"Can you tell I used ChatGPT?" How Perceived AI-Mediation Affects Workplace Email Persuasiveness— A Bayesian Approach

Shaked Karabelnicoff [1] (s.karabelnicoff@lse.ac.uk)

Jens Koed Madsen. (j.madsen2@lse.ac.uk)

[1]Department of Psychological and Behavioural Science, London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, UK.

Abstract

Large Language Models like ChatGPT are becoming everyday writing partners in the workplace. This study asked: how does simply knowing an email was "edited by ChatGPT" affect its persuasiveness and the perceived credibility of the sender? We collected data from 308 professionals using experimental vignettes that simulated realistic workplace emails. Some emails were described as entirely human-written, while others were labeled as AI-edited, with variations in the sender's reliability (who is sending the message) and strength of the argument (how well the content is constructed). A Bayesian Model of Argumentation provided normative predictions for how reliability and argument quality should influence persuasion. We found that when an email was labeled as "edited by ChatGPT," receivers saw it as less persuasive overall. However, AI-mediation did not diminish the relative influence of source reliability and argument quality. In other words, while the AI-edited label reduced overall persuasiveness, it didn't change how recipients inherently evaluated credibility. They still adjusted their beliefs primarily based on who sent the message and how strong the argument was. To our knowledge, this is the first study to apply a Bayesian framework to understanding how people process AI-mediated communication.

Keywords: AI-mediated communication; Bayesian argumentation; Workplace communication

Introduction

Today's technology-driven world relies on effective digital communication for successful collaboration and productivity (Sivunen and Laitinen, 2019; Turner et al., 2010). Emails play a central role in this process, with some employees spending nearly 60% of their workday on meetings, emails, and chats (Microsoft, 2023). Yet, the overwhelming volume of emails can lead to cognitive strain, or "email overload," with negative implications for employee well-being and organisational productivity (Giurge & Bohns, 2021; Reinke and Chamorro-Premuzic, 2014). Large language models (LLMs) like ChatGPT offer a potential solution by acting as virtual assistants to draft and edit personalized emails (Chang et al., 2023; Draxler et al., 2024). By generating human-like text, these tools may help reduce cognitive overload and boost productivity (Bastola et al., 2023). However, the growing use of these AI tools also raises concerns about their impact on communication and trust in the workplace. While they can make writing more efficient, they may also risk changing how colleagues are perceived and how their messages are interpreted.

Individuals often struggle to differentiate between messages created by humans and those generated by AI (Jakesch et al., 2023; Köbis & Mossink, 2021; Kreps et al., 2022; Spitale et al., 2023). Yet, when people know the content is AI-generated, they tend to mistrust the message (Ragot et al., 2020; Tiegen et al., 2023). Mistrust can also extend to suspicion towards the individual using it (Hohenstein et al., 2021; Jakesch et al., 2019; Liu et al., 2022; Wu & Kelly, 2020). This can vary based on individual differences, including a receiver's general attitude toward AI, whether positive or negative, which shapes how they evaluate AI-generated content. (Karinshak et al., 2023; Lim & Schmälzle, 2023; Jakesch et al., 2019). Still, prior studies found that, in general, people's trust in email senders decreased when they were told that ChatGPT was involved in the writing process (Liu et al., 2022). As future AI systems may be given greater autonomy in handling communication tasks without oversight (Hancock et al., 2020), email receivers might struggle to discern who they are communicating with and what inferences to draw. This could have implications for the way individuals maintain and develop relationships in the workplace, where trust and collaboration are essential.

Previous AI-MC research has indicated that factors such as trustworthiness, expertise, and message content influence interpersonal perceptions in AI-MC (Jakesch et al., 2019; Liu et al., 2022). However, the cognitive mechanisms underlying these effects have been explored mainly in theory. To our knowledge, no formal quantitative models have been empirically tested to validate these theoretical concepts. Therefore, this paper embeds AI-MC in a Bayesian Model of Argumentation (Madsen & Wong, 2023), which predicts belief-updating as a joint function of source reliability and argument quality.

Bayesian models offer a testable, quantitative framework for understanding cognitive processes (Hahn and Oaksford, 2006a). The Bayesian Model of Argumentation provides a formula for estimating how a 'rational agent 'should revise their belief in a hypothesis. The quantitative component of consists of ascribing conditional probability distributions for each variable in a Bayesian Network (BN). BN's provide the means to test causal models of scenarios and compare observed beliefs to the predicted normative standard (Lagnado et al., 2012; Madsen and Pilditch, 2018). The model used in this paper (Madsen and Wong, 2023), evaluates how argument quality and source reliability influence persuasiveness, accounting for initial subjective beliefs, the quality of the message content, and the trustworthiness and expertise of the source, as factors that influence belief change. The model

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

predicts positive updating when strong evidence comes from a trustworthy, expert source, and negative updating when weak evidence comes from an unreliable source.

The experiment uses vignettes that simulate workplace email exchanges in human versus AI-mediated conditions to explore the influence of trustworthiness, expertise, and argument quality on persuasiveness. The study then tests the moderating effects of these variables on belief change, and uses Bayes 'theorem to compare the observed results with predicted outcomes, thereby assessing the model's predictive power.

Taken together, these literatures motivate our approach. Prior AI-MC research suggests that disclosing AI involvement can reduce perceived trustworthiness. The specific Bayesian Model of Argumentation used in this study offers a structured framework for predicting how individuals revise their beliefs based on source reliability and argument quality. By integrating these perspectives, we aim to explore whether AI involvement influences persuasiveness, and whether it changes the extent to which people rely on source reliability and argument quality when updating their beliefs. We anticipate that AI-mediation will lower overall persuasiveness, and we examine whether, and to what extent, it alters the influence of a sender's reliability and the strength of their argument. Based on these theoretical foundations, we outline the following research questions and hypotheses:

1. How does perceived AI-mediation in workplace emails influence the persuasiveness of a message, accounting for the source's reliability and the quality of the argument?

2. To what extent can a Bayesian model predict the persuasive potential of AI-mediated communication?

H1. Source reliability will be positively correlated with perceived belief in the hypothesis. (More reliable sources will be more persuasive.)

H2. Argument quality will be positively correlated with perceived belief in the hypothesis. (Stronger arguments will be more persuasive.)

H3.1 Perceived AI-Mediation will have a moderating effect on influences of source reliability and argument quality on belief in the hypothesis.

H3. 2 Source reliability will have significant main effects in both the human and AI-mediated conditions

H3. 3. Argument quality will have significant main effects in both the human and AI-mediated conditions

H4. In the AI-Mediated condition, general attitudes towards artificial intelligence (GAAIS scale) will influence belief change. (Exploratory)

H5. Bayesian predictions modeled based on Madsen and Wong (2023) will be positively correlated with the observed responses.

This paper attempts to make two main contributions to academic literature. First, it seeks to formalise the effects of AI-MC by operationalising a Bayesian model. Second, it will test the predictive abilities of the Bayesian framework for AI-mediated communication. Beyond advancing theory, this study also offers practical guidance for organisations integrating AI into their daily communication workflows.

Method

The hypotheses, research design, and analyses for this study were pre-registered before data collection. Pre-registration can be found via the following link: https://osf.io/j8nk6

Participants: A priori power analysis was conducted using the software G*Power 3.1 (Faul et al., 2009), to determine the sample size needed to minimise false positive (type I) and false negative (type II) errors. Using effect size (f) = 0.25, significance level (α) = 0.05, and power level (1-β err prob) = 0.95, a sample of 251 was recommended. 308 participants were recruited using the research platform Prolific, providing a 23% cushion. Sampling criteria specified that participants should be native English speakers from the UK, U.S., or Canada and employed full-time at a knowledge-based organisation. This was intended to increase the likelihood of participants 'familiarity with email-based workplace communications.

Three distinct attention checks were included to ensure active engagement. Submissions failing one or more attention checks were rejected. Consequently, 33 participants were removed for incomplete responses and 35 participants were removed for failing one or more attention checks. Additionally, 8 participants were removed for having a prior belief rating of 'certainty '(0 or 1) as Bayesian belief updating relies on the ability or "openness" to revise beliefs as new evidence is presented (Hahn and Oaksford, 2006). Extreme priors imply 'absolute certainty 'and prevent this functioning.

The 232 remaining participants were evenly divided across conditions. Even distribution was visually inspected and revealed no apparent statistically significant differences in gender, education, and age between the two groups.

Design: The study employed a 2x2x2 factorial design. The independent variables were: Type of Mediation (human vs. AI-mediated), Source Reliability (reliable vs. unreliable), and Argument Quality (strong vs. weak). The type of mediation was a between-subjects variable, while source reliability and argument quality were within-subjects variables. The dependent variable, belief change, is the difference between prior and posterior assessments.

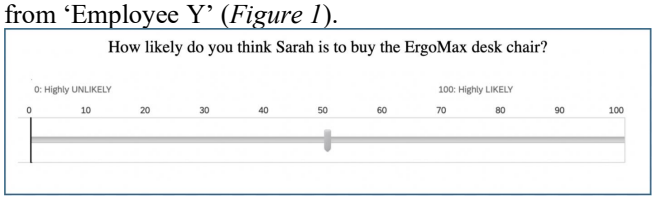
(*Figure 1:* Email exchange in 'Scenario A 'under the AI-mediated, reliable source, and strong argument quality condition)

As is common in Bayesian studies (e.g., Hahn et al., 2009; Oaksford and Hahn, 2004; Madsen and Wong, 2023), the independent variables were operationalised using vignettes, which have also been effectively utilised in management and organisational behaviour research through Experimental Vignette Methodology (Aguinis and Bradley, 2014). Following Madsen and Wong (2023), this setup employed a third-party argument evaluation paradigm, where participants evaluated the exchange rather than directly engaging in it (e.g., Oaksford and Hahn, 2004; Hahn et al., 2009).

All email exchanges were human-written, but in the AI-mediated condition, scenarios were labelled as AI-mediated. This employs a technique similar to the "Wizard of Oz" approach (Dahlbäck et al., 1993) as seen in various other AI-MC studies (Jakesch et al., 2019; Liu et al., 2022). We did

not use AI-generated messages because our goal was to understand whether simply believing a message was AI-mediated would influence how recipients evaluated the message and its sender. All emails were pre-tested in a pilot study (which we discuss in detail later) to ensure they did not sound distinctly human- or AI-written.

Participants were presented with four different scenarios set within a hypothetical mid-sized marketing agency in London. Each scenario followed a consistent structure, framed as an email exchange between 'Employee X 'seeking advice from 'Employee Y' (*Figure 1*).



(*Figure 2:* Eliciting Posterior Belief)

Prior to reading the exchange, participants provided a 'prior' degree of belief about the likelihood that 'Employee X' would follow 'Employee Y's advice. After reading the exchange, participants provided their 'posterior' degree of belief (*Figure 2*). Responses were recorded probabilistically on a scale from 0 to 100 (0 being "Highly Unlikely", and 100 being "High Likely"). There were no back buttons so participants could not modify their rated prior belief.

To assess participants 'perceptions of artificial intelligence, the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2023) was administered at the end of the survey in the form of a matrix table. The GAAIS is a validated measure used to gauge attitudes towards AI. Using a positive and negative subscale of 20 total items, attitudes are measured using a 5-point Likert scale, with answers ranging from 'Strongly Disagree 'to 'Strongly Agree '(Schepman & Rodway, 2023). This scale has been administered in similar studies within psychology (e.g. Teigen et al., 2023; Lim & Schmälzle, 2023).

Manipulations: Since Mediation Type was a between-subjects variable, each participant saw exclusively human or exclusively AI-mediated email exchanges. AI-mediation was disclosed within the emails as part of the sign-off as: "(Edited by ChatGPT)"

Source reliability is a combined measure of trustworthiness and expertise (Harris et al., 2015). Therefore, the manipulation for source reliability was two-fold, involving trustworthiness and expertise. Following Harris et al., (2015) and Madsen and Wong (2023), trustworthiness is defined in terms of helpfulness, while expertise is associated with knowledgeability.

Argument quality was manipulated following Madsen and Wong (2023), which adopted Petty et al.'s (1981) definition of strong messages as ones that "provided persuasive evidence (statistics, data, etc.)" to support their claim; whilst weak messages relied on subjective personal opinions to support their position.

Pilot study: A pilot study was conducted to assess the effectiveness of the manipulated variables. Specifically, manipulation checks for source reliability and argument quality were essential to confirm whether they were effectively distinguished as strong or weak. The pilot study consisted of 70 participants. Participants were asked to rate trustworthiness, expertise, and argument quality statements on a sliding scale from 0 to 100 . Strong and weak manipulations were considered valid if they achieved a mean rating above 70 for high and below 30 for low. The results from the pilot study were also used to parameterise the Bayesian model. The pilot also pre-tested participants' perceptions of whether each argument was human or AI-generated, using a 5-point Likert scale similar to the process outlined by Liu et al. (2022) and Jakesch et al. (2019). Participants were asked to rate each argument: "Do you think this email was (1) 'Definitely Human-written ' to (5) 'Definitely AI-generated'?" This pre-test was designed to ensure that the statements could be perceived as both human or AI- mediated. Results found the mean AI scores for the individual statements ranged from 1.69 to 3.41. The overall mean AI score across all statements was 2.55, with a standard error of 0.071, a 95% confidence interval [2.41 - 2.69], and a median of 2.00. These findings indicate that participants were generally uncertain of whether the statements were human or AI, tending to fall in the middle of the scale. This confirmed that the statements likely could be perceived as both human-written or AI-generated.

Procedure: Participants were assigned to either the human or AI-mediated condition. In each condition, they reviewed all four scenarios (A, B, C, D). The experimental conditions (weak/strong argument quality and reliable/unreliable source) and the order of scenarios were randomised using block randomisation to minimise carryover effects and ensure a balanced collection of responses (Cochran & Cox, 1957). The GAAIS scale was assessed at the end of the survey to further reduce any carryover bias.

Computing Bayesian Predctions: Parameter values for Argument Strength and Source Reliability were derived from the values obtained in the manipulation checks conducted during the pilot study, following Madsen and Wong (2023). The mean parameter values consist of the average ratings for strong and weak arguments, and reliable and unreliable sources, which aggregates the trustworthiness and expertise ratings. The Bayesian predictions for the current model were computed using the following formulas, replicating Madsen and Wong (2023) which operationalise Argument Quality and Source Relaibility within Bayes 'theorem:

P(rep|H) = P(rep|H, strong, rel)*P(strong)*P(rel) + P(rep|H, strong, unrel)*P(strong)*P(unrel) + P(rep|H, weak, rel)*P(weak)*P(rel) + P(rep|H, weak, unrel)*P(weak)*P(unrel)

P(rep|not-H) = P(rep|not-H, strong, rel)*P(strong)*P(rel) + P(rep|not-H, strong, unrel)*P(strong)*P(unrel) + P(rep|not-H, weak, rel)*P(weak)*P(rel) + P(rep|not-H, weak, unrel)*P(weak)*P(unrel)

Data Analysis: A 2x2x2 three-way mixed ANOVA (Analysis of Variance) was performed. First, a mixed linear regression model (lmer) was fitted to the dataset, specifying the fixed effects of the independent variables (mediation type, argument quality, source reliability) and the covariate (GAAIS

positive and negative scale) on belief change as the dependent variable, as well as relevant interactions. The random effects accounted for repeated measures, as each participant rated four scenarios, resulting in four unique observations per participant. In total, there were 928 observations in the dataset. Belief change was measured as the difference in participants 'prior and posterior. In other words, participants 'ratings of the likelihood of 'Employee X 'making a certain decision before and after receiving the advice from 'Employee Y. 'This was obtained by subtracting the corresponding prior belief rating from each posterior belief rating.

A type III ANOVA output with Satterthwaite's method was produced following the lmer. This provided F-values, p-values, and degrees of freedom for the fixed effects and their interactions, allowing for an assessment of the significance of each variable and interaction on belief change. To validate the model assumptions, the Shapiro-Wilk test was performed. The test indicated that the assumption of normality was violated. As a result, a Kruskal-Wallis test was conducted on each independent variable to provide a non-parametric alternative for analysing the differences between groups. Two separate two-way ANOVAs were performed to examine how the within-subjects variables interacted in the unique between-subjects condition and to assess the influence of the GAAIS scale as a covariate. The assumptions of normality were tested for each condition individually; while the human condition violated the normality assumption, the AI condition did not. Additionally, an ANCOVA (Analysis of Covariance) was conducted and found similar results to the initial mixed ANOVA, supporting the robustness of the findings despite the violation of the normality assumption. To assess whether the quantitative Bayesian predictions approximated the observed data, a linear regression analysis was conducted comparing the observed and predicted posterior beliefs. The resulting R2 value indicates the proportion of variance in the observed data that is explained by the model's predictions, thereby evaluating the accuracy and fit of the Bayesian predictions to the actual data. All statistical analyses were carried out at a significance level of $\alpha = 0.05$.

Results

A mixed linear regression found a statistically significant effect of mediation type on belief change (b=7.15, p=0.00027). This indicates that participants who perceived the emails to be human-written experienced more positive belief change compared to those who perceived the emails to be AI-mediated. In other words, participants who believed the emails were human- written were more persuaded than those who perceived the emails to be AI-mediated.

The mixed linear regression also found a statistically significant effect of argument quality on belief change (b=18.7019, p=<2e−16), and a statistically significant effect of source reliability on belief change (b=20.92, p=<2e−16). This indicates that higher-quality arguments and more reliable sources were strongly associated with more positive belief updating.

The ANOVA supported these findings, showing statistically significant effects for mediation type ($F(1, 225.02) = 13.6960$, $p = 0.0002703$), argument quality ($F(1, 836.19) = 204.9467$, $p < 2.2 \times 10$ -16), and source reliability ($F(1, 832.94) = 257.1303$, $p < 2.2 \times 10$-16). These results confirm that mediation type, argument quality, and source reliability independently account for substantial variance in the model. When interaction effects were added to the model, both the lmer and ANOVA found that the three-way interaction between mediation type, source reliability, and argument quality was not statistically significant ($F(1, 857.39) = 0.0002$, $p = 0.9896$). Individually, the two-way interactions between mediation type and source reliability ($F(1, 848.34) = 0.0132$, $p = 0.9084$) and between mediation type and argument quality ($F(1, 853.94) = 0.4613$, $p = 0.4972$) were also not statistically significant. However, the interaction between source reliability and argument quality was significant ($F(1, 850.85) = 7.7448$, $p = 0.005506$), suggesting that the combined effect of source reliability and argument quality lead to greater belief change.

To further investigate, two separate ANOVAs were conducted for each mediation type. Both the AI-mediated and human conditions found significant effects of source reliability and argument quality on belief change. In the AI-mediated condition, source reliability ($F(1, 411.96) = 133.7030$, $p <$ 2.2e-16) and argument quality ($F(1, 411.38) = 95.8762$, $p <$ 2.2e-16) were significant. Similarly, in the human condition, source reliability ($F(1, 426.00) = 127.3581$, $p <$ 2e-16) and argument quality ($F(1, 431.64) = 112.2311$, $p <$ 2e-16) were significant. When adding interaction effects to the model, both in the human and AI-mediated condition, ANOVAs revealed significant interactions between source reliability and argument quality on belief change (Human: $F(1, 437.57) = 7.7469$, $p = 0.005613$; AI-Mediated: $F(1, 423.32) = 8.4274$, $p = 0.0038892$).

Uniquely, the AI-mediated condition revealed a significant GAAIS Positive Mean ($F(1, 116.18) = 4.5306$, $p = 0.0354061$), whereas the human condition did not find this effect significant. This suggests that more positive attitudes towards AI (as measured by GAAIS) were uniquely associated with belief change when emails were perceived to be AI-Mediated rather than human. To further explore the direction of this effect, the equivalent lmer for the AI-mediated condition was assessed. It reported a significant positive GAAIS Positive Mean (b= 4.6692, p = 0.0354). This indicates that higher positive attitudes towards AI are significantly associated with greater belief change in the AI-mediated condition. In other words, individuals with more positive attitudes towards AI were more persuaded by the emails perceived to be AI-mediated.

Unexpectedly, the mixed linear regression for the full model found a statistically significant effect of education level on belief change (b=−2.5204, p=0.00388), indicating that higher education levels are associated with an average decrease in belief change. This suggests that individuals with higher levels of education were less persuaded by the mes-

sages overall. The ANOVA also found a statistically significant effect of education level on belief change ($F(1, 225.55) = 8.5123$, $p = 0.0038842$). However, it is possible that these results were influenced by the distribution of participants across education levels, with a substantial majority of the sample holding a bachelor's degree or higher.

The fit between Bayesian predicted posteriors and observed posteriors in both human and AI- mediated conditions were a good fit in both conditions. A linear regression analysis of the observed and predicted posterior beliefs was conducted to assess the model fit, yielding results of $R2 = 0.865$ for the human condition and $R2 = 0.856$ for the AI condition.

Hypothesis Testing: The above findings provide support for testing the hypotheses as follows.

H1. Source reliability will be positively correlated with perceived belief in the hypothesis.

The lmer and ANOVA results showed a statistically significant positive effect of source reliability on belief change ($b = 20.92$, $p < 2e{-}16$; $F(1, 832.94) = 257.1303$, $p < 2.2 \times 10{-}16$). This indicates that as source reliability increases, participants 'belief change also increases, supporting the hypothesis. Indeed, these results indicate that more reliable sources are more persuasive.

H1. Source reliability will be positively correlated with perceived belief in the hypothesis.

The lmer and ANOVA results showed a statistically significant positive effect of source reliability on belief change ($b = 20.92$, $p < 2e{-}16$; $F(1, 832.94) = 257.1303$, $p < 2.2 \times 10{-}16$). This indicates that as source reliability increases, participants 'belief change also increases, supporting the hypothesis. Indeed, these results indicate that more reliable sources are more persuasive.

H2. Argument quality will be positively correlated with perceived belief in the hypothesis. The lmer and ANOVA results demonstrated a statistically significant positive effect of argument quality on belief change ($b = 18.7019$, $p < 2e{-}16$; $F(1, 836.19) = 204.9467$, $p < 2.2 \times 10{-}16$). This indicates that as the quality of the arguments presented increases, participants 'belief change also increases, supporting the hypothesis. Therefore, we can assume that stronger arguments are more persuasive.

H3. Perceived AI-Mediation will have a moderating effect on influences of source reliability and argument quality on belief in the hypothesis.

The results from the lmer and ANOVA analyses indicated that the three-way interaction between mediation type, source reliability, and argument quality was not statistically significant. Additionally, the two-way interactions between mediation type and source reliability and between mediation type and argument quality were also not statistically significant. However, analyses did reveal a significant main effect of mediation type on belief change ($F(1, 225.02) = 13.6960$, $p = 0.0002703$), with participants in the human condition showing greater positive belief updating compared to those in the AI-mediated condition. This suggests that messages perceived to be AI-mediated were less persuasive than when perceived to be human written. Nevertheless, while AI-mediation significantly affects overall belief change, it does not appear to moderate the relationship between source reliability, argument quality, and belief change.

H3. 2. Source reliability will have main effects in both the human and AI-mediated conditions.

Separate ANOVAs for each mediation type were conducted to evaluate the main effects of source reliability on belief change. The results revealed significant effects of source reliability in both the AI-mediated condition ($F(1, 411.96) = 133.7030$, $p < 2.2e{-}16$) and the human condition ($F(1, 426.00) = 127.3581$, $p < 2e{-}16$). These findings indicate that reliable sources are more persuasive regardless of whether emails are perceived to be human-written or AI-mediated.

H3. 3. Argument quality will have main effects in both the human and AI-mediated conditions.

The separate ANOVAs for each mediation type indicated significant effects of argument quality in both the AI-mediated condition ($F(1, 411.38) = 95.8762$, $p < 2.2e{-}16$) and the human condition ($F(1, 431.64) = 112.2311$, $p < 2e{-}16$). These results demonstrate that higher-quality arguments are more persuasive regardless of whether they are perceived to be human-written or AI-mediated.

H4. In the AI-mediated condition, general attitudes towards artificial intelligence (GAAIS) will influence belief change. (Exploratory)

The ANOVA and lmer for the AI-mediated condition revealed a statistically significant effect of GAAIS Positive Mean on belief change ($b = 4.6692$, $p = 0.0354$; $F(1, 116.18) = 4.5306$, $p = 0.0354061$). This indicates that more positive attitudes towards AI are associated with greater belief change among participants in the AI-mediated context. This effect was not significant in the human condition, indicating that the influence of positive attitudes towards AI on belief change is uniquely relevant in the context of AI-mediation. Taken together, these results suggest that individuals with more favourable views of AI are more persuaded by AI-mediated emails than those with less favourable views of AI.

H5. Bayesian predictions modeled by Madsen and Wong (2023) will be positively correlated with the observed responses.

A linear regression analysis was conducted to assess the fit between the Bayesian predicted posteriors and the observed posteriors in both the human and AI-mediated conditions. The results showed a strong correlation, with $R2 = 0.865$ for the human condition and $R2 = 0.856$ for the AI condition. These high R2 values indicate that the Bayesian predictions are a good fit for the observed responses in both conditions. These findings support the hypothesis and suggest that the predictions derived from the previous model are effective in predicting belief change in both the human and AI-mediated contexts.

Discussion

Messages labelled as "edited by ChatGPT" were less persuasive than the same messages presented as entirely human-written. Consistent with previous research on AI-mediated communication which suggests that individuals experience increased uncertainty in AI-MC and may mistrust those they suspect of using it (Jakesch et al., 2019; Liu et al., 2022), this study demonstrates that individuals are less convinced to take advice from colleagues who have disclosed they are using ChatGPT to edit their emails. Therefore, disclosing the use of AI in an email made the sender less persuasive.

However, source reliability and argument quality had equivalent influences on belief change across both human and AI-mediated conditions. For example, credible senders delivering strong arguments produced the greatest belief change, regardless of AI involvement. Consistent with the Bayesian framework proposed by Madsen and Wong (2023), source reliability and argument quality interacted multiplicatively. The absence of interaction effects between mediation type and either source reliability or argument quality indicates that AI-involvement alters baseline persuasiveness but does not modify how recipients weigh these two factors.

As hypothesised, individuals 'general attitudes towards AI moderated responses in the AI-mediated condition. Participants with more favourable views of AI were comparatively more persuaded than those with less favourable views. This builds on previous research indicating that the persuasiveness of AI-generated content is influenced by general attitudes towards AI (Lim & Schmälzle, 2023), expanding this to include AI-mediated content. It also supports previous findings that attitudes towards AI influence trustworthiness in AI-MC (Jakesch et al., 2019). Additionally, it supports the Bayesian intuition that individuals will reach different reasonable conclusions based on their prior subjective beliefs.

This study also compared Bayesian predicted posteriors to observed posteriors using elicited conditional probabilities adapted from Madsen and Wong (2023). Predicted posteriors closely matched observed belief change in both conditions, indicating that the Bayesian framework captures belief revision under AI-MC. These results further evidence the predictive power of the Bayesian model tested, both within the context of digital communication and AI-MC, and invites further testing of Bayesian models to understand belief updating within computer and AI-mediated communication.

These findings offer guardedly positive implications for AI-MC. Participants did not abandon normative reasoning in the AI-condition: they still considered who was writing and how strong the arguments were. A trustworthy, expert sender with a strong argument retained considerable influence, even when ChatGPT was involved as an editor.

Practically, this underscores the importance of cultivating reliable relationships and high-quality content as AI becomes commonplace in organisational communication.

Limitations

The Shapiro-Wilk test revealed a violation of the assumption of normality in the results of the full lmer. Although non-parametric Kruskal-Wallis tests were employed to validate the findings, this violation limits how the results can be interpreted, particularly concerning interaction effects. Although the sample size was determined by a priori power analysis, a larger sample may have mitigated this distribution error. There are similar sample size concerns regarding the exploratory finding that higher education levels were associated with reduced belief change. This could have been influenced by the fact that most participants in this study held at least a bachelor's degree. This presents an opportunity for further investigation into how educational background influences perceptions of AI-mediated communication, and belief updating generally.

Concluding Remarks

This papers offers a first effort to apply a Bayesian framework for understanding the cognitive effects of AI-mediated communication. The key findings reveal that perceived AI-mediation and disclosure reduced the persuasiveness of emails compared with emails that were perceived to be human written. Despite this, AI-mediation did not entirely undermine the influence of source reliability and argument quality on relative persuasiveness. In practical terms, this means that when individuals believe an email was edited by ChatGPT, they tend to find it less convincing than if they believe it was written entirely by a colleague. However, this scepticism does not override the relative importance of who is sending the message or how well the arguments are constructed. If the sender of the email is trusted and the message is strong, these factors have a greater influence on the perception and persuasiveness of the message, even when AI involvement is acknowledged. This suggests that while AI can introduce a layer of doubt, it does not entirely discount the persuasiveness of the communication.

There are concerns in AI-mediated communication literature that warn of AI-MC fundamentally undermining trust in computer-mediated interactions all together (Hancock et al., 2020). However, the present findings offer a potentially reassuring perspective. They demonstrate that recipients do not completely discount AI-mediated emails. Instead, they continue to ground their judgements in who is sending the message and how strong the argument is.

As AI technologies become more deeply integrated into workplaces as tools for productivity and efficiency, these findings are relevant for informing how generative AI, and LLMs, are adopted within internal communications. Clearly, there is potential for such tools to be implemented in organisational communications without entirely undermining reliability and content. However, our findings suggest that building and maintaining interpersonal relationships among colleagues will become increasingly important in the AI-MC age. Considering that AI continues to advance and become more integrated into various aspects of communication, the study of AI-MC, and quantifiable frameworks like the Bayesian Model of Argumentation, offer vast potential for further research.

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Bastola, A., Wang, H., Hembree, J., Yadav, P., Gong, Z., Dixon, E., Razi, A., & McNeese, N. (2023). LLM-based Smart Reply (LSR): Enhancing collaborative performance with ChatGPT-mediated smart reply system. *arXiv*. https://doi.org/10.48550/ARXIV .2306.11980

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., & Wang, Y. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs*. Wiley.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems, 6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017- N

Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI- generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3637875

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149

Giurge, L. M., & Bohns, V. K. (2021). You don't need to answer right away! Receivers overestimate how quickly senders expect responses to non-urgent work emails. *Organizational Behavior and Human Decision Processes, 167*, 114–128. https://doi.org/10.1016/j.obhdp.2021.08.002

Hahn, U., & Oaksford, M. (2006a). A Bayesian approach to informal argument fallacies. *Synthese, 152*(2), 207–236. http://dx.doi.org/10.1007/s11229-005-5233-2

Hahn, U., & Oaksford, M. (2006b). Why a normative theory of argument strength and why might one want it to be Bayesian? *Informal Logic, 26*(1), 1–24. http://dx.doi.org/10.22329/il.v26i1.428

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*(3), 704–732. http://dx.doi.org/10.1037/0033-295X.114.3.704

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic, 29*(4), 337–367. https://doi.org/10.22329/il.v29i4.2903

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science, 40*(6), 1–38. http://dx.doi.org/10.1111/cogs.12276

Hohenstein, J., & Jung, M. (2018). AI-supported messaging: An investigation of human-human text conversation with AI support. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3170427.3188487

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior, 106*, 106190. https://doi.org/10.1016/j.chb.2019.106190

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., & Jung, M. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports, 13*, 5487. https://doi.org/10.1038/s41598-023-30938-9

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300469

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–29.

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI- generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Lagnado, D., Fenton, N., & Neil, M. (2012). Legal idioms: A framework for evidential reasoning. *Argumentation & Computation, 1*(1), 1–18.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. *The Economic Journal, 134*(658), 766–784. https://doi.org/10.1093/ej/uead056

Lim, S., & Schmälzle, R. (2023b). The effect of source disclosure on evaluation of AI-generated messages: A two-part study. a*rXiv Preprint arXiv:2311.15544.*

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. *CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3491102.3517731

Ma, X., Hancock, J. T., Lim, K. M., & Naaman, M. (2017). Self-disclosure and perceived trustworthiness of Airbnb

host profiles. *Proceedings of the 2017 ACM Conference on* Computer Supported Cooperative Work and Social Computing, 2397–2409. https://doi.org/10.1145/2998181.2998269

Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. *Proceedings of the Annual Meeting of the Cognitive Science Society, 38*. Retrieved from https://escholarship.org/uc/item/9v8245cs

Madsen, J. K., & Pilditch, T. D. (2018). A method for evaluating cognitively informed micro- targeted campaign strategies: An agent-based model proof of principle. *PLOS ONE, 13*(4), e0193909. https://doi.org/10.1371/journal.pone.0193909

Madsen, J. K., & Wong, M. (2023). Comparing predictions from the elaboration likelihood model and a Bayesian model of argumentation. *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*. Retrieved from https://escholarship.org/uc/item/75x869j9

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, *58*(2), 121-131. http://dx.doi.org/10.1037/h0085798

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847- 855. http://dx.doi.org/10.1037/0022-3514.41.5.847

Ragot, M., Martin, N., & Cojean, S. (2020). Ai-generated vs. Human artworks. A perception bias towards artificial intelligence? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1– 10.

Reinke, K., & Chamorro-Premuzic, T. (2014). When email use gets out of control: Understanding the relationship between personality and email overload and their impact on burnout and work engagement. *Computers in Human Behavior*, *36*, 502–509. https://doi.org/10.1016/j.chb.2014.03.075

Schepman, A., & Rodway, P. (2023). 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*, *39*(13), 2724– 2741.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sivunen, A., & Laitinen, K. (2019). Digital communication environments in the workplace. In L. Mikkola & M. Valo (Eds.), *Workplace communication* (1st ed., pp. 41–53). Routledge. https://doi.org/10.4324/9780429196881-4

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *arXiv Preprint arXiv:2301.11924*.

Teigen, C., Madsen, J. K., George, N. L., & Yousefi, S. (2024). Persuasiveness of arguments with AI-source labels. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*. Retrieved from https://escholarship.org/uc/item/6t82g70v

Turner, T., Qvarfordt, P., Biehl, J. T., Golovchinsky, G., & Back, M. (2010). Exploring the workplace communication ecology. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 841–850. https://doi.org/10.1145/1753326.1753449

Work Trend Index | Will AI Fix Work? (2023). Retrieved August 12, 2024, from https://www.microsoft.com/en-us/worklab/work-trend-index/will-ai-fix-work

Wu, Y., & Kelly, R. M. (2020). Online dating meets artificial intelligence: How the perception of algorithmically generated profile text impacts attractiveness and trust. *32nd Australian Conference on Human-Computer Interaction, 444–453*. https://doi.org/10.1145/3441000.3441074