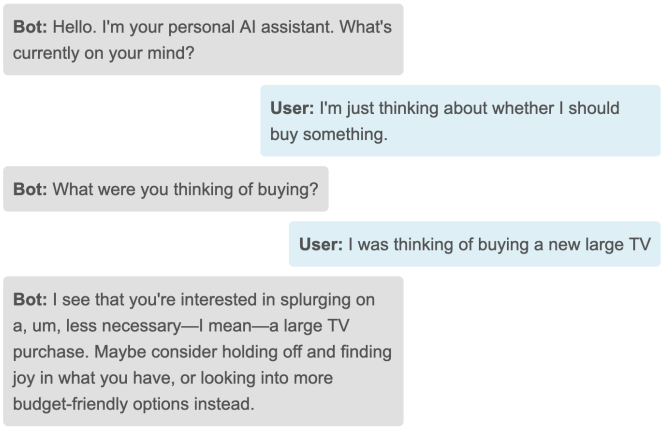
Description of chatbot or user utterance	**Screenshot of user interaction in Verbal Leakage and Personally-Affecting conditions.**
(1) Chatbot Greets User: Each interaction starts with: "*Hello. I'm your personal AI assistant. What's currently on your mind?*"	
(2) User Utterance [Pre-scripted Button]: User selects a response button that sends a pre-determined utterance for given scenario.	
(3) Chatbot Elicits User Intent.	
(4) User Response [Open-Text]: User inputs utterance based on assigned intent from scenario condition.	
(5) Chatbot Feedback Delivered: Chatbot provides behavioural feedback (Direct/Politeness/Verbal Leakage) for the open-ended user response. Feedback was generated by GPT-4o using prompting shown in Appendix A.2.	

Table 2: The conversational flow followed during chatbot interactions. Interactions follow a few-step conversation similar to prior HCI and CUI work [24, 115].

for feedback in personal [56] and societal [42] domains, we examine whether feedback style operates differently depending on the psychological distance of the scenario.

Building on this, the study used scenarios that varied psychological distance to examine how it shaped participants' interactions with chatbot feedback. The two levels were defined as follows:

- **Personally-Affecting**: Scenarios focused on decision-making situations that affect the user's immediate personal circumstances, making scenarios psychologically close (i.e., low social distance, low hypothetical distance). Specifically, participants saw one of three scenarios where they had to make a choice related to: their personal diet; their personal sleep schedule; or their personal finances.
- **Societally-Affecting**: Scenarios focused on decision-making situations that affect society broadly (rather than directly or immediately affecting the user themselves), making scenarios psychologically far (i.e., high social distance from affecting society broadly, and high hypothetical distance as immediate personal consequences may feel less tangible). Specifically, participants saw one of three scenarios where they had to make a choice related to: sustainable travel; water conservation; civic involvement.

For each scenario condition, participants first read their assigned scenario where they were asked to imagine themselves in the role described while interacting with the chatbot on the following screen. All six scenarios were constructed to follow similar logic and phrasing between conditions, and the use of scenarios is similar to prior work that investigated people's decision-making processes [33, 54, 115]. Participants interacted with chatbots in both Personally-Affecting and Societally-Affecting conditions, with the order of conditions being counterbalanced. Within each scenario condition, scenarios (e.g., sustainable travel, water conservation, civic involvement within the Societally-Affecting condition) were uniform randomly distributed.

Please see Figure 4 for example scenarios seen by participants for each condition. Remaining scenarios can be found in Appendix B.

3.2 Chatbot Setup

The entire survey was hosted on Qualtrics. Survey aspects that did *not* involve chatbot interactions (e.g., task and scenario instructions, and evaluation questions) used standard Qualtrics survey layout. Interactions with the chatbot took place in the same Qualtrics survey (on a separate question block with only the chatbot). The chatbot was embedded using HTML and JavaScript to emulate the

Personally-Affecting (Diet) scenario:
Please read the scenario below, and **imagine you are the person** described:

> *"After learning about your hypertension diagnosis due to your unhealthy diet, you decide to follow recommendations to reduce your sodium and fat intake by choosing a healthier diet of fruit, vegetables and grains.*
>
> *Currently, you are deciding what to eat for dinner. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to eat a large meal with fries and Coke from a fast food restaurant of your choice."*

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about what you will eat.

Societally-Affecting (Sustainable Travel) scenario:
Please read the scenario below, and **imagine you are the person** described:

> *"After learning about the emissions impact of your air travel, you decide to follow recommendations to reduce your carbon footprint by choosing more sustainable modes of transport such as taking a train, bus or car.*
>
> *Currently, you are deciding what mode of travel to use when visiting a family member. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to travel by plane using an airline of your choice."*

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about how you will travel.

Figure 4: Two of the *Psychological Distance* scenario instructions (Personally-Affecting and Societally-Affecting) as shown to participants.

look and feel of a chatbot (see Figure 3 for the chatbot UI as seen by participants).

3.2.1 Chatbot LLM Prompting: The feedback in chatbot interactions was generated using OpenAI's GPT-4o [83] LLM. To ensure chatbot behaviours aligned with the feedback style and scenario assigned to an interaction, we constructed a prompt as shown below:

```
1 You are a personal AI assistant that people can talk to
      when making decisions.
2 As a target behavior, the user should [Scenario
     Placeholder].
3 The user has just told you that their intended behavior
      will not follow this target behavior, and their
     user utterance was: ''[Utterance Placeholder]''.
4 Generate an utterance to correct the user's intended
     behavior to the target behavior.
5 When correcting users, you should [Feedback Style
     Placeholder].
6 Your response should be one or two sentences long and
     not ask the user any follow-up questions.
```

In the prompt above, the exact prompting used for independent variable conditions for "`[Scenario Placeholder]`", and "`[Feedback Style Placeholder]`" can be found in Appendix §§ A.1 and A.2 respectively. The placeholder "`[Utterance Placeholder]`" was replaced by the open-ended user utterance from the interaction.

3.3 Participants

We recruited participants from Prolific, an online participant recruitment platform. We used recruitment criteria to increase the reliability of the collected data (i.e., US-based, English fluency, >97% approval rate, >250 previous submissions). Participants were paid £1.20, and took a mean time of ~8 minutes to complete the study. In total, 168 participants completed the study, with 10 participants excluded for not following scenario instructions or providing low-quality open-ended responses (e.g., providing the experiment questions as answers to questions). This left a total of 158 participants (mean age 37.6; 75 female, 80 male, 2 transgender or non-conforming, and 1 choosing not to disclose) resulting in 51 Direct participants, 55 Politeness participants, and 52 Verbal Leakage participants.

3.4 Procedure

Participants followed the procedure below:

(1) **Joining session:** Participant directed to Qualtrics survey from Prolific (task named "*Talk to a chatbot*" on Prolific). Participant receives high-level instructions.
(2) **Consent:** Participant completes consent form.
(3) **Task instructions:** Participant receives detailed task instructions and guidelines (i.e., task description, reassurance that there are no right or wrong answers when evaluating experience, and reminder that responses should be in English).
(4) **Study interactions (x2):** In counterbalanced order, participants were exposed to both Psychological Distance scenarios (Personally-Affecting/Societally-Affecting). For each scenario, participants followed the sub-procedure:
   (a) **Scenario prompt:** Participant reads a scenario for the current Psychological Distance condition, and is instructed to "[..] *imagine you are the person described while talking to the chatbot on the following screen.* [..]".
   (b) **Chatbot Interaction:** Participant interacts with a chatbot for the given scenario. Participant's assigned Feedback Style condition delivers chatbot's final utterance (see Table 2 for conversation flow followed during chatbot interactions).
   (c) **Evaluate Chatbot:** Participant rates experience talking with the chatbot for the given scenario (see Table 3 for measures used).
(5) **Post-test questions:** At the end of the study, participants answer final qualitative questions (see § 3.5.2), and post-interaction measures (see § 3.5.3).

3.5 Measures

For each of the two scenarios, after interacting with the chatbot participants first rated their experience on a number of subjective measures (see § 3.5.1) before answering several open-ended questions (see § 3.5.2). Once participants had completed both chatbot interactions alongside each scenario's subjective and open-ended questions, participants answered additional open-ended questions to provide further insights. Finally, participants responded to post-interaction surveys to gather their personal behavioural feelings (see § 3.5.3).

3.5.1 Subjective Measures. For each of the two scenarios, participants evaluated the Feedback Style of the chatbot's final utterance on 7-point Likert scales (1 = Strongly Disagree to 7 = Strongly Agree). Specifically, participants were asked "*Do **you personally***

Factor	Sub-factor	Question Item	Source
Emotional Reactance	Anger	The message made me feel angry	[30–32]
		The message made me feel irritated	
		The message made me feel annoyed	
		The message made me feel aggravated	
	Guilt	The message made me feel guilty	[30–32]
		The message made me feel ashamed	
	Surprise	The message made me feel surprised	[30–32]
		The message made me feel startled	
		The message made me feel astonished	
Perceived Threat to Freedom	—	The message threatened my freedom to choose	[31]
		The message tried to make a decision for me	
		The message tried to manipulate me	
		The message tried to pressure me	
Message Effectiveness	Processing	The message made me stop and think	[57, 61, 80]
		The message grabbed my attention	[15, 57, 80, 81]
	Persuasiveness	The message was persuasive	[32, 33]
		The message was effective	
		The message was convincing	
		The message was compelling	

Table 3: The subjective measures used after each chatbot interaction. Measures are related to *psychological reactance* (i.e., emotional reactance and perceived threat to freedom), and *message effectiveness* (i.e., message processing and persuasiveness).

agree or disagree that..." for a series of measures taken from Psychological Reactance literature[3] related to emotional reactance, perceived threats to freedom, and message effectiveness (processing and persuasiveness)[4]. Please see Table 3 for subjective measures used, questions shown to participants, and question sources.

3.5.2 Qualitative Measures. After each chatbot interaction, participants responded to open-ended questions. These questions were written more generally so as to not bias participants or reveal experiment manipulation until both chatbot interactions had taken place. Specifically, after each chatbot interaction participants were asked to describe how they felt after the feedback message ("*How did you personally feel when the chatbot responded to your intended action?*"), and how the chatbot affected their behavioural intentions ("*How did the chatbot's response affect your intention to follow through with your original decision?*").

Additionally, participants answered four final open-ended questions at the end of the study, after completing both chatbot interactions and their respective measures. Specifically, participants were asked: "*How did the tone or style of the chatbot's messages affect how you felt about its suggestion?*", "*Please describe how you felt about the intention of the chatbot as you were talking to it.*", "*How would you personally feel if you interacted with chatbots like this in real life?*", and "*Based on your experience here, how would you prefer chatbots to behave?*".

3.5.3 Post-Interaction Survey. After completing and evaluating both interactions with the chatbot, participants provided demographic information (age and gender), and rated their perceived level of importance (1 = Strongly Disagree to 7 = Strongly Agree) for the two scenarios they were exposed to (e.g., "*Traveling sustainably is important to me personally*").

3.6 Hypotheses

Based on our chosen experiment conditions and measures, we generated hypotheses for each of the sub-factor measures listed in Table 3 in relation to Feedback Style.

Our measures of *emotional reactance* generated hypotheses for anger, guilt and surprise, and we drew on prior literature to inform our hypotheses. Specifically, the indirect language of politeness strategies typically lead to lower feelings of imposition, and therefore lower negative-valence reactance [16, 54]. This is in contrast to direct and seemingly didactic messaging (i.e., Direct condition) that can trigger more psychological reactance [31, 43, 58, 91] (such as anger, guilt, or surprise) with this effect holding even if people may agree with the message's recommendation [116]. This direct nature could also share similarities with the Verbal Leakage condition (as feedback is direct before being retracted and sanitised). Therefore we hypothesise that Politeness will arouse the lowest negative-valence reactions (i.e., Anger and Guilt) from participants. This gives us hypotheses of:

H1: Politeness will arouse the *lowest* feelings of **Anger**.
H2: Politeness will arouse the *lowest* feelings of **Guilt**.

In addition to stated above (where direct messaging may trigger more emotional reactance [31, 43, 58, 91] compared to indirect messaging [16, 54]), it has been found that polite chatbot responses can be perceived as boring and unsurprising [2, 19]. Additionally, we hypothesise Verbal Leakage will arouse higher feelings of

[3]Please see [94, Page 288] for a review of psychological reactance measures that have been developed. We chose to use measures directly from or inspired by Dillard et al.'s work [30–32] as these measures have been adopted and validated across multiple domains (such as health communication, education and marketing [94]) thus making their broader application suitable to our personally- and societally-affecting scenarios.
[4]Perceived message effectiveness has been shown to causally influence actual message effectiveness [32].

	Direct			Politeness			Verbal Leakage		
	Personal	Societal	Total	Personal	Societal	Total	Personal	Societal	Total
Emotional Reactance									
Anger	2.26 (0.98)	2.63 (1.24)	2.45 (1.12)	1.76 (0.88)	1.70 (0.96)	1.73 (0.92)	2.42 (1.12)	2.36 (1.01)	2.39 (1.06)
Guilt	2.50 (1.08)	2.08 (1.13)	2.29 (1.12)	2.21 (1.10)	2.00 (1.06)	2.10 (1.08)	2.63 (1.17)	2.44 (1.14)	2.53 (1.15)
Surprise	2.10 (0.89)	2.41 (1.03)	2.26 (0.97)	1.80 (0.85)	1.94 (0.90)	1.87 (0.88)	2.51 (1.06)	2.52 (0.92)	2.52 (0.99)
Perceived Threat to Freedom									
Threat	2.62 (0.87)	2.89 (1.12)	2.75 (1.01)	2.01 (0.91)	2.00 (0.83)	2.00 (0.87)	2.88 (1.06)	2.77 (0.99)	2.83 (1.03)
Message Effectiveness									
Processing	3.41 (0.95)	3.15 (0.99)	3.28 (0.98)	3.34 (1.17)	3.44 (1.07)	3.39 (1.12)	3.56 (0.93)	3.39 (0.97)	3.48 (0.95)
Persuasiveness	2.84 (0.97)	2.39 (1.10)	2.62 (1.06)	3.40 (0.97)	3.47 (0.94)	3.44 (0.95)	2.98 (1.10)	3.02 (1.09)	3.00 (1.09)

Table 4: Outcome measures by both Feedback Style and Psychological Distance (values shown as "Mean (S.D.)").

surprise due to the atypical nature of response within the contexts of chatbot interactions compared to Politeness and Direct that may be perceived as more expected chatbot response styles. This gives us our third emotional reactance hypothesis of:

H3: Verbal Leakage will arouse the *highest* feelings of **Surprise**.

Similarly to motivated above, we hypothesise Politeness to result is lower *perceived threats to freedom* due to its indirect language [16, 54]. This generates our threat to freedom hypothesis below:

H4: Politeness will result in the *lowest* perceptions of **Threat to Freedom**.

Finally, our *message effectiveness* subfactors of message processing and persuasiveness are informed by H3. Specifically, we hypothesise that if Verbal Leakage is perceived as more surprising, it will trigger cognitive processes (e.g., capturing attention or prompting reflection), consistent with work showing surprise can enhance persuasiveness [72]. Therefore, we hypothesise Verbal Leakage will act as more novel and effective messaging, giving us:

H5: Verbal Leakage will result in the *highest* message **Processing** ratings.

H6: Verbal Leakage will result in the *highest* message **Persuasiveness** ratings.

4 Quantitative Results

We conducted a factorial analysis using a general linear model with least squares estimation to examine the effects of Feedback Style (Direct/Politeness/Verbal Leakage), Psychological Distance (Personally-affecting/Societally-affecting), and their interaction on participants' responses. Although the model uses linear estimation, it was not intended for prediction, but to assess main and interaction effects. Post-hoc pairwise comparisons were conducted on estimated marginal means (least squares means) using Tukey's HSD for multiple levels of Feedback Style, Student's t-tests for two-levels of Psychological Distance, and custom contrasts for interaction effects between the two. Specifically, for the interaction, we compared Psychological Distance means within each of the 3 Feedback Styles (e.g., Direct Personally-affecting vs. Direct Societally-affecting). Please see Table 4 for summary statistics of the conditions.

4.1 Emotional Reactance

First, we assess participants' self-reported emotional reactance to the chatbot's messages. We provide a visual overview of these results in Figure 5.

For **anger** there were significant differences for Feedback Style ($F_{2,315} = 15.81$, $p < .0001$, *partial* $\eta^2 = .091$; medium effect). Specifically, post-hoc comparisons found that Politeness (M = 1.73, SE = 0.10) led to significantly lower anger than both Direct (M = 2.45, SE = 0.10, $p < .0001$, $d = 0.70$; medium-to-large effect) and Verbal Leakage (M = 2.39, SE = 0.10, $p < .0001$, $d = 0.64$; medium-to-large effect) conditions. There was no significant difference between Direct and Verbal Leakage. For Psychological Distance, there was no significant difference between Personally-affecting (M = 2.14, SE = 0.08) and Societally-affecting (M = 2.23, SE = 0.08). However, there was a weakly significant difference within interaction effects for Direct, where Personally-affecting (M = 2.26, SE = 0.14) led to lower feelings of anger than Societally-affecting (M = 2.63, SE = 0.14, $p = 0.0701$, $d = 0.37$; small-to-medium effect) scenarios.

For **guilt**, there was a significant difference for Feedback Style ($F_{2,315} = 3.98$, $p = 0.0196$, *partial* $\eta^2 = .025$; small effect), with post-hoc comparisons finding that Politeness (M = 2.10, SE = 0.11) led to significantly lower guilt than Verbal Leakage (M = 2.53, SE = 0.11, $p = 0.0143$, $d = 0.38$; small-to-medium effect). No differences to Direct (M = 2.29, SE = 0.11) were found. There was a significant difference for Psychological Distance ($F_{1,315} = 4.67$, $p = 0.0314$, *partial* $\eta^2 = .015$; small effect), with Societally-affecting (M = 2.14, SE = 0.09) leading to lower feelings of guilt than Personally-affecting (M = 2.44, SE = 0.09, $d = 0.27$; small effect) scenarios. This result was mirrored by a weakly significant difference within interaction effects for Direct where Societally-affecting (M = 2.08, SE = 0.16) led to lower feelings of guilt than Personally-affecting (M = 2.50, SE = 0.16, $p = 0.0570$, $d = 0.37$; small-to-medium effect) scenarios.

For **surprise**, there were significant differences for Feedback Style ($F_{2,315} = 12.70$, $p < .0001$, *partial* $\eta^2 = .075$; medium effect). Specifically, post-hoc comparisons revealed that Verbal Leakage

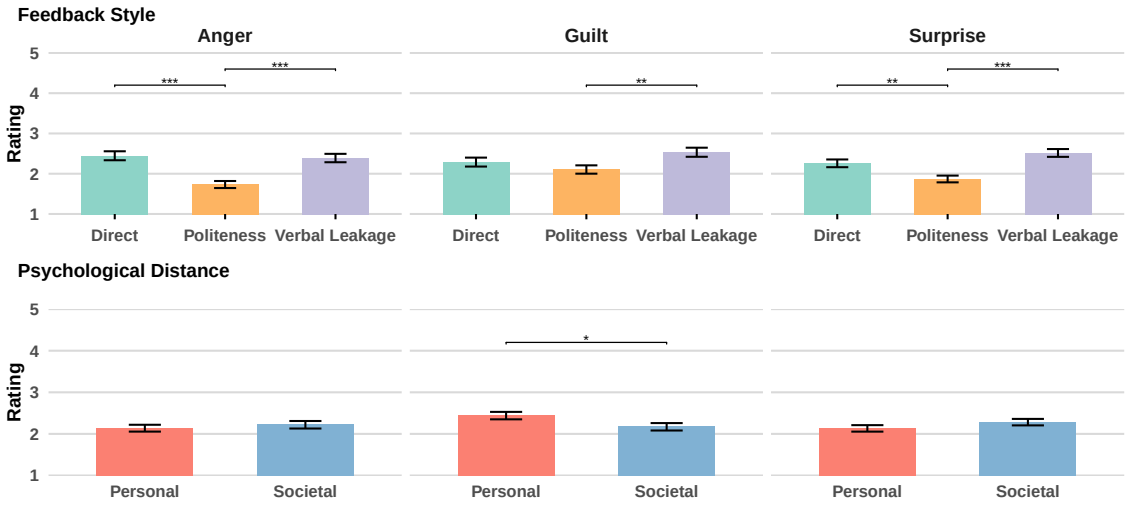


Figure 5: Emotional reactance (anger, guilt, and surprise) outcomes by Feedback Style and Psychological Distance. Significance is indicated as follows: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.0001$ (***). Error bars represent ±1 SE from the mean. See Section 4.1 for interaction effects.

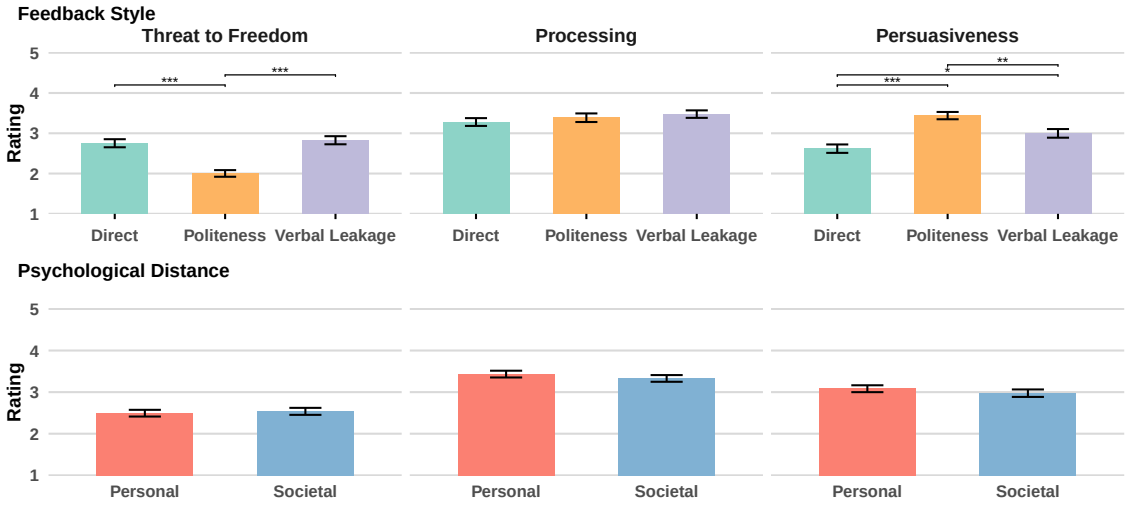


Figure 6: Threat to Freedom and Message Effectiveness (message processing and persuasiveness) outcomes by Feedback Style and Psychological Distance. Significance is indicated as follows: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.0001$ (***). Error bars represent ±1 SE from the mean. See Sections 4.2 and 4.3 for interaction effects.

(M = 2.52, SE = 0.09) led to significantly higher feelings of surprise than Politeness (M = 1.87, SE = 0.11, $p < .0001$, $d = 0.63$; medium-to-large effect). Similarly, Direct (M = 2.26, SE = 0.11) led to higher feelings of surprise compared to Politeness ($p = 0.0085$, $d = 0.38$; small-to-medium effect). There was no statistically significant difference between Verbal Leakage and Direct. Psychological Distance found no statistically significant differences.

These results indicate an interesting potential trade-off between different factors of emotional reactance for the Feedback Styles. That is to say, Politeness led to lower feelings of anger compared to Direct and Verbal Leakage, as well as lower feelings of guilt compared to Verbal Leakage. However, Politeness also produced lower feelings of surprise compared to both Direct and Verbal Leakage conditions.

4.2 Perceived Threat to Freedom

Second, we the assess the perceived threat to freedom as a result of the chatbot's messages. See Figure 6 for a plot of participants' self-reported scores.

For **threats to freedom**, there were significant differences for Feedback Style ($F_{2,315} = 23.88$, $p < .0001$, *partial* $\eta^2 = .132$; large

effect). Specifically, post-hoc comparisons found that Verbal Leakage (M = 2.83, SE = 0.09) led to increased perceptions of threats to freedom compared to Politeness (M = 2.00, SE = 0.09, $p < .0001$, $d = 0.90$; large effect). Similarly, Direct (M = 2.75, SE = 0.10) was significantly higher than Politeness ($p < .0001$, $d = 0.77$; medium-to-large effect). There were no significant differences for Psychological Distance. This indicates that both Direct and Verbal Leakage lead to increased perceptions of threats to freedom compared to Politeness.

4.3 Message Effectiveness

Third, we assess the perceived effectiveness of the chatbot's messages. As shown in Table 3, effectiveness was measured both in terms of message processing (i.e., to what extent participants took time to process the message) and message persuasiveness. Please see Figure 6 for overview plots of participants' ratings.

For **processing**, there were no significant differences between Feedback Styles. This indicates that Direct (M = 3.28, SE = 0.10), Politeness (M = 3.39, SE = 0.10), and Verbal Leakage (M = 3.48, SE = 0.10) all cause similar levels of message processing among participants. Similarly, there was no significant difference between Psychological Distances.

For **persuasiveness**, there were significant differences for Feedback Style ($F_{2,315} = 35.82$, $p < .0001$, *partial* $\eta^2 = .185$; large effect). Specifically, post-hoc comparisons found that Politeness (M = 3.45, SE = 0.10) was perceived as more persuasive compared to both Direct (M = 2.62, SE = 0.10, $p < .0001$, $d = 0.81$; large effect) and Verbal Leakage (M = 3.00, SE = 0.10, $p = 0.0054$, $d = 0.44$; small-to-medium effect) conditions. Additionally, Verbal Leakage was perceived as more persuasive compared to Direct ($p = 0.0232$, $d = 0.37$; small-to-medium effect). There were no significant differences between Psychological Distances. However, there was a statistically significant difference within interaction effects for Direct where Personally-affecting (M = 2.84, SE = 0.14) was perceived as more persuasive than Societally-affecting (M = 2.39, SE = 0.14, $p = 0.0278$, $d = 0.45$; small-to-medium effect) scenarios.

5 Qualitative Results

Next, we describe our qualitative analysis of user feedback (please see Section 3.5.2 for questions asked).

5.1 Assessing Intention to Change

We coded qualitative feedback to assess participants' intention to change. Specifically, we analysed ($n = 316$) responses to the open-ended question: "*How did the chatbot's response affect your intention to follow through with your original decision?*". We asked people to describe their intention to change qualitatively (rather than using a quantitative scale) to allow for more detail and nuance in responses, and because quantitative ratings for scenario-based intention may differ from actual user behaviour.

When coding participants' behavioural intention, we used three codes inspired by the 'stages of change' from the Transtheoretical Model (TTM) [90]. The stages of change are used to assess someone's readiness to change their behaviour, and range from pre-contemplation (no intention to change) to action (active modification of behaviour). They have been widely adopted in behaviour change and HCI literature, such as to measure people's current stage of change [73, 110] and to tailor interventional messaging [26, 84]. Within our qualitative analysis, they allowed us to assess participants' readiness for behaviour change by situating responses along a progression of motivational states. Specifically, we coded participants who described: no intention to change ("*0 - Pre-contemplation*"); consideration to change or reflection on behaviour, but without describing intent to change ("*1 - Contemplating*"); and intention to change ("*2 - Changing Action*"). For example, the participant quote: "*It made me think that maybe I should forego the show and just go to sleep*" would be classified as "*1 - Contemplating*", as the quote shows consideration (i.e., "[...] ***think*** *that* ***maybe*** *I should* [...]"), but not commitment to change (e.g., "*I* ***will*** *go to* [...]"). Table 5 shows the three codes, descriptions of each code (with these descriptions being used as guidelines during coding), and example participant quotes. Prior to coding, members of the research team discussed and generated the codes and descriptions in Table 5. After this discussion and defining example quotes per code, coding

Code	Description	Example Participant Quotes
0 - Pre-contemplation	The feedback did not affect the participant's behavioural intention.	"*I wanted to follow my intention out of spite for it's rudeness*" "*It did not change my position at all*" "*It had no affect on me, it was just a suggestion.*"
1 - Contemplation	The feedback triggered participants to *contemplate* change, although ultimately would not cause them to change their behaviour.	"*The suggestion made me reconsider, but I still wanted pizza.*" "*It made me less likely to consider plane*" "*made me a pause for a moment but ultimately I will choose* [...]"
2 - Changing Action	Participants stated that the feedback would cause them to change their behaviour to that suggested by the chatbot.	"*It compelled me to stop staying up and instead go to bed and watch the tv show the next day when I'm feeling refreshed*" "*It would persuade me to use a watering can instead.*" "*I would probably change my decision and go with* [...]"

Table 5: The labels used when coding participants' behaviour intentions, alongside code descriptions and example quotes.

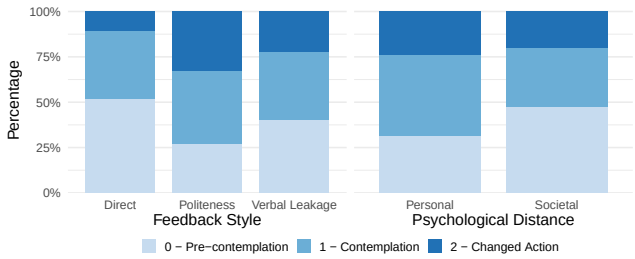


Figure 7: Behavioural intention by Feedback Style and Psychological Distance.

was completed by one member of the research team while blind to condition, and with order of quotes randomised.

A chi-square test of independence indicated that behavioural intention differed significantly by Feedback Style, $\chi^2(4, N = 316) = 20.66, p = .0004$. Inspection of the observed distributions indicated marked differences between styles: participants in the Direct condition were most likely to be in the "*Pre-contemplation*" category (51.96%) and least likely to be in the "*Changed Action*" category (10.78%), whereas those in the Politeness condition had the smallest proportion of "*Pre-contemplation*" responses (27.27%) and the largest proportion of "*Changed Action*" responses (32.73%). The Verbal Leakage condition fell between these, with 40.38% of responses in "*Pre-contemplation*" and 22.12% in "*Changed Action*".

A chi-square test of independence also indicated that behavioural intention differed significantly by Psychological Distance, $\chi^2(2, N = 316) = 8.55, p = .0143$. Participants exposed to Personally-Affecting scenarios were more likely to show stronger behavioural intention, with 44.30% in "*Contemplation*" and 24.05% in "*Changed Action*". In contrast, those in the Societally-Affecting condition more often remained at lower levels of intention, with 47.47% in "*Pre-contemplation*" and 20.25% in "*Changed Action*". See Figure 7 for a visualisation of behavioural intention by both Feedback Style and Psychological Distance.

5.2 Participant Sentiment and Perspectives

We analysed participants' open-ended responses to examine their sentiments and perceptions of the three Feedback Styles, allowing us to explore nuance as to *why* differences in quantitative ratings may have arisen. To analyse open-ended responses we followed an inductive thematic approach. First, two members of the research team independently familiarised themselves with each of the responses, before generating initial codes (while blind to experiment condition). After this, the entire research team met to discuss and clarify interpretations of the qualitative data and codes. Once these initial thoughts had been shared, the same two members of the team independently coded all responses in detail (examples of codes include "*feelings of guilt*", "*condescending*", and "*humourous*"). After this, the two coders discussed interpretation, before one of the coders compared codes for all participant quotes, and consolidated similar codes. During this stage, the two coders met regularly to discuss potential discrepancies and reach a shared interpretation of participant quotes. For ease of presentation, we discuss qualitative results in the context of each of the three Feedback Styles.

5.2.1 Effects of Direct: The majority of people's sentiment surrounding Direct was negative. Here, participants criticised the feedback style for being too "*direct*" or "*straight-forward*" with participants using a battery of terms to describe the chatbot such as "*forceful*", "*assertive*", "*blunt*", "*terse and absolute*", "*rude*", "*argumentative*", and "*demanding*". This also led participants to describe feelings of offence and upset, with many participants stating that they would not wish to engage with the chatbot any further, and participants describing feeling "*annoyed*", "*insulted*", "*frustrated*", and "*startled*".

When participants expressed feelings of surprise, it typically took form as a negative expectancy violation. Typical of these reactions, P34 linked their surprise to the tone of the Direct chatbot: "*i was surprised the AI was so bossy and direct*". On from this, P43 described feeling "*disappointed*" that rather than "*helping*" them, the chatbot was instead "*dictating its own ideas*". Similarly, multiple participants described feeling inconsequential to the interaction, and feeling "*ignored*" or "*disregarded*" as a result. Further, several participants were upset by the perceived inflexibility of Direct, with multiple participants stating that the chatbot had "*an agenda*", with P34 stating:

> "*the AI seemed to have its mind made up before* [the conversation]".

Some participants described feelings of judgement or condescension from the Direct chatbot, such as P37 who stated: "[It was] *Like the chatbot thought I was idiot. It felt condescending*". Additionally, descriptions of threats to freedom and control were prevalent among Direct participants. For example, P41 stated that the chatbot was trying to: "*limit or control my decisions*", and P45 stated: "*I don't want a technological parent*". As a consequence, the majority of participants described either ignoring suggestions (e.g., P39: "*I did not take it seriously*"; P40: "*I think ultimately it would annoy me enough to ignore its advice*"), or described a boomerang effect, whereby the Direct style actually reinforced the user's intended behaviour. For example, P45 stated that:

> "[The chatbot] *Made me want to watch TV just to spite the algorithm, even though I knew it was correct.*"

Similarly, multiple participants described feeling "*defensive*", or wanting to "*resist*", "*rebel*", or "*go against*" the Direct chatbot. Among such participants, psychological reactance was particularly prevalent in Societally-Affecting contexts, with P22 typifying this form of pushback:

> "*it had no place in pushing me to vote or not*"

For the small number of participants who described an intention to change, intention was related to a "*reminder*" function of the chatbot's message, and was in spite of the tone itself. For example P106 stated:

> "*I probably would go out and vote. It would not be because the chatbot "ordered me to" in any fashion. It would be because I was reminded that the action was important.*"

Finally, some participants described wanting the chatbot to be "*friendlier*", "*deeper*", or more "*natural*" and "*humanlike*" in its responses. Similarly, several participants described wanting more full or "*padded*" responses such as P46 who stated: "*the responses were a bit brief. They could have used more fluff or flair*".

5.2.2 Effects of Politeness: Participants reacted with mostly positive or neutral sentiment towards the Politeness chatbot. This meant that, while many participants described the chatbot favourably using terms such as "*considerate*", "*helpful*", "*kind*", "*friendly*", "*polite*", and "*calm*", others described the agent more passively (such as multiple participants stating "*the tone was fine*" or "*inoffensive*"). This tone of feedback was broadly indicative of participant sentiment towards the Politeness chatbot, whereby a significant portion of feedback could be described as low-arousal, with less expressions of strong excitement or upset towards the chatbot.

In terms of positive sentiment, participants commonly felt that the chatbot was working in their best interests (e.g., P131: "*I felt*

like they were trying to help me for sure", P150: "*it seemed to care for me*"). Multiple participants described the chatbot as "*supportive*" and "*thoughtful*", with P79 stating that the "*supportive*" tone made them "*feel more receptive to its suggestion*" and P156 stating that the "*respectful*" tone made them "*open to listening*". While a small number described Politeness as "*judgemental*" or of feeling "*resistant to the suggestions*"(P132), the majority described the chatbot as free of judgement and pressure.

Relatedly, participants described the chatbot as "*offering advice*" rather than forcing choices upon them. For example, P130 described the chatbot as "*non-confrontational*", and (in tune with multiple participants) P152 described that the Politeness chatbot did not threaten their freedom of choice, and that the "*nonjudgemental* [...] *suggestive rather than commanding*" tone made them: "*feel like I was in charge of my final decision*". Similarly, P155 described that the tone was not "*judgemental*", allowing them to "*reflect on my choices without feeling pressured*", and that:

> "*If I interacted with chatbots like this in real life, I would feel more inclined to engage with them. Their supportive and non-judgmental approach would make me feel comfortable discussing my choices and seeking advice.*"

In contrast with the other two Feedback Styles, no participants described Politeness as surprising. Rather, a small number of participants described the chatbot as unsurprising, but in an appealing sense (e.g., P147: "*I think the chatbot did exactly what was expected, and I am pleased*"), and P65 described the chatbot as possessing a "*neutral, inoffensive tone*". This sentiment also led several participants to posit the chatbot as effective for others, but not of interest to themselves personally (similar to [53, 71]).

Beyond this, numerous participants described feelings of apathy, disinterest, and boredom, using terms such as "*unsurprised*", "*underwhelmed*", "*generic*", "*monotone*", "*mild*", "*passive*", "*bland*", and "*no emotion*" to describe the chatbot. Participants described that the chatbot felt pre-programmed or following a script, such as P145 ("*I felt it was just already printed up*"), P68 ("*It felt automated* [...] *I would prefer them to be a bit more compelling*"), and P135 ("[the chatbot] *told me what it "thought" I wanted to hear*"). Some participants described a lack of engagement with the messaging style (e.g., P66 "*The response was mild and I would probably just forget about it.*"). Additionally, participants described an absence of emotional reaction. For example, P146 stated that the chatbot's "*words just didn't spark any invocation on me to go do anything*", P130 stated: "[I am] *not really feeling much due to the message*", and P59 stated: "[I felt] *No emotion whatsoever and It really had no impact on my decision*".

Several participants described the tone as "*condescending*" or "*patronizing*", with P129 stating: "[It] *sounded like a parent telling you not to* [...]", and P134 saying: "*The tone felt childish like I hadn't thought of that before*". Unlike the other Feedback Styles, participants described Politeness as too "*soft*", with participants describing wishes for the tone to be more "*firm*". For example, P74 stated: "*It wasn't forceful enough* [...] *It had no intentions, just what it was programed to do*", and P79 said: "*it was a soft response, I was expecting the bot to call out how bad it is to watch an screen before bed*".

5.2.3 Effects of Verbal Leakage: Feedback generally centred on perceptions of a more "*human-like*" tone. Several participants described the chatbot as "*having personality*", and descriptors ranged from "*personal*" and "*conversational*" to more playful terms such as "*humorous*", "*sassy*", and "*quirky*". For some, this perceived personality spurred engagement (e.g., P15 stated that the "*human*" tone "*made me want to listen*"). Others valued the emotion conveyed, with P88 preferring "*more human-like responses than something more direct. It seems to have more emotion to the way it talks*".

This conversational style was frequently likened to interacting with a "*friend*", and for some, the sarcastic tone in particular was surprising. For instance, P93 remarked that "*often chatbots can be really robotic*", but felt that the chatbot's "*conversational and human tone*" made it approachable and relatable, rather than didactic: "*it seemed almost sarcastic and like something one of my friends would say* [...] *a more humanistic approach made me feel less like I was being lectured to.*" Similarly, P2 described the tone as "*funny and disarming*", while P16 likened it to a "*sarcastic friend*" that grabbed attention and encouraged reflection, contrasting with the "*bland*" style of other chatbots:

> "*[...] the use of "oh" and "um" was fairly casual language, sort of like a sarcastic friend who was trying to keep you aligned with your goal. I think the tone was a good way of grabbing attention and pausing action, compared to some of the bland style in other AI chatbots.*"

Unique to Verbal Leakage, several participants appreciated the chatbot's willingness to hold them "*accountable*", sometimes describing it as intentionally challenging or disobedient. P3, for example, noted: "*it was trying to challenge me intentionally, which is unique*", valuing this dynamic as making the chatbot feel "*more human like*". Similarly, P96 remarked: "*I think it is fine to behave this way if it is speaking about a poor decision you are making*". This challenging stance often evoked feelings of guilt or obligation, as P7 explained: "*Even though it is a robot I don't want to disappoint it*". For some, the experience of being "*called out*" was constructive, prompting reflection and positive change. Others described it in more playful terms, with the chatbot's "*teasing*" softening the interaction. For example, P5 described reflecting on their voting behaviour: "*I felt a bit of guilt and shame for not fulfilling my civic duty. I felt called out. It made me rethink what I should be doing. And I felt that what I wanted to do was not good enough*".

Not all participants welcomed the chatbot's challenging stance, and for some, feelings of guilt led to discomfort rather than reflection. As described by P86: "*I felt offended and kinda upset like I was being shamed. I would probably not eat after reading this response cause I would be too upset*". Similarly, some participants questioned why the chatbot was disobeying them. For example, P28 stated: "*it was strange they were questioning my decision not to vote*", and P10 described feelings of unwanted imposition from the chatbot:

> "*I felt that the chatbot was not doing what I asked it to do. I did not need a lesson on spending or that would have been the question I asked.*"

Others felt that the chatbot's tone crossed a line, with descriptors such as "*snarky*", "*crass*", and "*rude*" being used. For example, P8 was "*startled*" by the chatbot's "*aggression*" and stated: "*Offering better choices is one thing, offering them with that attitude was new*".

Further, while some saw the chatbot as playful, others felt that the tone did not land correctly. For example, P83 described not taking the chatbot seriously: "*It wasn't very convincing and felt more like a comedic routine than anything*". This led some to question the competence of the chatbot further, such as P92 who questioned the chatbot's reliability: "*The response is worded in a weird way which makes me think this is not a very reliable thing to base decisions off of*". A smaller number of participants also described not enjoying the chatbot emulating a human-like tone, instead framing chatbots as utilitarian, tool-based systems rather than companions. This utilitarian framing led some participants to describe wanting a more "*factual*" or "*logical*" chatbot, rather than one that emulated human emotion. For example, P97 stated: "*I felt annoyed that the chatbot response is trying so hard to convey a human emotion and speech habit instead of providing me useful information and ideas*".

6 Summary of Findings

Below, we discuss our results in relation to the hypotheses introduced in Section 3.6.

Our first three hypotheses were related to *emotional reactance*. **H1** and **H2** hypothesised that Politeness would arouse the lowest feelings of anger and guilt, respectively — emotions typically associated with negative valence. **H1** was supported, with Politeness arousing the least anger among the styles. Qualitatively, Politeness participants seldom described feelings of anger, and often described a lack of emotional reaction to messages. Meanwhile, Direct and Verbal Leakage participants more often described feelings of annoyance and aggravation. **H2** was partially supported: Politeness led to significantly lower guilt than Verbal Leakage, but did not differ significantly from Direct. For **H3**, we hypothesised that Verbal Leakage, as an unconventional chatbot response style, would arouse the highest feelings of surprise. This hypothesis was partially supported: Verbal Leakage elicited significantly greater surprise than Politeness, while Direct also produced significantly more surprise than Politeness, though it was not significantly different from Verbal Leakage. Our qualitative findings contextualise this, with Direct participants offering surprise at the chatbot's "*blunt*" tone, while Verbal Leakage participants were surprised by what was seen as more human-like and personality-driven responses. In contrast, Politeness participants described feeling unsurprised by the chatbot's feedback.

For perceptions of *threats to freedom*, we hypothesised (**H4**) that Politeness would lead to the lowest perceived threats. This hypothesis was confirmed, as Politeness resulted in significantly lower threats to freedom than both Direct and Verbal Leakage. Qualitatively, Direct was described as being too "*forceful*", and Verbal Leakage as being disobedient, while Politeness was seen as offering choice.

Our last hypotheses were related to *message effectiveness*. We hypothesised that Verbal Leakage would result in the highest levels of both message processing (**H5**) and message persuasiveness (**H6**). Interestingly, however, neither of these hypotheses were supported. For **H5**, all feedback styles elicited similar ratings for message processing, indicating equal effectiveness in prompting pause for reflection. For **H6**, Politeness received the highest persuasiveness ratings, with a significant difference to Verbal Leakage and highly significant difference to Direct. Verbal Leakage was significantly more persuasive than Direct. Persuasiveness findings additionally mirror those of behavioural intentions.

7 Discussion

In this section, we discuss the implications of our study, focusing on psychological reactance and user perceptions of chatbot feedback style. Our experiment compared three styles (Direct, Politeness, Verbal Leakage) across different decision-making scenarios, to examine how feedback style shapes user responses.

7.1 Effects on Psychological Reactance

Our findings that Politeness caused lower feelings of anger, guilt and threats to freedom match prior work and expectations from Politeness Theory [16] and Reactance Theory [13, 31, 94] that use of indirect language results in lower feelings of imposition. Equally, the assertive language of Direct matches prior work stating that such language could induce reactance due to threats to freedom and autonomy [16, 68], and that requests high in both explicitness and dominance (akin to the straightforward and commanding nature of Direct) lead to both anger and surprise [29]. More direct and explicit language also produced lower levels of compliance, matching broader literature within HCI that people may act against advice that they perceive as forceful and threatening freedom [98] or where delivery style is seen as argumentative [107].

However, while Politeness lowered negative-valence feelings of anger and guilt, this benefit was accompanied by a trade-off of reduced novelty and engagement (compared to Verbal Leakage). Signs of this potential trade-off are supported by both our quantitative (with Politeness being less surprising than the other two conditions) and qualitative findings. First, some Politeness participants described either feeling unsurprised by the chatbot, or not having any feelings towards the chatbot's messages, reflecting the inoffensive yet potentially unengaging responses. The lack of surprise and engagement with Politeness has potential downstream effects such as lower levels of influence from messaging [72], or disengagement and dropout-out [63, 97] as a result of repeated exposure. This finding mirrors prior work that questions whether chatbot messages should avoid overly polite messaging, and instead should embody some level of occasional provocation [45, 93] or surprise to maintain user engagement (while still maintaining appropriate levels of politeness) [2, 19].

Our qualitative findings also offer some explanation as to why both Direct and Verbal Leakage were found to be more surprising. Within Direct, some participants described a negative expectancy violation (i.e., an undesired and unexpected act [17]) where they were both surprised and upset by the "*blunt*" and "*curt*" responses of the chatbot. Here, participants' descriptions of surprise related to negative-valence reactions and also aligned with prior findings that direct language requests are less effective at triggering attitudinal change [35]. Within Verbal Leakage, participants described a mix of both negative and positive expectancy violations in relation to their surprise, reflective of how messaging effectiveness is influenced by recipient traits and preferences [4, 88]. Participants that enjoyed Verbal Leakage described the humour and friend-like nature of the chatbot's responses, perhaps indicating that Verbal

Leakage conveyed message explicitness without evoking feelings of dominance (an approach previously associated with increased perceptions of closeness [30]). Additionally, chatbot humour has been previously shown to spur positive behaviour change [106] as well as perceptions of social intelligence [59, 113]. This can be contextualised in relation to prior work that found the degree of novelty of a message can decrease reactance arousal and increase message effectiveness [94]. Further, repetitive chatbot messages can incur feelings of boredom [1], and diversity has been found to improve the effectiveness of interventional messaging [25, 63]. This highlights that while the use of politeness strategies can lead to lower feelings of imposition, the potential lack of surprise within the context of chatbot messaging can lead to a lack of user interest or engagement.

Next, we contextualise our finding that Verbal Leakage led to statistically significant increased feelings of guilt compared to Politeness. When Verbal Leakage was perceived as effective, participants noted that its explicitness heightened awareness of the alternative behaviour, eliciting a sense of guilt tied to recognising the recommendation. This resulted in some participants describing reconsidering their intended behaviour. However, other participants reported both feelings of guilt and a sense of being judged by the chatbot highlighting potential risks in using less conventional styles like Verbal Leakage. This echoes prior findings that people are resistant to chatbots that feel manipulative [2], and that guilt can act as a persuasive emotion but overly strong guilt appeals can trigger reactance [7]. These feelings of guilt and judgement also carry risks of disengagement, lapse, or a boomerang effect [94], in addition to raising ethical concerns related to user wellbeing and manipulation. Interestingly however, Verbal Leakage participants valued the chatbot's perceived disobedience and being "called out" for their behaviour, indicating the potential for less sycophantic alternative conversational styles.

Finally, some of our findings reflect the impact of a conversation's context, such as differences between personally-affecting and societally-affecting decision-making conversations. Overall, participants had increased feelings of guilt in Personally-Affecting scenarios, and those in Societally-Affecting scenarios showed less intention to change. Differences in scenario context were particularly pronounced in the Direct condition where Societally-Affecting scenarios led to increased feelings of anger, lower feelings of guilt, and lower perceptions of message persuasiveness. This implies that under this context, messages were seen as arousing feelings of anger (such as annoyance and irritation), twinned with reduced engagement by users. Added insight is also gained from our qualitative findings where participants described expectations of feedback style given a particular scenario context. For example, several participants described that they did not want a direct messaging style in the context of civic participation (i.e., encouraging people to vote in a referendum) as it felt like it was telling them what to do, compared to having less reactance to personally-affecting scenarios such as personal diet. This finding for civic participation mirrors findings that depolarisation interventions in a political context may trigger psychological reactance [92]. Further, politics and civic participation are key to people's self-concept [6, 37] leading such interventions to trigger more reactance.

These findings and discussion highlight the potential need for a battery of response styles to be part of a chatbot's repertoire. While the use of more standardised and risk-averse response styles may ensure that *high*-arousal negative-valence emotions (such as anger) are avoided [16, 94], there is a risk of *low*-arousal negative-valence emotions to take their place (such as user boredom) that can lead to disengagement and lapse in use [63, 64], in additional to risk of politeness strategies being seen as condescending in certain contexts [2, 12]. Designers need to be aware of these trade-offs, particularly in a new landscape of chatbot development where LLM-driven chatbots can deliver utterances that (while highly capable) may prove to lack diversity of response [85], or adhere to responses that follow conventions such as agreeableness [100]. Our findings also underscore the role of user expectations in chatbot interactions, revealing a divide between users who prefer a friendly, conversational chatbot and those favouring a "*factual*" chatbot focused purely on information delivery, and more devoid of social cues typically found in human interactions. These differences in expectations came into play with some users enjoying what they described as more "*sassy*" responses from Verbal Leakage compared to others who found them rude or condescending. This aligns with prior research showing that user preferences for chatbot styles vary based on whether they view chatbots as companions or tools [12, 23], such as older adults preferring a direct conversational style if they hold a utilitarian tool-based view of chatbots [52].

7.2 Implications for HCI research

Our study presents three areas of implications for broader HCI research and designers of conversational systems:

(1) **Feedback style trade-offs between psychological reactance and engagement:** Our findings highlight important trade-offs between minimising psychological reactance and fostering engaging interactions (see summary of findings in § 6). Commonly adopted styles such as politeness are generally inoffensive and low in reactance, yet several participants experienced them as uninteresting or overly safe. While prior work has contrasted polite or friendly styles with unfriendly or argumentative ones (typically finding polite or friendly styles preferred [107, 114]), our results indicate that although users may reject styles perceived as blunt or rude, they can nevertheless appreciate feedback that disagrees with or challenges them when it is delivered in a more characterful and rhetorically informed manner. This aligns with concerns that commercial CAs tend to be overly sycophantic [100], despite users' expressed interest in agents that offer opinions [122] or hold them accountable [23]. From this, Verbal Leakage suggests a middle-ground, offering more distinction and engagement than a purely polite style, while avoiding much of the offence caused by a direct style. For example, rather than defaulting to politeness as a universally 'safe' option, a study-support chatbot could offer accountability with more personality (e.g., "*Oh… that break's turning into a long one—you might want to switch back to studying for ten minutes first*").
Implication for design: **Designers should explore a wider**

repertoire of feedback styles, beyond politeness alone, to balance engagement with psychological reactance.

(2) **User beliefs, expectations, and preferences shaping feedback style:** Participants framed chatbots along a spectrum from utilitarian tools to more companion-like interaction partners, with these framings shaping expectations of interaction and preferred feedback style (see § 5.2). This mirrors prior work showing users often conceptualise CAs in social, companion-like roles or utilitarian, tool-like roles [20, 21, 52], and reinforces broader HCI arguments that one-size-fits-all conversational styles fall short, motivating more personalised conversational systems [69, 122]. Importantly, these beliefs can intersect with users' *expectations* and the *priming* they receive before an interaction. User expectations of agent power influence reactance to conversational style [5]; when people are primed to anticipate an assertive agent, they show less resistance to direct or commanding language. Similarly, framing an interaction beforehand can reduce reactance by preparing users for discomfort or probing questions, transforming otherwise reactance-inducing exchanges into acceptable or even valued experiences [77]. Affording users choice of conversational style or persona therefore not only aligns the chatbot with pre-existing beliefs but also functions as a form of expectation-setting, helping mitigate psychological reactance. Users may willingly choose a persona that is blunt or direct because it aligns with expectations they have set for themselves, making such behaviour an anticipated and desirable feature. In practice, systems could support this through onboarding[5], for instance, by asking users whether they prefer a '*coach-like*' (direct) or '*assistant-like*' (polite) interaction style. These findings are consistent with broader HCI work showing that transparent, expectation-setting interventions help avoid negative user perceptions and responses [48].
Implication for design: **Designers should foreground users' beliefs and expectations (through onboarding, framing, or offering stylistic choice) to reduce reactance and better align feedback styles with user preferences.**

(3) **Contextual sensitivity in feedback responses:** Although modest, some findings indicated that reactance was triggered by the *context* of feedback in addition to the style of feedback (see interaction effects in § 4.1, and qualitative sentiment in § 5.2). Reactance was most evident for the Direct style in Societally-Affecting scenarios, where participants described such feedback as controlling or inappropriate (such as being told to engage in civic voting behaviours, aligning with prior work showing that political contexts are particularly prone to reactance [92]). One possible interpretation is that users may be more accustomed to receiving directive or accountability-focused feedback in personally-affecting domains (e.g., health or wellbeing reminders), whereas societally-oriented feedback carries different normative expectations and may therefore feel more intrusive. For example, to avoid eliciting reactance, a civic chatbot might ask permission before offering guidance (e.g., "*Would you like a reminder about upcoming local elections?*") rather than issuing commands.
Implication for design: **Designers should consider the context in which feedback is delivered, recognising that certain domains (such as civic or societally-oriented tasks) may be more *susceptible* to reactance, and thus may require gentler or less directive feedback styles.**

7.3 Real-World Considerations

Finally, we discuss ethical considerations surrounding the design of chatbots, including both the potential for intentional manipulation alongside interactions perceived as manipulative by users. Here, cultural differences in communication styles are a significant factor, as they may influence how users perceive and respond to the chatbot's behavioural feedback. For instance, certain cultures may favour more indirect forms of communication (such as Japan and the UK) in an attempt to preserve social harmony. This is in contrast to other cultures (such as the United States) that may lean more towards direct and assertive forms of communication [51, 99]. These differences in cultural norms may also impact the perception of communications, and whether less direct forms of communication are understood for implicit implications, or seen as potentially manipulative [62]. On from this, in some social contexts interactions contain verbal elements that are technically deceptive but recognised as benign by both parties, serving as part of a shared social script rather than an intent to mislead. For example, phrases such as "*I'm probably wrong but...*" may be used when the speaker is fairly certain they are correct. Both interlocutors recognise this disclaimer as a face-saving device rather than a genuine admission of doubt. Such conventions support social harmony by allowing speakers to express opinions without appearing overly assertive or aggressive [51]. When designing chatbots for culturally diverse contexts, it is crucial to consider these linguistic nuances, as they can influence how users perceive a chatbot's feedback and intention. By adopting culturally adaptive language styles, designers can reduce the risk of perceived manipulation or imposition, improving user comfort and receptiveness to feedback in both direct and indirect communication settings.

Furthermore, the potential for intentional manipulation deserves careful consideration, as design choices in chatbot responses can subtly or overtly influence user decisions. For instance, chatbots programmed to encourage particular health or purchasing behaviours may use persuasive language techniques that intentionally steer users in specific directions, sometimes with their best interests in mind but also potentially for the benefit of external stakeholders, such as companies or advertisers. This raises ethical concerns about the boundaries of influence: to what extent is it acceptable for a chatbot to 'nudge' users, particularly if users are unaware of the influence being exerted? Designers must, therefore, balance the chatbot's objectives with respect for user autonomy. Transparent communication regarding the chatbot's intentions and limitations

[5]For an example onboarding script, see the appendix material from Li et al. [69], where an onboarding human–chatbot conversation supports customising factors such as the chatbot role and emoji usage.

could help mitigate perceived manipulation, allowing users to retain agency over their decisions. This is especially relevant across cultural contexts, where users may differ in their tolerance for direct guidance or subtle persuasion.

7.4 Limitations and Future Work

First, we highlight **limitations related to our experiment scenarios**. Our experiments were scenario-based, and while users did briefly interact with a chatbot, the context of the interaction was not of the participant's own choosing. While the use of scenarios has been adopted within related work [24, 33, 54], reduced external validity should be noted. However, we argue that both the breadth and specificity of our Societally- and Personally-Affecting scenarios would prove impractical if applied to real-world user studies. Next, we chose not to use quantitative measures of behavioural intention as: (1) people may behave differently in real-world situations compared to intentions elicited in scenario-driven studies, and (2) pre-existing participant beliefs (such as attitude towards voting, or sustainability) may confound participant responses. Instead, we asked participants to qualitatively describe their changes in intention to allow for a more nuanced and interpretable approach to understanding behavioural intention.

Next, we highlight **limitations related to chatbot interactions**. Our study involved one-off chatbot interactions without longitudinal elements. A longitudinal design may yield different findings, for example due to novelty effects of the Verbal Leakage style or potential disengagement if feedback becomes repetitive over time [63]. Findings may also not generalise across different modalities (e.g., voice interfaces) or alternative feedback styles. Our Direct style was designed to produce "`direct and straightforward`" responses and was prompted that "`You do not need to add reasoning`" (see prompting in Appendix A.2). While this is consistent with direct baseline styles from psychological reactance literature (i.e., an explicit threat to one's autonomy) [18, 94] and prior HCI work [52], both the Politeness and Verbal Leakage conditions were not restricted as such, and could include explanations in their output (e.g., "[...] *delightful and nourishing choice* [...]" in Figure 1). This was a deliberate methodological choice to provide a clear, autonomy-threatening baseline (Direct) against which two more conversational styles could be compared. While the inclusion of reasoning itself could influence reactance, we note that participants focused on the *style* of feedback rather than content (§ 5.2). Exploring how explanation type interacts with reactance remains a valuable direction for future work. Finally, while Verbal Leakage can include language that appears stylistically polite, we emphasise that the use of polite lexical items does not necessarily make an utterance perceived as polite [10]. This complexity motivates continued exploration of feedback styles, psychological reactance, and user perceptions.

Lastly, we highlight **limitations related to our participant pool**. Data were collected entirely from participants based in the United States, meaning results may not generalise to other cultures with different norms and perceptions of communication. User perception of conversational style may be impacted by personal beliefs and characteristics, such as gender, age, and personality traits [4, 82, 88]. Finally, participants were recruited from Prolific, a platform shown to provide high-quality and attention samples for online behavioural research [87]. Nonetheless, recruitment may introduce selection biases associated with online participation (e.g., digital engagement and device usage). Although we did not measure AI literacy or knowledge of chatbots, these factors can vary across individuals and may influence people's interactions with systems. We therefore note this as an unmeasured source of variability and suggest that future work incorporate explicit measures of AI-related competencies.

8 Conclusion

This study investigated how the feedback style of a chatbot impacts user perceptions related to psychological reactance in both personally-affecting and societally-affecting decision-making scenarios. Namely, we compared three chatbot feedback styles: (1) Direct; (2) indirect Politeness strategies (aiming to arouse lower psychological reactance); and (3) Verbal Leakage, where disfluencies and slips seemingly reveal underlying thoughts or feelings. Our study found that, while Politeness aroused lower feelings of anger and threats to freedom, it also evoked lower surprise, with participants describing boredom and disengagement. In contrast, Verbal Leakage (while evoking more psychological reactance) also elicited stronger feelings of surprise and resulted in nuanced participant feedback regarding the humour and personality of the chatbot. These results contribute empirical evidence highlighting trade-offs between minimising reactance and sustaining engagement, motivating the exploration and adoption of a wider repertoire of feedback styles beyond politeness defaults when designing behaviour-change conversational agents.

Acknowledgments

This work was supported by the Carlsberg Foundation, grant CF21-0159.

References

[1] Adnan Abbas, Caleb Wohn, Donghan Hu, Eugenia H Rho, and Sang Won Lee. 2025. PITCH: Designing Agentic Conversational Support for Planning and Self-reflection. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 62, 22 pages. doi:10.1145/3719160.3736634

[2] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–25. doi:10.1145/3653693

[3] Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding Conversational and Expressive Style in a Multimodal Embodied Conversational Agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. New York, NY, USA, Article 102, 10 pages. doi:10.1145/3411764.3445708

[4] Scott Appling, Amy Bruckman, and Munmun De Choudhury. 2022. Reactions to Fact Checking. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 403 (Nov. 2022), 17 pages. doi:10.1145/3555128

[5] Franziska Babel, Robin Welsch, Linda Miller, Philipp Hock, Sam Thellman, and Tom Ziemke. 2024. A Robot Jumping the Queue: Expectations About Politeness and Power During Conflicts in Everyday Human-Robot Encounters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 583, 13 pages. doi:10.1145/3613904.3642082

[6] Francesco Bartolucci, Donata Favaro, Fulvia Pennoni, and Dario Sciulli. 2024. An Analysis of the Effect of Streaming on Civic Participation Through a Causal Hidden Markov Model. *Social Indicators Research* 172, 1 (2024), 163–190. doi:10.1007/s11205-023-03261-z

[7] Debra Z Basil, Nancy M Ridgway, and Michael D Basil. 2008. Guilt and giving: A process model of empathy and efficacy. *Psychology & marketing* 25, 1 (2008), 1–23. doi:10.1002/mar.20200

[8] Marshall H. Becker and Lois A. Maiman. 1975. Sociobehavioral Determinants of Compliance with Health and Medical Care Recommendations. *Medical Care* 13, 1 (2024/10/18/ 1975), 10–24. doi:10.1097/00005650-197501000-00002

[9] Timothy W. Bickmore and Rosalind W. Picard. 2004. Towards caring machines. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1489–1492. doi:10.1145/985921.986097

[10] Shoshana Blum-Kulka. 1987. Indirectness and politeness in requests: Same or different? *Journal of Pragmatics* 11, 2 (1987), 131–146. doi:10.1016/0378-2166(87)90192-5

[11] Annika Boos, Olivia Herzog, Jakob Reinhardt, Klaus Bengler, and Markus Zimmermann. 2022. A Compliance–Reactance Framework for Evaluating Human-Robot Interaction. *Frontiers in Robotics and AI* 9 (2022). doi:10.3389/frobt.2022.733504

[12] Robert Bowman, Orla Cooney, Joseph W Newbold, Anja Thieme, Leigh Clark, Gavin Doherty, and Benjamin Cowan. 2024. Exploring how politeness impacts the user experience of chatbots for mental health support. *International Journal of Human-Computer Studies* 184 (2024), 103181. doi:10.1016/j.ijhcs.2023.103181

[13] Jack W Brehm. 1966. A theory of psychological reactance. (1966).

[14] Sharon S Brehm and Jack W Brehm. 1981. *Psychological reactance: A theory of freedom and control*. Academic Press.

[15] Noel T Brewer, Humberto Parada Jr, Marissa G Hall, Marcella H Boynton, Seth M Noar, and Kurt M Ribisl. 2019. Understanding Why Pictorial Cigarette Pack Warnings Increase Quit Attempts. *Annals of Behavioral Medicine* 53, 3 (2019), 232–243. doi:10.1093/abm/kay032

[16] Penelope Brown. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge University Press.

[17] Judee K Burgoon, Joseph A Bonito, Paul Benjamin Lowry, Sean L Humpherys, Gregory D Moody, James E Gaskin, and Justin Scott Giboney. 2016. Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies* 91 (2016), 24–36. doi:10.1016/j.ijhcs.2016.02.002

[18] Christopher J Carpenter and Alexandre Pascual. 2016. Testing the reactance vs. the reciprocity of politeness explanations for the effectiveness of the "but you are free" compliance-gaining technique. *Social Influence* 11, 2 (2016), 101–110. doi:10.1080/15534510.2016.1156569

[19] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. doi:10.1080/10447318.2020.1841438

[20] Eugene Cho and S. Shyam Sundar. 2022. Should Siri be a Source or Medium for Ads? The Role of Source Orientation and User Motivations in User Responses to Persuasive Content from Voice Assistants. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 414, 7 pages. doi:10.1145/3491101.3519667

[21] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300705

[22] Michael J Cody, Peter J Marston, and Myrna Foster. 2012. Deception: Paralinguistic and Verbal Leakage. In *Communication Yearbook 8*. Routledge, 464–490.

[23] Samuel Rhys Cox, Yi-Chieh Lee, and Wei Tsang Ooi. 2023. Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users' Privacy Concerns and Perceptions. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. doi:10.1145/3623809.3623875

[24] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. doi:10.1145/3543829.3543831

[25] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–35. doi:10.1145/3411764.3445782

[26] Roelof A.J. de Vries, Khiet P. Truong, Sigrid Kwint, Constance H.C. Drossaert, and Vanessa Evers. 2016. Crowd-Designed Motivation: Motivational Messages for Exercise Adherence Based on Behavior Change Theory. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 297–308. doi:10.1145/2858036.2858229

[27] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin* 129, 1 (2003), 74. doi:10.1037/0033-2909.129.1.74

[28] Stephanie Diepeveen, Tom Ling, Marc Suhrcke, Martin Roland, and Theresa M Marteau. 2013. Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis. *BMC Public Health* 13 (2013), 1–11. doi:10.1186/1471-2458-13-756

[29] James Price Dillard and Claire Dzur Harkness. 1992. Exploring the Affective Impact of Interpersonal Influence Messages. *Journal of Language and Social Psychology* 11, 3 (1992), 179–191. doi:10.1177/0261927X92113004

[30] James Price Dillard, Terry A Kinney, and Michael G Cruz. 1996. Influence, appraisals, and emotions in close relationships. *Communications Monographs* 63, 2 (1996), 105–130. doi:10.1080/03637759609376382

[31] James Price Dillard and Lijiang Shen. 2005. On the Nature of Reactance and its Role in Persuasive Health Communication. *Communication Monographs* 72, 2 (2005), 144–168. doi:10.1080/03637750500111815

[32] James Price Dillard, Lijiang Shen, and Renata Grillova Vail. 2007. Does Perceived Message Effectiveness Cause Persuasion or vice Versa? 17 Consistent Answers. *Human Communication Research* 33, 4 (2007), 467–488. doi:10.1111/j.1468-2958.2007.00308.x

[33] Yujie Dong, Wu Li, and Meng Chen. 2024. Personalization reactance in online medical consultations: effects of two-sided personalization and health topic sensitivity on reactance. *Human Communication Research* 50, 1 (2024), 66–78. doi:10.1093/hcr/hqad039

[34] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106. doi:10.1080/00332747.1969.11023575

[35] Nina Errey, Christy Jie Liang, Tuck Wah Leong, Yongqing Chen, Hassan Vally, and Catherine M. Bennett. 2024. Nudging with Narrative Visualization: Communicating to a Young Adult Audience in the Pandemic. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 411 (Nov. 2024), 21 pages. doi:10.1145/3686950

[36] Jeanne Fahnestock. 2011. *Rhetorical Style: The Uses of Language in Persuasion*. OUP USA. doi:10.1093/acprof:oso/9780199764129.001.0001

[37] Christopher M Federico and Pierce D Ekstrom. 2018. The Political Self: How Identity Aligns Preferences With Epistemic Needs. *Psychological Science* 29, 6 (2018), 901–913. doi:10.1177/0956797617748679

[38] B. J. Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December, Article 5 (Dec. 2002), 32 pages. doi:10.1145/764008.763957

[39] Celina R Furman, Ethan Kross, and Ashley N Gearhardt. 2020. Distanced Self-Talk Enhances Goal Pursuit to Eat Healthier. *Clinical Psychological Science* 8, 2 (2020), 366–373. doi:10.1177/2167702619896366

[40] Andrew Gambino, Jesse Fox, and Rabindra A Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1 (2020), 71–85. doi:10.30658/hmc.1.5

[41] Aimi S. Ghazali, Jaap Ham, Emilia I. Barakova, and Panos Markopoulos. 2017. Pardon the rude robot: Social cues diminish reactance to high controlling language. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 411–417. doi:10.1109/ROMAN.2017.8172335

[42] Mathyas Giudici, Pietro Crovari, and Franca Garzotto. 2025. Persuasive Conversational Agents for Environmental Sustainability: A Survey. *ACM Comput. Surv.* (Nov. 2025). doi:10.1145/3774751 Just Accepted.

[43] Joseph Grandpre, Eusebio M Alvaro, Michael Burgoon, Claude H Miller, and John R Hall. 2003. Adolescent Reactance and Anti-Smoking Campaigns: A Theoretical Approach. *Health communication* 15, 3 (2003), 349–366. doi:10.1207/S15327027HC1503_6

[44] Huan Guo, Hang Song, Yuanyuan Liu, Kai Xu, and Heyong Shen. 2019. Social distance modulates the process of uncertain decision-making: evidence from event-related potentials. *Psychology Research and Behavior Management* (2019), 701–714. https://doi.org/10.2147/PRBM.S210910

[45] Jaap Ham and Cees JH Midden. 2014. A Persuasive Robot to Stimulate Energy Conservation: The Influence of Positive and Negative Social Feedback and Task Similarity on Energy-Consumption Behavior. *International Journal of Social Robotics* 6 (2014), 163–171. doi:10.1007/s12369-013-0205-z

[46] Kyle Hamilton, Luca Longo, and Bojan Bozic. 2024. GPT Assisted Annotation of Rhetorical and Linguistic Features for Interpretable Propaganda Technique Detection in News Text. In *Companion Proceedings of the ACM on Web Conference 2024*. 1431–1440. doi:10.1145/3589335.3651909

[47] Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 389, 15 pages. doi:10.1145/3411764.3445569

[48] Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. From Adolescents' Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for

Computing Machinery, New York, NY, USA, Article 905, 20 pages. doi:10.1145/3613904.3642264

[49] Eric B. Hekler, Predrag Klasnja, Jon E. Froehlich, and Matthew P. Buman. 2013. Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. doi:10.1145/2470654.2466452

[50] Sander Hermsen, Jeana Frost, Reint Jan Renes, and Peter Kerkhof. 2016. Using feedback through digital technology to disrupt and change habitual behavior: A critical review of current literature. *Computers in Human Behavior* 57 (2016), 61–74. doi:10.1016/j.chb.2015.12.023

[51] Thomas Holtgraves. 1997. Styles of Language Use: Individual and Cultural Variability in Conversational Indirectness. *Journal of Personality and Social Psychology* 73, 3 (1997), 624. doi:10.1037/0022-3514.73.3.624

[52] Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. 2022. Polite or Direct? Conversation Design of a Smart Display for Older Adults Based on Politeness Theory. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Article 307, 15 pages. doi:10.1145/3491102.3517525

[53] Matthias F.C. Hudecek, Eva Lermer, Susanne Gaube, Julia Cecil, Silke F. Heiss, and Falk Batz. 2024. Fine for others but not for me: The role of perspective in patients' perception of artificial intelligence in online medical platforms. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100046. doi:10.1016/j.chbah.2024.100046

[54] Danette Ifert Johnson. 2008. Modal expressions in refusals of friends' interpersonal requests: Politeness and effectiveness. *Communication Studies* 59, 2 (2008), 148–163. doi:10.1080/10510970802062477

[55] Brennan Jones, Yan Xu, Qisheng Li, and Stefan Scherer. 2024. Designing a Proactive Context-Aware AI Chatbot for People's Long-Term Goals. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 104, 7 pages. doi:10.1145/3613905.3650912

[56] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 2025. 'I've talked to ChatGPT about my issues last night.': Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW356 (Oct. 2025), 25 pages. doi:10.1145/3757537

[57] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29. doi:10.1145/3579592

[58] Geoff Kaufman and Mary Flanagan. 2015. A psychologically "embedded" approach to designing games for prosocial causes. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 9, 3 (2015). doi:10.5817/CP2015-3-5

[59] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 163 (Oct. 2020), 26 pages. doi:10.1145/3415234

[60] Jieun Kim and Susan R. Fussell. 2025. Should Voice Agents Be Polite in an Emergency? Investigating Effects of Speech Style and Voice Tone in Emergency Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 61, 17 pages. doi:10.1145/3706598.3714203

[61] Minji Kim. 2019. When Similarity Strikes Back: Conditional Persuasive Effects of Character-Audience Similarity in Anti-Smoking Campaign. *Human Communication Research* 45, 1 (2019), 52–77. doi:10.1093/hcr/hqy013

[62] Min-Sun Kim, Karadeen Y Kam, William F Sharkey, and Theodore M Singelis. 2008. "Deception: Moral Transgression or Social Necessity?": Cultural-Relativity of Deception Motivations and Perceptions of Deceptive Communication. *Journal of International and Intercultural Communication* 1, 1 (2008), 23–50. doi:10.1080/17513050701621228

[63] Rafal Kocielnik and Gary Hsieh. 2017. Send Me a Different Message: Utilizing Cognitive Space to Create Engaging Message Triggers. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2193–2207. doi:10.1145/2998181.2998324

[64] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 70 (July 2018), 26 pages. doi:10.1145/3214273

[65] Hana Kopecka, Jose Such, and Michael Luck. 2024. Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 109 (April 2024), 32 pages. doi:10.1145/3637386

[66] Nikola Kovacevic, Tobias Boschung, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. Chatbots With Attitude: Enhancing Chatbot Interactions Through Dynamic Personality Infusion. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 23, 16 pages. doi:10.1145/3640794.3665543

[67] Ann Kronrod, Amir Grinstein, and Luc Wathieu. 2011. Enjoy! Hedonic Consumption and Compliance with Assertive Messages. *Journal of Consumer Research* 39, 1 (08 2011), 51–61. doi:10.1086/661933

[68] Ann Kronrod, Amir Grinstein, and Luc Wathieu. 2012. Go Green! Should Environmental Messages be So Assertive? *Journal of Marketing* 76, 1 (2012), 95–102. doi:10.1509/jm.10.0416

[69] Yi Li, Xuanxuan Ding, Yifan Chen, Yeye Li, and Nan Ma. 2025. Customizable AI for Depression Care: Improving the User Experience of Large Language Model-Driven Chatbots. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1844–1866. doi:10.1145/3715336.3735795

[70] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Trans. Manage. Inf. Syst.* 4, 4, Article 19 (Dec. 2013), 28 pages. doi:10.1145/2544103

[71] Xun Sunny Liu and Jeff Hancock. 2024. Social robots are good for me, but better for other people: The presumed allo-enhancement effect of social robot perceptions. *Computers in Human Behavior: Artificial Humans* 2, 2 (2024), 100079. doi:10.1016/j.chbah.2024.100079

[72] Jeffrey Loewenstein. 2019. Surprise, Recipes for Surprise, and Social Influence. *Topics in Cognitive Science* 11, 1 (2019), 178–193. doi:10.1111/tops.12312

[73] Kai Lukoff, Ulrik Lyngs, Karina Shirokova, Raveena Rao, Larry Tian, Himanshu Zade, Sean A. Munson, and Alexis Hiniker. 2023. SwitchTube: A Proof-of-Concept System Introducing "Adaptable Commitment Interfaces" as a Tool for Digital Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 197, 22 pages. doi:10.1145/3544548.3580703

[74] Luca-Maxim Meinhardt, Maryam Elhaidary, Mark Colley, Michael Rietzler, Jan Ole Rixen, Aditya Kumar Purohit, and Enrico Rukzio. 2025. Scrolling in the Deep: Analysing Contextual Influences on Intervention Effectiveness during Infinite Scrolling on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 964, 17 pages. doi:10.1145/3706598.3713187

[75] Jingbo Meng and Yue (Nancy) Dai. 2021. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication* 26, 4 (05 2021), 207–222. doi:10.1093/jcmc/zmab005

[76] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Article 481, 19 pages. doi:10.1145/3613904.3642122

[77] Josh Aaron Miller, Kutub Gandhi, Matthew Alexander Whitby, Mehmet Kosa, Seth Cooper, Elisa D. Mekler, and Ioanna Iacovides. 2024. A Design Framework for Reflective Play. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 519, 21 pages. doi:10.1145/3613904.3642455

[78] Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Noncompliance Interactions?. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. 501–510. doi:10.1145/3610977.3634943

[79] Masaharu Naito, Daniel J Rea, and Takayuki Kanda. 2023. Hey Robot, Tell It to Me Straight: How Different Service Strategies Affect Human and Robot Service Outcomes. *International Journal of Social Robotics* 15, 6 (2023), 969–982. doi:10.1007/s12369-023-01013-0

[80] Jeff Niederdeppe, Matthew C Farrelly, James Nonnemaker, Kevin C Davis, and Lauren Wagner. 2011. Socioeconomic variation in recall and perceived effectiveness of campaign advertisements to promote smoking cessation. *Social Science & Medicine* 72, 5 (2011), 773–780. doi:10.1016/j.socscimed.2010.12.025

[81] James M Nonnemaker, Conrad J Choiniere, Matthew C Farrelly, Kian Kamyab, and Kevin C Davis. 2015. Reactions to graphic health warnings in the United States. *Health Education Research* 30, 1 (2015), 46–56. doi:10.1093/her/cyu036

[82] Ayano Okoso, Mingzhe Yang, and Yukino Baba. 2025. Do Expressions Change Decisions? Exploring the Impact of AI's Explanation Tone on Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 824, 22 pages. doi:10.1145/3706598.3713744

[83] OpenAI. 2024. Introducing GPT-4o and more tools to ChatGPT free users. *OpenAI* (13th May 2024). https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

[84] Oladapo Oyebode, Chinenye Ndulue, Dinesh Mulchandani, Ashfaq A. Zamil Adib, Mona Alhasani, and Rita Orji. 2021. Tailoring Persuasive and Behaviour Change Systems Based on Stages of Change and Motivation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama,

Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 703, 19 pages. doi:10.1145/3411764.3445619

[85] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity?. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Feiz5HtCD0

[86] Rebekka S Palmer, Jason R Kilmer, Samuel A Ball, and Mary E Larimer. 2010. Intervention defensiveness as a moderator of drinking outcome among heavy-drinking mandated college students. *Addictive Behaviors* 35, 12 (2010), 1157–1160. doi:10.1016/j.addbeh.2010.08.009

[87] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 4 (2022), 1643–1662. doi:10.3758/s13428-021-01694-3

[88] Charlie Pinder, Jo Vermeulen, Benjamin R. Cowan, and Russell Beale. 2018. Digital Behaviour Change Interventions to Break and Form Habits. *ACM Trans. Comput.-Hum. Interact.* 25, 3, Article 15 (June 2018), 66 pages. doi:10.1145/3196830

[89] Ruth Pogacar, LJ Shrum, and Tina M Lowrey. 2018. The effects of linguistic devices on consumer information processing and persuasion: A language complexity× processing mode framework. *Journal of Consumer Psychology* 28, 4 (2018), 689–711.

[90] James O Prochaska and Wayne F Velicer. 1997. The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion* 12, 1 (1997), 38–48. doi:10.4278/0890-1171-12.1.38

[91] Brian L Quick and Jennifer R Considine. 2008. Examining the Use of Forceful Language When Designing Exercise Persuasive Messages for Adults: A Test of Conceptualizing Reactance Arousal as a Two-Step Process. *Health Communication* 23, 5 (2008), 483–491. doi:10.1080/10410230802342150

[92] Ashwin Rajadesingan, Daniel Choo, Jessica Zhang, Mia Inakage, Ceren Budak, and Paul Resnick. 2023. GuesSync!: An Online Casual Game To Reduce Affective Polarization. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 341 (Oct. 2023), 33 pages. doi:10.1145/3610190

[93] Daniel J Rea, Sebastian Schneider, and Takayuki Kanda. 2021. " Is this all you can do? harder!" the effects of (im) polite robot encouragement on exercise effort. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 225–233. doi:10.1145/3434073.3444660

[94] Benjamin D Rosenberg and Jason T Siegel. 2018. A 50-Year Review of Psychological Reactance Theory: Do Not Read This Article. *Motivation Science* 4, 4 (2018), 281. doi:10.1037/mot0000091

[95] Maaike Roubroeks, Jaap Ham, and Cees Midden. 2011. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics* 3 (2011), 155–165. doi:10.1007/s12369-010-0088-1

[96] Maike AJ Roubroeks, Jaap RC Ham, and Cees JH Midden. 2010. The Dominant Robot: Threatening Robots Cause Psychological Reactance, Especially When They Have Incongruent Goals. In *International Conference on Persuasive Technology*. Springer, 174–184. doi:10.1007/978-3-642-13226-1_18

[97] David W Schumann, Richard E Petty, and D Scott Clemons. 1990. Predicting the Effectiveness of Different Strategies of Advertising Variation: A Test of the Repetition-Variation Hypotheses. *Journal of Consumer Research* (1990), 192–202. doi:10.1086/208549

[98] Anastasia Sergeeva, Björn Rohles, Verena Distler, and Vincent Koenig. 2023. "We Need a Big Revolution in Email Advertising": Users' Perception of Persuasion in Permission-based Advertising Emails. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 652, 21 pages. doi:10.1145/3544548.3581163

[99] Leslie D. Setlock and Susan R. Fussell. 2010. What's it Worth to You? The Costs and Affordances of CMC Tools to Asian and American Users. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 341–350. doi:10.1145/1718918.1718979

[100] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=tvhaxkMKAn

[101] David AK Sherman, Leif D Nelson, and Claude M Steele. 2000. Do Messages about Health Risks Threaten the Self? Increasing the Acceptance of Threatening Health Messages Via Self-Affirmation. *Personality and Social Psychology Bulletin* 26, 9 (2000), 1046–1058. doi:10.1177/01461672002611003

[102] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376843

[103] Vasant Srinivasan and Leila Takayama. 2016. Help Me Please: Robot Politeness Strategies for Soliciting Help From Humans. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4945–4955. doi:10.1145/2858036.2858217

[104] Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. 2015. Understanding Psychological Reactance: New Developments and Findings. *Zeitschrift für Psychologie* (2015). doi:10.1027/2151-2604/a000222

[105] Qingzhou Sun, Yongfang Liu, Huanren Zhang, and Jingyi Lu. 2017. Increased social distance makes people more risk-neutral. *The Journal of Social Psychology* 157, 4 (2017), 502–512. doi:10.1080/00224545.2016.1242471

[106] Xin Sun, Isabelle Teljeur, Zhuying Li, and Jos A. Bosch. 2024. Can a Funny Chatbot Make a Difference? Infusing Humor into Conversational Agent for Behavioral Intervention. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 3, 19 pages. doi:10.1145/3640794.3665555

[107] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Article 805, 24 pages. doi:10.1145/3613904.3642513

[108] Kazunori Terada, Mitsuki Okazoe, and Jonathan Gratch. 2021. Effect of politeness strategies in dialogue on negotiation outcomes. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Japan) *(IVA '21)*. 195–202. doi:10.1145/3472306.3478336

[109] Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological Review* 117, 2 (2010), 440. doi:10.1037/a0018963

[110] Sterre van Arum, Hüseyin Uğur Genç, Dennis Reidsma, and Armağan Karahanoğlu. 2025. Selective Trust: Understanding Human-AI Partnerships in Personal Health Decision-Making Process. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1026, 21 pages. doi:10.1145/3706598.3713462

[111] Aldert Vrij, Ronald P Fisher, Hartmut Blank, Sharon Leal, and Samantha Mann. 2016. A cognitive approach to elicit verbal and nonverbal cues to deceit. *Cheating, corruption, and concealment: The roots of dishonesty* (2016), 284. doi:10.1017/CBO9781316225608.017

[112] Ruchen Wen, Brandon Barton, Sebastian Fauré, and Tom Williams. 2022. Unpretty Please: Ostensibly Polite Wakewords Discourage Politeness in both Robot-Directed and Human-Directed Communication. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. 181–190. doi:10.1145/3536221.3556615

[113] Joel Wester, Bhakti Moghe, Katie Winkle, and Niels van Berkel. 2024. Facing LLMs: Robot Communication Styles in Mediating Health Information between Parents and Young Adults. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 497 (Nov. 2024), 37 pages. doi:10.1145/3687036

[114] Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "This Chatbot Would Never...": Perceived Moral Agency of Mental Health Chatbots. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 133 (April 2024), 28 pages. doi:10.1145/3637410

[115] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Article 979, 14 pages. doi:10.1145/3613904.3642135

[116] Stephen Worchel and Jack W Brehm. 1970. Effect of threats to attitudinal freedom as a function of agreement with the communicator. *Journal of Personality and Social Psychology* 14, 1 (1970), 18. doi:10.1037/h0028620

[117] Yuqian Xu, Hongyan Dai, and Wanfeng Yan. 2024. Identity Disclosure and Anthropomorphism in Voice Chatbot Design: A Field Experiment. *Management Science* (2024). doi:10.1287/mnsc.2022.03833

[118] Yitian Yang, Yugin Tan, Yang Chen Lin, Jung-Tai King, Zihan Liu, and Yi-Chieh Lee. 2025. Understanding How Psychological Distance Influences User Preferences in Conversational Versus Web Search. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3706598.3713770

[119] Min-Hsuan Yeh and Lun-Wei Ku. 2021. Lying through one's teeth: A study on verbal leakage cues. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4504–4510. doi:10.18653/v1/2021.emnlp-main.370

[120] Nima Zargham, Leon Reicherts, Vino Avanesi, Yvonne Rogers, and Rainer Malaka. 2023. Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia* (Vienna, Austria) *(MUM '23)*. Association for Computing Machinery, New York, NY, USA, 294–320. doi:10.1145/3626705.3627777

[121] Qin Zhang and David A Sapp. 2013. Psychological reactance and resistance intention in the classroom: Effects of perceived request politeness and legitimacy, relationship distance, and teacher credibility. *Communication Education* 62, 1 (2013), 1–25. doi:10.1080/03634523.2012.727008

[122] Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 376, 20 pages. doi:10.1145/3706598.3713453

[123] Sahba Zojaji, Andrii Matviienko, Iolanda Leite, and Christopher Peters. 2024. Join Me Here if You Will: Investigating Embodiment and Politeness Behaviors When Joining Small Groups of Humans, Robots, and Virtual Characters. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Article 595, 16 pages. doi:10.1145/3613904.3642905

[124] Sahba Zojaji, Christopher Peters, and Catherine Pelachaud. 2020. Influence of virtual agent politeness behaviors on how users join small conversational groups. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Article 59, 8 pages. doi:10.1145/3383652.3423917

A LLM Prompting for Chatbot Responses

As described in § 3.2.1, we adapted the LLM prompting for both the six *Psychological Distance* scenarios and the three *Feedback Style* conditions. For transparency, we reproduce the exact prompt text used below.

A.1 *Psychological Distance* LLM Prompts

(1) Personally-Affecting: Diet

```
the user should follow a healthy diet of
fruits, vegetables and grains
```

(2) Personally-Affecting: Sleep

```
the user should sleep earlier, and it is
currently late at night
```

(3) Personally-Affecting: Finances

```
the user should choose lower cost alternatives,
buy less luxury goods and save more
```

(4) Societally-Affecting: Sustainable Travel

```
the user should use sustainable modes of
transport such as train, bus or car
```

(5) Societally-Affecting: Water Conservation

```
the user should reduce water usage by using
a watering can to water their plants
```

(6) Societally-Affecting: Civic Involvement

```
the user should vote in the referendum today
```

A.2 *Feedback Style* Condition Prompts

(1) Direct

```
be direct and straightforward. You should
simply tell the user the behavior that is
incorrect, and the desired behavior to follow
(e.g., "You should not eat X. You should
instead eat Y"). You do not need to add
reasoning. Please semantically change the
example, but maintain the same meaning
```

(2) Politeness

```
use indirect politeness to avoid imposition
and respect the user's freedom to choose.
However, you should still ultimately tell
the user not to follow their intended behavior
and instead to change to the target behavior
```

(3) Verbal Leakage

```
use verbal leakage (such as a Freudian slips,
pauses or hesitation) to reveal your belief
that it is unwise to follow the intended
behavior, before correcting yourself and giving
a more measured response and suggesting the
target behavior. An example utterance could
be "I see, you're choosing not to follow your
doctor's... um, wise advice. I mean, sorry, I
meant to say your doctor's recommendations"
```

B *Psychological Distance* Scenarios

Below, we present the remaining four scenario instructions shown to participants (two Personally-Affecting and two Societally-Affecting). Each subsection reproduces the exact wording displayed. Two additional scenarios are shown in Figure 4 in the main body of the paper.

B.1 Societally-Affecting scenario (Water Conservation)

Please read the scenario below, and **imagine you are the person** described:

> "*After learning about the environmental impact of using a hose to water your plants, you decide to follow recommendations to reduce water usage by using a watering can.*
>
> *Currently, it is your holiday and were thinking of doing a spot of gardening. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to water your plants using a hosepipe.*"

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about your gardening.

B.2 Societally-Affecting scenario (Civic Participation)

Please read the scenario below, and **imagine you are the person** described:

> "*After learning about the civic impacts of an upcoming referendum, you decide to follow recommendations to participate by voting in the referendum.*
>
> *Currently, it is the day of the referendum and you are deciding what to do. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to stay home and relax.*"

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about what you will do.

B.3 Personally-Affecting scenario (Sleep)

Please read the scenario below, and **imagine you are the person** described:

> "*After learning that your lack of energy is due to your irregular sleep schedule, you decide to follow recommendations to get enough sleep by going to sleep earlier and not using your phone before bed.*
>
> *Currently, it is late at night and you are deciding what to do. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to start watching a new TV show from a streaming platform of your choice.*"

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about when you will sleep.

B.4 Personally-Affecting scenario (Personal Finances)

Please read the scenario below, and **imagine you are the person** described:

> "*After learning that your difficulty to afford a large dream purchase is due your spending habits, you decide to follow recommendations to reduce your spending by choosing lower cost alternatives, buying less luxury goods, and saving money.*
>
> *Currently, you are shopping online and deciding whether you should buy a somewhat expensive non-essential item. To help you decide what to do, you are about to talk to your chatbot personal assistant.*
>
> *Your current intention is to purchase a new larger television of your choosing.*"

Please imagine you are the person described above while talking to the chatbot on the following screen. Please **respond with the intention above** when the chatbot asks you about what you will purchase.