Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Explaining the Reputational Risks of AI-Mediated Communication: Messages Labeled as AI-Assisted Are Viewed as Less Diagnostic of the Sender's Moral Character

Pranav Khadpe*, **Kimi Wenzel***, **George Loewenstein**, **Geoff Kaufman**
Carnegie Mellon University
{pkhadpe, kwenzel, gl20, gfk}@andrew.cmu.edu

Abstract

When someone sends us a thoughtful message, we naturally form judgments about their character. But what happens when that message carries a label indicating it was written with the help of AI? This paper investigates how the appearance of AI assistance affects our perceptions of message senders. Adding nuance to previous research, through two studies ($N = 399$) featuring vignette scenarios, we find that AI-assistance labels don't necessarily make people view senders negatively. Rather, they dampen the strength of character signals in communication. We show that when someone sends a warmth-signalling message (like thanking or apologizing) without AI help, people more strongly categorize the sender as warm. At the same time, when someone sends a coldness-signalling message (like bragging or blaming) without assistance, people more confidently categorize them as cold. Interestingly, AI labels weaken both these associations: An AI-assisted apology makes the sender appear less warm than if they had written it themselves, and an AI-assisted blame makes the sender appear less cold than if they had composed it independently. This supports our *signal diagnosticity explanation*: messages labeled as AI-assisted are viewed as *less diagnostic* than messages which seem unassisted. We discuss how our findings shed light on the causal origins of previously reported observations in AI-Mediated Communication.

Introduction

How people communicate is one of the fundamental signals we use to judge their moral character. These judgments affect whom we choose to approach and whom we choose to trust (Fiske et al. 2018). Many of us instinctively trust the colleague who remembers to thank the cleaning staff unprompted each day, while questioning the warmth of the manager who thanks his staff only when reminded by HR. As society adapts to communication that is assisted by artificial intelligence (AI)—whether speeding up typing with autocomplete, streamlining exchanges with autoreply, or delegating thoughtfulness to ChatGPT—questions about how it will affect interpersonal relationships have become a matter of public debate and a focus of scholarly research (Jakesch et al. 2019; Liu et al. 2022; Weiss et al. 2022; Hohenstein and Jung 2020): if you received a thank you note labeled "written with the help of AI," would you perceive the sender as warmer or colder?

Empirical work on these perceptions consistently finds that when warmth-signalling messages such as greetings, thanks, and apologies are labeled as AI-assisted, they are associated with lower ratings of the sender's moral character (reduced perceived warmth) compared to identical messages presented without AI labels [1] (e.g. Jakesch et al. 2019; Liu et al. 2022; Weiss et al. 2022; Hohenstein and Jung 2020; Hohenstein et al. 2023; Mieczkowski et al. 2021). However, few explanatory mechanisms have been proposed for why this is the case. Two distinct mechanisms could explain this disparity in warmth judgments: either people view the AI-assisted message as negative evidence of warmth (as evidence that the sender is actually cold and calculated), or they see it as merely weaker positive evidence of warmth (accepting that the sender possesses warmth, but not intensely enough that its verbal expression flows without assistance).

In this paper, we describe, test, and ultimately find support for the *signal diagnosticity explanation*, which is that appearance of AI assistance makes messages weaker (less diagnostic) evidence rather than negative evidence. The implication of this is that AI labels *dampen* the effect a message would have otherwise had (as opposed to AI labels having a categorical negative effect). The sender of a thanks or apology is seen as less warm when AI-assisted, and the sender of brags and blames is seen as less cold (more warm) when AI-assisted. Our two studies provide evidence for this explanation by examining how people perceive senders when AI labels are attached to both warmth-signalling messages (thanks and apologies) and coldness-signalling messages (brags and blames). Study 1 directly tested message diagnosticity: apology messages without AI labels were viewed as more diagnostic of warmth than those with AI labels, while blame messages without AI labels were viewed as more diagnostic of coldness than those with AI labels. Study 2 tested effects on actual warmth judgments. AI-assisted thanks and apologies received significantly lower warmth ratings than non-AI-assisted messages. Yet strikingly, this drop disappeared for brags and blames, supporting the idea that AI labels have a dampening effect rather than a categorical neg-

*These authors contributed equally.

[1] *Warmth* is the standard measure of moral character in social cognition literature (Fiske, Cuddy, and Glick 2007).

ative effect.

By placing phenomena observed in AI-Mediated Communication (AIMC) on firm psychological foundations, we discuss how our work sheds light on previously reported phenomena, including how AIMC affects interpersonal trust (Jakesch et al. 2019) and serves as a moral crumple zone (Hohenstein and Jung 2020).

Contributions: Our contributions are (1) an explanation for how appearance of AI assistance affects judgments of a message sender's moral character; (2) empirical support for the explanation across two studies; (3) tracing previously reported empirical phenomena in AIMC to their psychological origins.

Approach

The Signal Diagnosticity Explanation: The diagnosticity approach (Skowronski and Carlston 1987) frames character judgments as a probabilistic categorization process. When evaluating others, people analyze signals—like the messages someone sends—to determine whether that person belongs in the "warm" or "cold" category. Signals contribute to categorization probabilistically rather than with absolute certainty (Skowronski and Carlston 1987). For instance, someone who sends a thoughtful message of thanks probably belongs in the warm category, although there is a real, though lesser, probability that they belong to the cold category.

Signals are defined as more diagnostic when they lead to higher perceived probabilities that a person belongs to one category (e.g. warm) and lower perceived probabilities that the person belongs to the alternate category (e.g. cold). Therefore, more diagnostic signals lead to more confident assignment of a person to either the warm or cold category.

Our signal diagnosticity explanation proposes that messages without AI labels are more diagnostic because people intuitively believe that genuinely warm individuals express warmth directly (through thanks or apologies) with greater ease, while genuinely cold individuals express coldness directly (through brags or blame) more naturally. We propose that when thanks and apologies appear unassisted, they serve as stronger probabilistic evidence that the sender belongs in the warm category. Similarly, when brags and blames appear unassisted, they function as stronger probabilistic evidence that the sender belongs in the cold category. We suggest (and show) that AI labels reduce this diagnosticity in both cases: AI-assisted thanks and apologies become weaker evidence that the sender belongs in the warm category, and AI-assisted brags and blames become weaker evidence that the sender belongs in the cold category.

The Current Work: In both studies, participants read a fictional scenario involving interactions between them and a fictional character. The scenario was manipulated across conditions to create warrant for the character to thank, blame, apologize to, or brag to the participant. Then, participants read a message from the character either thanking, apologizing, bragging or blaming, presented with or without an AI label ("parts of this message were written with the help of AI"). Participants then provided their judgments of the message and the fictional character. In Study 1, we focused only on an apologies and blames, and we elicited subjective probabilities from participants to compute message diagnosticity. In Study 2, we expanded the set of messages to also include thanks and brags, and we captured participants judgments of the sender's warmth, the sender's competence, and open-ended responses about their attributions of the sender's use of AI. This served as a confirmatory test of the explanation and provided richer insights into the basis of people's judgments.

Related Work

We begin by establishing the theoretical foundation for our investigation, drawing on research in social cognition that identifies warmth as a primary dimension of interpersonal judgment. We then make a distinction between judgments of messages versus their senders, before reviewing empirical work on AIMC that motivates our explanatory account.

Warmth: A Judgment of Moral Character

Research in social cognition shows that to navigate our social world safely and effectively, we evaluate others along two fundamental dimensions: warmth and competence (Fiske, Cuddy, and Glick 2007; Fiske et al. 2018). *Warmth* captures whether someone intends to help or harm us—distinguishing between those we perceive as warm (trustworthy, kind) versus cold (untrustworthy, hostile). It is typically understood as a judgment of moral character, encompassing traits such as trustworthiness, kindness, and friendliness, (Fiske, Cuddy, and Glick 2007; Russell and Fiske 2008). Meanwhile, *competence* captures whether they have the ability to act on those intentions—their perceived skill and effectiveness (Fiske, Cuddy, and Glick 2007; Russell and Fiske 2008). These two dimensions are shown to emerge reliably across stimuli, culture, and time (Fiske, Cuddy, and Glick 2007; Russell and Fiske 2008).

Of the two dimensions, judgments along the warmth dimension are primary. Warmth is judged first (Willis and Todorov 2006) and has a larger impact on individual's affective and behavioral reactions (Hack, Goodwin, and Fiske 2013; Fiske, Cuddy, and Glick 2007; Abele and Wojciszke 2014). A common explanation for this prioritization is that, from an evolutionary perspective, another person's disposition to help or harm has more consequential implications for one's survival than their ability to act on those intentions (Abele and Wojciszke 2007; Fiske, Cuddy, and Glick 2007; Cuddy, Fiske, and Glick 2008).

Our investigations focus on the warmth dimension. This is due to two intersecting reasons. First, the fact that warmth judgments are primary and consequential makes it an important dimension of study; most prior work on AIMC studies warmth-related traits (Jakesch et al. 2019; Hohenstein and Jung 2020; Hohenstein et al. 2023; Liu et al. 2022). Second, focusing on the broader dimension of warmth, rather than individual traits like friendliness or trustworthiness, allows us to explain findings across studies that examine different warmth-related characteristics. This unification provides a foundation for theory-building because while many prior studies on AIMC measure individual warmth-related traits (e.g., trustworthiness in Jakesch et al. 2019), few explicitly reference the broader warmth construct (Weiss et al. 2022; Mieczkowski et al. 2021), which has limited theoretical synthesis across individual studies.

Dissociated Responses to Message and Source

Character judgments—warmth judgments—are not assessments of the *message* per se; rather they are inferences about the *sender* drawn from the message. This distinction is important because character inferences about the source of communication can often dissociate from reactions to the message itself (Malle 2021). For example, you might find a friend's hastily written, typo-filled text message annoying or unclear while still viewing them as a caring and trustworthy person. Dissociated responses are, in fact, known to occur when social action is mediated by AI systems (Renier, Mast, and Bekbergenova 2021; Bower and Steyvers 2021; Liu and Moore 2022). For instance, people are just as persuaded by argument messages presented as AI-generated as they are when presented as human-written, even though they rate the AI source as less trustworthy than a human source (Aydin and Malle 2024; Gallegos et al. 2025).

Our focus in this paper is not on how people judge AI-assisted messages, but rather how they judge the *sender* of such messages. While there is research documenting that people view AI-assisted messages as morally inappropriate (Moradi and Levy 2024; Liu et al. 2022), the primary focus of our explanatory account and empirical investigation is on the inferences people make about the sender upon receiving the message. Focusing on how recipients judge senders helps develop a richer understanding of how AI reshapes interpersonal relationships and trust dynamics.

Empirical Investigations of AIMC

The term AI-mediated communication (AIMC) covers different kinds of mediated communication between people "in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals" (Hancock, Naaman, and Levy 2020; Jakesch et al. 2019). Instances of AIMC can be found directly embedded in communication tools (e.g. smart/auto reply[2] and writing support[3]) and can also be facilitated through third-party tools (e.g. drafting a message using a consumer-facing LLM, such as ChatGPT).

Most empirical investigations into the impact of AI mediation on interpersonal communication can be classified into two broad threads of work: a first thread studying *recipient judgments*, and a second thread studying *sender behaviors*. The first thread—and the thread we contribute to—investigates how awareness that a sender used AI affects recipients' judgments (Jakesch et al. 2019; Hohenstein and Jung 2020; Hohenstein et al. 2023; Liu et al. 2022; Purcell et al. 2023; Cheong et al. 2025). The second thread focuses on how AI assistance alters the sender's messaging behavior, including language use (Arnold, Chauncey, and Gajos 2020; Jakesch et al. 2023; Hohenstein et al. 2023; Mieczkowski et al. 2021; Agarwal, Naaman, and Vashistha 2025) and communication frequency (Hohenstein et al. 2023).

In this paper, we focus on the first thread: we study the recipient's judgments of the sender, when the message is known to be written with AI. Studies examining these perceptions reveal a consistent pattern: senders of warmth-signalling messages like greetings, thanks, and apologies receive lower warmth ratings when their messages carry AI assistance labels compared to when identical messages appear unassisted (Jakesch et al. 2019; Liu et al. 2022; Weiss et al. 2022; Hohenstein and Jung 2020; Hohenstein et al. 2023; Mieczkowski et al. 2021). For example, Jakesch et al. found that Airbnb hosts are perceived as less trustworthy when participants knew the host used AI (Jakesch et al. 2019) to write their profile greetings. Similarly, Liu et al. found that senders of invitations, and condolences—messages that by themselves would enhance perceptions of the sender's warmth—lead to lower warmth judgments when they appear AI-assisted (Liu et al. 2022). This pattern has been observed across both individual messages (Jakesch et al. 2019; Weiss et al. 2022; Liu et al. 2022) and extended communication episodes (Mieczkowski et al. 2021; Hohenstein and Jung 2020; Hohenstein et al. 2023).

Despite this consistent finding, few mechanisms have been proposed to explain why perceptions of AI assistance reduce warmth judgments in the above cases. Thus, we extend prior work by proposing that AI-labeled messages are viewed as *less diagnostic* of the sender's moral character. Crucially, our explanation also predicts when AI labels will **not** lower warmth judgments: for coldness-signalling messages like brags and blames. We provide empirical support for this explanation across the studies we present next.

Study 1

Our studies were designed to investigate whether the presence of AI labels affects message diagnosticity (Study 1), and to investigate the effect of AI labels on downstream judgments of the sender's warmth (Study 2). In the first study, participants rated the subjective probabilities that a particular message is associated with the trait category warm, and with the opposite trait category: cold. From this data, we computed perceptions of message diagnosticity for warmth and coldness. This study featured a 2 (message type: apologizing vs. blaming) $\times$ 2 (authorship label: AI label vs. no label) between-subjects design. Our studies used the method of hypothetical scenarios. The method is commonly used in studies of interpersonal perception (Caprariello, Cuddy, and Fiske 2009; Chaudhry and Loewenstein 2019), including those focused on AIMC (Liu et al. 2022; Jakesch et al. 2019). Hypothetical scenarios allow us to precisely control the variables of the setup—message type and authorship label—that are the focus of our investigation.

[2]https://blog.google/products/gmail/save-time-with-smart-reply-in-gmail/

[3]https://www.zdnet.com/article/you-can-now-use-gemini-in-google-messages-if-youre-among-the-lucky-few/

Method

Participants: Using the online crowdsourcing platform *Prolific*, we recruited 130 participants. This sample goal was based on a power analysis to detect a medium effect of size $f \geq 0.25$ (powered at 80% with a significance level of 0.05). After excluding participants who failed the attention check, or the comprehension check, we were left with 119 participants for analysis (29-30 in each condition; mean age = 40.8; 60 women and 59 men). Participants received $2 for completing the 10 minute survey.

Stimulus Development: We wanted to test how AI labels affect diagnosticity for both warmth-signalling (e.g. thanks and apologies) messages and coldness-signalling messages (e.g. brags and blames). Drawing on prior work that has established how apologies are warmth-signalling and blames are cold-signalling (Chaudhry and Loewenstein 2019), in this study we focus on apologies and blames are representative of those two categories. Specifically, we follow previous work (Chaudhry and Loewenstein 2019) and operationalize an apology message as the sender "*taking* blame for a negative outcome", and blame message as the sender "*giving* the recipient blame for a negative outcome".

To create social warrant for the participant to be receiving the message, we used a hypothetical scenario involving the participant and a friend ("John") who are members of a community organization planning neighborhood events. In the scenario, one character helps the other plan an event, which then receives harsh criticism at a community meeting. We manipulated who was the primary organizer versus helper to create contexts where either an apology or blame message from John would be situationally appropriate: an apology was warranted when the participant organized and John helped, while blame was warranted when John organized and the participant helped.

Both messages were 3-4 sentences long and generated using ChatGPT (GPT-4). The prompt included the scenario corresponding to the message, and asked the system to generate a short, 3-4 sentence long, text message from John. For participants assigned to see the message with AI labels, the message was additionally appended with a label that read "parts of this message were written with the help of AI". This resulted in four conditions: apology without label, apology with AI label, blame without label, and blame with AI label. (The scenario, the message content, and how they were shown to participants can be found in Supplementary Material.)

Trait-Behavior Probability Scale: To reason about the diagnosticity of each of the four messages, we need estimates of the perceived probabilities that the message is associated with the trait category *warm*, and with the opposite category: *cold*. For this, we draw on the trait-behavior probability scale (Skowronski and Carlston 1987) used to measure diagnosticity. In the scale, a particular behavior is paired with opposing traits in sentences of the form: "Would a (trait) person ever (behavior)?". For example, in Skowronski and Carlson's own study—which measured the diagnosticity of different behaviors for the trait of honesty (and dishonesty)—a pair of items shown to an individual participant were: "Would an honest person ever steal money from his roommate's wallet?" and "Would a dishonest person ever steal money from his roommates wallet?".

Participants respond to these questions on a 9-point scale, ranging from (1) *extremely unlikely to perform the behavior* to (9) *extremely likely to perform the behavior*, with (5) *moderately likely to perform the behavior* at the midpoint.

To reason about diagnosticity of a behavior, the responses are used to calculate a signal-validity score. This is calculated as a ratio of the perceived probabilities: "The numerator is the rated probability that an actor with a given trait will perform a specified behavior, and the denominator is the sum of this probability with the rated probability that an actor with the opposing trait will perform the same behavior." In the above example, if a participant rates the probability that an honest person would steal the money at 1, and the probability that a dishonest person would steal the money at 6, then the signal-validity of the behavior *steal money from roommates wallet* for the trait category of *honest* would be $0.14\ (1/[1+6])$, and the signal-validity of this behavior for the category *dishonest* would be $0.86\ (6/[1+6])$. We adapt the trait-behavior probability scale as we describe next.

Procedure and Measures: Participants were randomly assigned to one of four conditions in a 2 (message type: apologizing vs. blaming) $\times$ 2 (authorship label: no label vs. AI label) between-subjects design. After the consent procedures, participants read the scenario, which varied according to condition. Then they read the message from John. Following this they provided subjective probabilities in response to two questions: "Would a *warm* person ever communicate the way that John did?" and "Would a *cold* person ever communicate the way that John did?". Participants responded to each question on a 9-point scale, ranging from (1) *extremely unlikely* to (9) *extremely likely*, with (5) *moderately likely* at the midpoint. Finally, they provided information about age and gender.

Results

Message Diagnosticity: We calculated the signal-validity score of each message for the trait category of warm using the previously described formula (see the Trait-Behavior Probability Scale section). **A score close to 0.5 indicates that the message was *low diagnosticity***: it was not thought to discriminate between warmth and coldness. **Scores above 0.5 indicate the message was more associated with warmth than coldness, while scores below 0.5 indicated the opposite. Both of these are *high in diagnosticity*, just for opposite ends of the warmth spectrum**. By construction, the signal-validity statistic for warmth equals 1 minus the signal-validity statistic for coldness. For example, if an apology message has a signal-validity score of 0.8 for warmth, then its signal-validity score for coldness is automatically 0.2. Hence, for simplicity, we only report analysis on each message's signal-validity for the warmth trait.

Figure 1 shows the means of signal-validity scores of each message corresponding to the trait category of warm. In absence of AI labels, apologizing was viewed as more diagnos-

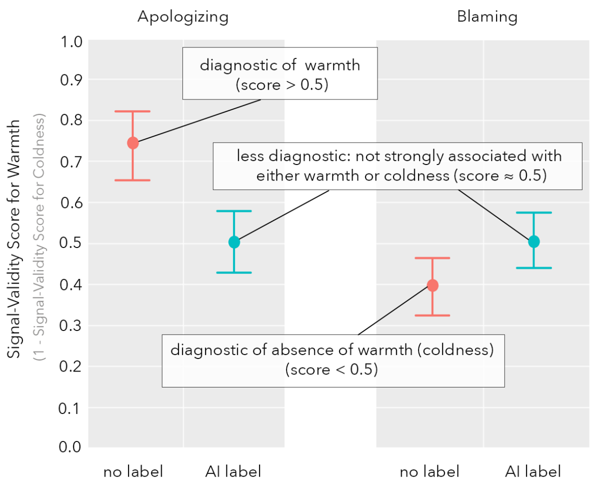

Figure 1: Study 1 shows that AI labels make both apology messages and blame messages less diagnostic (results in signal-validity scores closer to 0.5). In absence of AI labels, apology messages are diagnostic of warmth, and blame messages are diagnostic of an absence of warmth (diagnostic of coldness). Error bars are 95% CIs.

tic of warmth (Mean score $= 0.75$, $SD = 0.24$), and blaming was viewed as more diagnostic of absence of warmth (or more diagnostic of coldness) (Mean score $= 0.39$, $SD = 0.20$). The presence of AI labels, moves the validity score of both blaming (Mean score $= 0.50$, $SD = 0.19$) and apologizing (Mean score $= 0.50$, $SD = 0.21$) closer to 0.5, suggesting a drop in diagnosticity.

We analyzed the signal-validity score using a two-way analysis of variance (ANOVA). Signal-validity score (corresponding to the trait category warm) was the dependent variable, message type was the first factor (2 levels: apologizing, blaming), and authorship label was the second factor (2 levels: no label, AI label). The ANOVA revealed a significant interaction effect ($F(1, 115) = 20.10$, $p < 0.001$, $\eta^2 = 0.15$). We also found a large main effect of message type ($F(1, 115) = 19.83$, $p < 0.001$, $\eta^2 = 0.15$). This main effect is as expected: the signal-validity score for warmth was lower for blaming messages (Mean score $= 0.45$, $SD = 0.20$) than apologizing messages (Mean score $= 0.62$, $SD = 0.25$). Put simply, apologies are more strongly associated with warmth than blames. The main effect of authorship label was small and not significant ($F(1, 115) = 2.98$, $p = 0.09$, $\eta^2 = 0.03$). That the effect of AI labels is primarily an interaction effect (rather than a main effect) is consistent with the idea that the directional effect of AI labels depends on message type: as shown in Figure 1, for blames it *increases* the validity score for warmth and for apologies it *decreases* the validity score for warmth, ultimately reducing the diagnosticity of each message (moving the score closer to 0.5). A post hoc Tukey test revealed that the difference in validity score between AI-labeled and unlabeled apologies was significant ($p < 0.001$), but the difference between AI-labeled and unlabeled blames was not ($p = 0.21$).

Discussion

The unlabeled apology message was associated more strongly with warmth than with coldness, whereas the unlabeled blame message was more strongly associated with coldness than with warmth. However, AI labels weakened these associations in *both* cases: the AI-assisted apology and the AI-assisted blame both had signal-validity scores closer to 0.5. Taken together, these results provide preliminary evidence that appearance of AI assistance makes messages less diagnostic.

Based on these findings, we suggest that the lower diagnosticity of AI-labeled messages may explain patterns observed in prior work (Jakesch et al. 2019; Hohenstein and Jung 2020; Hohenstein et al. 2023; Liu et al. 2022; Purcell et al. 2023). For warmth-signalling messages like apologies, lower diagnosticity should produce drops in warmth judgments of the sender: when evidence of warmth is weaker, recipients ought to make more conservative judgments, rating the sender as less warm than they would based on stronger evidence. But, by the same token, for coldness-signalling messages like blames, lower diagnosticity should *prevent* such drops—because the message becomes less strongly associated with coldness (and more strongly associated with warmth). This second prediction has not been previously tested. Testing it would also clarify whether the primary effect of AI labels is one of lowering diagnosticity or whether the effect is categorically negative: if the effect were categorically negative, we would expect AI labels to lower warmth ratings even for brags and blames. Therefore, in Study 2, we tested these predictions. A secondary objective of Study 2 was to further investigate why people instinctively view AI-labeled messages as less diagnostic.

Study 2

The second study used the same scenario method as before, but we expanded the set of stimuli to also include thanking and bragging messages, resulting in two warmth-signalling messages (thanking, apologizing) and two coldness-signalling messages (bragging, blaming). Thus, the study featured a 4 (message type: thanking vs. apologizing vs. bragging vs. blaming) $\times$ 2 (authorship label: AI label vs. no label) between-subjects design. We included a measure for perceived warmth of the sender (for our confirmatory analysis), as well as exploratory measures to further understand the process through which people judge the sender's warmth. The main predictions being tested were that: (1) for thanking and apologizing messages, senders will be perceived as less warm when messages carry AI labels compared to no labels; and (2) for blaming and bragging messages, senders will not be perceived as less warm when messages carry AI labels compared to no labels. The study was preregistered. (Preregistration can be accessed at: `https://aspredicted.org/9QX_P8F`[4])

[4] For brevity, we do not report on exploratory variables that were collected for different analytical purposes unrelated to the hypotheses tested.

Method

Participants: We aimed for 35 participants in each of the four conditions to detect an effect size of $f \geq 0.25$ (powered at 95% and a significance level of 0.05). (We expected a medium effect size based on a pilot study with 80 participants.) Again, using Prolific, we recruited 290 participants. 9 participants did not complete the survey and 1 failed the attention check, leaving 280 valid responses for analysis (33-36 in each condition; mean age = 38; 132 women and 148 men). Participants received $2 for the 10 minute survey.

Stimulus Development: We used a scenario setup similar to Study 1. However, in Study 2, we also included thanks and brags in addition to the apologies and blames considered in Study 1. Thanks and apologies are known to be warmth-signalling and brags and blames are known to be coldness-signalling (Chaudhry and Loewenstein 2019). Again, drawing on prior work (Chaudhry and Loewenstein 2019), we operationalize a thanks message as the sender "*giving* the recipient credit for a positive outcome", and brag message as the sender "*taking* credit for a positive outcome".

Building on the scenario from Study 1, we expanded the design by additionally varying the event outcome (harshly criticized vs. effusively praised at the community meeting). This created a factorial structure where different combinations of participant role (organizer vs. helper) and outcome (positive vs. negative) naturally warranted different message types from John to the participant: thanks (participant helped John , positive outcome), apology (John helped participant, negative outcome), brag (John helped participant, positive outcome), and blame (participant helped John, negative outcome).

The content of the apologizing and blaming messages was the same as in Study 1. The thanking and bragging message were also generated in a similar manner, prompting ChatGPT (GPT-4) with the scenario setup and asking it to generate a 3-4 sentence long, text message from John. For participants assigned to see the message with AI labels, the message was additionally appended with a label that read "parts of this message were written with the help of AI". (The scenario, the message content, and how they were shown to participants can be found in Supplementary Material.)

Hypotheses: For our study context, the predicted effects of AI labels of warmth judgments translate to the following 5 hypotheses, for which we planned to conduct (and preregistered) confirmatory tests:

- **H1**: There **will be** an interaction effect between message type and authorship label on perceived warmth.
- **H2A**: For message type of thanking, perceived warmth **will be lower** when there is AI label than when there is no label.
- **H2B**: For message type of apologizing, perceived warmth **will be lower** when there is AI label than when there is no label.
- **H3A**: For message type of bragging, perceived warmth **will not be lower** when there is AI label than when there is no label.
- **H3B**: For message type of blaming, perceived warmth **will not be lower** when there is AI label than when there is no label.

Procedure and Measures Participants were randomly assigned to one of eight conditions in a 4 (message type: thanking vs. apologizing vs. bragging vs. blaming) $\times$ 2 (authorship label: no label vs. AI label) between-subjects design. After the consent procedures, participants read the scenario, which varied according to condition. Then they read the message from John. Participants then responded to two key measures. First, they rated John's **perceived warmth** on 8 items (kind, likeable, warm, trustworthy, assertive [reverse coded], competitive [reverse coded], cold [reverse coded], unfriendly [reverse coded]) using 7-point scales (1 = *not at all*, 7 = *extremely*), which we averaged into a composite score ($\alpha = 0.86$). Second, they rated John's **perceived competence** on 5 items (competent, intelligent, successful, hard working, skillful) using 7-point scales (1 = *not at all*, 7 = *extremely*), averaged into a composite score ($\alpha = 0.95$). Participants in AI label conditions additionally responded to an open-ended question that asked: "What do you think led John to use AI to write parts of the message?". This allowed us to capture the **causal attributions** they made for his AI use. We included perceived competence and the question about causal attributions as exploratory measures to further investigate the process through which recipients judge the sender. Finally, participants provided information about age and gender.

Results

Perceived Warmth As preregistered, we created a composite score for perceived warmth by taking the average of the corresponding items ($\alpha = 0.86$), and analyzed it using a two-way analysis of variance (ANOVA). Perceived warmth was the dependent variable, message type was the first factor (4 levels: bragging, thanking, apologizing, blaming), and authorship label was the second factor (2 levels: AI label, no label).

Figure 2 shows the mean perceived warmth in each study condition. The ANOVA revealed a significant interaction effect with a medium effect size ($F(3, 272) = 7.94$, $p < 0.001$, $\eta^2 = 0.08$). This provides support for **H1**. We also found a large main effect of message type ($F(3, 272) = 26.72$, $p < 0.001$, $\eta^2 = 0.23$) and a small main effect of authorship label ($F(1, 272) = 10.58, p < 0.01, \eta^2 = 0.04$).

Further, as preregistered, we carried out four planned comparisons, corresponding to each of the four remaining hypotheses. Each planned comparison (and each remaining hypothesis) considered a single message type and contrasted the group that saw the message type with AI label to the group that saw the same message type but with no label. For each planned comparison, we used an independent samples t-test[5]. This returns valid p-values (Seltman 2012; Midway et al. 2020), without requiring corrections, because the comparisons are: (1) planned, (2) orthogonal (each group ap-

[5]Analyzing the contrasts via a post hoc Tukey test also leads to the same findings.

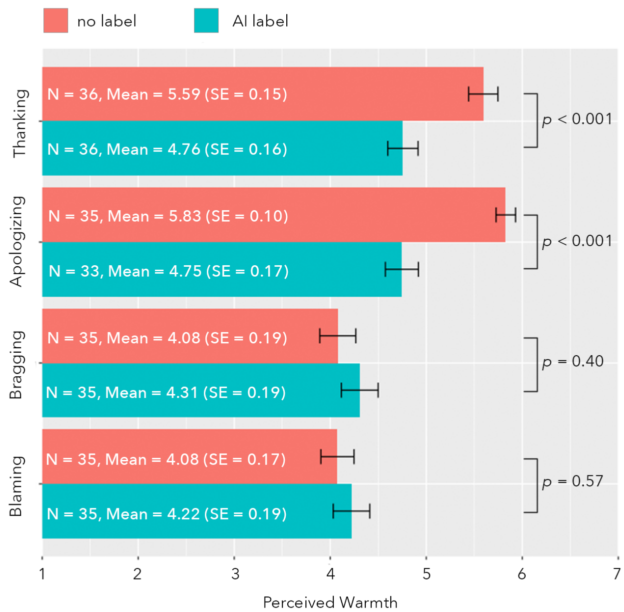

Figure 2: Study 2 shows that AI labels lead to lower warmth judgments in the case of thanking and apologizing but not for bragging and blaming. (S.E. and error bars denote standard error.)

pears in only one comparison), and (3) fewer in number than the total degrees of freedom (seven).

For message type of thanking (Figure 2), John's perceived warmth was significantly lower ($t(69.92) = -3.81$, $p < 0.001$) among participants who saw the message with AI label ($M = 4.76$, $SD = 0.95$) compared to participants who saw the message with no label ($M = 5.59$, $SD = 0.91$). This supports **H2A**.

Similarly, for message type of apologizing (Figure 2), John's perceived warmth was significantly lower ($t(52.20) = -5.42$, $p < 0.001$) among participants who saw the message with AI label ($M = 4.75$, $SD = 0.99$) compared to participants who saw the message with no label ($M = 5.83$, $SD = 0.60$). This supports **H2B**.

For message type of bragging (Figure 2), there was no significant difference in John's perceived warmth ($t(67.97) = 0.85$, $p = 0.40$) across authorship label, although the average rating was higher among participants with AI label ($M = 4.31$, $SD = 1.13$) than among participants with no label ($M = 4.08$, $SD = 1.11$). This supports **H3A**.

Finally, for message type of blaming (Figure 2), there was no significant difference in John's perceived warmth ($t(67.36) = 0.57$, $p = 0.57$) across authorship label, although the average rating was higher among participants with AI label ($M = 4.22$, $SD = 1.13$) than among participants with no label ($M = 4.08$, $SD = 1.02$). This supports **H3B**.

Perceived Competence We calculated a composite score ($\alpha = 0.95$) for perceived competence and analyzed it using a two-way ANOVA with perceived competence as the dependent variable, message type as the first factor and authorship label as the second factor. We find a main effect of authorship label ($F(1, 272) = 15.32$, $p < 0.001$, $\eta^2 = 0.05$) and a main effect of message type ($F(3, 272) = 9.74$, $p < 0.001$, $\eta^2 = 0.10$). The interaction effect was not significant and small ($F(3, 272) = 1.78$, $p = 0.15$, $\eta^2 = 0.02$).

A post hoc Tukey test showed that John's competence was judged significantly lower ($p < 0.001$) by groups that saw AI label ($M = 4.09$, $SD = 1.37$) as compared to groups that saw no label ($M = 4.70$, $SD = 1.35$). We develop this observation further in the discussion, where we integrate it with our causal attribution analysis.

Causal Attributions At the end of the survey, participants in the indication of AI use conditions were asked: "What do you think led John to use AI to write parts of the message?" There were a total of 139 responses (corresponding to the 139 participants across the indication of AI use conditions). These responses were inductively coded in two cycles to identify participants' interpretations of John's AI use, focused on structural and descriptive codes (Saldaña 2021). This resulted in five codes (see Table 1): *help verbalizing*, *sensitive topic*, *time*, *laziness*, and *character judgment*. Some participant responses received two tags, as participants explicitly mentioned multiple possibilities (ex. *"He either needed help to make it sound presentable or was simply too lazy to care"* was coded as *help verbalizing* and *lazy*.) In this way, the 139 responses resulted in a total of 145 tags. As a final step, after the inductive process above, we used a deductive process to group the codes based on whether they involve: (1) attributions to John's *warmth* (whether John intends good or ill) (2) attributions to John's *competence* (John's ability to act on those intentions) or (3) attributions to John's *situation*. Theories of attribution (Ross 1977; Gilbert and Malone 1995) (inferring the causes of someone's behavior) often describe attributions in terms of disposition and situation. The deductive grouping of attributions to warmth and competence, brings together responses where participants attributed John's use of AI to his disposition. Meanwhile, deductive grouping of attributions to situation brings together responses where participants attributed John's use of AI to his situation. Groupings are shown alongside the code in Table 1. The warmth group included one code (*character judgment*), the competence group included two codes (*help verbalizing* and *laziness*) and the situation group included two codes (*sensitive topic* and *time*).

Based on this deductive grouping, we found that John's use of AI was most often attributed to competence-related factors: 84 of the 145 tags included codes from the competence group (help with verbalizing and laziness). Another 52 tags attributed AI use to situational factors, such as the sensitivity of the topic or time constraints John likely faced. Only rarely did participants interpret AI use as reflecting John's poor character or lack of warmth (9 tags).

Discussion

Study 2 provides empirical support for the signal diagnosticity explanation. AI labels led to lower warmth judgments in the case of thanking and apologizing but not for bragging

Code *(Group)*	Frequency	Description	Quote
Help Verbalizing *(Competence)*	73	Use of AI to help find the right words to express oneself.	"Probably just because he wanted to say something clearly or he couldn't figure out the right words to use."
Sensitive Topic *(Situation)*	30	Use of AI to help manage emotional reactions of the sender or recipient.	"He was struggling with how to convey his emotions on a tough subject."
Time *(Situation)*	22	Use of AI to save time.	"I think it was more of a time issue. John wanted to say 'thanks,' but didn't really have time to compose a thoughtful message."
Laziness *(Competence)*	11	Use of AI due to laziness.	"I'd assume the use of AI was due to general laziness and a dislike of writing in general, rather than a personal slight directed toward myself."
Character Judgment *(Warmth)*	9	Use of AI due to a disregard for one's communication partner.	"Because he didn't really want to give the credit to his helper."

Table 1: Participants' causal attributions for why John used AI to write a message.

and blaming. These results align with a dampening rather than blanket negative effect: a categorically negative view of AI assistance would predict lower warmth ratings for all messages, including brags and blames

The qualitative analysis of causal attributions provides additional support for the signal diagnosticity explanation by revealing how participants interpret AI use. When asked why John used AI assistance, the most frequent attribution was "help verbalizing" (73 out of 145 tags), where participants believed John needed AI to find the right words to express himself. This aligns with the signal diagnosticity interpretation: participants aren't concluding that John lacks the underlying trait entirely, but rather that he possesses it without sufficient intensity or skill to articulate it independently. As one participant noted, John "wanted to say something clearly or he couldn't figure out the right words to use." This interpretation suggests that AI assistance signals a gap between having a sentiment and being able to express it effectively, making the message less diagnostic of the sender's true character. The predominance of competence-related attributions (84 tags) over warmth-related ones (9 tags) further supports this view—participants see AI use as reflecting limitations in expressive ability rather than absence of the underlying feeling. This interpretation is reinforced by the finding that AI labels reduced perceived competence ratings, suggesting participants view AI assistance predominantly as evidence of reduced communicative skill.

General Discussion

Across two studies, we found support for the *signal diagnosticity explanation*: that AI labels result in messages being viewed as less diagnostic of the sender's moral character. We now turn to discussing how our work sheds light on previously reported phenomena, including how AIMC serves as a moral crumple zone (Hohenstein and Jung 2020) and affects interpersonal trust (Jakesch et al. 2019). Finally, we discuss design opportunities, outline limitations of our work, and identify directions for future research.

AI as a Moral Dampener Rather Than a Moral Disqualifier

Rather than AI labels serving as categorical disqualifiers of moral character (marking senders as definitively cold or calculating), they seem to function as dampeners: AI labels don't just reduce the positive impact of warm messages, they also limit the negative impact of cold ones. The mere presence of AI assistance acts as a social buffer, absorbing interpersonal damage when conversations go poorly

This sheds light on the psychological basis for the previous observation of AIMC as a moral "crumple zone" (Hohenstein and Jung 2020)—a concept borrowed from automotive safety where certain parts of a car are designed to absorb impact during crashes, protecting the occupants. This previous work found that when recipients knew that AI was involved in crafting messages, they held the sender less responsible for negative conversational outcomes, prompting the researchers to conclude that AI mediation can strengthen relationships between communicators.

Our work provides support for this perspective but also reveals a more complex picture. The very feature that makes AI-assisted criticisms feel survivable makes AI-assisted thank you notes feel hollow. While dampening can prevent relationship damage in negative interactions, it can simultaneously undermine relationship building in positive ones.

In Search for Honest Signals

Previous research identified the *"Replicant Effect"* (Jakesch et al. 2019), where individuals merely suspected of relying on AI assistance for warmth-signalling messages like profile greetings were judged less warm. Our work shows that even when people have complete certainty about AI involvement (through explicit labels), the depression in warmth assessment persists.

People seem to inherently view unintermediated communication as a more "honest signal" (Donath 2007a,b; Zahavi 1975)—a signal that more reliably indicates underlying traits because it is costlier or more difficult to fake (Zahavi 1975). When someone expresses gratitude or remorse without assistance, many of us may instinctively read it as evidence that these feelings are strong enough to flow naturally into words. Here, the cognitive and emotional effort required to translate feeling into language serves to authenticate the emotion itself. By reducing or eliminating this effort, AI assistance strips away these authentication markers and recipients are left unable to distinguish between someone who deeply feels an emotion and someone who merely recognizes they should express it.

As AI assistance becomes ubiquitous, we risk creating a communicative race—where everyone must escalate their expressions of warmth just to achieve the same interpersonal impact that unassisted communication once delivered. This could result in paradoxical outcomes: as our messages become more polished and eloquent through AI, they simultaneously become less meaningful and less trusted. The handwritten note gains value precisely because it cannot be automated; the stuttered apology may repair relationships more effectively than the eloquent one. This suggests that preserving spaces for effortful expression isn't just nostalgic sentiment but a functional necessity for maintaining trust in human relationships.

Designs to Reduce Attributional Ambiguity

Rather than abandoning effortful expression entirely, we suggest that communication technologies could be designed to preserve and highlight markers of human investment (He, Houde, and Weisz 2025). There is an opportunity to investigate whether AI tools that reveal the sender's emotional and cognitive investment could help enhance relationship building and emotional connection. Future work could test whether effort-revealing designs (Kelly et al. 2017; Zhang et al. 2022) that highlight personal investment—such as showing that someone struggled to find the right words or spent significant time crafting their response—generate stronger feelings of warmth and appreciation than blanket AI-assistance labels. Understanding how these design choices affect relational satisfaction and emotional bonding could inform the development of AI communication tools that enhance rather than diminish human connection.

Limitations

Our study focusing on hypothetical scenarios offered experimental evidence of the signal diagnosticity explanation. However, this experiment did not allow us to test whether there are behavioral consequences of these effects. For instance, if someone is known to have used AI when blaming, is the recipient's response different than if there was no disclosure of AI use? Future work that investigates these message types in the context of live interactions can help identify the behavioral consequences.

Our focus in this work was on the recipient's perspective, rather than the sender's perspective. We were unable to investigate, for instance, whether senders would choose to use AI assistance for some message types over others. We were also unable to investigate whether the use of AI affected whether senders *felt* like they bragged, blamed, apologized, or thanked. Recent studies suggest that use of AI assistance can also affect the sender's self-perceptions about their agency and control (Kadoma et al. 2024; Hwang et al. 2025). Future investigation that considers the sender's perspective can advance our initial findings to strengthen the conceptual foundations of AI-mediated communication.

Finally, our work did not consider the ways in which actually composing a message with an AI tool can itself influence the sender's message (Arnold, Chauncey, and Gajos 2020; Jakesch et al. 2023). Consistent with prior work (Liu et al. 2022; Jakesch et al. 2019; Weiss et al. 2022), we opt for a controlled test of our hypotheses by keeping the content of the message constant across the indication of AI use and no indication of AI use conditions. However, prior work has documented how biases in generative AI (towards certain generation lengths or towards certain opinions) can influence what people write (Arnold, Chauncey, and Gajos 2020) and think (Jakesch et al. 2023). Future work is needed to understand how the technical underpinnings of current systems might interact with the effects we study.

Conclusion

When people see "written with AI," they do not conclude the sender is cold or calculating; they conclude the sender possesses their expressed sentiment, but the degree of the sentiment is obfuscated by AI. This dampening effect means AI labels weaken both positive and negative character impressions: AI-assisted apologies seem less warm, but AI-assisted blame seems less cold. These reactions provide credence to the idea that AI labels reduce the diagnosticity of messages, and help explain previously observed phenomena in AIMC. Our work offers a foundation for studying how perceptions of communication partners, rather than the content they share, change with AI integration.

Reflection on Ethics and Societal Impact

Positionality and Adverse Impact Statement

Our US-based team comprises individuals with diverse academic backgrounds, including human computer interaction (HCI), responsible AI, and psychology. Our backgrounds have informed how we frame and pursue our investigations.

Our research participants were all US-based, and so, they may share similar cultural assumptions about the relationship between expressive ability and moral judgments. The signal diagnosticity framework we employ reflects particular cultural values about individual agency and the importance of unmediated self-expression that may not generalize across all cultural contexts.

We also recognize that our focus on warmth and competence as primary dimensions of moral character reflects established Western psychological frameworks, which may not capture the full complexity of how different communities evaluate moral character or the role of technological mediation in social relationships.

Our goal has been to advance academic and public debates about AI's societal impact, while maintaining scientific rigor in our methodology and interpretation of results. Nevertheless, we invite readers to approach our findings with awareness of these contextual limitations.

Ethical Considerations

Both studies were approved by our university's Institutional Review Board. Our study included a minor psychological risk, given that some participants were exposed to a psychologically threatening (i.e. blaming) message. This risk was offset by the fictitious nature of the scenario, and further offset by the fact that participants were asked to reflect on the sender's motivation. Reflecting on an aggressor's motivation can help dampen negative psychological impacts (Gyurak, Gross, and Etkin 2011).

Acknowledgments

We thank our anonymous reviewers for their feedback on the paper. We also thank Chinmay Kulkarni for insightful discussions and valuable feedback that helped shape this work.

References

Abele, A. E.; and Wojciszke, B. 2007. Agency and communion from the perspective of self versus others. *Journal of personality and social psychology*, 93(5): 751.

Abele, A. E.; and Wojciszke, B. 2014. Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology*, volume 50, 195–255. Elsevier.

Agarwal, D.; Naaman, M.; and Vashistha, A. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–21.

Arnold, K. C.; Chauncey, K.; and Gajos, K. Z. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 128–138.

Aydin, Z.; and Malle, B. F. 2024. Dissociated Responses to AI: Persuasive But Not Trustworthy? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Bower, A. H.; and Steyvers, M. 2021. Perceptions of AI engaging in human expression. *Scientific reports*, 11(1): 21181.

Caprariello, P. A.; Cuddy, A. J.; and Fiske, S. T. 2009. Social structure shapes cultural stereotypes and emotions: A causal test of the stereotype content model. *Group Processes & Intergroup Relations*, 12(2): 147–155.

Chaudhry, S. J.; and Loewenstein, G. 2019. Thanking, apologizing, bragging, and blaming: Responsibility exchange theory and the currency of communication. *Psychological review*, 126(3): 313.

Cheong, I.; Zhang, A. X.; Lee, M.; Kadoma, K.; Chang, J. C.; Naaman, M.; Liao, S.; and Guo, A. 2025. The Role of AI Disclosure and Author Identity in Reader Perceptions: A Randomized Experimental Study.

Cuddy, A. J.; Fiske, S. T.; and Glick, P. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, 40: 61–149.

Donath, J. 2007a. Signals, cues and meaning. *Signals, Truth and Design*.

Donath, J. 2007b. Signals in social supernets. *Journal of computer-mediated communication*, 13(1): 231–251.

Fiske, S. T.; Cuddy, A. J.; and Glick, P. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2): 77–83.

Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, 162–214. Routledge.

Gallegos, I. O.; Shani, C.; Shi, W.; Bianchi, F.; Gainsburg, I.; Jurafsky, D.; and Willer, R. 2025. Labeling messages as AI-generated does not reduce their persuasive effects. *arXiv preprint arXiv:2504.09865*.

Gilbert, D. T.; and Malone, P. S. 1995. The correspondence bias. *Psychological bulletin*, 117(1): 21.

Gyurak, A.; Gross, J. J.; and Etkin, A. 2011. Explicit and implicit emotion regulation: A dual-process framework. *Cognition and emotion*, 25(3): 400–412.

Hack, T.; Goodwin, S. A.; and Fiske, S. T. 2013. Warmth trumps competence in evaluations of both ingroup and outgroup. *International Journal of Science, Commerce and Humanities*, 1(6): 99.

Hancock, J. T.; Naaman, M.; and Levy, K. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1): 89–100.

He, J.; Houde, S.; and Weisz, J. D. 2025. Which Contributions Deserve Credit? Perceptions of Attribution in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Hohenstein, J.; and 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.

Hohenstein, J.; Kizilcec, R. F.; DiFranzo, D.; Aghajari, Z.; Mieczkowski, H.; Levy, K.; Naaman, M.; Hancock, J.; and Jung, M. F. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1): 5487.

Hwang, A. H.-C.; Liao, Q. V.; Blodgett, S. L.; Olteanu, A.; and Trischler, A. 2025. 'It was 80% me, 20% AI': Seeking Authenticity in Co-Writing with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–41.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–15.

Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naaman, M. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Kadoma, K.; Aubin Le Quere, M.; Fu, X. J.; Munsch, C.; Metaxa, D.; and Naaman, M. 2024. The Role of Inclusion, Control, and Ownership in Workplace AI-Mediated Communication. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–10.

Kelly, R.; Gooch, D.; Patil, B.; and Watts, L. 2017. Demanding by design: Supporting effortful communication practices in close personal relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 70–83.

Liu, Y.; Mittal, A.; Yang, D.; and Bruckman, A. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–13.

Liu, Y.; and Moore, A. 2022. A Bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Malle, B. F. 2021. Moral judgments. *Annual Review of psychology*, 72(1): 293–318.

Midway, S.; Robertson, M.; Flinn, S.; and Kaller, M. 2020. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ*, 8: e10387.

Mieczkowski, H.; Hancock, J. T.; Naaman, M.; Jung, M.; and Hohenstein, J. 2021. AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–14.

Moradi, P.; and Levy, K. 2024. " A Fountain Pen Come to Life": The Anxieties of the Autopen. *International journal of communication (Online)*, 18: 784–793.

Purcell, Z. A.; Dong, M.; Nussberger, A.-M.; Köbis, N.; and Jakesch, M. 2023. Fears about AI-mediated communication are grounded in different expectations for one's own versus others' use. *arXiv preprint arXiv:2305.01670*.

Renier, L. A.; Mast, M. S.; and Bekbergenova, A. 2021. To err is human, not algorithmic–Robust reactions to erring algorithms. *Computers in Human Behavior*, 124: 106879.

Ross, L. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, volume 10, 173–220. Elsevier.

Russell, A. M. T.; and Fiske, S. T. 2008. It's all relative: Competition and status drive interpersonal perception. *European Journal of Social Psychology*, 38(7): 1193–1201.

Saldaña, J. 2021. The coding manual for qualitative researchers.

Seltman, H. 2012. Experimental design and analysis.

Skowronski, J. J.; and Carlston, D. E. 1987. Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of personality and social psychology*, 52(4): 689.

Weiss, D.; Liu, S. X.; Mieczkowski, H.; and Hancock, J. T. 2022. Effects of using artificial intelligence on interpersonal perceptions of job applicants. *Cyberpsychology, Behavior, and Social Networking*, 25(3): 163–168.

Willis, J.; and Todorov, A. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7): 592–598.

Zahavi, A. 1975. Mate selection—a selection for a handicap. *Journal of theoretical Biology*, 53(1): 205–214.

Zhang, L.; Chen, T.; Seow, O.; Chong, T.; Kratz, S.; Tham, Y. J.; Monroy-Hernández, A.; Vaish, R.; and Liu, F. 2022. Auggie: Encouraging Effortful Communication through Handcrafted Digital Experiences. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–25.