Utrecht University

Advocatus DiaBOTli:

Artificial Dissent, Task Domain & Conformity

Paul Ballot, BSc (0642878)

Master Thesis in Social, Health & Organizational Psychology (SHOP)

Social Influence Track | Social Cognition & Artificial Intelligence (SC&AI) Group

Supervisor: Dr. Manuel Barbosa de Oliveira

2nd Assessor: Dr. Ruud Hortensius

Date: 26.06.2022

Place: Utrecht, The Netherlands

Word Count: 8635

Manuscript can be made publicly accessible.

Abstract

With the rapid rise of artificial intelligence (AI), research interest has shifted to understanding its potential for social influence. In line with the "Computers Are Social Actors" (CASA) paradigm, computational agents have been proven capable of inducing conformity. Furthermore, previous studies demonstrated that dissenting social robots can reduce conformity. With their increased availability compared to social robots, however, the question remains about whether the latter also applies to AI. Therefore, the current study is investigating the impact of AI dissent and how it is moderated by task domain and the attitudes towards AI. To assess its effect on conformity, we conducted a pre-registered online experiment (N = 94) manipulating task type and whether the software agent dissented from or agreed with the confederate majority. Following our expectations, results indicated a medium-sized reduction of conformity in the presence of a dissenting AI agent. Contradicting our hypothesis, this did not depend on the individual's attitude towards AI. Additionally, task domain did not moderate the decrease in conformity, but it did increase the impact of AI dissent on accuracy for social tasks compared to analytical tasks. Thereby, our results indicate that while an AI agent's ability to break majority influence appears not to depend on the task, its capacity to exert minority influence might.

Keywords: Human-AI interaction, conformity, artificial intelligence, artificial influence, social influence, dissent, majority influence, minority influence

Introduction

When Joseph Weizenbaum (1976), an early pioneer in "artificial intelligentsia" (p. 184), came up with a script-based computer program capable of simulating rudimentary conversations, he hardly could have imagined, that his '*ELIZA*' would pave the way for what might be the "fourth industrial revolution" (Bock et al., 2020, p. 317) over half a century later. However, despite his invention being far less sophisticated than its spiritual successors in the likes of ChatGPT or Google's Bard, Weizenbaum made an observation that stood the test of time: Users perceived *ELIZA* to actually understand them. They interacted with the program the way they would with a social entity, even though they were factually aware of it being a machine (Weizenbaum, 1976). It would take many more years before insights like these were finally formalized as the 'Computers are Social Actors' paradigm (CASA; Nass & Moon, 2000; see also Reeves & Nass, 1996), which implies that individuals mindlessly rely on social conventions and expectations, when interacting with computers. A phenomenon that enables majorities of social robots or computers to induce conformity. However, so far, little research has been done on the influence of non-human minorities. Hence, the present study aims on investigating the link between conformity and artificial dissidence as well as the impact of perceived domain-actor fit and attitudes towards artificial intelligence (AI). But first, we review the current literature on the CASA paradigm as well as conformity to human and non-human agents.

Computers are social actors?

The CASA paradigm states that people treat robots and digital actors like real humans. An effect that has been found to induce behavioural and cognitive responses otherwise reserved for interpersonal communication (Krämer, 2005). This phenomenon can be explained by a strong reliance on social categories (e.g., gender or ethnicity), the use of overlearned social behaviours (e.g., politeness or reciprocity) and the exhibition of "premature cognitive commitments" (Nass & Moon, 2000, p. 82), the latter meaning that people will not update an assumption made about the artificial actor, even in the light of contradicting, new information. It is triggered by cues activating social scripts. Once done, individuals will stop searching for additional cues which might be capable of activating alternative, more appropriate scripts (Westerman et al., 2020).

For CASA to apply, however, two conditions concerning the nature of the artificial agents (e.g., a robot or a bot) need to be met: First, it must provide sufficient social cues for individuals to perceive it as "worthy of social responses" (Nass & Moon, 2000, p. 83). And while what constitutes 'sufficient' varies interpersonally and is also depending on situational factors (Waytz et al., 2010), some characteristics increasing 'humanness' for an artificial actor have been identified, including the act of replacing a human in specific roles (Nass & Moon, 2000; Westerman et al., 2020; Xu, 2019). The general significance of this prerequisite is, however, controversial, with results from human-computer interaction indicating that social behaviour might not depend on an actor's level of humanness after all (Hertz & Wiese, 2016).

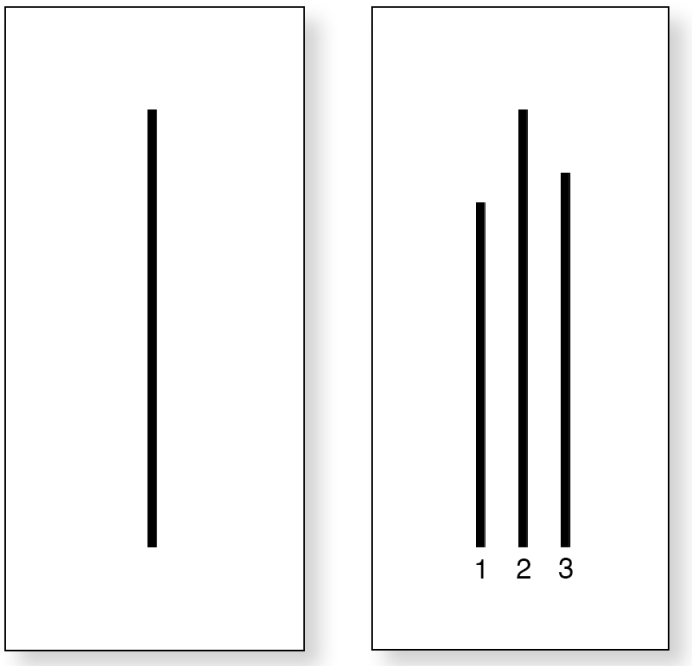
Secondly, the artificial agent must not be seen as a mere conduit for, but an autonomous source of communication (Nass & Steuer, 1993). This difference in orientation is what distinguishes human-computer interaction and computer-mediated communication (Sundar & Nass, 2000). It indicates that users need to assign at least some degree of agency to the actor, instead of perceiving it to just transmit messages from one person to another (Gambino et al., 2020). By fulfilling these requirements, the paradigm suggests an increase in likelihood of social interaction with any kind of computational actor.

Conformity and why it occurs

Conformity refers to "the act of changing one's behaviour to match the response of others" (Cialdini & Goldstein, 2004, p. 606). It was shaped by a series of groundbreaking experiments conducted by Solomon Asch in the 1950s. In his quest to understand "independence and submission to group pressure" (1951, p. 222), he developed the famous line judgement task prompting individuals to compare lines of different length to a reference standard, while being exposed to an obviously wrong, unanimous majority. The observations he made were striking: Participants would abandon their visual judgments, when encountering a group of confederates holding up a different opinion. In one third of the relevant trials, subjects followed the majority, and three out of four participants did so at least once. Further investigating these results, Asch (1955, 1956) modified the original experiment trying to identify relevant variables and their impact on conformity. Key findings include the importance of public judgement, the absence of cumulative peer pressure, the minimal increase of conformity once the

Figure 1

The Line Judgement Task developed my Asch (1951).



number of confederates exceeds three and – highly relevant for this thesis – the impact of dissent, which proved to significantly reduce conformity even with a single dissident. More recently, ambiguous situations with high levels of uncertainty regarding the correct answer were found to increase conformity (Brandstetter et al., 2014; Hertz & Wiese, 2016). Furthermore, research also learned that this is not limited to the physical world. Instead, the effects of conformity have been measured successfully both in virtual worlds (e.g., simulated environments like Second Life; Rayborn-Reeves et al., 2013) and digital settings (e.g., online classrooms or internet communities; Beran et al., 2015; Rosander & Eriksson, 2012).

In their search for an explanation for this observation, Deutsch and Gerard (1955) hypothesize two distinct types of social influence: Informational and normative influences. Informational influence describes the tendency to determine what is true based on what other people believe to be true (Cialdini, 2001). This concept, also known as social proof, stems from a need to "form an accurate interpretation of reality" (Cialdini & Goldstein, 2004, p. 606). Therefore, individuals conform because they rely on the judgments of others as a source of information to increase accuracy and identify

adaptive behaviour (Deutsch & Gerard, 1955). Or, more simply put, they conform "to be correct" (Morgan & Laland, 2012, p. 2) and might change their beliefs, because they assume the group to know more than them (Schnuerch & Gibbons, 2014).

Normative influence – on the other hand – aims to avoid rejection while obtaining social approval from peers (Cialdini & Goldstein, 2004). People thereby hope to receive normative rewards for agreeing and fitting in with their group (Morgan & Laland, 2012). On a biological level, this kind of pressure seems to be rooted in error and conflict processing as well as reward inhibition caused by disagreement (Botvinick et al., 2004; Morgan & Laland, 2012). It results from groups' aversion to deviates and individual's fear of exclusion and alienation, when not fulfilling expectations regarding their response (Kindcaid, 2010). These expectations can arise from oneself, being an internalized social process, or more commonly from the group. In the latter case and even with trivial and artificial groups, the impact of normative influence will be considerably increased (Deutsch & Gerard, 1955).

However, while normative and informational influence differ theoretically, separating them empirically can be challenging (Dávid & Turner, 2001). Additionally, neuroscientific findings suggest that they are in fact highly intertwined and – at least on a neural level – might not be distinct at all (Berns et al., 2005).

Conforming to artificial agents?

Coming up with their definition for conformity mentioned above, Cialdini and Goldstein (2004) abstained from specifying whom exactly the term "others" (p. 606) encompasses. Yet, with the application of AI tools becoming more common in areas like education, service and even therapy (Qin et al., 2021), it wasn't long before the question arose whether it could possibly include artificial agents. Early studies focusing on the potential of social robots to induce conformity failed to reproduce Asch's findings with non-human peers (Brandstetter et al., 2014; Shiomi & Hagita, 2016), apparently contradicting the CASA paradigm. And while Vollmer et al. (2018) confirmed these results for adults, they did find 7- to 9-year-old children to conform with humanoid robots. This is in line with previous research demonstrating that young children are considerably more vulnerable to the effects of social influence (Walter & Andrade, 1996; Pasupathi, 1999). However, all of these studies rely on the identical line judgement task used in Asch's (1951) original experiments. Therefore, as Hertz et al. (2019) speculate,

the stimuli might be too familiar to participants given its popularity in academia and culture, leading to an underestimation of conformity effects in those cases. And indeed: Applying alternative tasks, multiple studies were able to detect conformity with artificial agents as well as the conditions facilitating it (Hertz & Wiese, 2018; Riva et al., 2022; Salomons et al., 2018). Salomons et al. (2018), for example, observed that in 30% of critical trials, subjects conformed to a group of robots. This finding is remarkable as it closely resembles the results obtained in Asch's (1951) original study.

Importance of the task's domain

And yet another, relevant factor lies in the nature of the task, or more precisely the agent-task fit (Hertz & Wiese, 2018; Riva et al., 2022). This follows from Cialdini's (2001) principle of authority and its focus on expertise: While artificial agents should be less capable of inducing normative pressure, their informational influence might be increased depending on how competent they are perceived for a given task. Therefore, under the assumption that AI agents are perceived as highly competent in the analytical domain while lacking in social capacities, conformity should increase for the former and drop for the latter. To test this, Hertz & Wiese (2018) conducted an experiment including two different tasks – a calculation task and a task concerning the evaluation of facial expressions – with peer groups consisting exclusively of either humans, robots, or computers. Their findings indicate a significant interaction between agent type and task domain: For the analytical task conformity remained similar in all three agent conditions, while there were significant differences for the social task with computers and robots reaching lower levels compared to humans. Another study utilizing a similar task setup (subjective picture-concept fit vs. objective estimation) even found increased conformity for non-human actors given an objective task with high uncertainty (Riva et al., 2022). However, while these studies provide compelling evidence for non-human agent's potential in inducing conformity, the picture remains incomplete.

The Present Research

The aim of this study is to deepen our understanding of the role artificial agents can play in social influence processes. As hybrid systems will most likely consist of a human majority cooperating with an artificial minority, further research on artificially induced conformity should examine, how such situations could unfold (Qin et al., 2021).

For example, for the impact of robotic dissidence, Qin et al. (2021) found correct dissenters to reduce conformity and increase accuracy and incorrect dissenters to reduce conformity without increasing accuracy. However, with their heightened availability compared to social robots, the question remains about whether this effect also holds true for AI and software agents. As these are less likely to convey social cues, they fail to meet the requirements of the CASA paradigm potentially rendering them less effective regarding to normative pressure. Accordingly, robots were found to foster more compliance and trust than virtual agents (Bainbridge et al., 2008; Leyzberg et al., 2012). Furthermore, their physical presence has also been linked to being perceived as more informative and helpful (Kidd & Breazeal, 2004; Wainer et al., 2007), indicating a reduction in informational influence. This evidence suggests "fundamental differences between virtual agents and robots from a social standpoint" (Wainter at al., 2007, p. 872).

Additionally, task domain needs to be considered. Based on the findings provided above as well as the theoretical and empirical insights from the CASA paradigm, a comparable effect is plausible for AI agents as well as an interaction between task domain and AI dissidence. Presumably, the artificial dissident's impact on conformity will be stronger for analytical tasks compared to social tasks.

Lastly, none of the experiments on the interaction between task domain and conformity provided above considered personal beliefs towards artificial systems. This is problematic, as both liking and authority – principles that should moderate social influence (Cialdini, 2001; Cialdini & Goldstein, 2004) – are shaped by individual experiences and attitudes. Therefore, we expect these beliefs, as measured by the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2020; 2022), to moderate the link between AI dissidence and conformity.

Method

Participants

Participants were recruited via the Prolific platform (www.prolific.co) and compensated 9 £ per hour. In a first pilot experiment, 10 participants were recruited. After ensuring the absence of any technical issues with the data collection, another 90 participants were added. Finally, before analyzing the results, the data from the pilot study were incorporated in the overall data set.

In total, 100 individuals were recruited. Due to anomalies in the original data provided by Gorilla, one subject was excluded from the analyses. 5 more subjects failed the attention check. Thus, the final sample consisted of 94 participants (49% male, 47% female, 2% non-binary, 2% preferred not to answer; age range: 19 – 53 years, median age = 27 years, IQR = 10.75). However, based on the results of an additional power analysis with an increased number of simulations (n_{sim} = 1500), this does not pose an issue for the general analysis. Furthermore, for one subject duplicate observations were removed. Although the sample was predominantly international, it is noteworthy that substantial proportions of participants were from the European Union (42%) and South Africa (40%). A total of 16% of our subjects were found to have either a professional or academic background in AI and another 12% were familiar with the line judgement task.

Experimental Design

The study is based on a 2 (task domain: social vs. analytical) x 2 (AI dissent: incorrect majority vs. incorrect majority with correct AI minority) within-subjects design with conformity (yes or no) as the dependent variable. Based on Asch's (1956) experimental design and with the intention to build trust in the group, an additional level for AI dissent with a correct majority (including the AI agent; further referred to as neutral) was introduced. To ensure that the AI agent's position in the answer sequence did not influence the participants choice - for example, through the perception of the confederate's reaction to the AI – it was counterbalanced for between subjects.

Power considerations

Sample size was estimated based on an a priori power analysis for generalized linear mixed-effects model with domain, AI dissent and Attitudes towards AI modelled as fixed effects, and subject and trial modelled as random effects (intercepts only). Power was calculated to using a Monte Carlo power simulation. The relevant effect sizes were based on general effect size calculations for medium effects (Chen et al., 2010). This analysis suggested N = 100 as the optimal number of participants to detect a medium-sized three-way interaction of all three fixed effects (Odds Ratio = 3.47 ~

Cohen's *d* = 0.50) with 80% power at alpha = 0.05. The exact procedure can be found in the preregistration protocol (www.osf.io/8jsm2).

Stimuli

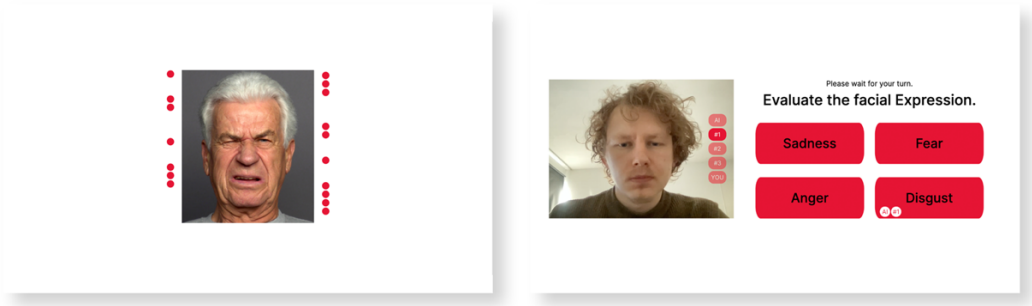
For this study, 18 unique target stimuli were created, each consisting of a portrait picture and a randomly generated series of dots left and right of it (see Figure 2). The portraits were of faces depicting negative emotional expressions randomly selected from the validated FACES database (set A; Ebner et al., 2010). Age and gender were counterbalanced for within-subject. The number of dots as well as their position relative to the image were randomly generated.

Procedure

The online experiment was designed and run using Gorilla (www.gorilla.sc). Individuals were deceived to partake simultaneously with other participants (confederates). To make this illusion more realistic, artificial loading times, a waiting screen for other participants and prerecorded and randomly glitched webcam footage of the confederates were introduced. Furthermore, the subject's webcam was activated,

Figure 2

Trial setup consisting of stimulus and answer display.



Note. The left image pictures the task stimulus layout consisting of a portrait from the FACES dataset (Ebner et al., 2010) framed by several dots on the left and right side. The right image shows the answer display. It consists of the question, the options, the sequence of answering, the choices made by the AI and the confederates and the prerecorded webcam footage of the agent choosing right now or – when it is the subject's turn to choose – live footage of the participant's webcam.

and their video output was visible to themselves while answering. Participants were told that the goal of the studies was the optimization of an AI system called "Neural Image Reasoning Network" or "NIRN". After giving consent, adjusting their webcam, reading the instructions, and waiting for other participants, the actual trials were initiated. Per run, individuals were presented with a target stimulus. As uncertainty was linked to an increase in conformity (Cialdini & Goldstein, 2004; Hertz & Wiese, 2016; Riva et al., 2022), this stimulus was visible for only 4 seconds per trial.

After being presented with the stimulus, subjects were instructed to either add or subtract the dots (analytical task, e.g., "Right Dots – Left Dots") or evaluate the facial expression (social task). For the latter, options were limited to the negative basic emotions (i.e., anger, fear, disgust, and sadness), as these are – in contrast to positive emotions – distinct in the way they are signaled (Ekman, 1992; see also, Ekman & Friesen, 1982). In every case, there were 4 choices available (see Figure 2). For the social task, these options were fixed with the four negative basic emotions. For the analytical task, options were selected randomly from a range of numbers surrounding the correct answer, clustering within a range that spans three numbers below and above the actual answer. Furthermore, direction (i.e., left to right or vice versa) and the operator for the calculation were randomly selected.

Before deciding on their answer, individuals would see three different (human) confederates and the AI agent choose an option. This number of confederates was selected because further expanding group size (e.g., from 3 to 4) was found to increase conformity only marginally (from 31.8% to 35.1%), while further lessening it (e.g., from 3 to 2) drastically reduced conformity (from 31.8% to 13.6%; Asch, 1955). Therefore, utilizing three confederates ensured that any effect measured was indeed a reduction of conformity due to the presence of a dissenter and not due to a smaller group size.

For the confederates, prerecorded webcam footage of them deciding was visible. For the NIRN AI agent, a loading animation was presented. The participants were made believe that the order of answering was randomly assigned once at the beginning of the experiment. However, in fact, participants were always assigned the last (i.e., the 5th) position in order. The choices of each agent remained visible until the end of the trial. Furthermore, after any agent's choice (including the subject's) became apparent, its video footage remained visible for another second. During their turn, the participant

was aware of the options chosen by every other entity when selecting their own answer. Post-trial, the subject was asked to rate how confident they were by their decision on a slider from 0 to 100.

In total, 18 trials per individual equally divided between social and analytical condition were conducted. In one third of these trials, the confederates and the AI agent unanimously choose the correct answer. These "neutral" trials were introduced by Asch (1956) in his original experiment and aim at building trust into the majority. Another third of the trials followed the "critical I" condition. In these cases, both the confederates and the AI agent chose a wrong option. The final third was "critical II" trials. In these cases, the confederate majority chose a wrong option and the AI agent dissented and chose the correct alternative. The trials were following a fixed pattern (n,n,c,c,n,c,c,c,c) that is repeated once. The order of the two types of critical trials as well as the order of social and analytical tasks was randomized. After completing the 18 relevant trials, individuals were asked to answer the items of the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2020; 2022). However, to reduce confusion related to the attention check, the corresponding item was slightly modified. Furthermore, participants were asked to answer demographic questions regarding age, gender, country of origin and professional or educational contact to AI and whether they ever saw the line judgement task, how in general they assessed AI's capacities in analytical and social tasks and whether they assumed any deception as part of this study. Afterwards, participants were debriefed and informed about the true nature of the study.

GAAIS Questionnaire

The GAAIS by Schepman and Rodway (2020; 2022) is a validated measure consisting of 20 items. It is capturing positive and negative attitudes towards AI, thereby reflecting both utility and concerns regarding the technology. For this study, both subscales were combined to form a general attitude score.

Data Analysis

Data were processed and analyzed using R (v.4.2.2; R Core Team, 2022). For the analysis and based on the OSF preregistration, a generalized linear mixed effects model was fitted with conformity (yes, no) as a response variable and trial type (C1, C2; N

trials were excluded from the analysis), domain (social, analytical) and attitude (as measured by the GAAIS) as fixed effects. Furthermore, based on the substantial proportion of participants expecting some kind of deception (see Manipulation check) and the differences in their response behaviour, perceived deception was also included in the model. To account for potential variability in subject and trial, these effects were considered random.

Results

Manipulation Check

As a manipulation check, subjects were asked whether they felt deceived by any part of the study. And indeed, 45% of the participants expected some sort of deception, especially regarding the other participants (32%) and the choices they made (34%) as well as the choices made by the AI agent (22%). Or, as put by one participant via prolific: "I knew immediately that these were recordings (I like to check it by doing stupid things in front of the camera)." However, while there was a significant difference (Welch Two-Sample t-test; $t_{1113} = 5.74$, $p < .001$, $d = 0.34$, 95% CI [0.23, 0.47]) in conformity behaviour between participants that believed they were being deceived ($M = 0.49$, $SD = 0.50$) and those that did not ($M = 0.60$, $SD = 0.49$), the relevant delta between C1 and C2 trials (felt deceived: $M = -0.19$, $SD = 0.25$; did not feel deceived: $M = -0.23$, $SD = 0.27$) was not significant ($t_{91} = -0.72$, $p = .48$, $d = -0.15$, 95% CI [-0.59, 0.23]).

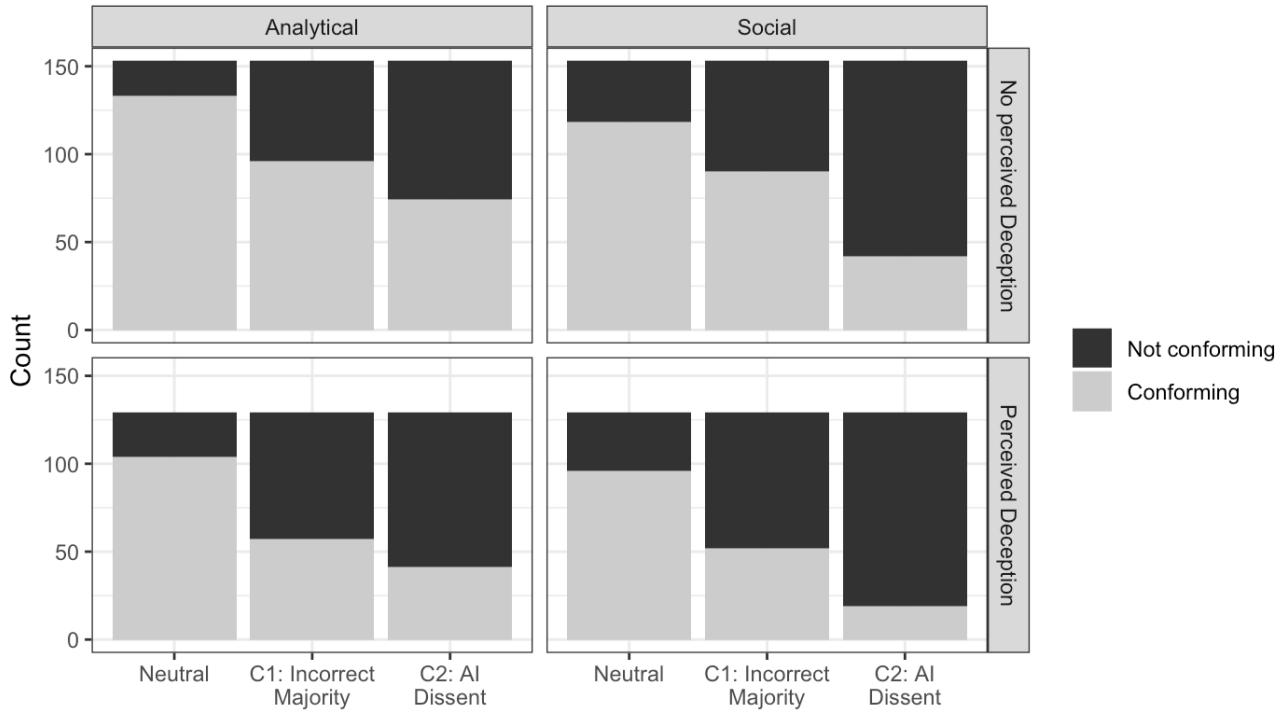
So, while overall conformity appeared to be reduced for participants that perceived deception, the general trend of the data remained manifest (see Figure 3). Therefore, respective participants were not excluded from the analysis. Instead, deception was included as a fixed effect in the model.

Model

Based on the binomial nature of our outcome variable and to account for multiple random effects, the data were analyzed with a generalized linear mixed-effects model (GLMM). In addition to the factors specified in the preregistration, perceived deception was also integrated into the model to account for the aforementioned effect. Following best-practice standards (Baayen et al., 2008; see also Charles et al., 2012), subjects and trials were included as crossed random effects.

Figure 3

Conformity for Trial, Task and Perceived Deception



For the model selection, a saturated model including all relevant fixed and random effects (Conformity ~ TrialType * Domain * GAAIS-Score + Deception + (1|ID) + (1|Trial)) was fitted first, utilizing the lme4 R package (Bates et al., 2015). Based on a binomial distribution it relied on the logit link function. Its maximum likelihood was estimated via Laplace approximation. This model, however, only detected one significant main effect for perceived deception with individuals assuming any kind of deception being less likely to conform in general (*OR* = 0.36, 95% CI [0.20, 0.64], *z* = -3.47, *p* < .001).

Following a stepwise approach, it was determined that including both subject and trial as random effects provided the most favorable structure for goodness of fit, as indicated by the Akaike information criterion (both AIC and AICc), the Bayesian information criterion (BIC), and the Bayes factor.

Subsequently, performing an automated model selection for the fixed factor combination based on the AICc, the dredge() function was applied to the global model (Barton, 2023). The resulting model demonstrated superior fit in terms of the AIC, AICc and BIC (see table 1). Furthermore, no overdispersion was detected.

Table 1

Model Comparison

			Goodness-of-Fit method			
Model	Fixed Effects	NPAR[1]	AIC	AICc	BIC	Deviance
Saturated:	AI Dissent * Domain * Attitude + Deception	11	1284.4	1284.6	1339.7	1262.4
Assumed:	AI Dissent * Domain * Attitude	10	1294.0	1294.2	1344.2	1274.0
Selected:	AI Dissent + Domain + Deception	6	1279.9	1279.9	1310.6	1367.9

Note: Random Effects (1|ID) + (1|Trial) were included in all models. [1] Number of parameters

AI Dissent

As expected, compared to trials with a unanimous majority (C1: M = 0.52, SD = 0.50), conformity was significantly reduced in the presence of AI dissent (C2: M = 0.31, SD = 0.46; OR = 0.29, *95%* CI [0.12, 0.72], z = -2.89, p = .004; selected GLMM, see Table 2). However, the wide range between confidence intervals for the OR highly suggests that the true effect could range anything from small to large.

Task Domain

Task domain was classified either social or analytical. A paired t-test confirmed the assumption that AI is perceived as more competent in the analytical domain (M = 78.09, SD = 18.70) compared to the social domain (M = 57.21, SD = 21.33, t_{93} = 7.63, p < .001, d = 0.79, 95% CI [0.56, 1.07)]). In total – including neutral trials – participants showed higher levels of conformity for analytical tasks (M = 0.60, SD = 0.49) compared to social tasks (M = 0.49, SD = 0.50). This is in line with the small and non-significant main effect for domain (OR = 0.49, 95% CI [0.20, 1.22], z = -1.66, p = .10) identified in the selected GLMM. It also matches previous findings on conformity effects for social and analytical tasks (Hertz & Wiese, 2018, Hertz et al., 2019). Within the analytical domain, subjects conformed more often in C1 trials (M = 0.54, SD = 0.50) compared to C2 trials (M = 0.41, SD = 0.49). Similarly, social tasks had higher conformity rates for C1 trials (M = 0.50, SD = 0.50) than for C2 trials (M = 0.22, SD = 0.41). With a Cohen's d of -0.50 (95% CI [-0.38, -0.64], t_{281} = -8.44, p < .001) the impact of the dissenter appears to

Table 2

Fixed Effect Estimates

Model	Fixed Effect	*Log Odds* (95% CI)	*SE*	*OR* (95% CI)	*z*	*p*
Saturated	Intercept	0.70 (-1.78, 3.19)	1.25	2.00 (0.17, 24.29)	0.56	.58
	AI Dissent[1]	-0.59 (-3.29, 2.08)	1.34	0.55 (0.04, 8.01)	-0.45	.66
	Domain[2]	0.93 (-2.03, 3.62)	1.34	2.54 (0.13, 37.34)	0.70	.49
	Attitude	0.01 (-0.67, 0.68)	0.34	1.00 (0.51, 1.97)	0.01	.99
	Deception[3]	-1.02 (-1.63, -0.45)	0.30	0.36 (0.20, 0.64)	-3.47	< .001***
	AI Dissent[1]:Domain[2]	-0.23 (-4.07, 3.60)	1.92	0.79 (0.02, 36.60)	-0.12	.90
	AI Dissent[1]:Attitude	-0.04 (-0.74, 0.66)	0.35	0.96 (0.48, 1.94)	-0.12	.91
	Domain[2]:Attitude	-0.33 (-1.03, 0.36)	0.35	0.72 (0.36, 1.43)	-0.95	.34
	AI Dissent[1]:Domain[2]:Attitude	-0.23 (-1.24, 0.78)	0.51	0.80 (0.29, 2.18)	-0.45	.66
Assumed	Intercept	0.05 (-2.44, 2.54)	1.27	1.05 (0.09, 12.68)	0.04	.97
	TrialType[1]	-0.58 (-3.21, 2.05)	1.34	0.56 (0.04, 7.77)	-0.43	.67
	Domain[2]	0.94 (-1.68, 3.56)	1.33	2.56 (0.19, 35.16)	0.70	.48
	Attitude	-.06 (-0.62, 0.73)	0.35	1.06 (0.54, 2.08)	0.16	.87
	AI Dissent[1]:Domain[2]	-0.23 (-3.98, 3.53)	1.91	0.80 (0.02, 252.14)	-0.12	.91
	AI Dissent[1]:Attitude	-0.05 (-0.73, 0.64)	0.35	0.95 (0.48, 1.90)	-0.13	.89
	Domain[2]:Attitude	-0.33 (-1.02, 0.35)	0.35	0.72 (0.36, 1.42)	-0.95	.34
	AI Dissent[1]:Domain[2]:Attitude	-0.23 (-1.22, 0.76)	0.50	0.80 (0.30, 2.14)	-0.45	.65
Selected	Intercept	0.96 (0.10, 1.51)	0.41	2.60 (1.11, 4.53)	2.31	.02*
	AI Dissent[1]	-1.24 (-2.16, -0.33)	0.43	0.29 (0.12, 0.72)	-2.89	.004**
	Domain[2]	-0.71 (-1.63, 0.20)	0.43	0.49 (0.20, 1.22)	-1.66	.10
	Deception[3]	-1.01 (-1.61, -0.43)	0.30	0.36 (0.20, 0.65)	-3.43	<.001***

Significance: * $p < .05$. ** $p < .01$. *** $p < .001$ Reference Level: [1]C1 [2]Analytical [3]No deception beliefs

be stronger for social tasks compared to analytical tasks (d = -0.22 ,95% CI [-0.09, -0.34], t_{281} = -3.06, $p < .001$). However, including an interaction for domain and AI dissent did not increase the model fit for the selected GLMM. Within the saturated model, the interaction had a non-significant effect (OR = 0.79, z = -0.12, p = .90) with very wide confidence intervals (95% CI [0.02, 36.60]). This lack of a significant interaction is surprising, as previous findings (Hertz & Wiese, 2018) indicated a relationship between domain and conformity – not the impact of AI dissent on conformity – for similar tasks.

However, as these results relied on a traditional ANOVA model, they failed to account for the variance of the stimuli. Similarly, when utilizing a generalized linear model (GLM) with an interaction term as well as main effects for trial type and domain, the results suggest a small, significant main effect for trial type (OR = 0.58, 95% CI [0.42,

0.81], z = -3.19, p < .001) and a small, significant interaction between AI dissent and domain (*OR* = 0.47, 95% CI [0.29, 0.77], z = -2.99, p = .003).

Attitudes

For the GAAIS, the scale mean was 3.45 (*SD* = 0.58), indicating a favorable attitude towards AI. However, it does not seem to be related to the impact of AI dissent, as there were no significant correlations with mean differences between C2 and C1 trials for either social (r_{92} = -.06, 95% CI [-0.26, 0.15], t = -0.57, p = .57) or analytical (r_{92} = -.02, 95% CI [-0.22, 0.19], t = -0.15, p = .88) tasks (see Figure 4). For the selected GLMM, the inclusion of attitude did not benefit the goodness-of-fit. For the saturated model, neither its main effect (*OR* = 1.00, 95% CI [0.51, 1.97], z = 0.01 p = .99) nor the interaction with AI dissent (*OR* = 0.96, 95% CI [0.48, 1.94], z = -0.12, p = .91) were significant. This general trend remained when excluding individuals assuming any form of deception from the analysis.

Perceived Deception

Perceived Deception was also included as a fixed effect. This resulted in a significant effect for the selected model (*OR* = 0.36, 95% CI [0.20, 0.65], z = -3.12, p = .002).

Loss of Trust over multiple Trials

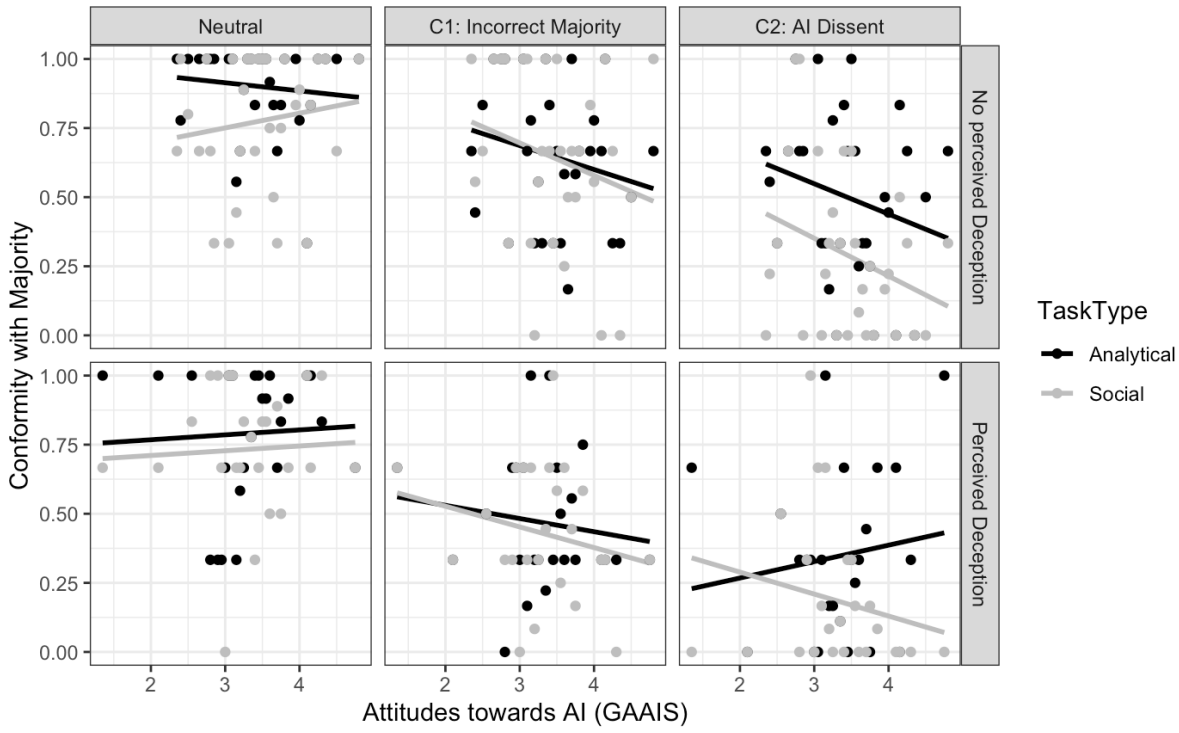
As previous findings suggested a loss of trust in the different agents due to repeated, incorrect answers (Salomons et al., 2018; Wiegmann et al., 2001), the event index – a numerical indicator for how many trials were done before the target trial – was also considered as a random effect. However, while there was a significant negative correlation of r_{1126} = -.10 (95% CI [-.15, -.04], t = -3.40, p < .001), the effect size was small and its implementation did not increase the goodness-of-fit for the model.

Accuracy

Looking at accuracy (i.e., whether subject chose the factually correct option) instead of conformity, a slightly different picture emerges. In C1 trials, accuracy levels were similar for the social (*M* = 0.35, *SD* = 0.48) and analytical (*M* = 0.38, *SD* = 0.49) domain.

Figure 4

Probability of Conformity for Trial, Task, Attitude and Perceived Deception



For C2, however, accuracy was increased for social tasks (M = 0.76, SD = 0.43) compared to analytical tasks (M = 0.46, SD = 0.50; see Figure 5). Visual inspection indicates a stronger contribution of domain (see Figure 6). For further exploration, another GLMM utilizing the same random and fixed effect structures as the saturated model but focusing on accuracy instead of conformity as a response was fitted and reduced (see table 3). Results suggest a medium-sized, significant interaction for domain and AI dissent (OR = 6.07, z = 2.64, p = .008. However, as 95% confidence intervals are wide (1.42, 26.31), these results need to be interpreted cautiously. Simple effects analyses utilizing multivariate t p-value correction for multiple comparisons revealed that accuracy in the presence of AI dissent was significantly lower for the analytical task compared to the social task (z = -3.39, SE = 0.49, p = .004). Furthermore, there was a significant difference between C1 and C2 conditions for the social task with higher accuracy in the presence of AI dissent (z = -4.52, SE = 0.49, p < .001).

Table 3

Fixed Effect Estimates for the selected model for Accuracy

Model	Fixed Effect	*Log Odds* (95% CI)	*SE*	*OR* (95% CI)	*z*	*p*
Selected	Intercept	-1.19 (-2.97, 0.58)	0.89	0.31 (0.05, 1.79)	-1.33	.18
	AI Dissent[1]	0.40 (-0.61, 1.43)	0.48	1.50 (0.54, 4.18)	0.84	.40
	Domain[2]	-1.64 (-3.60, 0.29)	0.98	0.19 (0.03, 1.34)	-1.59	.09
	Attitude	0.06 (-0.40, 0.53)	0.23	1.06 (0.67, 1.70)	0.26	.79
	Deception[3]	0.78 (0.31, 1.26)	0.23	2.17 (1.38, 3.53)	3.28	.001**
	Domain[2]:Attitude	0.43 (-0.05, 0.92)	0.24	1.54 (0.95, 2.51)	1.76	.07
	AI Dissent[1]:Domain[2]	1.80 (0.35, 3.27)	0.68	6.07 (1.42, 26.31)	2.64	.008**

Significance: * $p < .05$. ** $p < .01$. Reference Level: [1]C1 [2]Analytical [3]No deception beliefs

Confidence

Following Moscovici's (1980) conversion theory, we expected compliance effects for trials with a unanimous, incorrect human-AI majority (C1) and conversion effects for trials with the presence of a correct AI dissenter (C2). We assumed these to be reflected by differences in confidence in the decision. And while the latter appeared similar for C1

Figure 5

Accuracy for Trial, Task and Perceived Deception

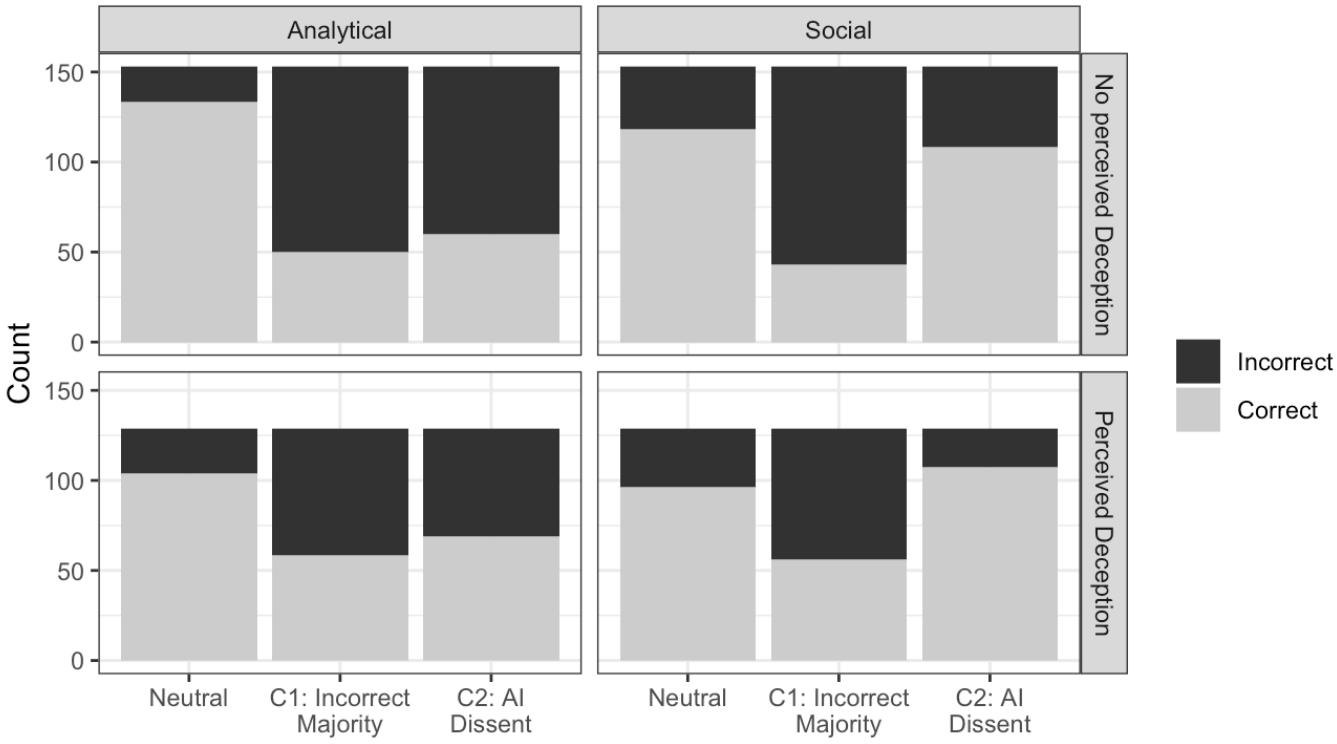
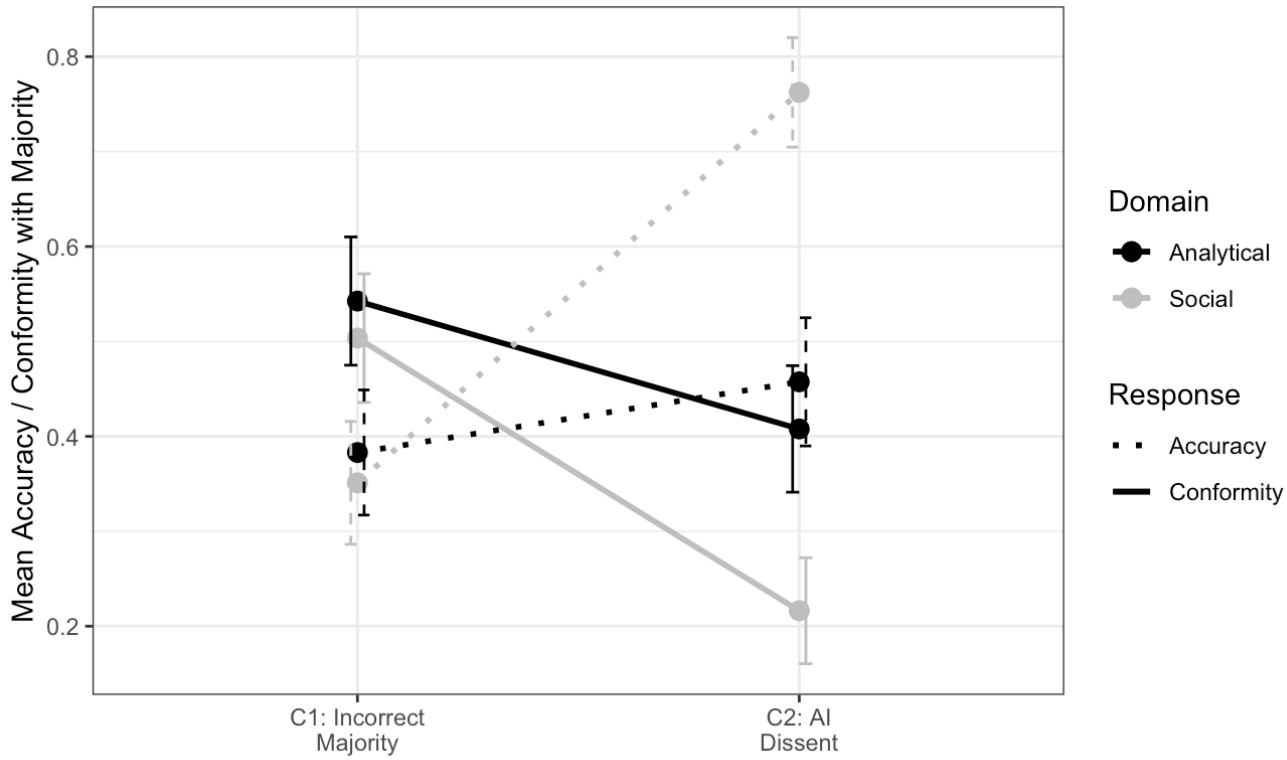


Figure 6

Interaction plot for Trial Type and Domain regarding Accuracy and Conformity with human Majority



(*M* = 70.68, *SD* = 24.61) and C2 (*M* = 70.90, *SD* = 24.14) trials, there were differences for domain with confidence being higher for social tasks (*M* = 74.60, *SD* = 20.54) than analytical tasks (*M* = 66.98, *SD* = 27.16). For analytical tasks, confidence decreases from C1 (*M* = 70.35, *SD* = 27.30) to C2 (*M* = 63.60, *SD* = 26.63). For social tasks, it increased from C1 (*M* = 71.00, *SD* = 21.63) to C2 (*M* = 78.20, *SD* = 18.75). However, based on the results from a GLMM (or more precisely a linear mixed effect model) following the same approach as before, no significant interaction was found.

Discussion

The aim of this study was to foster a better understanding of the capacities of AI induced dissent in hybrid human-AI conformity situations. By conducting an online experiment, we sought to investigate whether the presence of an AI agent dissenting against the majority would reduce conformity and whether this would be moderated by the nature of the task and the attitude towards AI of the subject. Additionally, we

explored how AI dissent impacts the decision accuracy as well as the confidence in the decision.

As previous research has demonstrated the capacity of software agents to promote conformity (Hertz & Wiese, 2018; Riva et al., 2022; Salomons et al., 2018) as well as the effectiveness of dissenting social robots in countering conformity effects (Qin et al., 2022), we expected the introduction of AI agents to produce a similar effect. In accordance with this hypothesis, conformity with the incorrect, human majority significantly dropped in the presence of AI dissent. An effect even held up for subjects perceiving deception in the manipulation.

The influence of task domain appears to be more complex. In line with previous studies (Hertz & Wiese, 2018, Hertz et al., 2019), conformity was higher for analytical tasks compared to social tasks. While a GLM not accounting for the variance of the stimuli did find a significant interaction with the effect of AI dissent increasing for social tasks compared to analytical tasks, the selected GLMM failed to detect any interaction between domain and AI dissent. This contradicts our expectations as well as previous studies identifying a positive relationship between perceived competence of artificial agents and subjects' susceptibility towards their influence (Zonca et al., 2023).

Furthermore, fitting a GLMM with accuracy as the response variable, a medium-sized interaction was found. In contrast to the absence of an interaction for conformity, this implies two distinct mechanisms at play: the mere presence of an artificial dissenter appears to be sufficient to break the normative influence exerted by the human group, regardless of task domain. However, the domain might influence whether individuals adopt the dissenter's information and follow their lead. Put differently: While an AI agent's capacity to break majority influence does not appear to rely on the given task, its potential to exert minority influence might. This is in line with Moscovici's (1980) insights suggesting an increased importance of informational influence for minorities. Nevertheless, the observed direction of the effect is counterintuitive, as the social-influential capacities of the dissenting AI agent were heightened for social tasks. This goes against the AI competence perception measured as well as previous results by Castelo et al. (2019) indicating algorithm appreciation for objective, quantitative tasks. One reason could be the mixture of algorithm aversion after repeated negative experience for the analytical tasks (Dietvorst et al., 2015), making the participants hesitant to agree to with the AI. In combination with the speculations about lower

difficulty for the social task (see above), this could be another plausible explanation for the interaction effect for accuracy. Then again, the analysis of the event index as a measure of lost trust did not support this. Therefore, additional research is necessary to further understand how task domain shapes the influence exerted by artificial agents. In this context, particularly the reactions to an incorrect dissenter could provide insights into the normative and informational processes and how these depend on task domain.

Lastly, attitudes towards AI do not seem to have a moderating effect on the influence of AI dissent on conformity. This outcome is contrary to the initial hypothesis and even more surprising, considering that the GAAIS covers both competence and warmth assessments, and these are both directly related to an actor's capacity for informational and normative pressure (Cialdini, 2001).

Limitations

While offering convenient recruitment and easy access to subjects outside WEIRD populations, it is important to acknowledge the limitations and flaws associated with research crowdsourcing platforms such as Prolific. Primarily, the userbase might differ systematically from the general population, resulting in a biased sample. Due to their above-average participance, these "professional" subjects might also be less susceptible to any form of experimental manipulation, which in turn could explain the high number of individuals exposing the confederate deception. Furthermore, as many of the platform's users understand prolific as an additional means of income, they might be incentivized to answer in line with researcher expectations to avoid rejection, resulting in some sort of demand characteristics.

The FACES database used for the experiment offers a variety of validated high-quality portraits including various ages and facial expressions. However, as it is heavily used in psychology, this might have led to unexpected familiarity effects. Additionally, it consists of Caucasian individuals only, potentially opening pandora's box regarding perception differences due to ethnicity (Cook & Over, 2021).

It is possible to speculate about differences in task difficulty as an explanation for the main effect of domain due to the link between uncertainty and conformity (Brandstetter et al., 2014; Hertz & Wiese, 2016). For instance, the interpretation of facial expressions – particularly negative ones – might be easier due to its evolutionary significance and frequent use in daily life. Since faces are highly effective at capturing

attention (Morrisey et al., 2019), the set time limit might also have had a greater impact on the analytical task, as participants engaged with the facial stimulus first and then ran out of time for the dot stimuli. To solve this issue, future studies should pretest for task difficulty or consider presenting the stimuli separately.

Lastly, since the questionnaire was not answered until after the experiment was completed, it cannot be ruled out that the interaction with the AI agent might have influenced the outcome of the GAAIS. This could also explain the general, positive attitude found, as it is in line with insights provided by Allen and Levine (1969), suggesting a positive evaluation of dissenters.

Conclusion

Based on the evidence provided above, we can confidently conclude that dissenting AI agents can reduce conformity effects similarly to humans and social robots. Surprisingly, this appears to be independent of the individual attitude towards AI in general, suggesting that the mere presence of a dissenting opinion suffices to counteract the majority pressure. However, while there were no relevant interaction effects of domain on conformity, there was an interaction for accuracy, indicating that it moderates whether individuals choose to trust the AI agent. In general, these results suggest that AI agents can be used to reduce unwanted conformity effects, which in turn can be used to benefit decision making, oppose group polarization processes or prevent illicit activities.

References

Allen, V. L., & Levine, J. M. (1969). Consensus and conformity. J*ournal of Experimental Social Psychology, 5*, 389–399. https://doi.org/10.1016/0022-1031(69)90032-8

Asch, S. E. (1951). *Effects of group pressure upon the modification and distortion of judgments. In Groups, leadership and men; research in human relations.* (pp. 177–190). Carnegie Press.

Asch, S. E. (1955). *Opinions and Social Pressure. Scientific American, 193*, 31–35. JSTOR.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*, 1–70. https://doi.org/10.1037/h0093718

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 701–706. https://doi.org/10.1109/ROMAN.2008.4600749

Barton, K. (2023). *Package 'mumin'*. Version, 1.47.5.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*. https://doi.org/10.18637/jss.v067.i01

Beran, T., Drefs, M., Kaba, A., Al Baz, N., & Al Harbi, N. (2015). Conformity of responses among graduate students in an online environment. *The Internet and Higher Education, 25*, 63–69. https://doi.org/10.1016/j.iheduc.2015.01.001

Berns, G. S., Chappelow, J., Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., & Richards, J. (2005). Neurobiological Correlates of Social Conformity and Independence During Mental Rotation. *Biological Psychiatry, 58*, 245–253. https://doi.org/10.1016/j.biopsych.2005.04.012

Bock, D. E., Wolter, J. S., & Ferrell, O. C. (2020). Artificial intelligence: Disrupting what we know about services. *Journal of Services Marketing, 34*, 317–334. https://doi.org/10.1108/JSM-01-2019-0047

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*, 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brandstetter, J., Racz, P., Beckner, C., Sandoval, E. B., Hay, J., & Bartneck, C. (2014). A peer pressure experiment: Recreation of the Asch conformity experiment with robots. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 1335–1340. https://doi.org/10.1109/IROS.2014.6942730

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research, 56*, 809–825. https://doi.org/10.1177/0022243719851788

Chen, H., Cohen, P., & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation, 39*, 860–864. https://doi.org/10.1080/03610911003650383

Cialdini, R. B. (2001). *Influence: The psychology of persuasion.* Collins.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cook, R., & Over, H. (2021). Why is the literature on first impressions so focused on White faces? *Royal Society Open Science*, *8*, 211146. https://doi.org/10.1098/rsos.211146

David, B., & Turner, J. C. (2001). Majority and minority influence: A single process self-categorization analysis. In C. K. W. De Dreu & N. K. De Vries (Eds.), *Group consensus and minority influence: Implications for innovation* (pp. 91–121). Blackwell Publishing.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*, 629–636. https://doi.org/10.1037/h0046408

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*, 114–126. https://doi.org/10.1037/xge0000033

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods, 42*, 351–362. https://doi.org/10.3758/BRM.42.1.351

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior, 6*, 238–252. https://doi.org/10.1007/BF00987191

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication, 1*, 71–86. https://doi.org/10.30658/hmc.1.5

Hertz, N., Shaw, T., de Visser, E. J., & Wiese, E. (2019). Mixing It Up: How Mixed Groups of Humans and Machines Modulate Conformity. *Journal of Cognitive Engineering and Decision Making, 13*, 242–257. https://doi.org/10.1177/1555343419869465

Hertz, N., & Wiese, E. (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*, 313–317. https://doi.org/10.1177/1541931213601071

Hertz, N., & Wiese, E. (2018). Under pressure: Examining social conformity with computer and robot groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 60*, 1207–1218. https://doi.org/10.1177/0018720818788473

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54–69. https://doi.org/10.1037/a0028347

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), 4*, 3559–3564. https://doi.org/10.1109/IROS.2004.1389967

Kincaid, D. L. (2004). From Innovation to Social Norm: Bounded Normative Influence. *Journal of Health Communication, 9*, 37–57. https://doi.org/10.1080/10810730490271511

Krämer, N. C. (2005). Social Communicative Effects of a Virtual Program Guide. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Intelligent Virtual Agents* (Vol. 3661, pp. 442–453). Springer. https://doi.org/10.1007/11550617_37

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34*. https://escholarship.org/uc/item/7ck0p200

Morgan, T. J. H., & Laland, K. N. (2012). The Biological Bases of Conformity. *Frontiers in Neuroscience, 6*. https://doi.org/10.3389/fnins.2012.00087

Morrisey, M. N., Hofrichter, R., & Rutherford, M. D. (2019). Human faces capture attention and attract first saccades without longer fixation. *Visual Cognition, 27*, 158–170. https://doi.org/10.1080/13506285.2019.1631925

Moscovici, S. (1980). *Toward A Theory of Conversion Behavior. In Advances in Experimental Social Psychology* (Vol. 13, pp. 209–239). Elsevier. https://doi.org/10.1016/S0065-2601(08)60133-1

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., & Steuer, J. (1993). Voices, Boxes, and Sources of Messages.: Computers and Social Actors. *Human Communication Research, 19*, 504–527. https://doi.org/10.1111/j.1468-2958.1993.tb00311.x

Pasupathi, M. (1999). Age differences in response to conformity pressure for emotional and nonemotional material. *Psychology and Aging, 14*, 170–174. https://doi.org/10.1037/0882-7974.14.1.170

Qin, X., Chen, C., Yam, K. C., Cao, L., Li, W., Guan, J., Zhao, P., Dong, X., & Lin, Y. (2022). Adults still can't resist: A social robot can induce normative conformity. *Computers in Human Behavior, 127*, 107041. https://doi.org/10.1016/j.chb.2021.107041

R Core Team. (2022). *R: A language and environment for statistical computing* (4.2.2).

Rayburn-Reeves, R., Wu, J., Wilson, S., & Kraemer, B. (2013). Do As We Do, Not As You Think: The Effect of Group Influence on Individual Choices in a Virtual Environment. *Journal For Virtual Worlds Research, 6*. https://doi.org/10.4101/jvwr.v6i1.7002

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* (pp. xiv, 305). Cambridge University Press.

Riva, P., Aureli, N., & Silvestrini, F. (2022). Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica, 229,* 103681. https://doi.org/10.1016/j.actpsy.2022.103681

Rosander, M., & Eriksson, O. (2012). Conformity on the Internet – The role of task difficulty and gender differences. *Computers in Human Behavior, 28*, 1587–1595. https://doi.org/10.1016/j.chb.2012.03.023

Salomons, N., van der Linden, M., Strohkorb Sebo, S., & Scassellati, B. (2018). Humans conform to robots: Disambiguating trust, truth, and conformity. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 187–195. https://doi.org/10.1145/3171221.3171282

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, *1*, 100014. https://doi.org/10.1016/j.chbr.2020.100014

Schepman, A., & Rodway, P. (2022). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction,* 1–18. https://doi.org/10.1080/10447318.2022.2085400

Schnuerch, R., & Gibbons, H. (2014). A Review of Neurocognitive Mechanisms of Social Conformity. *Social Psychology, 45*, 466–478. https://doi.org/10.1027/1864-9335/a000213

Shiomi, M., & Hagita, N. (2016). Do Synchronized Multiple Robots Exert Peer Pressure? *Proceedings of the Fourth International Conference on Human Agent Interaction*, 27–33. https://doi.org/10.1145/2974804.2974808

Sundar, S. S., & Nass, C. (2000). Source Orientation in Human-Computer Interaction: Programmer, Networker, or Independent Social Actor. *Communication Research, 27*, 683–703. https://doi.org/10.1177/009365000027006001

Vollmer, A.-L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics, 3*, eaat7111. https://doi.org/10.1126/scirobotics.aat7111

Wainer, J., Feil-Seifer, D., Shell, D. A., & Matarić, M. J. (2007). Embodiment and Human-Robot Interaction: A Task-Based Perspective. *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 872–877.

Walker, M. B., & Andrade, M. G. (1996). Conformity in the Asch Task as a Function of Age. *The Journal of Social Psychology, 136*, 367–372. https://doi.org/10.1080/00224545.1996.9714014

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science, 5*, 219–232. https://doi.org/10.1177/1745691610369336

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* Freeman.

Westerman, D., Edwards, A. P., Edwards, C., Luo, Z., & Spence, P. R. (2020). I-It, I-Thou, I-Robot: The Perceived Humanness of AI in Human-Machine Communication. *Communication Studies, 71*, 393–408. https://doi.org/10.1080/10510974.2020.1749683

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*, 352–367. https://doi.org/10.1080/14639220110110306

Xu, K. (2019). First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media & Society, 21*, 2522–2547. https://doi.org/10.1177/1461444819851479

Zonca, J., Folsø, A., & Sciutti, A. (2023). Social influence under uncertainty in interaction with peers, robots and computers. *International Journal of Social Robotics, 15*, 249–268. https://doi.org/10.1007/s12369-022-00959-x