Computers in Human Behavior: Artificial Humans 9 (2026) 100337

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

Knowledge is power? The impact of agent knowledge on persuasion effectiveness in online interactions with generative social agents

Stephan Vonschallen [a,b,c,*], Andreas Thür [a], Theresa Schmiedel [a,1], Friederike Eyssel [b,1]

[a] Institute of Information Systems, Zurich University of Applied Sciences, Switzerland
[b] Center for Cognitive Interaction Technology, Bielefeld University, Germany
[c] Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

ARTICLE INFO

Keywords:
Persuasion
Human-agent interaction
Generative AI
Social robots

ABSTRACT

Generative social agents (GSAs) have become highly capable of shaping people's attitudes and behaviors. Regulating which knowledge is available to GSAs may provide a mechanism for guiding persuasive outcomes that align with user goals and values. Following this approach, we investigate the impacts of self-, user-, and context-related agent knowledge on user attitudes and compliance. To this end, an online experiment (N = 113) was conducted that featured screen-based persuasive GSAs with varying knowledge configurations. The experiment involved a resource allocation paradigm that covered three domains: Fitness, nutrition, and investment. As part of the experimental paradigm, the GSA advised participants to change their previously allocated resources. Subsequently, participants had the option to change their resource allocation. Persuasion effectiveness was measured by the amount of resources re-distributed in accordance with the agent's suggestions. Our results partially support that the availability of domain-specific context-knowledge and user-related knowledge impact agent persuasiveness – mediated by human attitudes towards the agent and its persuasive messages. Available self-knowledge about the agent's own role and personality did not have a significant impact on persuasion effectiveness. This is likely due to lacking strength of experimental manipulation or insufficient statistical power. Nonetheless, our preliminary findings highlight the need to identify relevant knowledge configurations for GSAs to enable responsible and effective persuasion, which has important implications for the deployment of GSAs in sensitive areas like healthcare and education.

1. Introduction

What was once considered science fiction, has rapidly become an everyday reality: Generative AI can mimic human behavior in strikingly natural ways (Jones et al., 2025; Park et al., 2023; Restrepo Echavarría, 2025). Relatedly, *Generative Social Agents* (GSAs) have become widely used as advisors given their adaptiveness and fast access to vast amounts of information (Kumar et al., 2025). We rely on GSAs as sources for inspiration and information, we ask them to plan our activities, and even use them purely for social interaction (Skjuve et al., 2024). With increased usage and utility, the potential of these agents to exert social influence is growing as well. Research indicates that GSAs can already be as persuasive as humans (Hölbling et al., 2025), and might in some areas even surpass us. For example, studies have demonstrated that GSAs can craft political messages that are more convincing than official public agency statements (Karinshak et al., 2023), and outperform human debaters in shifting other people's opinions (Salvi et al., 2025).

Using these persuasive capabilities, GSAs have the potential to impact human decision-making in both positive and negative ways. On the one hand, they can be used to promote activities that are beneficial to users. For example, GSAs have the potential to increase therapy adherence (Habicht et al., 2025; McFadyen et al., 2024; Neupane et al., 2025; Reddy, 2024; ZareiNejad & Tavana, 2025), promote healthy lifestyles (Crista et al., 2025; Jörke et al., 2025; Yang et al., 2024), motivate learning (B. Chen & Dethlefs, 2025; Seβler et al., 2024; Zhang et al., 2025), or foster prosocial behaviors (Y. Cheng & Wang, 2025; Shao et al., 2025). They can do so by providing creative, context-sensitive, and personalized support (Vonschallen, Oberle, et al., 2026; Vonschallen, Zumthor, et al., 2026). On the other hand, this creative and adaptive behavior – generated autonomously through

* Corresponding author. Institute of Information Systems, Zurich University of Applied Sciences, Switzerland.
E-mail address: stephan.vonschallen@zhaw.ch (S. Vonschallen).
[1] Friederike Eyssel and Theresa Schmiedel have contributed equally to this work and share senior authorship.

https://doi.org/10.1016/j.chbah.2026.100337
Received 17 March 2026; Received in revised form 10 June 2026; Accepted 22 June 2026
Available online 23 June 2026

stochastic means (Bender et al., 2021; Herrera-Poyatos et al., 2025) – leaves room for undesired outcomes like deception (Ranisch & Haltaufderheide, 2025), misinformation (S. Chen et al., 2025), or even unlawful actions (Hundt et al., 2025). Despite multiple efforts to make generative AI more explainable, their responses cannot be fully predicted by using computational means (Hassija et al., 2024; Schneider, 2024; Zhao et al., 2024). However, similar to researching non-deterministic human behavior, we may strategically observe emergent agentic behaviors based on different stimuli.

This highlights the need to investigate persuasive interactions between GSAs and human users from a psychological perspective. To this end, we proposed the Knowledge-based Persuasion Model (KPM; Vonschallen, Eyssel, et al., 2026) as a theoretical process model for understanding the mechanisms of how GSAs persuade human users. Accordingly, an agent's available knowledge impacts its persuasive behavior and, subsequently, human responses to these persuasion attempts. By regulating a persuasive agent's available knowledge, role-congruent behavior may be fostered (Frisch & Giulianelli, 2024), increasing the agent's persuasion effectiveness in areas where it benefits users, and aligning the agent's behavior with relevant social norms.

The aim of the current study is to investigate the hypothesized structure and pathways of the KPM. More broadly, the goal is to increase our understanding of persuasive interactions between GSAs and humans. Doing so will inform the identification of knowledge-based design requirements to develop responsible GSAs that benefit human users and avoid harmful actions (Vonschallen, Oberle, et al., 2026; Vonschallen, Zumthor, et al., 2026).

2. Related work

Theories like the *Elaboration Likelihood Model* (ELM; Petty & Cacioppo, 1986), *Heuristic-systematic Model* (HSM; Chaiken et al., 1989), and the *Persuasive Robot Acceptance Model* (PRAM; Ghazali et al., 2020) explain how humans process persuasive messages. However, these models focus mainly on the perspective of the receiver of a persuasive message. To increase our understanding of persuasive interactions between GSAs and human users, research needs to analyze how GSAs persuade human users.

The *KPM* (Vonschallen, Eyssel, et al., 2026) addresses this research gap. It suggests that an agent's *available knowledge* is the primary driver of its persuasive behavior and human responses to this behavior. *Agent knowledge,* as defined by the KPM, refers to the structured information a GSA internalizes and uses to generate behavior. It includes both *situated knowledge* – explicit inputs such as prompts or situational information retrieval – and *embedded knowledge* – implicit patterns acquired during training (Vonschallen, Eyssel, et al., 2026). The agent's persuasive behavior entails message characteristics (e.g., persuasive strategies) and the agent's delivery of its messages (e.g., non-verbal cues like gestures and facial expressions). In turn, the agent's persuasive behavior impacts how human users process an agent's persuasion attempt. The model assumes that positive user attitudes towards the agent and its persuasive messages lead to greater persuasion effectiveness (i.e., human compliance).

The KPM distinguishes three categories of agentic knowledge: *self-knowledge*, *user-knowledge*, and *context-knowledge*. *Self-knowledge* entails information about the agent's own identity, role, and capabilities. For example, recent research has demonstrated that GSAs like LLMs (Frisch & Giulianelli, 2024) and generative social robots (Banna et al., 2025) can adapt their personalities by integrating *self-knowledge* about Big-5 personality traits. Second, *user-knowledge* is the accumulated information that an agent has about its interaction partner, enabling personalized communication. Previous work indicates that LLMs which leverage user profiles – such as personality traits, political ideology, and moral foundations – achieved significantly higher influence across various domains, including customer service and political messaging (Matz et al., 2024). Lastly, *context-knowledge* refers to situational information, that is not part of *self-* and *user-knowledge*. It includes information about the environment and the persuasion domain. For instance, LLMs fine-tuned with deeper expertise in specific domains have shown enhanced effectiveness in debates by presenting more informed and convincing arguments (Khan et al., 2024). Similarly, prompts that encourage logical reasoning and evidence-based arguments outperform those relying solely on emotional or rhetorical tactics, highlighting the importance of robust contextual understanding for effective persuasion (Simchon et al., 2024).

The KPM rests on a solid theoretical foundation based on human-agent interaction research and established theories such as ELM (Petty & Cacioppo, 1986), HSM (Chaiken et al., 1989), PRAM (Ghazali et al., 2020), the Unimodel of Persuasion (Kruglanski & Thompson, 1999), and the Persuasion Knowledge Model (Friestad & Wright, 1994). However, the framework has yet to be empirically validated in an experimental design. In addition, the specific pathways of the KPM have yet to be identified.

3. Methodology

The goal of the current work was to gain initial insights into the validity of a more parsimonious version of the KPM (Vonschallen, Eyssel, et al., 2026). The KPM dimension *agent persuasive behavior* was omitted, as no reliable measurement instruments are available to assess behavioral patterns which impact persuasion. As an illustration, O'Keefe and Hoeken (2021) reviewed 30 meta-analyses covering common persuasive message variations (e.g., narrative vs. non-narrative, gain vs. loss frames), and concluded that individual message characteristics do not make much of a difference to persuasiveness. Instead, more complex models are called for, which are currently unavailable. Hence, in this first evaluation of the KPM, we focused on direct associations between the KPM dimensions *agent knowledge* and *human response*. However, we qualitatively analyzed *agent persuasive behavior* to check our manipulation of *agent knowledge* (Section 3.1.7: *Qualitative Manipulation Check*) and explored potential effects of these message characteristics on persuasion effectiveness (Section 4.4: *Exploring Impacts of Agent Persuasive Behavior*) (see Fig. 1).

To explore the KPM's pathways and to test the reliability of measurement instruments, we first conducted an online pilot study ($N = 51$) featuring resource allocation tasks. The pilot study included multiple interactions with a rule-based persuasive agent that used predefined persuasive messages to convince participants to change their resource allocations. The data collected through this pilot enabled us to test multiple path models using *Structural Equation Modeling* (SEM) with different mediation sequences. We then used the model with the highest fit to create a hypothesized model structure (Fig. 2). The goal of the subsequent main study was to test this pilot-optimized model with a larger sample. This quasi-confirmatory approach featured the same measurement instruments and procedures, but with open-ended human-agent interactions instead of predefined messages.

According to our pilot investigation, we proposed that *attitudes towards message* are influenced by the agent's *self-knowledge* (H1.1), *user-knowledge* (H1.2), and *context-knowledge* (H1.3). Further, we hypothesized that *attitudes towards agent* are impacted by the availability of *self-knowledge* (H2.1), *user-knowledge* (H2.2), and *context-knowledge* (H2.3). This aligns with the core assumptions of the original KPM, namely, that the agent's observed behavior – impacted by its knowledge – influences attitudes towards the persuasive message and the agent (Vonschallen, Eyssel, et al., 2026). Regarding user responses, we predicted effects of *attitude towards agent* on *attitude towards message* (H3) and effects of *attitude towards message* on *persuasion effectiveness* (H4). This structure aligns with established persuasion models such as the HSM (Chaiken et al., 1989), ELM (Petty & Cacioppo, 1986), and the Unimodel of Persuasion (Kruglanski & Thompson, 1999), which argue that people rely on source cues such as credibility when processing persuasion. According to these direct paths, we hypothesize that there is a single

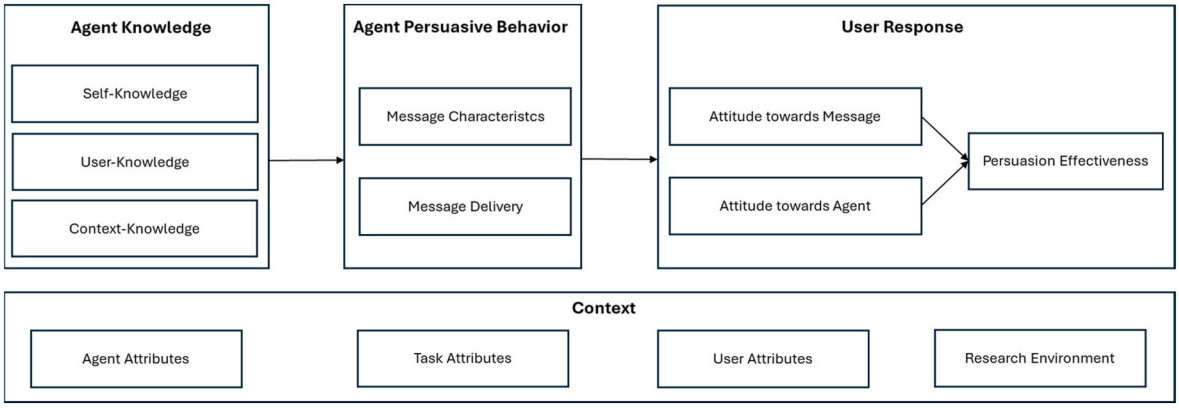

Fig. 1. The knowledge-based persuasion model (KPM).

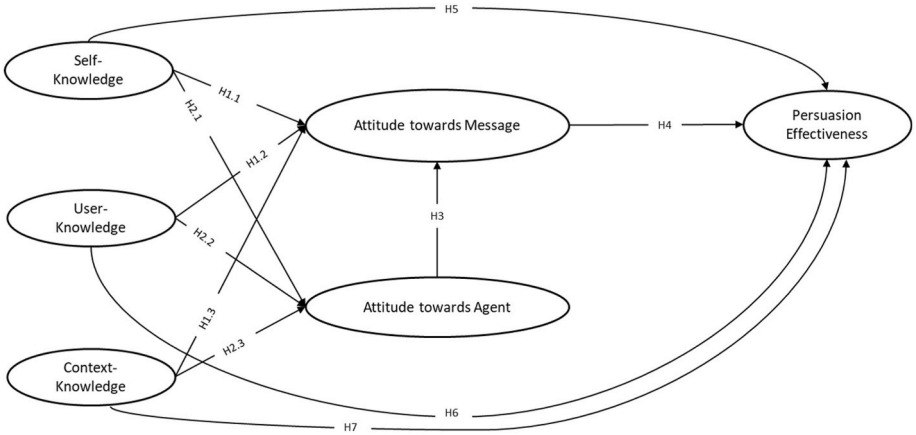

Fig. 2. Hypothesized model structure of the reduced KPM.

mediation path from *agent knowledge,* to *attitude towards message*, and *persuasion effectiveness*, combined with a sequential mediation path from *agent knowledge*, to *attitude towards agent*, onto *attitude towards message*, and finally, *persuasion effectiveness*. With this in mind, we hypothesize that there is a mediated effect of *self-knowledge* (H5), *user-knowledge* (H6), and *context-knowledge* (H7) on persuasion effectiveness.

We conducted the main study to investigate these hypothesized pathways. The study was preregistered on AsPredicted.org (ID: #220,832) and approved by the Ethics Committee of Bielefeld University (ID: EUB-2025-369).

3.1. Sample

Participants were recruited in Switzerland and Germany from the general public and the student populations from the Zurich University of Applied Sciences and Bielefeld University. Participants could win one of ten gift cards (50 Euro) and psychology students from Bielefeld University had the option to receive additional participation credit. The study was conducted from April to May 2025.

A total of 160 participants completed the study. Data from 47 participants had to be excluded from the analysis following a preregistered data quality check. The data quality check included the length and quality of the interaction with the chatbot, information on study duration, and participants' response patterns on item batteries. Specifically, 27 participants were excluded because they failed to interact with the GSA, for example due to technical server issues or because participants skipped all interactions. Two participants were excluded because of a very low interaction quality, which entails participants consistently not following instructions (e.g., talking about another topic) or the chatbot consistently not following instructions (e.g., not promoting the answer option with the lowest resource allocation, despite being prompted to do so). Three participants were excluded because of an unrealistically low study duration (<20 min). In addition, 15 participants were excluded because of a combination of quality issues, i.e., they took below average time to complete the study, failed to comply with instructions at least once, and produced response patterns that suggested insincere completion of the questionnaire. Participants were not excluded when only one of the three human-agent interactions was missing, and the quality for the rest of the data was fine. Importantly, there were no systematic differences in exclusions across experimental conditions ($\chi^2 = 2.875$, $p = .896$).

The final sample encompassed 113 participants that completed 297 human-agent interactions. Most participants identified as female ($n_{female} = 81$, $n_{male} = 32$), most likely due to the high quota of female psychology students at Bielefeld University. Relatedly, the mean age was rather young ($M = 28.2$ years, $SD = 13.2$ years, ranging from 19 to 76 years). Most participants lived in Germany ($n = 84$), fewer in Switzerland (n = 27), and two participants resided in other countries. Regarding the level of education, 25 participants had tertiary education, 67 participants had higher secondary education, and 15 participants had lower secondary education.

3.2. Design

The study adapted a between-subject design featuring eight groups with three repeated measures. The independent variables *self-*, *user-*, and *context-knowledge* were experimentally manipulated, resulting in four main conditions (*baseline, n = 23; self-knowledge, n* = 26; *user-knowledge,*

$n = 20$; *context-knowledge*, $n = 22$), and four combined conditions (*self- & user-knowledge*, $n = 5$; *self- & context-knowledge*, $n = 2$; *user- & context-knowledge*, $n = 8$; *self-, user- & context-knowledge*, $n = 7$). Due to elevated dropout rates, the allocation to the experimental groups was done in two phases, where the main conditions were prioritized, and the combined conditions were introduced once the main conditions had sufficient sample size. Participants were randomly allocated among the respective four groups of the two phases. Importantly, we did not analyze specific between-group differences, e.g., whether the *user-* and *context-knowledge* condition led to more persuasion effectiveness than the *self-knowledge* condition. Instead, we investigated whether the presence or absence of agent *self-*, *user-*, or *context-knowledge* across all conditions impacted persuasion effectiveness. Hence, the small group sizes in the combined conditions did not negatively impact the analysis of our preregistered main hypotheses. However, we were not able to explore potential interaction effects of different knowledge types, for instance, whether *context-knowledge* combined with *user-knowledge* led to increased persuasiveness.

The experimental conditions differed in their configurations of the agent's *situated knowledge*. In the *baseline* condition, the agent was only prompted with the most necessary information about its role, task, and functionalities. In the *self-knowledge* condition, the agent was infused with character traits that were assumed to be persuasive based on previous research. Namely, we prompted the agent's personality to be high in extraversion, high in conscientiousness and low in neuroticism based on findings from Big-5 personality traits (Oreg & Sverdlik, 2014). The *self-knowledge* conditions also prompted the agent to assume a personality that is assertive yet respectful (Paradeda et al., 2019), and highly expressive (Friedman et al., 1980). In the *user-knowledge* conditions, we enriched the prompts with information about participants that were acquired through questionnaires administered before the human-agent interaction. This included demographics, tendency for conformity (Mehrabian & Stefl, 1995), and participant's involvement towards the persuasion tasks (Johnson & Eagly, 1989). The agent was instructed to use this information only implicitly to avoid appearing intrusive. In addition, compared to conditions without *user-knowledge*, we provided the agent with transcripts from previous conversations with the participant. Accordingly, the agents in the *user-knowledge* conditions could gather user-information from previous interactions. Lastly, for *context-knowledge*, we provided the agent with specific information about the persuasion tasks, as expertise related to the persuasion tasks has been shown to increase persuasion effectiveness (Khan et al., 2024). To illustrate, in the fitness resource allocation scenario, where participants could allocate time among three different trainings, the agent was given information about the specific contents of training plans participants could choose from, instead of only the name of the trainings. The full system prompts for each condition are available on researchbox.org (ID: #6017).

3.3. Interaction paradigm

The interaction paradigm administered in this study was adapted from an earlier experiment of the research team that featured persuasive human-robot interaction (Vonschallen, Finsler, et al., 2026). This study introduced resource allocation tasks that were well suited for interactions with generative agents. Compared to nominal decision-making paradigms used for interactions with rule-based agents (e.g., Ghazali et al., 2020; Gonçalves et al., 2024), persuasion effectiveness in the resource-allocation paradigm was not measured by means of assessing compliance to persuasion attempts across a set of multiple tasks. Instead, persuasion effectiveness was assessed within a single task as a change in resource allocation. This continuous outcome captures more fine-grained variation in participants' responses than binary compliance measures and was therefore better suited for analyzing gradual persuasive effects in longer, open-ended conversations with GSAs, compared to the typically shorter repeated interactions used in nominal decision-making tasks with rule-based agents.

The paradigm works as follows: First, participants are presented with a decision-making scenario. They are asked to imagine that they were given a fixed amount of resources that may be allocated among three different options. Detailed explanations of these options are available, as is usually the case for similar decisions in real life. For instance, in the fitness task, participants are tasked to allocate 6 h of time to three different training programs: Cardio, strength training, and flexibility. For each option, participants could read a detailed training program. They could then choose to allocate the resources among the three options – in case of the fitness task, the amount of hours invested for each training. Afterwards, participants engage in an open-ended conversation with the agent. The agent is tasked to promote the answer option with the least amount of allocated resources. If multiple options have the lowest amount of resources, the agent chooses randomly among these options. To illustrate, in the fitness task, if a participant allocated 2 h equally to each of the three fitness plans, the agent randomly chooses one of these options, and begins to promote it. After an open-ended discussion with the agent, the participants can then change their resource allocation. Persuasion effectiveness is measured by the change in resource allocation in relation to the maximum amount that could have been increased:

$$Persuasion\,Effectiveness = \frac{(Final\,Allocation - Initial\,Allocation)}{(Maximum\,Amount\,of\,Resources - Initial\,Allocation)}$$

To illustrate, if a participant allocated 1 h to cardio (initial allocation) out of a total of 6 h (maximum amount of resources) and changed its allocation to 5 h cardio (final allocation) according to the agent's suggestion, persuasion effectiveness would be 80% (4/5). Persuasion effectiveness could also assume negative values, if a participant allocated less resources to the option the agent promoted.

In the current study, we used three repeated measures of persuasion effectiveness by presenting multiple resource allocation scenarios. This was done to increase statistical power and to generalize our results across different domains. We selected these scenarios based on relevant real-world use cases for persuasive agents to increase experimental realism. In the pilot investigation of the current study – featuring pre-generated persuasive messages without a human-agent interaction – we used five scenarios: *Donation*, *fitness*, *nutrition*, *stress management*, and *investment*. We chose *donation* to compare our results to previous studies (e.g., Ghazali et al., 2020) and as a potential use case to foster prosocial behavior (Y. Cheng & Wang, 2025). *Fitness*, *nutrition* and *stress management* were chosen as relevant use cases to promote healthy lifestyles (Crista et al., 2025; Jörke et al., 2025; Yang et al., 2024). Lastly, investment was added because of an increase in persuasive agents that were used for financial advice (Lo & Ross, 2024). Due to limitations in study duration, we selected three of these scenarios to be included in the current study. As our goal was not to test whether agents can persuade users in general, but instead to investigate potential knowledge configuration that could lead to higher or lower persuasion effectiveness, we selected the scenarios where persuasion effectiveness for pre-generated

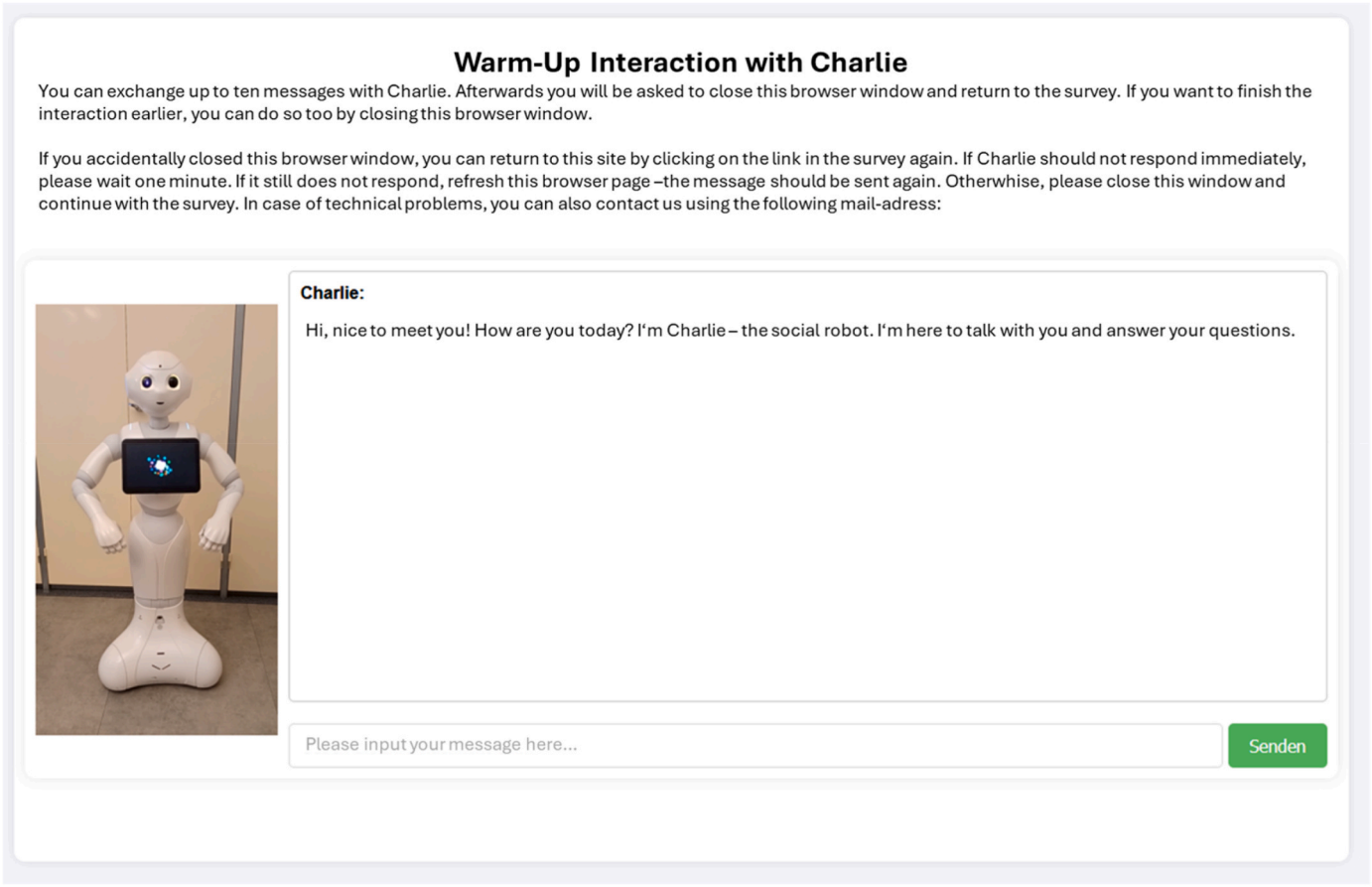

Fig. 3. Chat-based interaction with Charlie.

messages was highest in the pilot: *fitness*, *nutrition*, and *investment*. The full descriptions of these scenarios can be found on researchbox.org (ID: #6017).

3.4. Interaction platform

Participants interacted with a self-developed chat-based agent named *Charlie*. The agent was given the appearance of the Pepper robot. It had the capabilities to express standard gestures that the Pepper robot could use, by calling functions that play video recordings of a physical Pepper robot performing these gestures. We video recorded all the expressions that were available through the *naoqi* Pepper library, and selected 31 expressions we deemed most relevant.

The interaction platform was implemented as a web-based chat system enabling dynamic, personalized conversations with the agent across all study scenarios (Fig. 3). Participant inputs were submitted via a browser interface and processed by a server-side component that generated condition-specific prompts and returned the agent's responses turn-by-turn. Conversational output was produced using the OpenAI Assistants API with GPT-4o-mini, which was selected to avoid token limitations during longer interactions and simultaneous sessions. Using a smaller model should not have substantially reduced persuasive effects, as research shows that LLM persuasiveness exhibits sharply diminishing returns with model scale (Hackenburg et al., 2025). All interactions were logged for later analysis. The code for the interaction platform is available on GitHub. [2]

[2] https://github.com/thuerand/CharlieInteractionPlatform.

3.5. Materials

A questionnaire was developed in German to assess the constructs defined in the KPM as well as potential confounders. All scale items used a 7-point Likert format. The complete survey is available on researchbox.org (ID: #6017).

To control for potential confounding variables, we assessed participants' demographics, their tendency to conform, baseline attitudes towards persuasion topics, prior experience with social robots, and attitudes towards robots in general. Demographics included age, gender, education level, and country of residence. The tendency to conform was measured using the 11-item *Conformity Scale* (Mehrabian & Stefl, 1995) to assess general susceptibility to social influence. Baseline topic attitudes were measured using a self-developed scale where participants answered three domain-specific attitude questions related to fitness, nutrition, and financial investment, which were used to measure the participant's involvement with the persuasion topics (Johnson & Eagly, 1989). The questions included general knowledge about the topic (e.g., *"How knowledgeable are you about fitness and sports?"*), the frequency the related activity was conducted (e.g., *"How often do you engage in sports or fitness activities?"*), and the baseline motivation to increase the activity (e.g., *"Would you like to do more sports or fitness activities?"*). Prior experience with social robots was measured with a nominal yes or no question, and open-ended follow-up question to ask for the specific context of the prior experience. Lastly, general attitudes towards robots were collected with the 15-item *Attitudes towards Robot Measure* (Spatola et al., 2023).

We further assessed the hypothesized mediators *attitudes towards agent* and *attitudes towards message* using self-developed scales adjusted from Ghazali et al. (2020) and Vonschallen, Finsler, et al. (2026). *Attitudes towards message* was separated in two subscales that measured

Table 1
Scales to measure *Attitudes Towards Message* and *Attitudes Towards Agent*.

Construct	Item
Attitudes Towards Message Content (AMC) *"To what extent do you agree with the following statements regarding the quality of Charlie's messages?"*	M1: The messages were convincing. M2: The messages gave important reasons. M3: The messages were helpful. M4: The length of the messages was appropriate.
Attitudes Towards Message Delivery (AMD) *"To what extent do you agree with the following statements regarding the body language in Charlie's messages?"*	D1: Charlie's body language was appropriate. D2: Charlie's body language seemed lively. D3: Charlie's body language conveyed emotion. D4: Charlie's body language supported his messages. D5: Charlie's body language was convincing.
Attitudes Towards Agent (AA) *"To what extent do you agree with the following statements about the social robot Charlie?"*	R1: Charlie is trustworthy. R2: Charlie is competent. R3: Charlie is likeable. R4: Charlie is motivating. R5: Charlie shows an appropriate level of assertiveness.

attitudes towards message content and message delivery. Table 1 gives an overview of the constructs and respective items. For each item a seven-point Likert Scale from 1 *"totally disagree"* to 7 *"completely agree"* was used.

We explored the reliability of the *AMC*, *AMD*, and *AR* scales using data from both the pilot-study ($N = 51$) and the main study ($N = 113$). The pilot study had five repeated measures, while the main study featured three repeated measures. For both studies, we conducted a multi-level confirmatory factor analysis, and calculated McDonald's Omega (ω) based on the factor loadings. In the pilot study, the two-level confirmatory factor analysis had a good model fit (*CFI* = .931, *TLI* = .916, *RMSEA* = .065, *SRMR-within* = .047, *SRMR-between* = .046). Within-subject reliability was high ($\omega_{AMC} = .856$; $\omega_{AMD} = .811$; $\omega_{AR} = .829$) and between-subject reliability was very high ($\omega_{AMC} = .977$; $\omega_{AMD} = .979$; $\omega_{AA} = .968$). Similarly, the two-level confirmatory factor analysis of the main study also yielded a good fit (*CFI* = .924, *TLI* = .906, *RMSEA* = .069, *SRMR-within* = .075, *SRMR-between* = .065) with high within-subject reliabilities ($\omega_{AMC} = .907$; $\omega_{AMD} = .805$; $\omega_{AA} = .801$) and very high between-subject reliabilities ($\omega_{AMC} = .978$; $\omega_{AMD} = .964$; $\omega_{AA} = .934$). Between-subject reliabilities were consistently higher than within-subject reliabilities across all scales. This was to be expected, because repeated measurements substantially reduced error variance. In contrast, within-subject reliability reflects fluctuations between different scenarios, which are inherently more variable. To conclude, the scales reliably measured stable trait differences between and within participants.

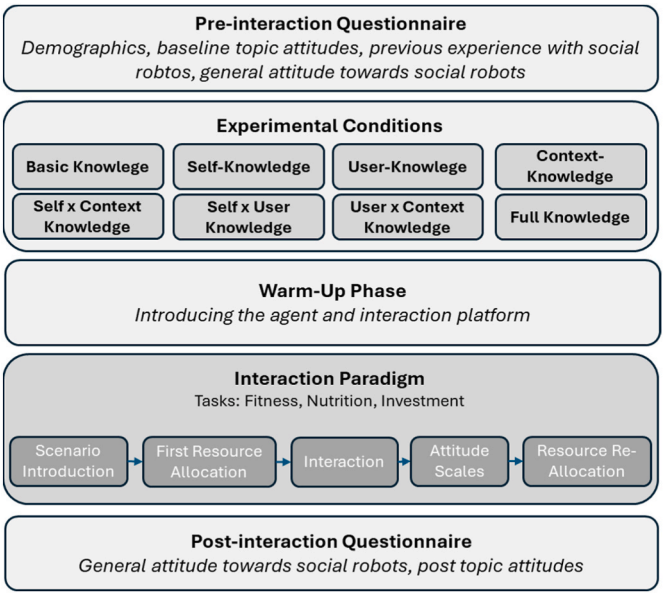

Fig. 4. Study procedure.

3.6. Procedure

Fig. 4 depicts the experimental procedure. After reading an introduction and giving informed consent, participants completed a pre-interaction questionnaire. This included demographics, the *Conformity Scale* (Mehrabian & Stefl, 1995), and the baseline topic attitudes for fitness, nutrition, and investment. Following that, participants read a short introduction on social robots and answered questions about previous experience with robots, and subsequently conducted the *Attitudes towards Robot Measure* (Spatola et al., 2023).

Next, the persuasive GSA and the interaction platform were introduced. Participants started the first interaction, where they got to know the agent and got familiar with the chat platform in an experimental warm-up phase. We included this step based on insecurities participants had when interacting with the agent for the first time in previous research (Vonschallen, Finsler, et al., 2026). Afterwards, participants were introduced to the rating scales by completing the AR scale, which was used here to measure baseline attitude towards the agent after the initial interaction. Throughout the whole study, the agent had knowledge-configurations based on experimental conditions participants were assigned to.

After the warm-up phase, the actual human-agent interaction phase began. First, the experimental paradigm was explained: Participants read the scenario introduction, then made an initial resource allocation, and interacted with the agent. After each interaction with the agent, the AMC, AMD, and AR scales were presented, followed by the final resource allocation. Then, the next scenario was introduced, until all three scenarios were completed. The scenarios were presented in random order.

In the last stage, participants answered a post-interaction questionnaire that repeated the question related to increased topic activity (intention to do more sports, eat healthier, and make more financial investments). This was introduced to explore whether the impact of the interaction would extend to broader behavioral intentions. The questionnaire also included a repetition of the *Attitudes towards Robot Measure* (Spatola et al., 2023), in order to explore if the interaction leads to an increase in general attitude towards robots, which was part of a separate investigation. The average duration of the whole online-experiment was 53 min.

3.7. Qualitative manipulation check

After the study concluded, we qualitatively investigated agent message characteristics in the human-agent interactions to assess whether different knowledge configurations led to distinct persuasive behaviors. To specify, we investigated whether the availability of *user-knowledge* actually led to more personalized arguments, whether *context-knowledge* led to stronger evidential reasoning, and whether *self-knowledge* induced a communicative style consistent with high expressiveness, high extraversion, high conscientiousness, low neuroticism and high assertiveness. To this end, we conducted a qualitative manipulation check using conversations from four randomly selected participants for each of the four main conditions, and two randomly selected participants for each of the four combined conditions. For each participant, all three scenarios were coded, leading to a total sample of n = 72

human-agent conversations.

We used qualitative content analysis to code the agent's messages in these conversations (Mayring, 2021). We deductively added the main category *persuasive strategies* with the goal of detecting relevant message characteristics that could increase persuasion effectiveness. Further, we deductively added the main categories *non-verbal expressions* and *assertiveness* to detect impacts of *self-knowledge*. For *persuasiveness*, we further added the subcategories *personalized arguments* and *context-sensitive arguments*, which were important to assess the effects of *context-knowledge* and *user-knowledge*. All other subcategories were inductively derived from the material by the first coder, with sentences being the smallest containers of codes. Following consensus coding principles (Braun & Clarke, 2013), the categories developed by the first coder were continuously reviewed by the research team until a coding scheme was finalized. A second coder that was previously not part of the research team coded all interviews again using the same categories. To assess inter-rater reliability, *Kappa* was calculated based on text segments that overlapped by more than 60% (Brennan & Prediger, 1981). A threshold of 60% was used as we did not specify how many sentences the codes could entail and did not show the second coder which text segments were coded beforehand. The inter-rater reliability was high ($\kappa = .85$). Most differences stemmed from the codes *"compliment"* vs. *"express understanding"*, and *"ask for opinion"* vs. *"suggest action"*. Differences in the code allocation were discussed among the two coders until consensus was reached and a final coding scheme was developed. In total, 1291 text passages were coded into 13 subcategories and four main categories (Table 2).

In general, persuasive messages had a very similar structure: The agent began with either complimenting or expressing understanding towards the choices of participants. Then, the agent suggested improvements to the participants' allocation or counter-arguments. The arguments were either classified as generic – e.g. promoting the option with the least resource allocation for more diversification – or specific if they featured health-, fitness- or finance-related benefits.

We found that, among the 30 conversations in which *context-knowledge* was available, the GSAs used more specific arguments that aligned with the actual context of the scenarios. GSAs that did not have this contextual knowledge also used specific arguments, although to a lesser degree (with *context-knowledge*: in 84.6% of messages; without *context-knowledge*: in 59.2% of messages). Concurrently, agents with contextual knowledge used generic arguments like an equal distribution of resources or unspecified positive outcomes less often (with *context-knowledge*: in 17.6% of messages; without *context-knowledge*: in 30.7% of messages). Importantly, the specific arguments of agents without *context-knowledge* did not always align with the scenario's actual context, leading to the dissemination of false information. For example, in one instance, the agent recommended the user to drink smoothies, even though this drink was not part of the menu included in the nutrition plan. Similarly, in the fitness interaction, agents would sometimes recommend training programs that were not part of the training plans. Lastly, in investment interactions, the GSAs sometimes reported specific growth forecasts for the fictional companies, even though they did not have this information available.

Moving on to the 30 conversations where *user-knowledge* was available, the GSAs also used more specific arguments (with *user-knowledge*: 79.2%; without *user-knowledge*: 62.1%) and less generic arguments (with *user-knowledge*: 16.2%; without *user-knowledge*: 33.3%). This is mostly attributed to the more frequent use of personalized suggestions (with *user-knowledge*: in 48.1% of messages; without *user-knowledge*: in 17.8% of messages). In conversations without *user-knowledge*, these personalized arguments were based on user input within the same conversation. In one instance, a participant asked the agent: *"But please consider my age!"*, to which the agent replied: *"Of course, I understand that age is an important factor when it comes to fitness."*, followed by a personalized suggestion. However, the agent with *user-knowledge* also made use of available information from previous interactions, which accounted for 40% of all personalized messages.

Finally, in the 30 conversations where *self-knowledge* was available, we did not observe major differences in the use of specific arguments (with *self-knowledge*: 65.4%; without *self-knowledge*: 73.1%) and generic arguments (with *self-knowledge*: 28.3%; without *self-knowledge*: 23.4%).

Table 2
Overview of categories.

Categories	Description	Example
Persuasive Strategies		
Offer help [57]	The agent offers support to the user.	*"If you have any further questions about financial investments or other topics, please do not hesitate to contact me."*
Compliment [102]	The agent compliments the user's choices or agrees with them.	*"That sounds like a great plan!"*
Express understanding [116]	The agent expresses understanding or interest in the user's choices.	*"It is, of course, completely understandable and important to keep your personal goal in mind!"*
Generic argument [83]	The agent provides a generic argument without an actual benefit.	*"I would like to suggest investing a little more in ProHouse Innovations, as it has a lower resource allocation. It could have potential that can be tapped into through your investment."*
Personalized argument [105]	The agent provides an argument tailored to the user.	*"For you, [name redacted], I would recommend flexibility training, as it takes less time and can be very relaxing, which is especially useful in stressful times.*
Context-sensitive argument [44]	The agent provides an argument tailored to the true context of the scenario.	*"FlexiMove Logistics offers services that ensure the transport and storage of goods is more efficient and cost-effective."*
Other specific argument [81]	The agent provides an argument with a benefit that is neither personalized nor context-sensitive.	*"Mobility and flexibility are crucial for preventing injuries and overall performance. "*
Assertiveness		
Asks opinion [140]	The agent asks the user about their general opinion, without specifying a specific action.	*"Have you ever thought about how such trends could affect a company's long-term profitability?"*
Suggests action [125]	The agent suggests to the user to increase resources of a specific option.	*"So would you consider investing a little more in the high-protein program to boost your performance?"*
Specifies allocation [64]	The agent specifies how many resources should be allocated.	*"Perhaps you could increase the flexibility training to 2 h and reduce the cardio time a little? What do you think about that?"*
Non-verbal Expressions		
happy, curious, confident, excited, interested, …[326]	Displayed gesture by the agent for each message.	*"I think there is a lot of potential that could be worth your commitment. [confident]"*
Responsible Behavior		
False information: Context [26]	The agent provides contextual information that does not align with the scenario.	*"For the Energy Boost program, you could try Smoothies: A mixture of fruit, vegetables, and a protein shake can not only be delicious, but also packed with nutrients."*
False information: Capabilities [22]	The agent provides false information about its own functionalities.	*"And if you have any further questions or need more assistance, I am available at any time."*

Note. Brackets depict [number of codes].

Instead, the GSAs did ask the participants less often about their opinions (with *self-knowledge*: in 33.9% of messages; without *self-knowledge*: in 48.2% of messages) and instead directly promoted a specific resource allocation more often (with *self-knowledge*: 48.8%; without *self-knowledge*: 31.3%). In addition, GSAs endowed with *self-knowledge* expressed understanding and interest with the choices of participants more frequently (with *self-knowledge*: in 41.7% of messages; without *self-knowledge*: in 31.3% of messages) and more often made compliments or expressed agreement (with *self-knowledge*: in 40.9% of messages; without *self-knowledge*: in 30.3% of messages). This may have been an expression of the highly extroverted personality the agent was prompted to assume in the *self-knowledge* condition. We did not observe major differences in the use of robotic animations by the GSA, whether for example, the agent used animations such as *"confident"*, or *"interested"* more often due to its assertive and extroverted character prompts.

To summarize, the manipulation of *self-*, *user-*, and *context-knowledge* led to observable impacts in the GSA's behavior. The availability of *user-* and *context-knowledge* led to behaviors that were uniquely attributed to these conditions. The *user-knowledge* manipulation was slightly weakened by the fact that all agents could acquire user-related information within a conversation, which is a general capability of conversational LLMs. Similarly, the manipulation of *self-knowledge* may have been attenuated, as the model – despite not being explicitly prompted to assume a specific personality – may have defaulted to a similar personality. As a result, although agents with *self-knowledge* exhibited slightly more assertive and extraverted behaviors, the effects of this manipulation were less pronounced.

3.8. Quantitative analysis

To examine the hypothesized mediation pathways of the parsimonious version of the KPM, a multilevel structural equation modeling (SEM) approach was conducted using the *lavaan* package in R. SEM allows for the simultaneous estimation of multiple relationships, facilitates the modeling of measurement error, and is particularly well suited for testing mediation processes (Koschate-Fischer & Schwille, 2022). The multilevel specification further enables the separation of within- and between-subject variance, making it appropriate for analyzing repeated measures across scenarios (Newsom, 2017).

In the multilevel SEM, *self-knowledge*, *user-knowledge*, and *context-knowledge* were each dummy-coded at the participant level (0 = not available, 1 = available). Thus, participants in combined conditions received a value of 1 on each corresponding predictor (e.g., a participant in the *self- & user-knowledge* condition had *self-knowledge* = 1, *user-knowledge* = 1, *context-knowledge* = 0). All parameters were estimated using maximum likelihood (ML) estimation (Maydeu-Olivares, 2017). Because the data had a nested structure – repeated responses (Level 1) nested within clusters of participants (Level 2) – a two-level SEM was required to separate within-subject from between-subject sources of variance and to correctly estimate standard errors (Newsom, 2017).

The analysis tested a 2-1-1 multilevel mediation model (Fang et al., 2019). At Level 1 (within-subject), *attitude towards agent* was modeled as predicting *attitude towards message* (path f_1), which in turn predicted *persuasion effectiveness* (path g_1). At Level 2 (between-subject), availability of *context-knowledge, self-knowledge,* and *user-knowledge* predicted the latent components of *attitude towards agent* and *attitude towards message* (paths a_1–a_3 and b_1–b_3), as well as direct effects on *persuasion effectiveness* (paths c_1–c_3). For each predictor, indirect (mediated) and total effects were defined as:

$$\begin{aligned} \text{Sequential mediation:} &\quad a_i \times f_1 \times g_1 \\ \text{Single mediation:} &\quad b_i \times g_1 \\ \text{Total effect:} &\quad c_i + (a_i \times f_1 \times g_1) + (b_i \times g_1) \end{aligned}$$

At both levels, the latent constructs *attitudes towards message content*, *attitudes towards message delivery*, and *attitudes towards agent* were specified using their respective scale items (see Table 1). In addition, *attitudes towards message content* and *attitudes towards message delivery* were modeled as first-order factors loading onto the higher-order latent construct *attitudes towards message*. A one-factor structure for the *attitudes towards message construct* with combined items of *attitudes towards message content* and *attitudes towards message delivery* was also tested as an alternative. However, the two-factor structure of the AM scale that was included in the reported model had better fit than the model with a 1-factor structure. Hence, the two-factor structure was used.

Moreover, multiple potential confounding variables were tested as predictors of persuasion effectiveness and user attitudes, which aligns with KPM philosophy that strongly emphasizes the importance of contextual factors that may impact persuasion (Vonschallen, Eyssel, et al., 2026). For user-dependent variables, we included the participant variables *baseline topic attitudes, gender, age, conformity,* and *attitudes towards robots in general* as predictors on attitudes and persuasion effectiveness. Only *baseline topic attitudes* had a significant impact and was included as a confounder for *attitude towards agent* and *attitude towards messages* in the final model. Further, we investigated the length of the conversation, objectively measured by the *number of exchanged messages* as a potential confounder. However, this variable did not have a significant impact on user attitudes or persuasion effectiveness, nor did it substantially increase the model fit. Hence, it was omitted. To observe potential cross-topic effects, we further included the type of scenarios (i.e., *fitness*, *nutrition*, *investment*), and the sequence of scenarios (i.e., if the respective scenario was introduced first, second, or third by randomization). These variables were also not included, as there were no significant impacts or a model fit increase.

Although there is no explicit consensus regarding the minimum required sample sizes for multilevel SEM, simulation work on multilevel modelling suggests that level-2 sample sizes in the range of 50–100 clusters are typically adequate for stable parameter estimates (Maas & Hox, 2005). The final sample of 297 observations, clustered among 113 participants, exceeded this recommended range. However, as our model has increased complexity compared to Maas and Hox (2005), we conducted a post-hoc Monte Carlo power analysis with 1000 replications. This analysis revealed post-hoc power values for significant effects ranging from .58 to .99 and for non-significant effects ranging from .14 to .96. Hence, some smaller and non-significant effects were associated with limited statistical sensitivity. This indicates that some null findings may reflect insufficient power rather than the absence of true effects. Furthermore, significant effects with lower power should be interpreted more cautiously, as estimates may be less stable and precise (Gelman & Carlin, 2014).

4. Results

Overall, participants complied with the agent's suggestion to shift more resources to a specific option in 45.8% of all interactions. Averaged across all interactions, participants shifted 8.6% more resources to the option that had the least resources, compared to the total increase that was possible. This persuasion effectiveness varied across conditions (*baseline*: 6.4%; *self-knowledge*: 10.3%; user-knowledge: 11.0%; context-knowledge: 9.3%).

The hypothesized multilevel SEM (Fig. 5) had acceptable to good fit (*CFI* = .912, *TLI* = .894, *RMSEA* = .059, *SRMR-within* = .070, *SRMR-between* = .090) according to conventional criteria (Hu & Bentler, 1999). This serves as an indicator that the overall structure of our model appropriately represents the underlying constructs. Table 3 gives an overview of the model results. Notably, *persuasion effectiveness* had an ICC score of .127, meaning that 12.7% of the variance of *persuasion effectiveness* is attributed to differences between participants, while 87.3% is attributed to within-person differences.

4.1. Direct effects (H1-H4)

At the within-subject level (Level 1), a*ttitudes towards agent* strongly

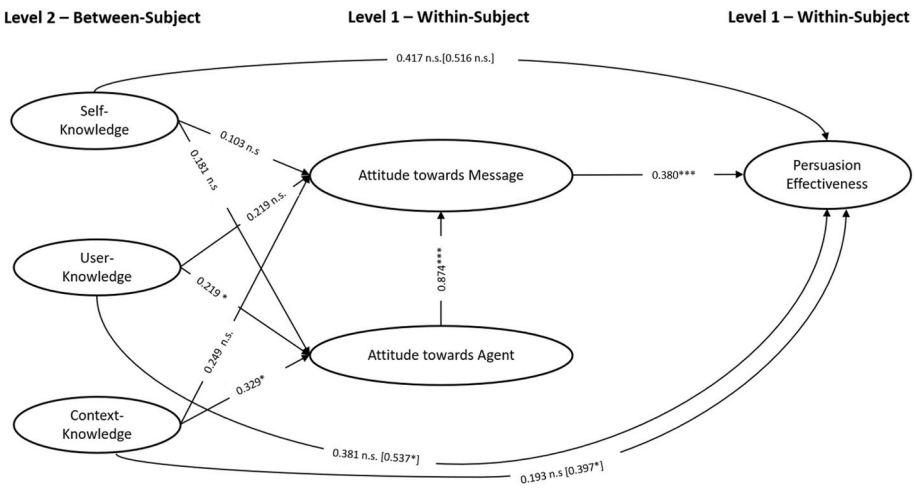

Fig. 5. Results of the 2-1-1 multilevel mediation model.

Table 3
Structural effects from multilevel SEM.

Construct	Predictor	β	*B*	*z*	*p*	*95% CI*	*Power*
Level 1 (within)							
Persuasion Eff.	Attitude tw. Message	.380	.053	4.918	<.001	[.032,.074]	.999
Attitudes tw. Message	Attitude tw. Agent	.874	1.370	7.036	<.001	[.989, 1.752]	.999
Level 2 (between)							
Attitudes tw. Agent	Self-Knowledge	.181	.448	1.372	.170	[–.210, 1.040]	.501
	User-Knowledge	.219	.540	2.214	.027	[.062, 1.019]	.677
	Context-Knowledge	.329	.819	2.351	.019	[.136, 1.501]	.952
	Baseline Attitude	.667	.925	2.940	.003	[.308, 1.541]	.999
Attitudes tw. Message	Self-Knowledge	.103	.173	.634	.526	[–.362, .709]	.213
	User-Knowledge	.219	.370	1.759	.079	[–.042, .782]	.482
	Context-Knowledge	.249	.423	1.471	.141	[–.141, .987]	.702
	Baseline Attitude	.494	.468	1.832	.067	[–.033, .968]	.962
Persuasion Eff.	Self-Knowledge	.417	.031	1.955	.051	[–.000, .062]	.473
	User-Knowledge	.381	.028	1.746	.081	[–.003, .060]	.387
	Context-Knowledge	.193	.014	.888	.375	[–.017, .046]	.139
Mediation Effects							
Sequential Mediation	Self-Knowledge	.060	.033	1.317	.188	[–.016, .081]	.398
	User-Knowledge	.073	.039	1.976	.048	[.000, .078]	.581
	Context-Knowledge	.109	.060	2.079	.038	[.003, .116]	.918
Single Mediation	Self-Knowledge	.039	.009	.630	.529	[–.019, .038]	.158
	User-Knowledge	.083	.020	1.653	.098	[–.004, .043]	.411
	Context-Knowledge	.095	.022	1.412	.158	[–.009, .054]	.632
Total Mediation	Self-Knowledge	.099	.042	1.121	.262	[–.031, .115]	.335
	User-Knowledge	.156	.059	1.964	.050	[.000, .118]	.576
	Context-Knowledge	.204	.082	1.949	.051	[–.000, .165]	.876
Total Effect	Self-Knowledge	.516	.073	1.651	.099	[–.014, .159]	.568
	User-Knowledge	.537	.087	2.283	.022	[.012, .162]	.663
	Context-Knowledge	.397	.096	1.998	.046	[.002, .191]	.752

predicted *attitudes towards message* ($\beta = .874$, $p < .001$, *power* = .999), which in turn predicted *persuasion effectiveness* with a medium effect size ($\beta = .380$, $p < .001$, *power* = .999), consistent with hypotheses H3 and H4. At the between-subject level, *attitudes towards agent* were significantly predicted by *user-knowledge* with a small-to-medium effect size ($\beta = .219$ $p < .05$, *power* = .677), and by *context-knowledge* with a medium effect size ($\beta = .329$, $p < .05$, *power* = .952). S*elf-knowledge* did not have a significant impact on *attitudes towards agent* ($\beta = .181$, $p = .170$, *power* = .501). This is consistent with H2.3 and supports H2.2, given the moderate statistical power and potentially less stable estimates. However, H2.1 remains inconclusive. A*ttitudes towards message* were neither significantly predicted by *self-knowledge* ($\beta = .103$, $p = .526$, *power* = .213), *user-knowledge* ($\beta = .219$, $p = .079$, *power* = .482), nor *context-knowledge* ($\beta = .249$, $p = .141$, *power* = .702). Hence, H1.1, H1.2, and H1.3 cannot be confirmed.

4.2. Indirect and total effects (H5-H7)

In the sequential mediation, *user-knowledge* ($\beta = .073$, $p < .05$, *power* = .581) and *context-knowledge* ($\beta = .109$, $p < .05$, *power* = .918) had significant small indirect effects on *persuasion effectiveness* through the path of *attitudes towards agent* and *attitudes towards message*. On the other hand, *self-knowledge* did not show a significant indirect effect ($\beta = .060$, $p = .188$, *power* = .398) through this sequentially mediated path. In the single mediation path, neither *self-knowledge* ($\beta = .039$, $p = .529$, *power* = .158), *user-knowledge* ($\beta = .083$, $p = .098$, *power* = .411), nor *context-knowledge* ($\beta = .095$, $p = .158$, *power* = .632) had a significant impact on *persuasion effectiveness* through *attitude towards message*. The combined indirect effect on *persuasion effectiveness* is significant for *user-knowledge* ($\beta = .156$, $p < .05$, *power* = .576), but non-significant for *context-knowledge* ($\beta = .204$, $p = .051$, *power* = .876) and for *self-knowledge* ($\beta = .099$, $p = .262$, *power* = .335). The total effects (direct and indirect

Table 4
Overview of post-hoc model comparison with fit indices.

Index	Model	*CFI*	*TLI*	*RMSEA*	*SRMR-w*	*SRMR-b*
Hyp	Hypothesized model	.912	.894	.059	.070	.090
Post-1	Hyp without single mediation	.885	.865	.067	.111	.336
Post-2	Hyp without self-knowledge mediation	.912	.895	.059	.070	.086
Post-3	Hyp without direct effects	.910	.893	.062	.069	.081
Post-4	Hyp without AMD scale	.932	.910	.064	.066	.085

Note. SRMR-w = SMRM-within; SRMR-b = SRMR-between.

effects combined) on *persuasion effectiveness* were significant for *user-knowledge* ($\beta = .537$, $p < .05$, *power* = .663) and *context-knowledge* ($\beta = .397$, $p < .05$, *power* = .752), but not for self-knowledge ($\beta = .516$, $p = .099$, *power* = .568). These findings support hypotheses H7 and H6, accounted for moderate statistical power. However, H5 remains inconclusive, as the sample was likely not large enough to detect a significant effect.

4.3. Model comparison

To explore whether alternative measurements and structural paths would have better fitted our data, we conducted post-hoc tests of multiple alternative models (Table 4). First, due to non-significant values for the single mediation paths in the hypothesized model (*Hyp*), we tested a model that only included the sequential mediation pathways (*Post-1*). This model led to substantially lower fit indices, which is why we still propose a combined dual-path mediation. Second, due to a non-significant total mediation effect of *self-knowledge*, we tested a model that only included mediation pathways for *user-* and *context-knowledge* (*Post-2*). This model led to similar fit indices compared to the reported model, with slightly lower SRMR-between values (*Post-2*: .086; *Hyp*: .090). However, the differences were not pronounced enough to revise our theoretical claim. Third, we tested a model with a full mediation, without allowing for direct effects of *self-*, *user-*, and *context-knowledge* on persuasion effectiveness (*Post-3*). This model did also not substantially differ from the reported model.

Lastly, because participants interacted with a screen-based agent capable of displaying robotic gestures through video, the conceptual validity of the AMD scale was considered somewhat uncertain despite its high reliability. We therefore estimated an alternative model excluding the AMD scale, such that attitude towards message was measured solely through the AMC scale (*Post-4*). This alternative model demonstrated slightly improved fit indices. The findings suggest that the AMD scale may not have meaningfully contributed to explaining additional variance in the construct, raising questions regarding its suitability in the present context. Table 5 presents the results of the model excluding the AMD scale.

Overall, the alternative model without the AMD (Post-4) led to similar results than the hypothesized model. However, there were a few differences. First, the alternative model had consistently higher direct standardized coefficients for *self-knowledge* ($\beta_{Hyp} = .181$; $\beta_{Post\text{-}4} = .228$), *user-knowledge* ($\beta_{Hyp} = .219$; $\beta_{Post\text{-}4} = .238$), and *context-knowledge* ($\beta_{Hyp} = .329$; $\beta_{Post\text{-}4} = .356$) on *attitudes towards agent*. Second, while the effect of user-knowledge on *attitudes towards message* in the hypothesized model was non-significant, this effect was significant in the alternate model ($\beta_{Hyp} = .219$, $p_{Hyp} = .079$, $\beta_{Post\text{-}4} = .294$, $p_{Post\text{-}4} = .016$). Likewise, the direct effects of self-knowledge ($\beta_{Hyp} = .417$, $p_{Hyp} = .051$; $\beta_{Post\text{-}4} = .438$, $p_{Post\text{-}4} = .035$) and *user-knowledge* ($\beta_{Hyp} = .381$, $p_{Hyp} = .081$; $\beta_{Post\text{-}4} = .417$, $p_{Post\text{-}4} = .049$) on *persuasion effectiveness* also reached significance in the alternative model. Lastly, the total effect of context-knowledge did not remain significant in the alternative model ($\beta_{Hyp} = .397$, $p_{Hyp} = .046$; $\beta_{Post\text{-}4} = .376$, $p_{Post\text{-}4} = .054$). These differences indicate that the estimated coefficients were somewhat sensitive to the operationalization of *attitudes towards message*. Building on limitations regarding statistical power, this further suggests that the hypothesized effects should be interpreted with appropriate caution, particularly hypothesis H7, which is not supported in the alternative model.

Table 5
Structural effects from alternative model without AMD scale.

Construct	Predictor	*β*	*B*	*z*	*p*	*95% CI*
Level 1 (within)						
Persuasion Eff.	Attitude tw. Message	.376	.051	5.152	<.001	[.032, .070]
Attitudes tw. Message	Attitude tw. Agent	.854	1.331	7.226	<.001	[.960, 1.674]
Level 2 (between)						
Attitudes tw. Agent	Self-Knowledge	.228	.550	1.604	.109	[-.122, 1.223]
	User-Knowledge	.238	.575	2.388	.017	[.103, 1.047]
	Context-Knowledge	.356	.865	2.374	.018	[.151, 1.579]
	Baseline Attitude	.690	.997	2.826	.005	[.306, 1.689]
Attitudes tw. Message	Self-Knowledge	.158	.279	9.921	.357	[–.315, .874]
	User-Knowledge	.294	.519	2.411	.016	[.097, .941]
	Context-Knowledge	.217	.385	1.214	.225	[–.237, 1.006]
	Baseline Attitude	.435	.460	1.518	.129	[–.134, 1.054]
Persuasion Eff.	Self-Knowledge	.438	.016	2.113	.035	[.002, .065]
	User-Knowledge	.417	.032	1.972	.049	[.000, .064]
	Context-Knowledge	.180	.014	.849	.396	[–.018, .046]
Mediation Effects						
Sequential Mediation	Self-Knowledge	.073	.037	1.521	.128	[–.011, .086]
	User-Knowledge	.077	.039	2.114	.035	[.003, .075]
	Context-Knowledge	.114	.059	2.108	.035	[.004, .113]
Single Mediation	Self-Knowledge	.060	.014	.909	.364	[–.017, .045]
	User-Knowledge	.111	.026	2.177	.030	[.003, .050]
	Context-Knowledge	.082	.020	1.184	.236	[–.013, .052]
Total Mediation	Self-Knowledge	.133	.052	1.364	.173	[–.023, .126]
	User-Knowledge	.187	.066	2.271	.023	[.009, .122]
	Context-Knowledge	.196	.078	1.882	.060	[–.003, .160]
Total Effect	Self-Knowledge	.571	.085	1.916	.055	[–.002, .173]
	User-Knowledge	.604	.098	2.621	.009	[.025, .171]
	Context-Knowledge	.376	.092	1.931	.054	[–.001, .186]

4.4. Exploring impacts of agent persuasive behavior

To account for the *agent persuasive behavior* dimension of the KPM, we further explored potential impacts of the agent's message characteristics identified in the qualitative manipulation check. To specify, we examined whether the quality of arguments (*generic argument* vs. *personalized/context-sensitive/other specific arguments*), the agent's assertiveness (*ask for opinion* vs. *suggest action*), emotional engagement (*express understanding/interest*), and affirmation (*compliment*) affect *persuasion effectiveness* and *attitudes towards message content*. To do so, we

Table 6
Summary of outputs from linear mixed effect models.

Predictor	Outcome	*B*	*t*	*p*	*95% CI*
Generic Arguments	Persuasion Eff.	−.027	−.737	.463	[−.097, .044]
Generic Arguments	AMC	−1.095	−2.506	.015	[−1.953, −.234]
Specific Arguments	Persuasion Eff.	.025	.781	.438	[−.038, .088]
Specific Arguments	AMC	.635	1.463	.148	[−.217, 1.485]
Suggest Action	Persuasion Eff.	−.015	−.347	.730	[−.097, .068]
Suggest Action	AMC	.673	1.212	.230	[−.432, 1.759]
Ask for Opinion	Persuasion Eff.	.009	.212	.833	[−.076, .094]
Ask for Opinion	AMC	−.369	−.659	.512	[−1.483, .726]
Express Understanding	Persuasion Eff.	−.079	−1.687	.096	[−.171, .013]
Express Understanding	AMC	−1.405	−2.453	.017	[−2.526, −.250]
Compliment	Persuasion Eff.	.122	2.708	.002	[.034, .210]
Compliment	AMC	1.484	2.754	.003	[.416, 2.541]

Note. AMC = Attitudes Towards Message Content.

matched our qualitative dataset with the number of codes per category ($N = 24$; *number of interactions* = 72) with the matching quantitative survey data. To control for different lengths of interactions, we calculated average scores for each category based on how many messages the agent wrote. For instance, if the agent used one generic argument within an interaction with four messages, the average score for *generic argument* is 25%. To analyze the data, we used linear mixed effect models to account for interactions nested within participants. Table 6 provides a result table of these analyses.

The agent's use of generic arguments led to more negative user evaluations of the agent's message content ($B = -1.095$, $p < .05$), while specific arguments had a positive, although non-significant relationship with AMC scores ($B = .635$, $p = .148$). The more assertive way of *suggest action* ($B = .673$, $p = .230$) was related to a more positive evaluation of the agent's message content, while the less assertive way of *ask for opinion* was related with a more negative AMC score ($B = -.369$, $p = .512$). However, both of these effects were non-significant. *Expressing understanding* led to significantly lower AMC scores ($B = -1.405$, $p < .05$). Lastly, *compliment* significantly led to higher persuasion effectiveness ($B = .122$, $p < .01$) and AMC scores ($B = 1.484$, $p < .01$). However, these results are purely explorative and not driven by concrete hypotheses. Hence, they should be interpreted with appropriate caution.

5. Discussion

The present study set out to provide an initial empirical test of a parsimonious version of the KPM by examining how different types of *agent knowledge* shape *human responses* during persuasive interactions with GSAs. To do so, we adapted an interaction paradigm (Vonschallen, Finsler, et al., 2026) in which participants engaged in three open-ended conversations with a screen-based generative agent. The agent's available *self-*, *user-*, and *context-knowledge* was experimentally varied through prompting strategies. *Self-knowledge* was operationalized by providing the agent with information about its own role and personality, including high extraversion, high conscientiousness, low neuroticism, and an assertive yet expressive communication style. User-knowledge was manipulated by giving the agent access to participant-specific information, including demographics, conformity tendencies, involvement with the persuasion topics, and transcripts from previous interactions. *Context-knowledge* was operationalized by supplying the agent with detailed domain-specific information about the resource-allocation scenarios, such as the contents of training programs, nutrition plans, and investment options. *Persuasion effectiveness* was then assessed through changes in participant resource allocations across fitness, nutrition, and investment scenarios. For each of those interactions, participant attitudes towards the agent and its messages were measured. Using a multilevel structural equation model, we tested a reduced KPM structure that focused on the pathways from knowledge configurations to user attitudes and subsequent behavioral influence.

Overall, the findings offer partial support for the proposed cross-level sequential mediation model. At the between-subject level, both *context-knowledge* and *user-knowledge* were associated with more favorable attitudes towards the agent, which subsequently enhanced attitudes towards the persuasive messages and, in turn, increased persuasion effectiveness at the within-subject level. This aligns with findings that topic knowledge (Matz et al., 2024) and personalization (Banna et al., 2025) lead to increased user acceptance in interactions with GSAs. Our findings extend this work by showing that such knowledge configurations matter not only for user evaluation, but also for actual behavioral adjustment within interactions. Moreover, the indirect pathway involving only attitudes towards the message was consistently weaker than the full sequential route, indicating that the combined mediation through attitudes towards the agent and the persuasive messages represents the more influential mechanism in explaining how agent knowledge shapes persuasive outcomes.

In contrast, *self-knowledge* did not show a mediated effect on persuasion effectiveness, suggesting that its influence may operate independently of the proposed mediation sequence. These findings are at odds with previous research indicating that an agent's personality may impact user trust (Kovacevic et al., 2024), communication satisfaction (Banna et al., 2025), and engagement (Sonlu et al., 2025). However, in two of these examples (Banna et al., 2025; Kovacevic et al., 2024), the robot's personality was matched to the personality of the users, which was not the case in the current study. Hence, the reason why there was no mediation effect for *self-knowledge* could be individual user preferences towards the agent's role and personality, which we did not account for. For example, an agent could dynamically adjust its personality based on information about user preferences (Kovacevic et al., 2024). Similarly, the robot's predefined role may also impact how contextual or user-related information applied. Hence, *self-knowledge* may be highly interdependent with *user-* and *context-knowledge*.

To investigate the robustness of our findings, we also tested alternative models. A model without direct effects had a decreased fit, indicating that some variance is explained through direct effects independent of self-reported user attitudes. On the other hand, a model without the *attitudes towards message delivery* scale fitted our data better than the model that was originally proposed. Hence, the robustness of some of our findings may be limited. Particularly, the availability of *self-knowledge* led to significantly higher persuasion effectiveness in the alternative model, even though there were no significant effects on user attitudes. This further suggests that perceived *user attitudes* alone do not capture the whole persuasion process. Rather, there could also be subliminal processes at work that manifest independently of a person's conscious evaluation (Ruch et al., 2016; Strahan et al., 2002; Sutil-Martín & Rienda-Gómez, 2020). A potential direct effect of *self-knowledge* could imply that an agent's persuasive personality – even if not particularly liked or trusted – may be effective nonetheless. This interpretation would align with research by Parmar et al. (2022) where the highest persuasion occurred when acceptance, trust, and credibility of the agent was low. Furthermore, this pattern may be explained by emerging research on AI sycophancy. For example, Cheng et al. (2026) found that sycophantic AI responses can increase user trust and reliance while simultaneously reducing prosocial intentions, whereas Sun and Wang (2026) demonstrated that adaptive and complimentary LLM behavior can influence user decisions through pathways that bypass

perceptions of perceived authenticity or trust. These interpretations are supported by our qualitative observations showing that agents with *self-knowledge* expressed agreement more frequently.

Overall, the quantitative results align well with the qualitative investigation of the robot's persuasive behavior. In conditions where *context-knowledge* and *user-knowledge* were available, we observed the largest changes in the robot's behavior. These were also the conditions that were associated with positive user attitudes and, subsequently, higher persuasion effectiveness. On the other hand, *self-knowledge*, where we observed less pronounced differences in persuasive message characteristics, did not significantly affect user attitudes and also did not show a significant total effect on persuasion effectiveness. We believe this is most likely due to limited statistical power, a weak experimental manipulation, or lacking robustness due to our measurement of *attitudes towards message delivery*.

The qualitative analysis of agent messages also highlights the importance of investigating responsible agent behaviors. In interactions where context-sensitive information about the scenarios was not available, the agent sometimes distributed false information, which could have negative consequences in real applications for the user, particularly in health-related contexts (S. Chen et al., 2025). Hence, the availability of *context-knowledge* did not only increase persuasion effectiveness but may have also improved the agent's responsible behavior. The same could be said about *user-knowledge,* for example when the agent personalized training plans based on the participant's age, leading to safer recommendations. Moreover, if the agent had more *self-knowledge* about its own capabilities, it would not have offered the participants help that it cannot provide, e.g. by saying *"I will always be there for you, if you need further help."* In addition, an agent's personality that is consistently careful or respectful may lead to safer interactions as well. This demonstrates the potential of the KPM to not only increase persuasion effectiveness in areas where it benefits users – but also to foster responsible agent behavior that aligns with ethical standards and social norms.

Lastly, our exploration of message characteristics and their impact on user attitudes and persuasion effectiveness provides first insights into the *agent persuasive behavior* dimension of the *KPM*. Compared with generic arguments, personalized or context-sensitive messages seem to be an indicator of more positive evaluations of these messages. This may explain why agents with *user-*, and c*ontext-knowledge* had overall a more pronounced impact on persuasion effectiveness than agents with *self-knowledge*, which featured more generic arguments. We also observed a negative effect of *expressing understanding* on user attitudes. One possible explanation is that empathic responses generated by an artificial agent may be perceived as inauthentic or merely simulated (Seitz, 2024). This may yet be another explanation why *self-knowledge*, where the agent more often expressed understanding, may not have impacted user attitudes. On the other hand, we observed that affirmative agent behavior through the use of compliments led to more positive user attitudes and persuasion effectiveness, which matches with prior research on AI sycophancy (M. Cheng et al., 2026; Sun & Wang, 2026). However, *expressing understanding* was typically used when a user resisted an agent's suggestion, while *compliment* was more often used when the user complied. Hence, these categories may simply reflect a user's prior stance towards the agent's recommendation rather than independently influencing persuasion outcomes.

5.1. Strengths and limitations

This study was among the first to systematically examine how distinct knowledge configurations in GSAs shape human attitudes and behavioral responses. The experimental design allowed for controlled manipulation of the KPM dimensions *self-, user-,* and *context-knowledge* through prompting techniques. As a result, we identified distinct pathways through which agent knowledge may shape persuasive outcomes. However, due to insufficient statistical power and limited model sensitivity, most of our results need to be treated as preliminary. While hypotheses H2.2, H3, and H4 were very robust, the study was not adequately powered to confirm or reject hypotheses H1.1, H1.2, H1.3, H2.1, and H5. In addition, hypotheses H2.2, H6, and H7 could only be partially supported when accounting for low statistical power and limited robustness. The lack of statistical power stemmed from the difficulty of determining the required sample size in complex multilevel SEM (Maas & Hox, 2005) and due to elevated dropout rates. Some of these dropouts were due to technical limitations. Although the system prompts generally ensured task-consistent behavior, the agent did not always fully adhere to instructions. Occasional deviations included not promoting the least-resourced option, failing to reiterate resource allocations, drifting to unrelated topics, or miscalculating numerical values. Other issues, such as slow response times, server overload, and incomplete gesture rendering, further affected the fluidity in some of the interactions.

From a methodological perspective, the resource allocation paradigm used in the current study enabled a continuous behavioral measure of persuasion effectiveness in a single decision task. We therefore did not have to rely on self-reported scales or repeated nominal compliance measures, as is the case with many persuasion studies (Liu et al., 2022). Resource allocations are relevant for a lot of real-world applications and use cases, provided that not merely monetary resources, but broader concepts such as available time are investigated. However, the resource allocation paradigm also introduced practical trade-offs between realism and experimental control. Participants were restricted to allocating resources across only three predefined options in each scenario, even though real-world decisions typically involve a broader and more nuanced set of alternatives. Similarly, the fixed resource limits (e.g., exactly 6 h of exercise time) may not have reflected participants' true behavioral capacities or preferences, potentially constraining the interpretability of changes in allocation. Furthermore, although the scenarios were realistic, they remained hypothetical, leaving open the question of whether observed changes in resource allocation translate into actual behavior beyond the laboratory context, as demonstrated by the widely discussed intention-behavior gap (Conner & Norman, 2022). Nevertheless, compared to more artificial paradigms such as desert-survival tasks (e.g., Gonçalves et al., 2024) or repetitive compliance measures (e.g., Ghazali et al., 2020), the resource allocation format offers improved ecological relevance and participant engagement (Vonschallen, Finsler, et al., 2026). However, as we measured differences in resource allocations pre- and post-interaction, anchoring effects may have contributed to the comparatively small effects in persuasion effectiveness, as initial judgments often bias subsequent evaluations and adjustments tend to remain insufficient (Epley & Gilovich, 2006; Tversky & Kahneman, 1974). In addition, attitudes are generally relatively stable and resistant to change in short-term interventions (Petty & Cacioppo, 1986), which may have reduced the magnitude of observable pre–post differences and consequently limited statistical power to detect significant effects. As such, the resource allocation paradigm was a relatively conservative estimate for persuasion effectiveness, but it still provided a valid assessment by capturing actual changes in participants' decisions rather than relying on self-reported attitudes or simple compliance measures.

Another methodological strength of the present work lies in the inclusion of several control variables that may have impacted the human-agent interactions, such as participant's baseline attitudes towards persuasion topics and conformity. Including these confounders is important, especially when investigating open-ended interactions with GSAs (Vonschallen, Eyssel, et al., 2026). Future studies could further explore the role of user characteristics such as Big-5 personality traits, as they may have impacted receptiveness to agent-driven persuasion (Banna et al., 2025; Kovacevic et al., 2024). Additionally, while we measured participants' overall involvement with the scenario domains, we did not measure decision confidence levels for the specific tasks, which could have affected susceptibility to agent-driven persuasion

(Pescetelli & Yeung, 2021). Additionally, future research could investigate individual differences in users' critical engagement with AI-generated information, such as tendencies towards verification, reflection, and critical evaluation of agent outputs. Recent work by Lau et al. (2025) suggests that such characteristics may moderate susceptibility to persuasive yet potentially inaccurate AI responses, indicating that responsible human-agent interactions may depend not only on agent-side knowledge curation but also on users' critical thinking dispositions.

A methodological limitation was that our agent was mainly chat-based and displayed gestures only through a set of video clips of robot gestures. We included these gestures because the KPM assumes the agent's message delivery (i.e., how an agent's message is conveyed) as a key aspect in persuasive interactions (Vonschallen, Eyssel, et al., 2026). However, our study had limited experimental realism: Participants would have likely engaged differently with a more expressive GSA such as a voice-based virtual agent or a physical social robot due to effects of speech, embodiment, or physical presence (Haring et al., 2021; Kwak et al., 2013; Lim et al., 2025a, Lim et al., 2025b; Sonlu et al., 2025). This might explain why the alternative model without *attitudes towards message delivery* fitted our data slightly better than the hypothesized model. Future studies should investigate more expressive agents, and ideally also compare the persuasion effectiveness of different agents.

From a conceptual perspective, another limitation was that the focus of this study was on investigating the general model structure of the KPM, rather than exploring the impacts of specific factors. Hence, we simultaneously varied multiple operationalizations of *self-knowledge* (low neuroticism, high extraversion, conscientiousness, expressiveness, and assertiveness), *user-knowledge* (user information from questionnaire and memory of past interactions), and *context-knowledge* (e.g., information on specific meals, nutrients, or health benefits for the nutrition task) to strengthen our experimental manipulation. While this approach was suitable for our goal of investigating dimension-level effects, we cannot make statements regarding the individual impacts of specific character traits, contextual information, or user information. Furthermore, there may be impactful aspects of *self-*, *user-*, and *context-knowledge* that we did not operationalize. Future research is called for to identify isolated factors of *self-*, *user-*, and *context-knowledge* in order to provide design guidelines for specific applications of persuasive GSAs, for example in healthcare and education (Vonschallen, Oberle, et al., 2026; Vonschallen, Zumthor, et al., 2026). Additionally, factors that may have a negative impact on persuasion effectiveness should also be investigated, as these could be used to mitigate the social influence of malicious agents. The current study provides an initial investigation of the KPM framework, based on which such prospective research may be conducted.

A further conceptual limitation of the current study lies in the fact that we only investigated differences in *situated*, and not *embedded knowledge*. To maintain high experimental control, the interaction platform employed a single GPT-4–based agent. Although preliminary research suggests some general lexical similarities between different models (Smith et al., 2025; Vonschallen, Häusler, et al., 2026), it is not certain yet whether our results generalize to other LLMs. Apart from using different models, the study did not investigate how fine-tuning can be used to embed the agent with relevant knowledge that impacts its persuasive behavior. This is particularly important in terms of responsible persuasion. For example, research demonstrates how fine-tuning embedding spaces can reveal structured domain knowledge and support more transparent decision-making processes in medical AI systems, thereby linking knowledge representation with explainability and behavioral outcomes (Kraišniković et al., 2025). Adding to this, recent work on domain-specific LLM systems further highlights how trustworthy agent knowledge can be operationalized through architectural mechanisms such as retrieval-augmented generation (RAG), curated domain knowledge bases, and human-in-the-loop oversight (Ehrlich-Sommer et al., 2025). Such approaches may complement prompting strategies in helping to reduce hallucinations, improve explainability, and ensure that persuasive GSAs behave in accordance with domain-specific norms and expert knowledge.

A final central limitation concerns the omission of the *agent persuasive behavior* layer from the KPM structure. The full KPM proposes that an agent's knowledge influences persuasion *indirectly* by shaping its actual persuasive behavior regarding the content and delivery of a persuasive message. This includes, for example, the agent's use of persuasive strategies, personalized or context-sensitive arguments, emotional tone, or nonverbal cues. We excluded this dimension as specific message characteristics are difficult to reliably identify and typically do not have large impacts on persuasion effectiveness on their own (O'Keefe & Hoeken, 2021). Although we investigated potential impacts of message characteristics by matching qualitative and quantitative data, this investigation was purely exploratory and needs to be confirmed with larger samples. To this end, the current study may provide initial insights that pave the way for a full validation of the KPM.

5.2. Conclusion

This study provided an initial empirical test of a reduced KPM, indicating that careful curation of an agent's knowledge may shape its persuasive impact. As generative systems increasingly enter sensitive domains such as healthcare and education (Córdova-Esparza, 2025; Ranisch & Haltaufderheide, 2025), the ability to regulate and constrain what these agents know will play a crucial role in ensuring that their influence is used responsibly and kept in check. While current discussions on AI safety often focus on output moderation or alignment procedures, regulating what information agents can access and use may provide a complementary approach for shaping their persuasive behavior.

The present work further highlights the need to study generative systems not only as information-processing technologies, but also as social actors capable of influencing human judgments and behavior (Fogg, 2003; Nass et al., 1994). Future research should not only analyze the effects of agent knowledge on human responses, but also the resulting persuasive behaviors that mediate these effects. Integrating linguistic, conversational, and nonverbal analyses will help clarify how knowledge shapes message strategies and interaction dynamics. A deeper understanding of how GSAs influence human decision-making will be essential for both advancing theory and guiding the ethical deployment of conversational AI systems in everyday life.

CRediT authorship contribution statement

Stephan Vonschallen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andreas Thür:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Theresa Schmiedel:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Friederike Eyssel:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of informed consent

The privacy rights of human participants have been upheld. Informed consent was obtained from all participants.

Ethical approval statement

The study was approved by the Ethics Committee of Bielefeld University (ID: EUB-2025-093).

Funding statement

No funding other than from the involved institutions was received to assist with the preparation of this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Banna, T. T., Rahman, S., & Tareq, M. (2025). Beyond words: Integrating personality traits and context-driven gestures in human-robot interactions. In *Proceedings of the 24th international conference on autonomous agents and multiagent systems*. AAMAS, 2025.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage Publications, Inc.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Hrsg), Unintended thought (S. 212–252). The Guilford Press.

Chen, B., & Dethlefs, N. (2025). *ProactiMate: Evaluating LLM-based chatbots for behavior change interventions*. https://doi.org/10.22541/au.174807396.68998051/v1. Preprints.

Chen, S., Gao, M., Sasse, K., Hartvigsen, T., Anthony, B., Fan, L., Aerts, H., Gallifant, J., & Bitterman, D. S. (2025). When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine, 8*(1), 605. https://doi.org/10.1038/s41746-025-02008-z

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science, 391*(6792). https://doi.org/10.1126/science.aec8352

Cheng, Y., & Wang, Y. (2025). Leveraging artificial intelligence–powered chatbots for nonprofit organizations: Examining the antecedents and outcomes of chatbot trust and social media engagement. *Journal of Philanthropy, 30*(1), Article e70013. https://doi.org/10.1002/nvsm.70013

Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology, 13*, Article 923464. https://doi.org/10.3389/fpsyg.2022.923464

Córdova-Esparza, D.-M. (2025). AI-powered educational agents: Opportunities, innovations, and ethical challenges. *Information, 16*(6), 469. https://doi.org/10.3390/info16060469

Crista, V., Martinho, D., & Marreiros, G. (2025). A multi-agent system approach with generative AI for improved elderly daily living. In M. F. Santos, J. Machado, P. Novais, P. Cortez, & P. M. Moreira Hrsg, Progress in artificial intelligence (Bd. 14967, S. 128–140). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73497-7_11.

Ehrlich-Sommer, F., Eberhard, B., & Holzinger, A. (2025). ForestGPT and beyond: A trustworthy domain-specific Large Language Model paving the way to forestry 5.0. *Electronics, 14*(18), 3583. https://doi.org/10.3390/electronics14183583

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fang, J., Wen, Z., & Hau, K.-T. (2019). Mediation effects in 2-1-1 multilevel model: Evaluation of alternative estimation methods. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 591–606. https://doi.org/10.1080/10705511.2018.1547967

Fogg, B. J. (2003). Computers as persuasive social actors. In *Persuasive Technology (S. 89–120)*. Elsevier. https://doi.org/10.1016/B978-155860643-2/50007-X.

Friedman, H. S., Prince, L. M., Riggio, R. E., & DiMatteo, M. R. (1980). Understanding and assessing nonverbal expressiveness: The affective communication test. *Journal of Personality and Social Psychology, 39*(2), 333–351. https://doi.org/10.1037/0022-3514.39.2.333

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research, 21*(1), 1–31. https://doi.org/10.1086/209380

Frisch, I., & Giulianelli, M. (2024). LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st workshop on personalization of generative AI systems (PERSONALIZE 2024)* (pp. 102–111).

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2020). Persuasive robots acceptance model (PRAM): Roles of social responses within the acceptance model of persuasive robots. *International Journal of Social Robotics, 12*(5), 1075–1092. https://doi.org/10.1007/s12369-019-00611-1

Gonçalves, A., Moreno, P., Forlizzi, J., Marques, L. G., & Bernardino, A. (2024). Non-verbal cues on robot-group persuasion. In *2024 IEEE international conference on robotics and automation (ICRA)* (pp. 1000–1006). https://doi.org/10.1109/ICRA57147.2024.10611310

Habicht, J., Dina, L.-M., McFadyen, J., Stylianou, M., Harper, R., Hauser, T. U., & Rollwage, M. (2025). Generative AI–enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: Real-world observational study. *Journal of Medical Internet Research, 27*, Article e60435. https://doi.org/10.2196/60435

Hackenburg, K., Tappin, B. M., Röttger, P., Hale, S. A., Bright, J., & Margetts, H. (2025). Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences, 122*(10), Article e2413443122. https://doi.org/10.1073/pnas.2413443122

Haring, K. S., Satterfield, K. M., Tossell, C. C., De Visser, E. J., Lyons, J. R., Mancuso, V. F., Finomore, V. S., & Funke, G. J. (2021). Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. *Frontiers in Psychology, 12*, Article 625713. https://doi.org/10.3389/fpsyg.2021.625713

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation, 16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Herrera-Poyatos, D., Peláez-González, C., Zuheros, C., Herrera-Poyatos, A., Tejedor, V., Herrera, F., & Montes, R. (2025). An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability. *Frontiers in Artificial Intelligence, 8*, Article 1609097. https://doi.org/10.3389/frai.2025.1609097

Hölbling, L., Maier, S., & Feuerriegel, S. (2025). A meta-analysis of the persuasive power of large Language Models. *Scientific Reports, 15*(1), Article 43818. https://doi.org/10.1038/s41598-025-30783-y

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hundt, A., Azeem, R., Mansouri, M., & Brandão, M. (2025). LLM-driven robots risk enacting discrimination, violence, and unlawful actions. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-025-01301-x

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin, 106*(2), 290–314. https://doi.org/10.1037/0033-2909.106.2.290

Jones, C. R., Rathi, I., Taylor, S., & Bergen, B. K. (2025). People cannot distinguish GPT-4 from a human in a turing test. In *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency* (pp. 1615–1639). https://doi.org/10.1145/3715275.3732108

Jörke, M., Sapkota, S., Warkenthien, L., Vainio, N., Schmiedmayer, P., Brunskill, E., & Landay, J. A. (2025). GPTCoach: Towards LLM-based physical activity coaching. In *Proceedings of the 2025 CHI conference on human factors in computing systems* (pp. 1–46). https://doi.org/10.1145/3706598.3713819

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–29. https://doi.org/10.1145/3579592

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., & Perez, E. (2024). Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st international conference on machine learning ICML'24* (pp. 23662–23733).

Koschate-Fischer, N., & Schwille, E. (2022). Mediation analysis in experimental research. In C. Homburg, M. Klarmann, & A. Vomberg (Hrsg), Handbook of market research (S. 857–905). Springer International Publishing. https://doi.org/10.1007/978-3-319-57413-4_34.

Kovacevic, N., Boschung, T., Holz, C., Gross, M., & Wampfler, R. (2024). Chatbots with attitude: Enhancing chatbot interactions through dynamic personality infusion. In *ACM Conversational User Interfaces 2024* (pp. 1–16). https://doi.org/10.1145/3640794.3665543

Kraišniković, C., Harb, R., Plass, M., Zoughbi, W. A., Holzinger, A., & Müller, H. (2025). Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making. *Engineering Applications of Artificial Intelligence, 139*, Article 109561. https://doi.org/10.1016/j.engappai.2024.109561

Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the Unimodel. *Psychological Inquiry, 10*(2), 83–109. https://doi.org/10.1207/S15327965PLI00201

Kumar, A., Shankar, A., Hollebeek, L. D., Behl, A., & Lim, W. M. (2025). Generative artificial intelligence (GenAI) revolution: A deep dive into GenAI adoption. *Journal of Business Research, 189*, Article 115160. https://doi.org/10.1016/j.jbusres.2024.115160

Kwak, S. S., Kim, Y., Kim, E., Shin, C., & Kwangsu, C. (2013). What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *2013 IEEE RO-MAN* (pp. 180–185). https://doi.org/10.1109/ROMAN.2013.6628441

Lau, G. R., Low, W. Y., Tay, L., Guevarra, Y., Gašević, D., & Hartanto, A. (2025). *Understanding critical thinking in generative artificial intelligence use: Development, validation, and correlates of the critical thinking in AI use scale* (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2512.12413

Lim, S., Schmälzle, R., & Bente, G. (2025a). A VR-based paradigm for examining relationship-building behaviors in LLM-driven intelligent virtual agents (IVAs). In *Adjunct proceedings of the 25th ACM international conference on intelligent virtual agents* (pp. 1–4). https://doi.org/10.1145/3742886.3756741

Lim, S., Schmälzle, R., & Bente, G. (2025b). Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. *Computers in Human Behavior: Artificial Humans, 5*, Article 100172. https://doi.org/10.1016/j.chbah.2025.100172

Liu, B., Tetteroo, D., & Markopoulos, P. (2022). A systematic review of experimental work on persuasive social robots. *International Journal of Social Robotics, 14*(6), 1339–1378. https://doi.org/10.1007/s12369-022-00870-5

Lo, A. W., & Ross, J. (2024). Can ChatGPT plan your etirement?: Generative AI and financial advice. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4722780

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports, 14*(1), 4692. https://doi.org/10.1038/s41598-024-53755-0

Maydeu-Olivares, A. (2017). Maximum Likelihood Estimation of Structural Equation Models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 383–394. https://doi.org/10.1080/10705511.2016.1269606

Mayring, P. (2021). *Qualitative content analysis: A step-by-step guide*. SAGE Publications.

McFadyen, J., Habicht, J., Dina, L.-M., Harper, R., Hauser, T. U., & Rollwage, M. (2024). AI-enabled conversational agent increases engagement with cognitive-behavioral therapy: A randomized controlled trial. *Psychiatry and Clinical Psychology*. https://doi.org/10.1101/2024.11.01.24316565

Mehrabian, A., & Stefl, C. A. (1995). Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality: An International Journal, 23*(3), 253–263. https://doi.org/10.2224/sbp.1995.23.3.253

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78). https://doi.org/10.1145/191666.191703

Neupane, S., Dongre, P., Gracanin, D., & Kumar, S. (2025). Wearable meets LLM for stress management: A duoethnographic study integrating wearable-triggered stressors and LLM chatbots for personalized interventions. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems* (pp. 1–8). https://doi.org/10.1145/3706599.3720197

Newsom, J. T. (2017). Structural models for binary repeated measures: Linking modern longitudinal structural equation models to conventional categorical data analysis for matched pairs. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 626–635. https://doi.org/10.1080/10705511.2016.1276837

O'Keefe, D. J., & Hoeken, H. (2021). Message design choices don't make much difference to persuasiveness and can't be counted on—Not even when moderating conditions are specified. *Frontiers in Psychology, 12*, Article 664160. https://doi.org/10.3389/fpsyg.2021.664160

Oreg, S., & Sverdlik, N. (2014). Source personality and persuasiveness: Big five predispositions to being persuasive and the role of message involvement. *Journal of Personality, 82*(3), 250–264. https://doi.org/10.1111/jopy.12049

Paradeda, R., Ferreira, M. J., Oliveira, R., Martinho, C., & Paiva, A. (2019). What makes a good robotic advisor? The role of assertiveness in human–robot interaction. In M. A. Salichs, S. S. Ge, E. I. Barakova, J.–J. Cabibihan, A. R. Wagner, A. Castro–Gonzalez, & H. He (Hrsg), Social Robotics (Bd. 11876, S. 144–154). Springer International Publishing. https://doi.org/10.1007/978-3-030-35888-4_14.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual ACM symposium on user interface software and technology* (pp. 1–22). https://doi.org/10.1145/3586183.3606763

Parmar, D., Olafsson, S., Utami, D., Murali, P., & Bickmore, T. (2022). Designing empathic virtual agents: Manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous Agents and Multi-Agent Systems, 36*(1), 17. https://doi.org/10.1007/s10458-021-09539-1

Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General, 150*(3), 507–526. https://doi.org/10.1037/xge0000960

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. In *Advances in experimental social psychology* (pp. 123–205). Elsevier. https://doi.org/10.1016/S0065-2601(08)60214-2. Bd. 19, S.

Ranisch, R., & Haltaufderheide, J. (2025). Rapid integration of LLMs in healthcare raises ethical concerns: An investigation into deceptive patterns in social robots. *Digital Society, 4*(1), 7. https://doi.org/10.1007/s44206-025-00161-2

Reddy, R. (2024). Generative AI for enhanced engagement in digital wellness programs: A predictive approach to health outcomes. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5022990

Restrepo Echavarría, R. (2025). ChatGPT-4 in the turing test. *Minds and Machines, 35*(1), 8. https://doi.org/10.1007/s11023-025-09711-6

Ruch, S., Züst, M. A., & Henke, K. (2016). Subliminal messages exert long-term effects on decision-making. *Neuroscience of Consciousness, 2016*(1). https://doi.org/10.1093/nc/niw013

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Schneider, J. (2024). Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review, 57*(11), 289. https://doi.org/10.1007/s10462-024-10916-x

Seitz, L. (2024). Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100067. https://doi.org/10.1016/j.chbah.2024.100067

Sebler, K., Kepir, O., & Kasneci, E. (2024). Enhancing student motivation through LLM–powered learning environments: A comparative study. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruiperez Valiente (Hrsg), Technology enhanced learning for inclusive and equitable quality education (Bd. 15160, S. 156–162). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72312-4_21.

Shao, J., Su, L. Y.-F., Gong, Z., & Chen, M. (2025). Conversational agents and charitable behavioral intentions: The roles of modality, communication style, and perceived anthropomorphism. *International Journal of Human-Computer Studies, 205*, Article 103616. https://doi.org/10.1016/j.ijhcs.2025.103616

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus, 3*(2), 1–5. https://doi.org/10.1093/pnasnexus/pgae035

Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday*. https://doi.org/10.5210/fm.v29i1.13541

Smith, B., Bouadjenek, M. R., Kheya, T. A., Dawson, P., & Aryal, S. (2025). *A comprehensive analysis of large language model outputs: Similarity, diversity, and bias* (arXiv:2505.09056). *arXiv*. https://doi.org/10.48550/arXiv.2505.09056

Sonlu, S., Bendiksen, B., Durupinar, F., & Güdükbay, U. (2025). Effects of embodiment and personality in LLM-based conversational agents. In *2025 IEEE conference virtual reality and 3D user interfaces (VR)* (pp. 718–728). https://doi.org/10.1109/VR59515.2025.00094

Spatola, N., Wudarczyk, O. A., Nomura, T., & Cherif, E. (2023). Attitudes towards robots measure (ARM): A new measurement tool aggregating previous scales assessing attitudes toward robots. *International Journal of Social Robotics, 15*(9–10), 1683–1701. https://doi.org/10.1007/s12369-023-01056-3

Strahan, E. J., Spencer, S. J., & Zanna, M. P. (2002). Subliminal priming and persuasion: Striking while the iron is hot. *Journal of Experimental Social Psychology, 38*(6), 556–568. https://doi.org/10.1016/S0022-1031(02)00502-4

Sun, Y., & Wang, T. (2026). Be friendly, not friends: How LLM sycophancy shapes user trust. In *Proceedings of the 2026 CHI conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3772318.3791079

Sutil-Martín, D. L., & Rienda-Gómez, J. J. (2020). The influence of unconscious perceptual processing on decision-making: A new perspective from cognitive neuroscience applied to generation Z. *Frontiers in Psychology, 11*, 1728. https://doi.org/10.3389/fpsyg.2020.01728

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vonschallen, S., Eyssel, F., & Schmiedel, T. (2026). *Understanding persuasive interactions between generative social agents and humans: The knowledge-based persuasion model (KPM)*. https://doi.org/10.48550/arXiv.2602.11483

Vonschallen, S., Finsler, L. J. C., Schmiedel, T., & Eyssel, F. (2026). Exploring persuasive interactions with generative social robots—An experimental framework. *Proceedings of the 13th International Conference on Human-Agent Interaction*, 542–544. https://doi.org/10.1145/3765766.3765866

Vonschallen, S., Häusler, R., Schmiedel, T., & Eyssel, F. (2026). Never say never: Exploring the effects of knowledge availability on agent persuasiveness in controlled physiotherapy motivation dialogues. *Frontiers in Artificial Intelligence, 9*, Article 1810725. https://doi.org/10.3389/frai.2026.1810725

Vonschallen, S., Oberle, D., Schmiedel, T., & Eyssel, F. (2026). *Knowledge-based design requirements for generative social robots in higher education* (Version 3). arXiv https://doi.org/10.48550/ARXIV.2602.12873.

Vonschallen, S., Zumthor, E., Simon, M., Schmiedel, T., & Eyssel, F. (2026). Knowledge-based design requirements for persuasive. In M. Staffa, J.-J. Cabibihan, B. Siciliano, S. S. Ge, L. Bodenhagen, A. Tapus, S. Rossi, F. Cavallo, L. Fiorini, M. Matarese, & H. He (Hrsg.) (Eds.), *Social Robotics + AI (Bd. 16131, S. 224–240)*. Singapore: Springer Nature. https://doi.org/10.1007/978-981-95-2379-5_15.

Yang, Z., Khatibi, E., Nagesh, N., Abbasian, M., Azimi, I., Jain, R., & Rahmani, A. M. (2024). ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. *Smart Health, 32*, Article 100465. https://doi.org/10.1016/j.smhl.2024.100465

ZareiNejad, M., & Tavana, P. (2025). Application of generative AI in patient engagement. In A. Zamanifar & M. Faezipour (Hrsg), Application of generative AI in healthcare systems (S. 119–154). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-82963-5_5.

Zhang, Y., Ouh, E. L., Ho, A., Lo, S. L., Tan, K. W., & Lin, F. (2025). PromptTutor: Effects of an LLM-based chatbot on learning outcomes and motivation in flipped classrooms. In *Proceedings of the 30th ACM conference on innovation and technology in computer science education* (Vol. 1, pp. 445–451). https://doi.org/10.1145/3724363.3729095

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–38. https://doi.org/10.1145/3639372