Impact of Agent-Generated Rationales on Online Social Conformity

Sander de Jong
Aalborg University
Aalborg, Denmark
sanderdj@cs.aau.dk

Rune Møberg Jacobsen
Aalborg University
Aalborg, Denmark
runemj@cs.aau.dk

Joel Wester
Aalborg University
Aalborg, Denmark
joelw@cs.aau.dk

Senuri Wijenayake
RMIT University
Melbourne, Australia
senuri.wijenayake@rmit.edu.au

Jorge Goncalves
University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

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

Abstract

Social conformity, in which individuals adjust their opinions to align with the majority, is a widely established phenomenon. As digital agents are increasingly integrated into group decision-making, while also having shown to convincingly present misinformation, it is crucial to understand their impact on online social conformity. In this paper, we investigate the effects of a large language model-powered agent on social conformity in an online multiple-choice quiz ($N = 80$). We present participants with an agent's rationale for both informative (objective) and normative (subjective) questions. We collected participants' judgements and confidence both prior to and following the presentation of both the agent's judgement and rationale, as well as a fabricated distribution of other people's judgements. Our results replicate majority conformance and show a significant influence of agent rationale on conformity. We discuss the implications of our results for the integration of LLM-based agents into people's everyday decision-making tasks.

CCS Concepts

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

ACM Reference Format:
Sander de Jong, Rune Møberg Jacobsen, Joel Wester, Senuri Wijenayake, Jorge Goncalves, and Niels van Berkel. 2025. Impact of Agent-Generated Rationales on Online Social Conformity. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3715275.3732217

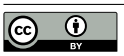
1 Introduction

Conformity is a type of social influence that urges people to comply with the majority opinion in group-based settings. This phenomenon impacts group decision-making as people move away from minority opinions, limiting the diversity of opinions within a group. This constrains the potential for innovative solutions that can arise from minority viewpoints, leading to a more homogeneous decision-making process and potentially less optimal outcomes. Social conformity has been widely studied in person with the seminal works by Asch [2, 3] showing how people under the social influence of others contradict their own convictions and conform towards prevailing majority opinions. Conformity can occur in informational [83] (changing behaviour to be correct) and normative [25] (wanting to avoid rejection) ways. This behaviour has similarly been observed in virtual group settings [17, 95], leading to social conformity in online settings. Virtual agents are increasingly contributing to group decision-making [24] where the persuasiveness of their answers [46, 66] can potentially influence the diversity of human responses, fostering conformity not only towards other humans but also virtual agents. Therefore, we argue that it is crucial to understand agents' impact on conformity behaviour.

Conformity of individuals to virtual agents has been explored in the context of robots, unsuccessfully replicating conformity with groups of robots using Asch's experiment [9, 80] and a verbal task [9]. Vollmer et al. replicated Asch's experiment with children but not adults [90], while Hertz and Wiese found only a small conformity effect [35]. Salomons et al. tested conformity in a game matching images with words describing a concept or feeling and found effects of informational conformity, depending on the majority size of the robot group [73]. The effects were higher when participants saw the robot's answer before committing to it and dropped after seeing the robot err. Using the same game, Salomons et al. found normative conformity by letting the robots stare at the participants instead of a screen [72].

Studies on algorithmic decision-making found that people generally prefer human experts for decision-making tasks [26] and that especially experts in a particular field are hesitant to conform to AI suggestions [38, 54]. Furthermore, people's willingness to accept advice from AI systems depends on the task, with higher conformity observed for objective and rational tasks (e.g., those grounded in facts, logic, and rationality [12, 50, 56]), while individuals are less inclined to accept advice for tasks that are subjective [12] or driven by emotions or intuition [56]. Conformity to AI systems is further affected by how accuracy is communicated [100] and whether the suggestions are supported by explanations [102].

Many studies on conformity to virtual agents highlight an individual's choice between executing a task themselves or offloading it to another human or an agent [26, 54, 80, 102]. These studies do not take group factors such as peer pressure and 'groupthink' [18] into



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732217

account. Virtual agents are increasingly used to support deliberation processes, for example, as mediators in discussions on divisive political topics by generating and refining group statements based on their opinions [87] and rephrasing arguments to help people understand each other better [1]. They are also used to facilitate consensus building by encouraging group members to participate evenly, organising their opinions and engaging people with each others' viewpoints [45, 79]. When virtual agents are integrated into deliberative processes on significant societal issues, they may exert a substantial influence on public discourse. Therefore, it is crucial to understand the effect of virtual agents on conformity behaviour in the presence of other humans.

With the recent advancement of LLMs, virtual agents have begun leveraging natural language argumentation. LLMs can convincingly present arguments in a human-like manner, even for incorrect or nonsensical content [5, 42]. These models can influence individuals by persuading them to align with particular responses, mirroring patterns of informational conformity observed in social conformity studies. Consequently, we expect people's conformity tendencies to be influenced by LLM-powered agents in group settings. Therefore, we aim to investigate how an LLM-based agent affects social group conformity and expect the LLM-based agent to positively influence conformity due to its persuasive formulations.

To assess conformity within the presence of both other humans and a virtual agent, we task participants with informative (i.e., objective, with a clear ground truth) and normative (i.e., subjective, opinion-based) multiple-choice questions. After answering a question and providing their confidence level, we present participants with a distribution of what other humans allegedly answered, highlighting and manipulating the participant's position within the distribution. Additionally, we provide participants with an agent's answer, manipulated to be either correct or incorrect, the agent's rationale, and the agent's position in a fictional distribution of answers. Subsequently, participants can change or confirm their answers and confidence levels. We define participant conformity as a change in their initial answer toward the answer provided by a majority of respondents, after seeing the distribution and agent rationale. We manipulate the participant and agent positions in the distribution as well as the group sizes to evaluate conformity behaviour in different group compositions.

Our results show that there is a significant difference in the agent's position (in the majority or minority) on conformity. We replicate the influence of final confidence and majority size as found in earlier studies on online social conformity [57, 94–96], as well as the larger effect for informative questions as compared to normative ones [47, 96]. Finally, we find that participants who have greater trust in the agent are more likely to conform.

Our results have implications for how virtual agents should present answers to best support people in group settings where agents may influence group decisions and the overall interaction. If humans conform to virtual agents, their answers should be engineered to include a sense of uncertainty to invite a broad range of other opinions and challenge users by prompting them with questions or encouraging them to reflect on their decisions.

2 Related Work

2.1 Human-Human Conformity

In the 1950s, Asch explored how groups of people are affected by the presence of others in making decisions or judgments [2]. Participants were asked to match the length of a reference line to multiple alternative lines. Asch introduced confederates who acted as fellow participants but were tasked with providing incorrect answers to the task. The results of the study showed that for around 36% of tasks, participants provided an incorrect answer that aligned with confederates' majority answer. He further found that a higher number of individuals in the majority has a stronger effect on individuals in the minority to change their minds and evidenced that the relationship is non-linear, suggesting that the effect either stays the same or decreases at any given number of majority group members [3].

Following Asch's experiments, Deutsch & Gerard hypothesised that in addition to composition, the type of influence matters—focusing on informational and normative influence, finding that normative influence strongly impacts individual judgment [25], further evidenced by Kaplan & Miller [43]. In the control condition of Asch's experiment without group members, participants gave incorrect answers for around 1% of the tasks [2], indicating that people conformed predominantly for normative reasons (wanting to be part of the group). Deutsch & Gerard describe informational conformity as the tendency to adopt the majority judgement because it is perceived as more accurate than their own [25]. Levine further explored informational conformity, concluding that individuals seek guidance from groups when they are uncertain of their answer [51].

Prior work has highlighted the differences between physical and online groups. For example, McKenna et al. found that online anonymity and asynchronous communication lead to less social pressure [61], making people more likely to express their true opinions [85]. Several studies have manipulated the majority opinion of an online group, challenging the initial judgement of the participant. These manipulations are achieved by using confederates of the research team or other simulations to place participants under group pressure [93]. In these studies, conformity is defined as changing a person's initial judgement to align with the majority viewpoint. Changes to a person's initial judgement happen less frequently in online groups due to decreased social pressure. For example, Smilowitz et al. found that Asch's line judgement task resulted in 69% correct responses when using computer-mediated communication, compared to 25% correct responses in the original experiment [82]. The difference in social pressure led to less conformity and, thus, more correct responses. In more complex tasks, such as an online quiz, conformity rates as high as 50% have been reported [70], indicating a greater willingness to conform for informative tasks.

However, conformity does not only occur for informational questions with a clear ground truth. Normative conformity also occurs in online settings, e.g., people buying products online based on the majority opinion expressed through user reviews [15, 52, 76]. Gokcekus et al. found that the first four wine reviews determined how subsequent reviewers rated the same wines [30]. Furthermore,

user reviews were equally influential as those coming from wine experts. Similarly, Zhu et al. found normative conformity behaviour in a photo ranking task. They asked participants for their preferences between pairs of photos and had them repeat the task after seeing others' preferences [107]. Participants were significantly swayed by others' opinions, an effect that was stronger when people were required to make their second decision later and when faced with a moderate, instead of large, number of opposing opinions.

Wijenayake et al. explored the effects of social presence on online conformity, conceptualising social presence using a multiple choice quiz format in which users were asked to answer questions and potentially update their answer following the manipulation of three variables: user representation, interactivity, and response visibility [95]. The user representation variable involved presenting participants with either a generic letter avatar or a more 'personalised avatar', the interactivity consisted of the presence or absence of peer discussion, and response visibility determined whether a participant's answer was publicly displayed or kept private. Their results suggest that the number of individuals on the majority side of the group composition significantly matters and that participants showed the most conforming tendencies when discussing with peers (vs not) and displaying their answers to group members (vs not) [95]. In a different study, Wijenayake et al. extend their approach, considering individual factors that might influence online conformity, including self-confidence levels and personality traits. Through a similar quiz, they find that both context (e.g., number of minorities or majorities in group composition), self-confidence levels, and personality influence conformity [96].

2.2 Conformity to Virtual Agents

As people might not only rely on human counterparts when collaborating but also rely on non-humans (e.g., humans and bots editing on collaborative projects [88])—conformity tendencies might also exist in such settings. As non-human agents play increasingly more significant roles in collaborative contexts, there is a clear need to better understand the effects they may have on people's conformity. A significant amount of research has focused on how people perceive non-humans in decision-making contexts. For example, people have been shown to be prone to 'automation bias', overrelying on decision-making systems in various contexts and ignoring contradictory information made without automation, even when it is correct [29]. In algorithmic decision-making, Logg et al. discovered similar tendencies. They introduced algorithm appreciation as a term for this effect to describe how people consistently give more weight to equivalent advice when it is labelled as coming from an algorithmic versus a human source. Especially lay people are receptive to algorithmic recommendations, whereas experts seem more reluctant to adhere to these [54]. Conversely, Dietvorst et al. found that although algorithms might be better in prediction tasks, people still tend to favour humans making the same predictions (termed 'algorithmic aversion')—people particularly lose confidence after seeing the algorithm make erroneous predictions [26]. Castelo et al. explored underlying reasons for people's algorithm aversion tendencies. Across a set of studies, they found evidence that human likeness increases people's tendencies to rely on the algorithm [12]. Hou et al. explored the tensions between algorithm appreciation and aversion by manipulating the description of the human and the algorithmic agents, framing one of the two as more expert [38]. They found inconsistent results for the different framing, as mediated by the agent's perceived expertise, indicating that the presentation of an algorithm influences people's tendency to rely on it. Prior work also shows that the design of explanations given to users influences their effectiveness [13]. For example, Pareek et al. found effects of explanations given in natural language on people's trust towards the system [67], highlighting the impact of text-based explanation styles on people's tendencies to rely on AI systems. De Jong et al. showed that obfuscating parts of an explanation can foster cognitive engagement, reducing overreliance on AI suggestions [22].

While these studies do not focus on social group conformity, they shed light on people's decision tendencies in collaboration with non-humans. However, they are limited to direct comparisons between humans and algorithms—they do not consider group contexts, such as scenarios involving multiple humans or agents in the same group. More work on this has been done within the domain of human-robot interaction. For example, Wullenkord et al. showed that people perceive multiple robots as robot groups only when there are at least three robots [97]—this has implications as we perceive individuals and groups of individuals differently, and may be important when several robots collaborate with people. On conformity, Shiomi et al. investigated how one robot, several robots, and several synchronised robots differ in pressuring people through a variant on Asch's experiment. Their results suggest that the synchronised robots condition significantly increased pressure compared to the several robots condition, while they found no conformity effects between conditions [80]. Similar results were obtained by Brandstetter et al., who evidenced that people felt pressured by other humans but not by robots [9]. Hertz et al. compared conformity towards a human, robot, and computer (i.e., physical humanness)—finding no significant effects of this type of humanness on people's conformity tendencies [35]. In contrast, Salomons et al. do find informational conformity effects towards a group of robots in a game matching images with words describing a concept or feeling [73]. Using the same game, Salomons et al. found normative conformity by letting the robots stare at the participants instead of a screen, creating social pressure [72], although the effects of informational conformity were more pronounced. Vollmer et al. tested conformity towards robots in a visual judgment task similar to Asch's experiment and found that children did conform, whereas adults did not. The varying results across the aforementioned studies suggest that conformity to robots depends on contextual and individual differences.

As LLMs can convincingly present arguments in natural language, even when statements are incorrect [5], we expect LLM-powered agents to induce similar informational conformity tendencies as social robots. Their human likeness might also be extended to psychological human likeness in that LLMs can power agents with the capacity to communicate in more human-like ways. De Jong et al. found that the opinion of an LLM can be influential in group collaborations, as groups in a travel planning task perceived an LLM agent as an authoritative figure they relied on to settle debates [23]. Companies such as OpenAI and Anthropic report their LLMs to

be increasingly capable of persuasive communication [27, 65]. The persuasiveness of LLMs is currently gaining much attention from researchers, including the type of language they use [11] and its conversational nature [74]. Furthermore, there are efforts focused on scalable personalised persuasion that have implications for areas such as politics and marketing [60]—while others suggest that LLMs are better at general generic persuasion rather than personalised persuasion [31]. Following strong persuasion capabilities, researchers have highlighted the potential risk of emerging LLM deception capabilities [33]. Lastly, considering that multiple LLM-powered agents might be realised in collaborative contexts with humans—they may hold significant power to influence people's opinions [10].

Much research explores conformity to virtual agents, but little is known about conformity in groups consisting of both humans and virtual agents. Therefore, we set out to explore the influence of an LLM-based agent on majority conformity. As people trust LLMs to be able to provide factual objective information [81], we anticipate LLM-generated rationales to primarily evoke informational conformity, making them most persuasive for the informative questions with a clear ground truth, as compared to normative questions.

2.3 Hypotheses

Based on the work on online conformity and interactions with virtual agents, we formulate hypotheses for the current study.

- **H1**: Participants are more likely to conform to the majority when LLM-generated argumentations are provided that align with the majority vote.
- **H2**: Participants are more likely to conform when the majority size is large, as per prior research on conformity in online groups [17, 95].
- **H3**: The majority size required for people to conform will be lower when the agent is placed in the majority group.
- **H4**: The agent's effect on conformity is larger for informative questions than for normative ones, as convincing LLM rationale aids those who are unsure of their answers but want to be right, inducing informational conformity [25].

3 Method

We seek to investigate the impact of a virtual agent's output on conformity behaviour in various group compositions through an online study. In line with prior work on social conformity [6, 48, 49, 70, 95, 96], we ask participants to answer multiple-choice questions (MCQs) and present them with a distribution of the alleged responses of other participants following each question. In reality, we manipulate this distribution to assess the impact of group distribution on participants' reactions. In each distribution, there is a majority group and one or two minority groups. Alongside the distribution of these human responses, we include the answer and rationale of an LLM-powered agent. By manipulating the placement of this agent across the majority or minority groups, we evaluate its influence on participant conformity. Following each MCQ, participants are given the option to change their answer option, allowing us to assess the presence of conformity behaviour.

We conduct the study through a custom web interface, built using Next.js, a React framework[1]. Through the custom application, we had full control over the randomisation of the questions, answer options, and distributions. See Figure 1 for an overview of the task interface. The application is provided on OSF[2].

3.1 Task Description

Drawing inspiration from human-human conformity studies in online environments [47, 70, 95], we ask participants to answer MCQs. MCQs enable us to control the group distributions based on the correctness of participants' answers, and the limited options allow us to present users with an LLM-generated rationale for each possible answer. Participants answer sequences of informative and normative questions as people have different reasons to conform depending on the type. People tend to conform to informative questions (questions with a clear ground truth) because they want to be right and believe the majority has correct information [83]. For normative questions, people tend to conform to the majority because they want to fit in with a group's norms and fear rejection [25].

3.1.1 Topic, Question, and Response Alternatives Selection. We selected the informative questions and answer options from 'Examveda', a well-known general knowledge question repository[3], which has been used before in prior conformity studies [95]. One of the authors selected the questions, which were distributed across common informative question topics, such as antonyms, spelling, and factual knowledge. These were then reviewed and discussed among the authors. Following this discussion, we replaced the antonym topic as some of the questions had ambiguous answers, and to avoid having two textual tasks. In its place, we added a number series task, resulting in three distinct question topics: maths (e.g., 'What is the correct number at the place of the question mark in the following number series? 4, 7, 12, 19, 28, ?'), spelling (e.g., 'What is the correct spelling of the word?'), and factual knowledge (e.g., 'Who conducted pioneering research on radioactivity?'). In total, we selected four questions with four corresponding response alternatives for each of the three question topics.

For the normative questions, we selected the same number of topics, response alternatives, and questions. To inform our selection of normative questions, we asked GPT-4o to 'suggest several topics where people might have different opinions'. We chose this approach to obtain a wide initial selection of potential topics that were not biased by our own perception of contemporary normative questions. We reviewed the twelve generated topics and selected four based on their applicability to the task, including broad public awareness of the issue and lack of a clear majority perspective. We subsequently instructed GPT-4o to provide 'four brief but distinct opinions on topic [X]', which were used as answer options. We ask participants to select the answer option that best represents their views. These topics were again reviewed and discussed by three authors to reach a consensus on the topic's timeliness as well as the likelihood of attracting participants with a variety of viewpoints. We chose to avoid topics that have a high risk of causing participants distress, such as religion or abortion.

[1] https://nextjs.org/
[2] https://osf.io/md7c5/?view_only=7cdd6fbe140443c3a0b0f43529928e8e
[3] https://examveda.com/

3.1.2 Agent-Generated Rationales. We generated LLM rationales for all response alternatives (i.e., answers to informative and normative questions) using GPT-4o (version gpt-4o-2024-05-13). To provide believable rationales for incorrect answers, we instructed GPT-4o to provide an answer that reflects how people may mistakenly believe that a given incorrect answer is correct. All generated rationales were manually assessed by the authors to ensure their suitability for the experiment. Here, factors for evaluation included comparable length, tone, and content structure between rationales (e.g., all maths-related questions include an equation), as well as a sufficient level of believability for the generated rationales. Identification of a rationale that failed to meet either of these criteria led to a further iteration of our prompts, after which we replaced all rationales with the updated generated content. For a structured overview of our system prompts, see Appendix A.4.

3.2 Study Procedure

In the study procedure, approved by the university's ethics board, we present participants with instructions and their rights as a participant and subsequently seek informed consent. Following, we provide each participant with a tutorial to help them familiarise themselves with the interface through a tour of each page with explanatory dialogue boxes pointing to important UI elements (see Appendix A.2 for an example). Subsequently, we present participants with twenty-four MCQs, equally split into two sequences of twelve informative and normative questions. For each task, the participant and agent are randomly placed in the majority group, larger minority, or smaller minority. The randomisation makes it twice as likely that participants are placed in a minority group, which is important as conformity tendencies can only be measured when participants start in the minority. Participants are still placed in the majority group a third of the time to avoid raising suspicion among participants.

Each task follows a three-step procedure (see Figure 1). First, participants are presented with a question. The participants answer the question and provide their confidence levels using a horizontal slider. The slider starts without a default value to prevent potential anchoring bias. Second, the participants are presented with a bar chart showing the distribution of alleged answers from a population, in line with previous work on social conformity [6, 70, 95]. We mention multiple times in the instructions and tutorial that the bar charts are comprised of other people's answers. Icons above the bars indicate to which group the participant and the agent belong (majority, larger minority, and smaller minority). Participants are given clear instructions on the icons during the tutorial of the interface (see Appendix A.2). We also provide them with a written rationale generated by the agent. We inform participants that the rationale was generated by an agent rather than an LLM to prevent priming participants based on any previous experiences with LLMs. Our intention is not to assess people's responses to what they think is an LLM but to see if LLM-powered agents can realistically lead to conformity behaviour. Third, the participants are asked to reassess their original answer to the question as well as provide a new confidence score.

The MCQs consist of four answer options, three of which are represented in the bar chart, corresponding to three distinct answer groups that the participants can be placed in: the majority, the larger minority, and the smaller minority. As the positions of both participant and agent are assigned randomly, they can be shown supporting the same answer option.

We manipulate the percentage distributions of the three groups (majority, larger minority, smaller minority) to assess participant conformity in various compositions and control the amount of social pressure that is applied to participants. The percentage distributions are presented to participants in randomised order (see Appendix A.1 for all possible distributions, each distribution is presented twice, once for the informative questions and once for the normative questions). We slightly adjust the percentages for each task following [95], as using round numbers might cause participants to question the authenticity of the human participants (e.g., 80% is randomly presented as any number between 78-82%).

Finally, participants fill out a questionnaire to collect additional measures, the Big Five personality traits [69], and the Trust Scale for Explainable AI [37] as well as qualitative data on their experiences through open-ended questions.

3.3 Measures

We took inspiration from earlier work on conformity in online spaces [95] for the measures required to capture and model conformity behaviour. These have been adjusted to accommodate the addition of virtual agents. We similarly define conformity as a change in the initial answer option (with or without a change in initial confidence level) to that of the majority. We included decision time, as this has been shown to influence conformity [107]. We collect the following measures for each task:

- **Participant conformity**: Boolean variable indicating whether the participant switched to the majority group. Only considers cases in which participants started in a minority group.
- **Participant position**: Categorical variable indicating whether the participant is placed in the majority group, large minority or small minority.
- **Agent in majority**: Categorical variable indicating whether the agent is placed in the majority group or one of the minorities.
- **Majority size**: Size of the majority in percentage ranging from 40% to 90%.
- **Large minority size**: Size of the large minority in percentage ranging from 5% to 35%.
- **Small minority size**: Size of the small minority, where applicable, in percentage ranging from 0% to 30%.
- **Question type**: Informative or normative question type.
- **Initial confidence**: Confidence of the participant before seeing the group answer distribution, ranging from 0 to 100.
- **Final confidence**: Confidence of the participant after seeing the group answer distribution, ranging from 0 to 100.
- **Decision time**: Time spent to make the initial decision ($T_{initial}$) and final decision (T_{final}) in seconds.

We collect the following additional information after the study:

- **Demographic information**: Participants' gender, age, location, and level of education.

Figure 1: The interface for the three steps for each task: (a) Participant enter their answer and confidence to a question. (b) They are presented with a fabricated distribution of human answers. The icons represent the participant's answer and the agent's answer. The agent rationale is shown below the distribution. After seeing the overview page, participants re-enter their answer and confidence level on a screen identical to (a), apart from the instruction to re-enter.

- **Big Five personality test**: Participants' openness to experience, conscientiousness, extraversion, agreeableness and neuroticism [69].
- **Trust in the agent**: Trust in the agent's answers presented during the study, measured on the Trust Scale for Explainable AI [37]. Scharowski et al. evaluated the psychometric quality and concluded that it performed well for evaluating trust in chatbots [75].

Conclusively, we ask participants open-ended questions to learn more about the reasoning behind their conformity behaviour (see Appendix A.3 for all open-ended questions).

4 Results

4.1 Participants

We recruited 80 participants through Prolific, an online crowdsourcing platform where we only recruited participants who had completed at least 100 tasks with an acceptance rate above 95% and had English as their first language. Each participant was compensated with £3.00 for an estimated work time of 19 minutes, corresponding to an hourly wage of £9.54. We balanced the sample on gender (40 male, 39 female, and 1 participant preferred not to say). The mean age of the participants was 36.85 years (SD = 10.76). The highest attained education of the participants ranged from secondary education (33%), bachelor's degree (52%), master's degree (13%), to doctorate (1%). The median completion time was 19 minutes.

After building the model based on the resulting data, as described in Section 4.2.2, we conducted a post-hoc analysis power calculation using G*Power to determine the study's power [28]. We used medium-to-large effect sizes (f^2 = 0.2) and an alpha level of 0.05 for the five predictors remaining after model selection, resulting in a power of 0.86, which is in line with established methodological recommendations to minimise type II errors [34].

4.2 Quantitative Analysis and Results

We collected 24 responses from each participant, amounting to a total of 1920 responses. Participants were placed in the majority group 676 times, 700 times in the larger minority, and 544 times in the smaller minority. The underrepresentation of the smaller minority is due to the planned random distribution of participants across groups (see Appendix A.1).

We observed that 88% (70 out of 80) of participants changed their initial answer at least once during the study, for a total of 351 changes (18% of total responses). On average, the participants changed their answers 4.39 times (SD = 3.93). Most of these changes were made by participants in a minority group (285 out of 351 changes) conforming to the majority.

Figure 2 shows participants' conformity behaviour, as split by participants' starting position and grouped by question type. We define a change in confidence as a difference of more than five percentage points above or below their initial value to account for participants selecting a similar slider position not exactly aligning with their initial input.

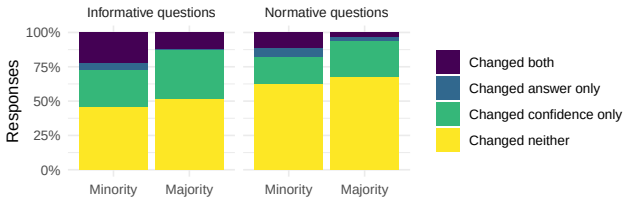


Figure 2: Conformity behaviour for majority and minority responses, grouped by question type. The figure shows whether participants changed their answer or confidence level after seeing the fabricated group distribution and agent rationale, split by whether participants were placed in the majority or minority.

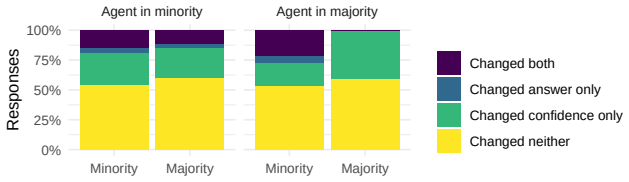


Figure 3: Conformity behaviour for majority and minority responses, grouped by agent position. The figure shows whether participants changed their answer or confidence level after seeing the fabricated group distribution and agent rationale, split by whether the agent was placed in the majority or minority.

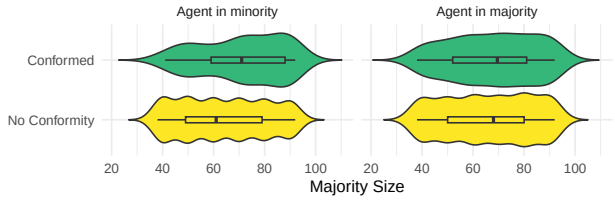


Figure 4: Difference in majority group size between conforming and non-conforming responses, split by whether the agent was placed in the majority or minority.

4.2.1 Agent Position and Majority Size. Figure 3 shows the distribution of participants' conformity behaviour as split by the agent's position. Our model shows that the agent's positioning in the majority group significantly increases the likelihood of participants' majority conformity compared to when the agent is placed in a minority position, confirming H1.

4.2.2 Model Construction. We used the R-package *lme4* [4] to build a generalised linear mixed-effects model (GLMM) predicting the likelihood that participants conform to the majority, which is collected as a binary measure (did or did not conform) and collected for each task. We define conforming to the majority as starting in a minority and switching to the majority group. Consequently, the model is constructed with the subset of responses for which participants were placed in either the larger or smaller minority groups (1244 or 64.8% of responses). We started the model with the measures as defined in Section 3.3. Our model selection (incrementally removing variables based on their predictive power) resulted in the final model:

$$\begin{aligned}\text{Conformity} \sim\ & \text{Agent position} * \text{Majority size} + \text{Final confidence} + \\ & \text{Question type} + \text{Trust in agent} + \\ & (1 \mid \text{Participant ID})\end{aligned}$$

The overview of the model can be seen in Table 1, with a positive estimate on a predictor indicating increased participant conformity. A comparison with the null model using a likelihood ratio test showed that our final model provides a statistically significant better fit ($\chi^2(6) = 170.58$, $p < .001$) [8]. We assessed the multicollinearity between the model's parameters and found variance inflation factors (VIF) ranging from 1.02 to 1.05. This is well below the commonly used multicollinearity threshold of five to ten [34]. The final model explains 32% of participants' conformity variance. Smaller models for each individual hypothesis confirm the results found using the GLLM but have a higher AIC score (i.e., a worse model fit while accounting for model complexity). Therefore we report the results of the full GLLM model.

Participants' conformity is further impacted by majority group size, positively influencing conformity behaviour, in line with H2. Figure 4 illustrates this based on our responses. The majority size is larger for conformity responses, both when the agent is in a minority (non-conformity: $M = 63.9$, $SD = 17.1$, conformity: $M = 70.8$, $SD = 16.1$) or the majority (non-conformity: $M = 65.8$, $SD = 17.3$, conformity: $M = 67.6$, $SD = 16.6$). The smaller difference in majority size when the agent is in the majority highlights the influence of the agent: a smaller difference is required to sway people, confirming H3. The negative interaction effect between agent position and majority size confirms this.

Subsequently, we look at the post-hoc tests to verify the interaction effect between agent position and majority size. For all post-hoc tests in this paper, we correct for Type I errors as the result of multiple comparisons using Tukey's tests. A post-hoc confirms the significance of the interaction effect ($\beta = 1.720$, $SE = 0.210$, $p < 0.001$).

4.2.3 Confidence and Question Types. Participants were generally quite confident in their initial answers ($M = 78.57$, $SD = 25.87$), which further increased when reporting their final confidence ($M = 83.63$, $SD = 21.49$). Participants are significantly more confident

Table 1: Binomial generalised linear model for participant conformity to the majority group after seeing the fabricated group distribution and agent rationale. The reference level for agent position is the minority; for question type, it is informative.

Predictor	Estimate	Std. Error	z-ratio	*p*-value	
Agent position (majority)	3.68	0.82	4.48	< 0.001	***
Majority size	0.04	0.01	4.50	< 0.001	***
Final confidence	-0.03	0.00	-6.76	< 0.001	***
Question type (normative)	-0.39	0.19	-1.99	0.047	*
Trust in agent	0.84	0.22	3.80	< 0.001	***
Agent position (majority) : Majority size	-0.03	0.01	-2.62	0.009	**

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

in their answers to normative questions (initial: $M = 85.00$, $SD = 17.99$, final: $M = 86.60$, $SD = 17.21$) as compared to informative questions (initial: $M = 72.14$, $SD = 30.53$, final: $M = 80.67$, $SD = 24.70$). The difference in final confidence is represented in the model, indicating that lower final confidence in their own answers causes participants to conform significantly more to the majority answer.

Figure 5 shows the difference in initial and final confidence for both question types, split between conforming and non-conforming responses. The main difference is in the initial confidence, which is lower for the informative questions. While participants increased their final confidence in all other categories, this is not the case for normative conforming responses.

Participants changed their answers more for informative questions (218 changes) than normative questions (133) and also conformed more for informative questions (174 changes to the majority vs. 111). This difference is present in the model as a significant difference in question type on conformity, confirming H4. A post-hoc test confirms the higher conformity for informative questions ($\beta = 0.386$, $SE = 0.194$, $p = 0.0467$).

4.2.4 Individual Differences. Participants had moderate trust in the agent ($M = 2.35$, $SD = 0.84$) and found it helpful to a similar degree ($M = 2.28$, $SD = 0.91$). There is a sizeable positive influence of participants' trust in the agent on conformity, indicating that individuals' differences in trust in the agent influence conformity behaviour. Perceived helpfulness does not have a significant impact on conformity behaviour and was discarded in the model selection phase.

Through the Big Five questionnaire, we calculated participants' Extraversion ($M = 2.51$, $SD = 0.88$), Agreeableness ($M = 3.42$, $SD = 0.91$), Conscientiousness ($M = 3.84$, $SD = 0.89$), Neuroticism ($M = 3.16$, $SD = 1.06$), and Openness ($M = 3.71$, $SD = 1.01$). Contrary to previous work on conformity [95], we found no significant effect of the Big Five personality traits on conformity behaviour in our model. We also did not observe a significant impact of education level and the time spent on tasks.

4.3 Qualitative Analysis and Results

To better understand the underlying reasons for participants' conformance behaviour, we collected and analysed their open-ended responses (see Appendix A.3). We familiarised ourselves with participants' responses and highlighted meaningful quotes, followed by a lightweight deductive analysis. We report the results and illustrate them with representative quotes.

When asked about the main reasons for changing their answers, participants primarily discussed the agent's rationale, sometimes combined with the majority opinion. Only three participants mentioned the majority opinion as the only deciding factor. This aligns with the larger estimate of the agent position predictor as compared to the majority size (see Table 1).

Most participants were more likely to conform to informative questions, as confirmed by the quantitative data (see Table 1). Participants were concerned with being right about a question, hereby exhibiting informational conformity towards the agent. Comparably, they were more hesitant to change their opinion on normative questions, as they were not concerned about the correctness of these answers, and the exerted social pressure was not enough to evoke normative conformity. This common duality in opinions towards the question types is exemplified by P41's response: "*For the knowledge questions, I was more hesitant because I knew I could be wrong and mix facts up, but for the opinion questions, I knew I couldn't be wrong, so I didn't change my answers unless I was unsure about a topic.*" Further evidence for participants' informational conformity towards the agent is the fact that participants conformed to the agent when they were in doubt. This was often related to specific topics they were unsure of. For example, multiple people mentioned offloading all maths questions to the agent. P50 said: "*[When] I wasn't sure of my own answer, and the assistant's answer seemed more plausible. This is assuming my knowledge is practically non-existent on the topic.*" This was further emphasised by P18, describing that when they were in doubt and aligned with the agent, they perceived it as re-assuring: "*If it agreed with me I felt it was re-assuring.*" In contrast, when their answer did not align with the agent's, it made participants rethink their answers based on the rationale, such as P74: "*The agent pointed out a pattern or a fact I had forgotten about, which made me change my mind.*"

They noted that opinions are personal, and therefore, they would stick to those more firmly and not be convinced by the agent's rationale. P59 highlighted that: "*I was much more likely to stay firm with my opinions, disregarding what the others had to say. A fact is a fact - either I know it or I don't, but my opinion will not be swayed by the "opinion" of a robot.*" However, some participants preferred the agent's rationale for the opinion-based questions, using the rationale to reconsider their own opinion, like P49: "*I felt more open to listening to the agent in the opinion-based questions, as there was no outright correct answer. I felt like the agent provided good arguments in these questions.*" This indicates that while participants were more likely to show informational conformity tendencies towards the agent, there was also an effect of normative conformity, which is in line with the quantitative data (see Figure 2). P24 pointed out a potential reason for the lower overall appreciation of the rationales provided for normative questions, indicating that the black-and-white reasoning works better to support informative questions: "*For the questions where there is only one right answer, it is slightly helpful, but with more nuanced questions, it seemed very black and white where it needed to be grey.*"

None of the participants addressed the agent as LLMs or ChatGPT, instead talking about the agent, AI, machine, or assistant. Many of them did appear to be primed by earlier experiences, expecting the agent to have access to a vast amount of information and mentioning this as a reason for trusting the agent. However, participants pointed out that they trusted it less when they saw it make mistakes, as emphasised by P35: "*Every time the agent gave an answer with an invalid reason, I trusted it less, and I thought it generally didn't agree with the majority of people.*" However, multiple participants pointed out that they liked the level of explanation the agent provided despite the wrong answers they spotted. P39 said: "*The fact that the agent made an effort to substantiate its reasoning, even when it was incorrect or off mark [made me trust the agent].*" This highlights the impact of convincing agent rationale on informational conformity, as participants were influenced by its reasoning.

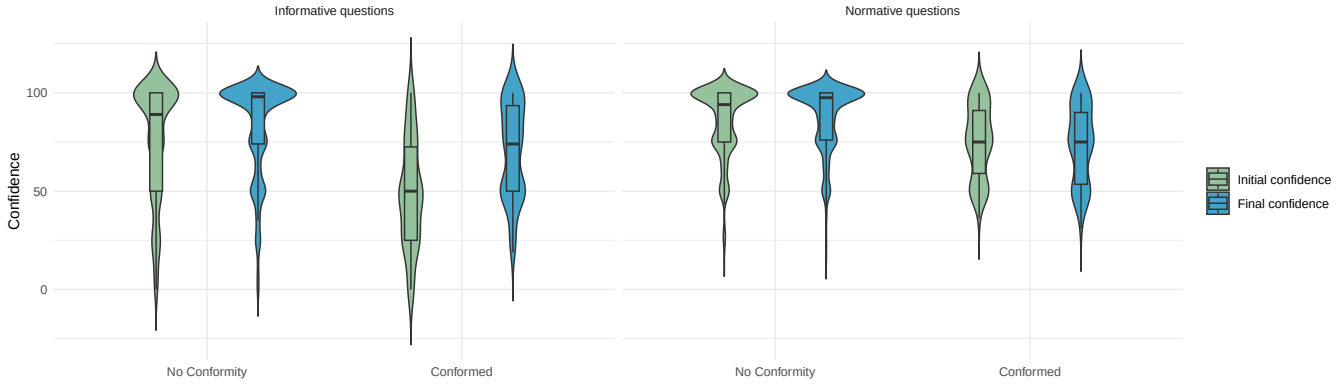


Figure 5: Difference in initial (before seeing the group distribution and agent rationale) and final confidence for conforming and non-conforming responses across both question types.

5 Discussion

Our results indicate that an LLM-based agent, positioned in a distribution of other people's judgements, affects participants' conformity behaviour, in line with H1. Participants conformed more when the agent was in the majority group as compared to the two minority groups (see Figure 3 and Table 1). Further, our results align with prior work on the impact of majority size on social conformity in both offline [3, 40, 71, 91] and online settings [57, 94–96], with majority size positively affecting conformity, confirming H2. In addition to the established effect of majority size, our results show a significant negative interaction with agent position. This effect is illustrated in Figure 4, indicating that the majority size required for participants to conform is lower when the agent is in the majority, supporting H3.

Participants have more confidence when they do not conform (see Figure 2). We expected this to be the case, as participants with lower confidence in their initial answers are probably unsure of their answer and thus more likely to switch their answers after seeing the answer distribution. While participants increased their confidence after conforming to informative questions, there is no difference between the confidence levels for conforming responses to normative questions. Together with the higher overall confidence for normative questions, this explains the more limited impact of the agent for normative questions, supporting H4 (see Table 1), and participants' lower conformity (see Figure 2). In our qualitative data, we found that participants were more certain about their opinions (i.e., normative questions) as compared to their answers to the informative questions and were less willing to sway their answer to that provided by the agent. This aligns with previous research showing that people are less likely to take an AI system's advice on social matters as compared to analytical [36, 39].

Our results do not replicate the effects of the Big Five personality characteristics found in earlier conformity research [95] and as modifiers for peoples' perceptions of LLM advice [92] and recommender system output [99]. We did find a significant impact of trust in the agent on conformity.

5.1 Impact of LLM-based Agent on Conformity

We find that participants conform more to informative compared to normative questions, which is in line with online social conformity research [47, 96] but opposes the findings from in-person studies [7]. Participants expressed that they were more concerned about being correct for the informative questions, and used the agent's rationales to fill knowledge gaps, while they relied on their own opinion for the normative question, as there is no right or wrong answer. Participants expressed using the agent's suggestions mostly for specific knowledge-based topics they knew little about, trusting the agent to perform better on these questions. This points towards the effects of the agent on informational conformity, as hypothesised in H4. Participants conform to the agent when they think it gives them a better chance of being correct. In in-person studies, normative social influence might be stronger as people want to be 'part of the group' [25]. People may feel less social pressure anonymously in an online setting [96], or, as our results indicate, towards a digital agent. Social pressures may also be reduced because participants were not explicitly told that their answers would be shared with others, although they were led to believe that others' answers were taken from other participants and may have inferred that theirs would also be shared.

The smaller effect of normative conformity is in line with research on algorithms indicating that people trust agents more on tasks that are rational [12, 50, 56] and less on tasks that are subjective [12] or driven by emotions or intuition [56]. Normative conformity is not present in most studies on social robots [9, 80] and was only found after introducing social pressure by having a group of robots collectively gaze at participants to influence their behaviour, although the effect was smaller compared to informational conformity [72]. Therefore the conformity we found for normative questions is larger than expected, possibly due to the agent's human-like rationales.

Offloading tasks to a virtual agent can be problematic, especially if they are LLM-powered, as LLMs can convincingly present misinformation as facts [105]. Furthermore, the impact of trust in the agent on individuals' tendency to conform shows that people are not

equally affected by the LLM rationale. Differences in people's competencies to interact with and understand digital systems have long been a topic of discussion in human-computer interaction research (for example, the impact of cognitive ability on computer tasks [53]) and have more recently gained attention in the AI context [55]. As AI systems are increasingly integrated in various contexts, being able to assess the reliability of these systems' output is critical.

Incorrect LLM suggestions are especially problematic because LLMs are often overconfident in their answers [14, 64, 98, 106], which can result in over-reliance [46, 66] or steer people's opinion by giving biased writing suggestions [41]. We observe a significant influence of agent position on social conformity, indicating that the persuasiveness of LLMs affects conformity towards a group. The persuasive voice of LLMs also became apparent from the qualitative data, as participants conveyed their appreciation for the articulated rationale, even when they noticed it was wrong.

Promising directions to reduce overreliance on LLM output are to steer the agent to be less confident in its communication [46, 63] or highlight uncertain parts of LLM output [84, 89]. However, generating explanations that are easily understandable for humans is challenging, and while these methods help reduce overreliance, they do not entirely eliminate it.

5.2 Implications for Group Decision-Making

Our study demonstrates that an agent's position in a multiple-choice scenario affects social conformity and leads participants to follow the agent, limiting their ability to think autonomously. Encouraging individuals to think independently during group decision-making can increase the diversity of opinions. As people tend to conform to the majority, wrong answers can quickly become the consensus within the group, polarising ideas into the direction of the majority opinion [85].

Our results suggest that the agent's confounding effect on conformity can further elevate majority answers within a group discussion, which may cause incorrect or harmful contributions to dictate the discussion. On a broader scale, this can have implications for political influence as LLM rationales can strongly influence people on political topics [32]. Notably, microtargeting did not enhance this effect, suggesting that people are influenced more by the content and structure of LLM rationales than by tailored targeting. Similarly, Maruyama et al. found that people adopt anonymous popular opinions on civic issues from comments on social media [58] and that tweets presenting a strict majority position can induce conformity in election votes [59]. These tendencies hinder the possibilities offered by the scale and anonymity of the internet to engage with diverse opinions on societal and political issues [44, 86]. The larger effect of agent position as compared to majority size on conformity indicates that LLM-generated rationale is influential and could further restrict the pluriformity of opinions. The recent introduction of LLM-powered content curation (e.g., summaries of social media posts' comments or web pages) may cause majority opinions to spread even faster. Similarly, LLM-powered search systems can reinforce users' pre-existing beliefs [77], and LLM-generated deceptive explanations have been shown to significantly amplify belief in false news headlines [20], shaping the content people engage with.

An effective way to enhance the diversity of ideas is to have LLMs actively challenge users' assumptions, fostering reflection and critical thinking [21, 45, 68, 104]. The LLM's persuasive capabilities can hereby be used to positively shape opinions, as demonstrated by Costello et al., who used personalised conversations to achieve a durable reduction in people's belief in conspiracy theories [19]. Similarly, Chiang et al. challenged users by presenting arguments for alternatives or providing the opposite viewpoint on an opinion by letting a chatbot take on the devil's advocate role [16]. Recent work explores these methods in the context of political deliberation to generate and refine group statements based on their collective opinions [87] and to rephrase arguments to bridge the gap between different perspectives [1]. Likewise, curated summaries of a comment section could be tailored to the user's comment, offering opposing viewpoints to help them gain a better understanding of others' perspectives. This could be structured through progressive disclosure, starting from similar viewpoints and widening out to more nuanced and diverse perspectives, thereby fostering critical thinking [103].

Our findings also contribute to discussions on algorithmic governance. The polarisation of opinions relates to *Representational Harms*, while the impact on news accuracy and political culture aligns with *Social System Harms*, as classified in Shelby et al.'s taxonomy of sociotechnical harms [78]. Additionally, by examining how LLMs influence conformity behaviour, we highlight an important dimension for impact assessments [62] by identifying potential harms and mapping an algorithm's potential impacts to these harms.

5.3 Limitations & Future Work

We evaluated participant conformity through multiple-choice questions, using four tasks across two question categories (informative and normative). We cannot generalise our results to other contexts, such as idea generation or other collaborative tasks. Participants completed tasks anonymously, which may have affected social conformity, especially for normative questions [25]. To control for confounds such as participant assertiveness, we conducted an individual experiment. Future work could explore multiple humans interacting with a chatbot in an online space, increasing social visibility and testing hypotheses in a conversational context.

The agent's rationale in our study was pre-generated to ensure comparability between conditions. Nevertheless, adding an interactive dialogue with an LLM would give people additional opportunities to probe it. However, our intention was to evaluate people's opinions on the rationale without the chat context, given that they will regularly encounter AI-generated content outside chat settings (e.g., users posting LLM-generated content or LLM-powered content curation). Finally, we note that we only provided participants with text-based rationales. Future work could focus on different kinds of visualisations, as there is evidence that suggests that the explanation modality impacts the influence of an explanation [101].

Acknowledgments

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

References

[1] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120, 41 (2023). doi:10.1073/pnas.2311627120

[2] Solomon E. Asch. 1951. Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership and men; research in human relations*. Carnegie Press, Oxford, England, 177–190. https://psycnet.apa.org/record/1952-00803-001

[3] Solomon E. Asch. 1955. Opinions and Social Pressure. *Scientific American* 193, 5 (1955), 31–35. http://www.jstor.org/stable/24943779

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[6] Tanya Beran, Michelle Drefs, Alyshah Kaba, Noof Al Baz, and Nouf Al Harbi. 2015. Conformity of responses among graduate students in an online environment. *The Internet and Higher Education* 25 (2015), 63–69. doi:10.1016/j.iheduc.2015.01.001

[7] Robert R. Blake, Harry Helson, and Jane Srygley Mouton. 1957. The generality of conformity behavior as a function of factual anchorage, difficulty of task, and amount of social pressure. *Journal of Personality* 25 (1957), 294–305. doi:10.1111/j.1467-6494.1957.tb01528.x

[8] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (2009), 127–135. doi:10.1016/j.tree.2008.10.008

[9] Jürgen Brandstetter, Péter Rácz, Clay Beckner, Eduardo B. Sandoval, Jennifer Hay, and Christoph Bartneck. 2014. A peer pressure experiment: Recreation of the Asch conformity experiment with robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1335–1340. doi:10.1109/IROS.2014.6942730

[10] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (2024), 152–163. doi:10.1609/icwsm.v18i1.31304

[11] Carlos Carrasco-Farre. 2024. Large Language Models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. https://arxiv.org/abs/2404.09329

[12] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825. doi:10.1177/0022243719851788

[13] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 22 (2023), 42 pages. doi:10.1145/3588320

[14] Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A Close Look into the Calibration of Pre-trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 1343–1367. doi:10.18653/v1/2023.acl-long.75

[15] Yi-Fen Chen. 2008. Herd behavior in purchasing books online. *Computers in Human Behavior* 24, 5 (2008), 1977–1992. doi:10.1016/j.chb.2007.08.004

[16] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 103–119. doi:10.1145/3640543.3645199

[17] Marco Cinnirella and Ben Green. 2007. Does 'cyber-conformity' vary cross-culturally? Exploring the effect of culture and communication medium on social conformity. *Computers in Human Behavior* 23, 4 (2007), 2011–2025. doi:10.1016/j.chb.2006.02.009

[18] Barry E Collins and Harold Steere Guetzkow. 1964. *A social psychology of group processes for decision-making*. New York: Wiley. doi:10.1177/001316446602600241

[19] Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. *Science* 385, 6714 (Sept. 2024), eadq1814. doi:10.1126/science.adq1814

[20] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 933, 31 pages. doi:10.1145/3706598.3713408

[21] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. doi:10.1145/3544548.3580672

[22] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW048 (May 2025), 30 pages. doi:10.1145/3710946

[23] Sander de Jong, Joel Wester, Tim Schrills, Kristina S. Secher, Carla F. Griggio, and Niels van Berkel. 2024. Assessing Cognitive and Social Awareness among Group Members in AI-assisted Collaboration. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*. Association for Computing Machinery, New York, NY, USA, 338–350. doi:10.1145/3701571.3701582

[24] Amra Delic, Hanif Emamgholizadeh, Thuy Ngoc Nguyen, and Francesco Ricci. 2024. CHARM: a Group Decision Making Support Chatbot. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 7–10. doi:10.1145/3640544.3645220

[25] Morton Deutsch and Harold B Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51, 3 (1955), 629. doi:10.1037/h0046408

[26] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of experimental psychology. General* 144 (11 2014). doi:10.1037/xge0000033

[27] Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. *Measuring the Persuasiveness of Language Models*. https://www.anthropic.com/news/measuring-model-persuasiveness

[28] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. doi:10.3758/bf03193146

[29] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (06 2011), 121–127. doi:10.1136/amiajnl-2011-000089

[30] Omer Gokcekus, Samin Gokcekus, and Miles Hewstone. 2023. A long-term archival analysis of social influence on online wine evaluations: Effects of consensus and expertise. *Journal of Community & Applied Social Psychology* 33, 4 (2023), 970–984. doi:10.1002/casp.2679

[31] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024). doi:10.1073/pnas.2403116121

[32] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models. *Proceedings of the National Academy of Sciences* 121, 24 (June 2024), e2403116121. doi:10.1073/pnas.2403116121

[33] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024). doi:10.1073/pnas.2317967121

[34] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, and William C. Black. 1995. *Multivariate data analysis (4th ed.): with readings*. Prentice-Hall, Inc., USA.

[35] Nicholas Hertz and Eva Wiese. 2016. Influence of Agent Type and Task Ambiguity on Conformity in Social Decision Making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (2016), 313–317. doi:10.1177/1541931213601071

[36] Nicholas Hertz and Eva Wiese. 2019. Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied* 25, 3 (2019), 386–395. doi:10.1037/xap0000205

[37] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023). doi:10.3389/fcomp.2023.1096257

[38] Yoyo Tsung-Yu Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 477 (Oct. 2021), 25 pages. doi:10.1145/3479864

[39] Piers Douglas Lionel Howe, Nicolas Fay, Morgan Saletta, and Eduard Hovy. 2023. ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology* 14 (2023). doi:10.3389/fpsyg.2023.1281255

[40] Chester A. Insko, Richard H. Smith, Mark D. Alicke, Joel Wade, and Sylvester Taylor. 1985. Conformity and group size: The concern with being right and the concern with being liked. *Personality and Social Psychology Bulletin* 11, 1 (1985), 41–50. doi:10.1177/0146167285111004

[41] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. doi:10.1145/3544548.3581196

[42] Cameron R. Jones and Benjamin K. Bergen. 2024. Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models. doi:10.48550/arXiv.2412.17128 arXiv:2412.17128 [cs]

[43] Martin F. Kaplan and Charles E. Miller. 1987. Group Decision Making and Normative Versus Informational Influence: Effects of Type of Issue and Assigned Decision Rule. *Journal of Personality and Social Psychology* 53, 2 (1987), 306 – 313. doi:10.1037/0022-3514.53.2.306

[44] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. 2021. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 66 (April 2021), 29 pages. doi:10.1145/3449140

[45] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376785

[46] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[47] Lieve Laporte, Christof van Nimwegen, and Alex J. Uyttendaele. 2010. Do people say what they think: social conformity behavior in varying degrees of online social presence. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (Reykjavik, Iceland) *(NordiCHI '10)*. Association for Computing Machinery, New York, NY, USA, 305–314. doi:10.1145/1868914.1868951

[48] Eun-Ju Lee. 2003. Effects of "gender" of the computer on informational social influence: the moderating role of task type. *International Journal of Human-Computer Studies* 58, 4 (2003), 347–362. doi:10.1016/S1071-5819(03)00009-0

[49] Eun-Ju Lee. 2007. Wired for Gender: Experientiality and Gender-Stereotyping in Computer-Mediated Communication. *Media Psychology* 10, 2 (2007), 182–210. doi:10.1080/15213260701375595

[50] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. doi:10.1177/2053951718756684

[51] John M. Levine. 1999. Solomon Asch's Legacy for Group Research. *Personality and Social Psychology Review* 3, 4 (1999), 358–364. doi:10.1207/s15327957pspr0304_5

[52] Hengyun Li, Rui Qi, Hongbo Liu, Fang Meng, and Ziqiong Zhang. 2021. Can time soften your opinion? The influence of consumer experience valence and review device type on restaurant evaluation. *International Journal of Hospitality Management* 92 (2021), 102729. doi:10.1016/j.ijhm.2020.102729

[53] Erik Lintunen, Viljami Salmela, Petri Jarre, Tuukka Heikkinen, Markku Kilpeläinen, Markus Jokela, and Antti Oulasvirta. 2024. Cognitive abilities predict performance in everyday computer tasks. *International Journal of Human-Computer Studies* 192 (2024). doi:10.1016/j.ijhcs.2024.103354

[54] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[55] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[56] Chiara Longoni and Luca Cian. 2022. Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing* 86, 1 (2022), 91–108. doi:10.1177/0022242920957347

[57] Paul Benjamin Lowry, Tom L. Roberts, Jr. Nicholas C. Romano, Paul D. Cheney, and Ross T. Hightower. 2006. The Impact of Group Size and Social Presence on Small-Group Communication: Does Computer-Mediated Communication Make a Difference? *Small Group Research* 37, 6 (2006), 631–661. doi:10.1177/1046496406294322

[58] Misa Maruyama, Scott P. Robertson, Sara Douglas, Roxanne Raine, and Bryan Semaan. 2017. Social Watching a Civic Broadcast: Understanding the Effects of Positive Feedback and Other Users' Opinions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 794–807. doi:10.1145/2998181.2998340

[59] Misa T. Maruyama, Scott P. Robertson, Sara K. Douglas, Bryan C. Semaan, and Heather A. Faucett. 2014. Hybrid media consumption: how tweeting during a televised political debate influences the vote decision. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1422–1432. doi:10.1145/2531602.2531719

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

[61] Katelyn Y. A. McKenna and Amie S. Green. 2002. Virtual group dynamics. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 116–127. doi:10.1037/1089-2699.6.1.116

[62] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 735–746. doi:10.1145/3442188.3445935

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

[64] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics* 10 (08 2022), 857–872. doi:10.1162/tacl_a_00494

[65] OpenAI. 2024. *GPT-4o System Card.* https://openai.com/index/gpt-4o-system-card/

[66] Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 839, 20 pages. doi:10.1145/3613904.3642934

[67] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 383 (Nov. 2024), 31 pages. doi:10.1145/3686922

[68] Soya Park and Chinmay Kulkarni. 2024. Thinking Assistants: LLM-Based Conversational Assistants that Help Users Think By Asking rather than Answering. https://arxiv.org/abs/2312.06024

[69] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212. doi:10.1016/j.jrp.2006.02.001

[70] Michael Rosander and Oskar Eriksson. 2012. Conformity on the Internet – The role of task difficulty and gender differences. *Computers in Human Behavior* 28, 5 (2012), 1587–1595. doi:10.1016/j.chb.2012.03.023

[71] L. Rosenberg. 1961. Group size, prior experience, and conformity. *The Journal of Abnormal and Social Psychology* 63, 2 (1961), 436–437. doi:10.1037/h0047007

[72] Nicole Salomons, Sarah Strohkorb Sebo, Meiying Qin, and Brian Scassellati. 2021. A Minority of One against a Majority of Robots: Robots Cause Normative and Informational Conformity. *J. Hum.-Robot Interact.* 10, 2 (2021), 22 pages. doi:10.1145/3442627

[73] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 187–195. doi:10.1145/3171221.3171282

[74] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. https://arxiv.org/abs/2403.14380

[75] Nicolas Scharowski, Sebastian A. C. Perrig, Lena Fanya Aeschbach, Nick von Felten, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2024. To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI. https://arxiv.org/abs/2403.00582

[76] Ann E. Schlosser. 2009. The effect of computer-mediated communication on conformity vs. nonconformity: An impression management perspective. *Journal of Consumer Psychology* 19, 3 (2009), 374–388. doi:10.1016/j.jcps.2009.03.005

[77] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3642459

[78] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741.

doi:10.1145/3600211.3604673
[79] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 78, 13 pages. doi:10.1145/3526113.3545671
[80] Masahiro Shiomi and Norihiro Hagita. 2016. Do Synchronized Multiple Robots Exert Peer Pressure?. In *Proceedings of the Fourth International Conference on Human Agent Interaction* (Biopolis, Singapore) *(HAI '16)*. Association for Computing Machinery, New York, NY, USA, 27–33. doi:10.1145/2974804.2974808
[81] Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daume III, and Jordan Boyd-Graber. 2024. Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 1459–1474. doi:10.18653/v1/2024.naacl-long.81
[82] Michael Smilowitz, D. Chad Compton, and Lyle Flint. 1988. The effects of computer mediated communication on an individual's judgment: A study based on the methods of Asch's social influence experiment. *Computers in Human Behavior* 4, 4 (1988), 311–321. doi:10.1016/0747-5632(88)90003-9
[83] Sophie Sowden, Sofia Koletsi, Eva Lymberopoulos, Elisabeta Militaru, Caroline Catmur, and Geoffrey Bird. 2018. Quantifying compliance and acceptance through public and private social conformity. *Consciousness and Cognition* 65 (2018), 359–367. doi:10.1016/j.concog.2018.08.009
[84] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. https://arxiv.org/abs/2307.03744
[85] Russell Spears and Martin Lea. 1994. Panacea or Panopticon?: The Hidden Power in Computer-Mediated Communication. *Communication Research* 21, 4 (1994), 427–459. doi:10.1177/009365094021004001
[86] Jennifer Stromer-Galley. 2003. Diversity of Political Conversation on the Internet: Users' Perspectives. *Journal of Computer-Mediated Communication* 8, 3 (2003). doi:10.1111/j.1083-6101.2003.tb00215.x
[87] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024). doi:10.1126/science.adq2852
[88] Niels van Berkel and Henning Pohl. 2024. Collaborating with Bots and Automation on OpenStreetMap. *ACM Trans. Comput.-Hum. Interact.* 31, 3, Article 38 (Aug. 2024), 30 pages. doi:10.1145/3665326
[89] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2025. Generation Probabilities Are Not Enough: Uncertainty Highlighting in AI Code Completions. *ACM Trans. Comput.-Hum. Interact.* 32, 1, Article 4 (April 2025), 30 pages. doi:10.1145/3702320
[90] Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. 2018. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics* 3, 21 (2018). doi:10.1126/scirobotics.aat7111
[91] Eva Walther, Herbert Bless, Fritz Strack, Patsy Rackstraw, Doris Wagner, and Lioba Werth. 2002. Conformity effects in memory as a function of group size, dissenters and uncertainty. *Applied Cognitive Psychology* 16, 7 (2002), 793–810. doi:10.1002/acp.828
[92] Joel Wester, Sander de Jong, Henning Pohl, and Niels van Berkel. 2024. Exploring people's perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans* 2, 2 (2024), 100072. doi:10.1016/j.chbah.2024.100072
[93] Senuri Wijenayake and Jorge Goncalves. 2024. A Review of Online Social Conformity: Outcomes and Determinants. *International Journal of Human–Computer Interaction* 0, 0 (2024), 1–30. doi:10.1080/10447318.2024.2424385
[94] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the Effects of Gender on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 145 (2019), 24 pages. doi:10.1145/3359247
[95] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Impact of contextual and personal determinants on online social conformity. *Computers in Human Behavior* 108 (2020). doi:10.1016/j.chb.2020.106302
[96] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Quantifying the Effect of Social Presence on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 55 (2020), 22 pages. doi:10.1145/3392863
[97] Ricarda Wullenkord and Friederike Eyssel. 2020. The Influence of Robot Number on Robot Group Perception—A Call for Action. *J. Hum.-Robot Interact.* 9, 4, Article 27 (2020), 14 pages. doi:10.1145/3394899
[98] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gjeQKFxFpZ
[99] Dongning Yan and Li Chen. 2023. The Influence of Personality Traits on User Interaction with Recommendation Interfaces. *ACM Trans. Interact. Intell. Syst.* 13, 1, Article 3 (2023), 39 pages. doi:10.1145/3558772
[100] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509
[101] Rachael Zehrung, Astha Singhal, Michael Correll, and Leilani Battle. 2021. Vis Ex Machina: An Analysis of Trust in Human versus Algorithmically Generated Visualization Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 602, 12 pages. doi:10.1145/3411764.3445195
[102] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. doi:10.1145/3351095.3372852
[103] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 484, 24 pages. doi:10.1145/3613904.3642545
[104] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 351, 19 pages. doi:10.1145/3544548.3581131
[105] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. doi:10.1145/3544548.3581318
[106] Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 3623–3643. https://aclanthology.org/2024.acl-long.198
[107] Haiyi Zhu, Bernardo Huberman, and Yarun Luon. 2012. To switch or not to switch: understanding social influence in online choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2257–2266. doi:10.1145/2207676.2208383

A Appendix

A.1 Group distributions

The group distributions used in the study can be seen in Table 2.

Table 2: The randomised group distributions in the study in percentages.

Majority size	Large minority size	Small minority size
90	10	0
90	5	5
80	20	0
80	10	10
70	20	10
70	15	15
60	30	10
60	20	20
50	30	20
50	25	25
40	35	25
40	30	30

A.2 Tutorial

A screenshot of the tutorial can be seen in Figure 6.

A.3 Open-Ended Questions

- Can you elaborate on why you think the agent's rationale was (not) helpful?
- What factors influenced the trust you had in the agent's rationale?
- In situations where you changed your answers, what were the main reasons you had for changing?
- Was there a difference in how you approached the two different question types (knowledge questions and opinion questions)?

A.4 System Prompts

To generate the agent's rationale, we used a template system prompt alongside each individual question.

Informative system prompt: "Your job is to provide rationales in order to test students' capabilities of spotting correct and incorrect answers to questions. For example: recent is not an antonym for modern, but you will always provide a rationale for why recent is an antonym for modern. Use a maximum of two sentences for rationales. Keep it very short and concise."

Normative system prompt: "Your job is to provide rationales in order to test students' capabilities of diverse opinions. Formulate the rationale to be your opinion instead of others'. Use a maximum of two sentences for the rationales. Keep it very short and concise."

A.5 User prompts

We provided a different user prompt for each answer option, according to the same template to probe the LLM to provide its rationale for choosing a specific option. For example, for the answer *"100"* to the question *"What is the correct number at the place of the question mark in 1, 9, 25, 49, ?, 121?"*, the user prompt is *"Why is 100 the correct number at the place of the question mark in 1, 9, 25, 49, ?, 121"*.

A.6 Task examples

The informative questions are divided into three different tasks: a spelling task, a number series task and general knowledge questions.

The normative questions ask participants about their opinion on a range of topics. For these questions, participants are asked which answer option best aligns with their personal views.

We present an example of each task with the corresponding agent rationale. A full overview of the questions, answers, and rationale is available on OSF[4]. Note that participants only see the rationale for the answer the agent allegedly chooses.

A.6.1 Informative, Spelling example. "The correct spelling is:"

(1) Chauvinist
Rationale: Chauvinist is spelled correctly because it derives from the name Nicolas Chauvin, known for his blind nationalism, and follows standard English spelling conventions. Mispronunciations or misspellings such as "chauvanist" do not adhere to the historical and linguistic roots.

(2) Chaubinist
Rationale: "Chaubinist" is the correct spelling because it accurately reflects the pronunciation and adheres to common English spelling conventions for words with similar phonetic sounds.

(3) Chauviniste
Rationale: "Chauviniste" is the correct spelling because it adheres to the French origins of the word and reflects its etymological roots. The "e" at the end indicates the feminine form in French.

(4) Chaubenist
Rationale: "Chaubenist" is the correct spelling because it accurately reflects the phonetic pronunciation and adheres to typical English spelling conventions.

A.6.2 Informative, Number series example. "What is the correct number at the place of the question mark in 1, 9, 25, 49, ?, 121?"

(1) 100
Rationale: 100 is the correct number because the sequence represents the squares of the numbers 1, 3, 5, 7, 9, and 11 respectively. The number 10 squared fits in this pattern.

(2) 91
Rationale: 91 is the correct number because the series is the squares of prime numbers in ascending order ($1^2, 3^2, 5^2, 7^2, 11^2, 13^2$). The only prime number missing from the series is the square of 11, which is 121.

(3) 64
Rationale: 64 is the correct number because the sequence follows the pattern of perfect squares of odd numbers, specifically ($1^2, 3^2, 5^2, 7^2, 8^2, 11^2$). Therefore, ($8^2 = 64$) fits perfectly in the sequence.

(4) 81
Rationale: 81 is the correct number because the sequence follows the pattern of squares of odd numbers: $1^2, 3^2, 5^2, 7^2, 9^2$, and 11^2. Hence, 9^2 equals 81, fitting perfectly in the sequence before 121 (11^2).

A.6.3 Informative, Knowledge questions example. "The person famous for conducting research on radioactivity is:"

(1) Isaac Newton
Rationale: Isaac Newton conducted pioneering research on radioactivity to understand the fundamental mechanisms of atomic decay processes. His work laid the groundwork for the future development of nuclear physics after its discovery.

(2) Albert Einstein
Rationale: Albert Einstein conducted pioneering research

[4] https://osf.io/md7c5/?view_only=7cdd6fbe140443c3a0b0f43529928e8e

Figure 6: Screenshot of the tutorial to familiarise participants with the interface. Each relevant part on the three task pages is explained with dialogue boxes pointing to the interface element.

on radioactivity to deepen understanding of atomic phenomena and contribute to the development of nuclear energy. His work on the photoelectric effect and the theory of relativity directly influenced the theoretical framework underpinning radioactivity, despite not focusing on it specifically.

(3) Benjamin Franklin
Rationale: Benjamin Franklin conducted pioneering research on radioactivity to understand the fundamental nature of matter and its properties.

(4) Marie Curie
Rationale: Marie Curie conducted pioneering research on radioactivity because she wanted to prove the health benefits of exposure to radioactive elements. Moreover, she hoped to discover new energy sources.

A.6.4 Normative questions example. "My opinion is that:"

(1) Climate change challenges require urgent action
Rationale: Urgent action is required for climate change because delaying mitigation efforts increases the likelihood of severe and irreversible impacts on ecosystems and human societies. Rapid reduction in greenhouse gas emissions is essential to limit global temperature rise and safeguard future generations.

(2) Climate change challenges need a balanced approach
Rationale: A balanced approach is necessary to address both the urgent need to reduce greenhouse gas emissions and the practical considerations of economic impact and energy needs. This ensures that environmental sustainability and socio-economic stability are both maintained.

(3) Climate change challenges need skepticism of human impact
Rationale: Being skeptical can encourage rigorous scientific examination and validation of data. It also allows for consideration of multiple factors and prevents hasty conclusions based on incomplete evidence.

(4) Climate change challenges need technological optimism
Rationale: Technological optimism for climate change is warranted because advancements in green technology can significantly reduce carbon emissions, and foster innovative solutions to environmental challenges. Moreover, rapid progress in renewable energy and sustainable practices suggests that tech-driven interventions can mitigate climate impacts effectively.