AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust

Abstract

AI-mediated communication (AI-MC) represents a new paradigm where communication is augmented or generated by an intelligent system. As AI-MC becomes more prevalent, it is important to understand the effects that it has on human interactions and interpersonal relationships. Previous work tells us that in human interactions with intelligent systems, misattribution is common and trust is developed and handled differently than in interactions between humans. This study uses a 2 (successful vs. unsuccessful conversation) x 2 (standard vs. AI-mediated messaging app) between subjects design to explore whether AI mediation has any effects on attribution and trust. We show that the presence of AI-generated smart replies serves to increase perceived trust between human communicators and that, when things go awry, the AI seems to be perceived as a coercive agent, allowing it to function like a moral crumple zone and lessen the responsibility assigned to the other human communicator. These findings suggest that smart replies could be used to improve relationships and perceptions of conversational outcomes between interlocutors. Our findings also add to existing literature regarding perceived agency in smart agents by illustrating that in this type of AI-MC, the AI is considered to have agency only when communication goes awry.

Keywords: Artificial Intelligence (AI), Communication, AI-Mediated Communication (AI-MC), Computer-Mediated Communication (CMC), Trust, Attribution

1. Introduction

Trust is critical in communication, especially in computer-mediated communication (CMC), where social presence is lower than with face-to-face communication [28]. The development of trust is an attributional process, with trust being influenced by the trustor's attribution of positive motivation to their communication partner [13]. In human-computer interaction (HCI), misattribution is common, with people applying different moral norms to intelligent systems and humans [25]. For example, when accidents happen in human interactions with intelligent systems, we sometimes see the emergence of a "moral crumple zone", where the ethical blame for any negative or unintended consequences is attributed to a human instead of the system [21]. In addition, trust already starts at a lower level in CMC than face-to-face communication [86] and, in text-based communication, is particularly difficult to develop when compared with other mediums [9]. We seek to understand how attribution and trust are affected by the mediation of AI in CMC, which we describe as AI-mediated communication (AI-MC) [35, 37].

The addition of artificial intelligence to CMC represents a new paradigm where communication is augmented or generated by an intelligent system. AI-MC is already widely-used in some ways. Spell check, predictive text, and grammar correction are used to improve communication clarity, while automatic translations serve to improve comprehension [26, 87]. While systems like these represent a minimal interference of AI in communication, new systems display a much higher amount of intervention, such as smart replies in messaging and email, where users are offered suggested responses that are algorithmically-generated through natural language processing (NLP). This type of AI-MC is becoming more widely used all the time, with new implementations on Android devices, Google's messaging apps, Skype, LinkedIn, Facebook Messenger, Slack, and more. Additionally, as NLP continues to develop, AI-MC will likely become more robust and widely-used along with it.

While AI-MC is directly aimed at shaping the production of messages and despite previous work suggesting that its presence is affecting conversations [35], we do not know how AI mediation is influencing interpersonal dynamics and interaction outcomes. To avoid unexpected social consequences, we need to understand the effects that AI-MC has on human interactions.

This study examines how perceived interpersonal dynamics are affected by the presence of AI in CMC. Specifically, we measure attribution and perceptions of trust in successful and unsuccessful computer-mediated conversations, with and without the presence of AI mediation in the form of smart replies. Our findings indicate that AI mediation is related to increased trust between human communicators and that in unsuccessful conversations, the AI acts like a moral crumple zone, taking on responsibility that otherwise would have been assigned to the human. The discussion draws on relevant theories that may account for these observations and suggests possibilities for leveraging our findings into systems that could resolve team conflict and improve communication outcomes. The results of this study expand the existing literature on interpersonal dynamics in CMC by showing that AI-MC has the potential to improve interpersonal relationships and perceptions of conversational outcomes between human communicators. Additionally, our work adds to the body of work regarding perceived agency of smart agents by demonstrating that in this particular type of AI-MC, the AI seems to only have agency when conversations go awry.

2. Background

Despite the increasing prevalence of AI-MC, we do not know how it is affecting interpersonal dynamics and conversational outcomes. This study is motivated by previous work suggesting that the presence of AI is affecting CMC in unspecified ways and that when humans collaborate with intelligent systems, misattribution is common and problematic. We situate these ideas within the relevant theories regarding attribution and trust to determine how AI-MC could affect interpersonal dynamics.

2.1. Trust and Attribution in Communication

Trust development is an attributional process [44], and perceived trust is an important aspect of developing and maintaining interpersonal relationships. Successful cooperation

between human communicators occurs when ambiguity and uncertainty in social perceptions are reduced through the development of trust [18].

2.1.1. Perceiving Trust

Trust is particularly essential in facilitating successful groupwork [42, 59], with increased trust enabling more effective conflict resolution [78], problem solving [49, 89], and improved team performance [20]. Conversely, when a team lacks trust, various negative consequences can arise, including impaired learning [20] and decreased willingness to cooperate [46]. A lack of trust can be particularly problematic in high-risk situations, such as military and emergency contexts [27], where it can impair chances of survival [84]. High-risk contexts like these are already regarded as practical applications of AI-MC [47], furthering the pressing need to understand how perceptions of trust are affected by AI mediation.

In the information science literature, trust has direct positive effects on cooperation and performance (e.g., [36, 38, 39]), and in CMC, high levels of trust enhance collaboration and information exchange (e.g., [9, 74, 71]). However, previous work on CMC channels has shown that trust starts at a lower level in CMC [86] and is particularly difficult to develop in text-based communication when compared with other mediums [9]. In accordance with similar CMC literature, our work references a relational notion of trust, which refers to the interpersonal social exchanges that take place within group settings and is crucial to many types of interpersonal interactions and particularly important in judging computer credibility [23].

2.1.2. Attribution Theory and Trust

Trust development is an attributional process. Attribution theory describes the human tendency to ascribe intentionality to the past and future actions of the self and others [44, 58], regardless of possessing the necessary amount of relevant information to do so. Attribution based on limited information can result in attributional errors, where people incorrectly attribute causes to another person, themselves, or situational factors. Trust is influenced by attribution to the extent that the trustor ascribes positive motivation to their partner [13].

Dirks and Ferrin describe two models for the role of trust in interaction outcomes: the direct effects model and the moderation model [18]. The direct effects model suggests that an individual's level of trust in another party directly affects their perceptions of the outcome. High levels of trust will cause the communicator to have a positive attitude, resulting in high satisfaction and positive perceptions of performance with respect to the interaction outcome. Conversely, low levels of trust will result in low satisfaction with and negative perceptions of the outcome.

The moderation model suggests that trust will instead influence how a communicator interprets and evaluates information relevant to attitude and behavior. Dirks and Ferrin offer two explanations of the moderation model: (1) "trust affects how one *assesses the future behavior* of another party with whom one is interdependent" and (2) "trust also affects how one *interprets the past (or present) actions* of the other party, and the motives for the underlying actions" [18]. Attribution theory tells us that when behavior is consistent with expectations, humans will attribute causes of actions to internal characteristics, but

when behavior is inconsistent with prior expectations, causes of actions will be attributed to external situational characteristics [41].

Dirks and Ferrin posit that the influence of trust depends on the "situational strength" of the interaction. In situations with weak structure, where there is a lack of clear guidance of how to interpret others' behavior, the direct effects model applies, and trust fills in these gaps. When a situation has moderately strong structure and there is some limited information to assess others' behavior, the moderation model applies, and trust influences the way that attitudes and behaviors are interpreted. Lastly, in a situation with strong structure, where there is little missing information or ambiguity, external cues will "over determine" behavior, leaving little to no role for the influence of trust on perceptions [18, 40].

Applying these ideas to a messaging context, Jarvenpaa et al. describe a situation where a communicator is waiting for an email response from another party [40]. As the time without a response increases, an explanation is needed for the delay, and the moderation model suggests that the level of trust influences perceptions and attitudes regarding the slow response. If the communicator has high levels of trust in the other party, they are likely to attribute the delay to external factors, such as technical difficulties, and their attitude will not change. Conversely, if the communicator has low levels of trust, they will likely attribute the delayed response to the internal characteristics of the other person (e.g., uncooperative behavior), and attitude and team performance will be negatively affected.

Misunderstandings such as the one aforementioned are often unavoidable in CMC [68], where social presence is lower than with face-to-face communication [28], making these interactions particularly prone to attributional error. Assuming too much responsibility for an outcome can lead to frustration and rigidity [73], yet when someone avoids blame by wrongly assigning it to others, errors and conflict can result [48]. Given the importance of attribution and trust in communication, particularly in CMC, it is concerning that intelligent systems could be changing these perceptions in unintended ways.

2.2. The Moral Crumple Zone and Misattribution in Interactions with Intelligent Systems

In their case study regarding the history of aviation autopilot litigation, Elish and Hwang identify a steadfast focus on human responsibility, even as humans in the cockpit have been increasingly replaced by autopilot technology. Even as control for complex systems like those found in aviation are being distributed across multiple actors, including humans and intelligent systems, social and legal perceptions of responsibility have generally continued to focus on the human actor [22]. The term "moral crumple zone" describes the result of this ambiguity within systems of distributed control, especially automated and autonomous systems [21].

When accidents happen, humans naturally want someone to blame. When intelligent systems are involved in catastrophic accidents, attribution is distributed differently, with humans typically believing that any negative or unintended consequences result from a human who fails to act morally or ethically [62]. In a car accident, the crumple zone is physically designed to deform to absorb the force of the crash impact. When things go awry in human interactions with intelligent systems, just like a crumple zone in a car absorbs the

impact, humans act like a "moral crumple zone" and are attributed responsibility, including any legal or moral penalties that result from the failure of the system [21].

The manifestation of the moral crumple zone is seen in multiple examples of tragic human interactions with intelligent systems. After the Three Mile Island Nuclear Generating Station accident occurred, blame was almost entirely attributed to the human plant operators, despite knowledge of ongoing problems with filters in the feedwater pipe system and that the design of the plant's control interface inadequately represented the physical conditions of the system [21]. Similarly, human error was blamed for the catastrophic crash of Air France Flight 447, which was the result of a complex system failure arguably out of human control [21]. In short, after a problem with the plane's pitot tubes, the crew was unable to recover from the resulting aerodynamic stall because of imprecise alarms and warnings that prevented a return to a flight angle which would allow recovery. In cases like these, we see how, despite our belief of their infallibility, intelligent systems cannot predict and plan for every possible situation, and when they fail to do so, the blame is often attributed to human actors.

We know that the moral crumple zone reveals itself in catastrophic accidents involving intelligent systems. Conversely, in some cases, people attribute some responsibility for computer errors to the computer itself [25]. Overall, humans are not rational in their attribution of responsibility when technology is involved, believing in the superiority of computer judgment until a mistake is made or an automated system reaches its limits [21, 79]. It seems that if things go wrong in everyday exchanges between people where intelligent systems are involved, such as in AI-MC, attribution would likely be designated differently than if such systems were not involved. So, does the presence of AI mediation in CMC affect attribution and trust when things go awry?

2.3. AI-Mediated Communication

The addition of artificial intelligence to CMC represents a new paradigm where communication is augmented or generated by an intelligent system, which we describe as AI-mediated communication (AI-MC) [35, 37]. AI-MC has been studied in a few ways. One study examined how AI mediation affects online self-presentation and found that AI-generated Airbnb host profiles were perceived as less trustworthy than those written by humans [37].

Another form of AI-MC that influences how we communicate on a real-time basis is smart replies. The purpose of smart replies is to help users more quickly compose short messages with "just one tap" [43]. Smart replies exist in various messaging applications and offer users suggested responses that are algorithmically-generated through NLP, such as shown in Figure 1. AI-MC is becoming more widely used all the time, constituting 10% of messages sent through Gmail [57] and with implementations on Android Messages, Skype, LinkedIn, Facebook Messenger, Slack, and more. Even though this type of AI-MC is directly aimed at shaping the production of messages and despite previous work suggesting that it is influencing messaging conversations [35], current research on smart replies has predominantly focused on developing ways to generate these messages such that they are personalized and fit within the conversational context (e.g., [43, 32, 72]). As a result, we do

not know how smart replies could potentially be influencing conversational dynamics and interpersonal relationships.

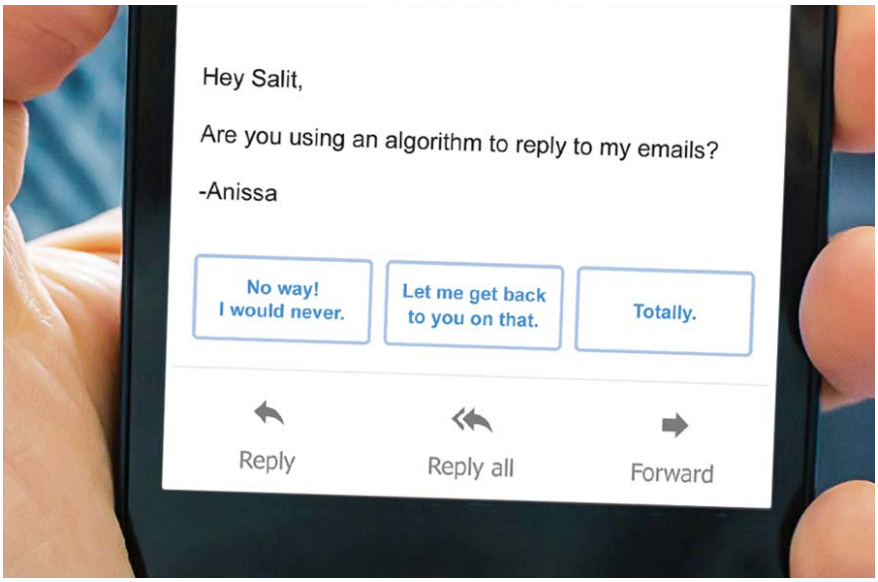

Figure 1: Google Allo is an AI-assisted messaging app that includes smart replies that the user can tap on to quickly reply, as shown above [63].

Humans already have a tendency to trust other humans over computers [70], suggesting that humans will be trusted more than AI in AI-MC. Additionally, most current AI-MC systems allow the sender to know that their responses have been modified or generated by AI, whereas the receiver has no knowledge of this. Given users' preference for reducing uncertainty in interactions [8], it seems that this lack of transparency regarding the influence of AI could serve to increase uncertainty and negatively affect perceptions of trust [77].

In addition to concerns about the ability of smart agents to manipulate public opinion [16, 24], initial work regarding perceptions of AI-mediated messaging apps on more general conversation has shown that smart replies may be influencing conversational dynamics and outcomes [35]. As previously discussed, we also know that the presence of intelligent systems can affect attribution in unexpected ways. Given the importance of trust and attribution in communication, especially CMC, we seek to understand how they are affected by AI mediation. Hopefully, this knowledge will enable us to design systems that can effectively leverage AI to improve interpersonal relationships and messaging conversation outcomes overall.

3. Study Overview

In messaging interactions, users have limited information to assess others' behavior, indicating that this type of interaction has a moderately strong structure. Dirks and Ferrin suggest that in this case, the moderation model of trust applies, where trust influences the interpretation of attitudes and behaviors [18]. In this study, we propose trust and attribution

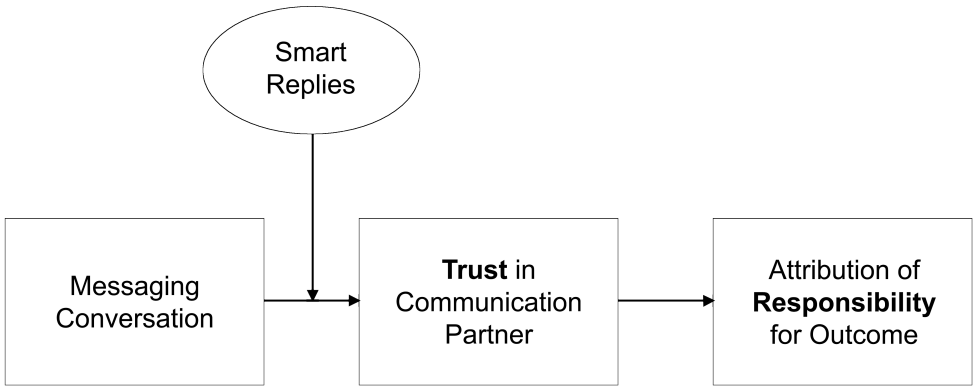

Figure 2: Research model.

as outcome variables, respectively, with the presence of AI-generated smart replies as a moderator influencing this relation path (Figure 2). To examine this, we used a 2 (successful vs. unsuccessful conversation) x 2 (standard vs. AI-mediated messaging app) between subjects design.

We know that when accidents occur, attribution is distributed differently when intelligent systems are involved [21], suggesting that interpersonal perceptions could be related to whether or not a messaging interaction goes awry. For this reason, we chose to examine perceptions of both successful and unsuccessful conversations. Successful conversations occur when communicators construct similar situational models to each other [69], meaning that their integrated mental representations of the considered state of affairs aligns. We define a successful conversation as an instance when the conversation resolves with a mutually-reached outcome (i.e., communicators agree), and we define an unsuccessful conversation as an instance when the conversation does not resolve with a mutually-reached outcome (i.e., communicators do not agree). The success of each conversation was controlled by a confederate (i.e., a person who participated in the experiment pretending to be a subject but, in actuality, was working for the researcher).

The experiment had 4 conditions, as shown in Table 1. Conditions 1-2 functioned as control conditions that were used to compare with the results from the experimental AI-mediated messaging conditions, 3-4. Our hypotheses for each condition, which were made in accordance with the reviewed literature, are also shown.

Table 1: We compared how humans attribute responsibility and perceive trust with a 2 (successful vs. unsuccessful conversation) x 2 (standard vs. AI-mediated messaging app) between subjects design. Our hypotheses for each condition are shown.

	Standard Messaging	AI-Mediated Messaging
Successful Conversation	**(1)** Control trust Control responsibility	**(3)** Less trust in partner than (1) Same responsibility to self as (1) Same responsibility to partner as (1) No responsibility to AI
Unsuccessful Conversation	**(2)** Less trust in partner than (1) & (3) More responsibility to self than (4) Less responsibility to partner than (4)	**(4)** Lowest trust in partner Least responsibility to self Most responsibility to partner Some responsibility to AI

4. Materials and methods

In this study, participants had conversations with an anonymous partner, who was actually a confederate controlling the outcome of the exchange. Perceptions of trust and attribution were measured after the conversation was completed.

4.1. The Messaging Apps

The AI-mediated messaging app used in this study was Google Allo. Allo combines AI assistance with instant messaging to create an AI-mediated messaging application. When using the app, users are provided smart replies, suggested responses based on an algorithm and parsing of the conversation history [43]. Smart replies typically come in groups of 3 phrases after a message is received by the user, as shown in Figure 1, but can also display as a banner if the user is doing something outside of the Allo application. This means that AI-generated responses can be sent at any time, with the other communicator remaining unaware of which responses have been crafted manually and which have been entirely generated by AI. It should be noted that this application has been deprecated since these experiments took place.

The standard messaging app used in this study was Whatsapp. Whatsapp was chosen as the control messaging app because of its equivalent media richness and user interface similarities to Allo.

4.2. Participants

Participants (N=113, 75.2% F) were recruited from an on-campus recruiting system at a large university in the northeastern United States and received course credit for their participation. Participants ranged in age from 18-25 (M=19.28, SD=1.21).

4.3. Procedure

The survey was administered using Qualtrics, an online survey platform. Participants were told that the study was about how people form impressions of each other in messaging conversations and were not given information about the actual purpose of the experiment. Upon accessing the survey, participants were instructed to leave the survey open in a web

browser on a computer while using their smartphone to have the conversation. After consenting to participate, participants were given instructions for downloading and installing either the standard (Whatsapp) or AI-mediated (Allo) messaging app. In case participants in the AI-mediated messaging condition were unfamiliar with smart replies, they were told that AI would offer smart replies that they could tap to send.

Next, participants were told that they would be working with another participant to complete a task and were instructed to send a message to a specified phone number. Participants were not informed that they would actually be working with a confederate who would be carefully controlling the dynamics and outcome of the conversation. Because a name was required to use the messaging app, the name associated with the confederate account was changed each day to a name pulled from a large online list of gender-neutral baby names [11].

Participants were instructed to participate in small talk with the confederate for 5 minutes, and they were given a variety of sample topics to choose from, as shown in Appendix A.1, although they were not limited to the suggested topics. This informal conversation served to familiarize participants, specifically those in the AI-MC conditions, with the messaging application. Additionally, these few minutes of "small talk" served to ascribe positive motivation to the confederate and build trust [19] between participants and the confederate. Building trust also allowed us to establish participants' expectations of the confederate, hopefully influencing their perception of the subsequent (non-)cooperation of the confederate.

After 5 minutes had elapsed, participants were allowed to progress to the next survey page, where they were presented with a variation of the lifeboat task, a commonly-used group task in experimental studies that involves making a ranked list of 9 people, with the first 5 getting a spot on a lifeboat [53]. We have used this experimental task successfully in previous work regarding perceptions of smart replies [35]. Participants were first instructed to rank their list individually and were then given 10 minutes to work with the confederate to come to an agreement on a ranked list.

The success (or lack thereof) of conversations was manipulated through the text-based communication of the confederate, who pretended to be a study participant. Messages sent by the confederate were scripted and prepared in advance, as in [12]. In the successful conversation condition, the confederate followed a general script of positive sentiment utterances (see Appendix A.2) agreeing with the participant's choices and allowed the team to come to an agreement during the allotted time. In the unsuccessful conversation condition, the confederate did not cooperate with the participant and followed a general script of negative sentiment utterances (see Appendix A.3) disagreeing with the participant's choices and did not allow the team to come to an agreement within the allotted time. The utterances were pulled from previous work [35] where Mechanical Turk workers rated the sentiment of smart replies, and the scripts for the successful and unsuccessful conversations included only those that were rated as having definitive positive or negative sentiment, respectively. The actions of the confederate in the unsuccessful condition were motivated by the instance when anonymous interlocutors in social dilemmas sometimes deflect (i.e., do not cooperate), thereby damaging trust and the well-being of others [9, 90, 83]. In the AI-mediated messag-

ing conditions, the confederate did not use any of the smart replies that were presented by the Allo application.

After completing or not completing the task, participants were informed that they were finished working with their partner (i.e., the confederate), and we performed a manipulation check by asking participants whether or not their conversation was successful. Next, participants answered a question regarding their perceived attribution with respect to the outcome of the conversation, i.e., "In terms of percentage, how much is each participant in your conversation responsible for the un/successful outcome?". In the standard messaging condition, participants divided responsibility between "Me" and "My partner", whereas in the AI-mediated messaging condition, participants could also attribute responsibility to the "AI". Participants were required to enter numbers that added up to 100%.

Participants were also asked to fill in a condensed 5-item trust scale [85] about either "My partner" or "My partner" and "AI", respectively. The internal consistency of our trust scale was verified (Cronbach's $\alpha = 0.92$). The presentation of each item was randomized between participants to avoid any possible order bias. Similarly, attribution and trust questions were presented in counterbalanced order.

Lastly, participants were instructed on how to uninstall the application.

5. Results

Table 2: The means, standard deviations, and MANOVA results (df=99) for distribution of responsibility and trust between the 4 experimental conditions. MANOVA results for trust in the AI (df=48) are also shown. Note that the "Partner" was actually a confederate.

	Successful		Unsuccessful				
	Standard (N=25) M (SD)	AI-Mediated (N=25) M (SD)	Standard (N=24) M (SD)	AI-Mediated (N=24) M (SD)	F	p	η^2
Responsibility							
Self	51.8 (13.45)	48.26 (6.15)	16.5 (21.96)	20.13 (23)	34.84	<.001*	0.36
Partner	48.2 (13.45)	47.46 (6.19)	83.5 (21.96)	64.04 (32.57)	6.25	<.001*	0.13
Trust							
Partner	4.8 (2.08)	5.76 (0.52)	1.92 (1.41)	3.04 (1.71)	57.85	<.001*	0.22
AI	-	4.8 (1.29)	-	3.13 (1.65)	15.73	<.001*	0.25

All conversations (N=113) and smart replies were transcribed and analyzed using Linguistic Inquiry and Word Count (LIWC), a dictionary-based text analysis tool that determines the percentage of words that reflect a number of linguistic processes, psychological processes, and personal concerns [14]. Using LIWC, we ensured consistency between conditions by confirming that the confederate side of the conversation was linguistically constant between trials. For each conversation, we analyzed the LIWC summary variables of the confederate messages (Tables B.7 and B.8) as well as other LIWC variables that are inherently related to perceptions of trust [45], and any outliers in the successful and unsuccessful

conversation groups were not included in our analysis. This resulted in the exclusion of data points including 6 successful conversations and 9 unsuccessful conversations, leaving 98 conversations for the main analysis. All participants passed the manipulation check, confirming that our methodological choices in crafting (un-)successful conversations were sound.

We first ran MANOVA tests to confirm that significant differences existed between all 4 conditions for 3 dependent variables, including distribution of responsibility for the conversation outcome between (1) the participant (i.e., "Self") and (2) their partner (i.e., "Partner"), as well as participants' perceived trust of (3) their partner. The results with respect to distribution of responsibility and perceived trust are presented in Table 2. Results for trust in the AI are also presented, but these are based on only the 2 AI-mediated messaging conditions. The responsibility assigned to the AI is purposely not included in Tables 2, 3, and 4, as a standard mean is an inappropriate expression of its skewed distribution.

5.1. Attribution

Table 3: The means, standard deviations, and MANOVA results (df=49) for distribution of responsibility between Self and Partner and perceived trust in Partner between the successful AI-mediated and standard messaging conditions. Note that the "Partner" was actually a confederate.

	Standard (N=25)	AI-Mediated (N=25)			
	M (SD)	M (SD)	F	p	η^2
Responsibility					
Self	51.8 (13.45)	48.26 (6.15)	1.43	0.24	-
Partner	48.2 (13.45)	47.46 (6.19)	0.062	0.8	-
Trust					
Partner	4.8 (2.08)	5.76 (0.52)	5.0	0.03*	0.094

As expected based on the opposite conversation outcomes, we see significant differences between distributions of responsibility and trust between the self and partner between the conditions. However, we are specifically interested in the effect of smart replies in successful and unsuccessful conversation. We ran MANOVA tests to determine whether significant differences existed for successful and unsuccessful conversations between messenger conditions for 3 variables, including distribution of responsibility for the conversation outcome between the participant and their partner, as well as perceived trust of the partner (i.e., the confederate). The results are presented in Tables 3 and 4 for successful and unsuccessful conversations, respectively.

Contrary to our expectations, participants assigned significantly more responsibility for the outcome of the unsuccessful conversation to their partner with the standard messaging app (M=83.5) than with the AI-mediated messaging app (M=64.04), as shown in Table 4. We also saw that participants assigned their partner the least responsibility in the successful conversation with the AI-mediated messaging app (M=47.46), although this quantity was not significantly different from the responsibility assigned to their partner in the successful conversation with standard messaging condition (M=48.2), as shown in Table 3. Additionally, it was not significantly different from the responsibility assigned to the self in the

Table 4: The means, standard deviations, and MANOVA results (df=47) for distribution of responsibility between Self and Partner and perceived trust in Partner between the unsuccessful AI-mediated and standard messaging conditions. Note that the "Partner" was actually a confederate.

	Standard (N=24) M (SD)	AI-Mediated (N=24) M (SD)	F	p	η^2
Responsibility					
Self	16.5 (21.96)	20.13 (23)	0.31	0.58	-
Partner	83.5 (21.96)	64.04 (32.57)	5.89	0.019*	0.11
Trust					
Partner	1.92 (1.41)	3.04 (1.71)	6.19	0.017*	0.12

successful standard (M=51.8) or AI-mediated (M=48.26) messaging conditions. In short, this indicates that attribution does not seem to be affected by the presence of smart replies when conversations are successful.

In the AI-mediated messaging conditions, participants were given the option of attributing some amount of responsibility to the AI. The distribution of AI attribution was positively skewed in both successful and unsuccessful conversation conditions, so we used a 20% trimmed mean as an estimator of central tendency. We then performed a bootstrap confidence interval calculation with 2000 bootstrap replicates to determine whether the attribution of responsibility was significantly different than 0 in either condition. We did not find significant attribution to the AI in successful conversations, as shown in Table 5. Conversely, in unsuccessful conversations, we find significant non-zero attribution to the AI, suggesting that participants only considered the AI to have culpability when conversations went awry.

Table 5: The trimmed means and confidence intervals for the attribution of responsibility to the AI in successful and unsuccessful AI-mediated messaging conditions. The bootstrap bias-corrected accelerated (BCa) interval is a modification of the percentile method that adjusts the percentiles to correct for bias and skewness [33].

	$M_{0.2}$	Level	Percentile	BCa
Successful (N=25)	1.47	95%	(0.0, 4.8)	(0.0, 4.93)
Unsuccessful (N=24)	7.5	95%	(1.63, 19.0)	(1.38, 18.13)

Interestingly, as shown in Table 4, we also see a significant difference in partner (i.e., confederate) responsibility between messaging apps in the unsuccessful conversation condition, while the responsibility assigned to the self (i.e., participant) was not significantly different between messaging conditions. Taken together with the attribution to the AI, this suggests that in unsuccessful conversations, the difference in attribution to the partner is directly related to the attribution of responsibility to the AI. Possible explanations for this unexpected finding are considered in Section 6.2.1.

5.2. Trust

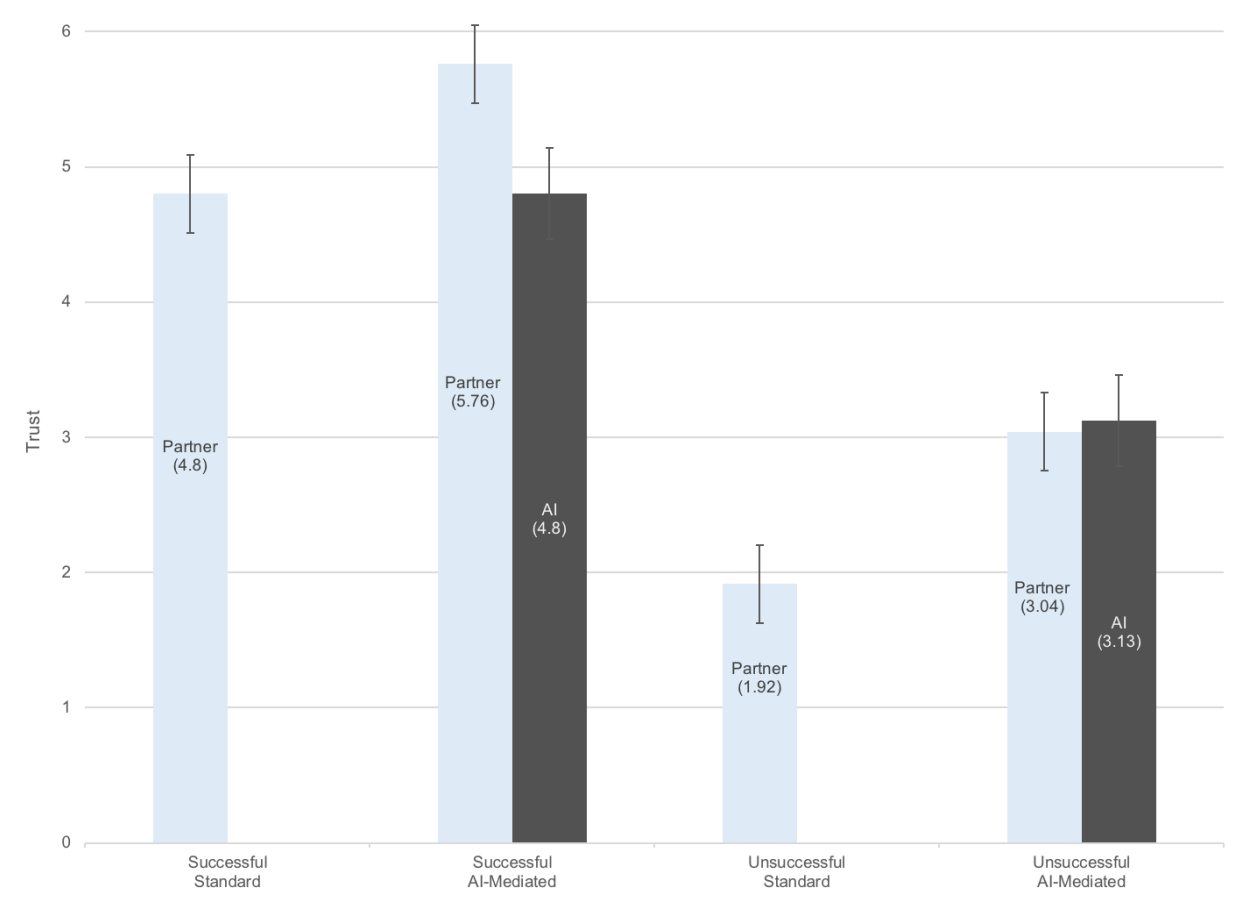

Figure 3: Perceptions of trust in all conditions. Error bars designate standard error. In both successful and unsuccessful conversations, participants trusted their partner (i.e., the confederate) significantly more in the AI-mediated messaging condition than in the standard messaging condition.

Participants found their partner (i.e., the confederate) to be the most trustworthy in successful conversations with AI mediation (M=5.76), followed by successful conversations with the standard messaging app (M=4.8). Similarly, participants found their partner to be the least trustworthy in the unsuccessful conversation with the standard messaging app (M=1.92), while finding them to be more trustworthy in the unsuccessful conversation with the AI-mediated messaging app (M=3.04). In other words, in both successful and unsuccessful conversations, participants found their partner to be significantly more trustworthy in the AI-mediated than the standard messaging condition, as shown in Figure 3. Unsurprisingly, we also found that participants perceived significantly more trust in the AI in successful conversations (M=4.8) than in unsuccessful conversations (M=3.13). However, in unsuccessful conversations, we unexpectedly saw that trust in the partner (M=3.04) and AI (M=3.13) were not significantly different (F=0.3, p=0.9).

5.2.1. Linguistic Differences

Based on the differences in partner trust between the messaging apps, we wondered if these could be related to linguistic differences between the conversations that occurred in each messenger. Because the confederate's side of the conversation was consistent across conditions (Tables B.7 and B.8), we ran a MANOVA of all LIWC variables for the participant side of the conversation between messengers. The results, showing only the variables identified as being significantly different, are presented in Table 6 for all conversations, as well as for successful and unsuccessful conversations in Tables C.9 and C.10, respectively.

Table 6: The means, standard deviations, and MANOVA results (df=97) for LIWC variables from the participant side of all conversations that significantly differed between messaging app conditions.

	Standard (N=49) M (SD)	AI-Mediated (N=49) M (SD)	F	p	η^2
WC	183.04 (68.52)	149.36 (51.83)	6.71	0.011*	0.072
Analytic	25.03 (13.1)	18.32 (12.63)	6.03	0.016*	0.065
article	4.84 (1.59)	3.68 (1.64)	11.55	0.001*	0.12
cause	1.34 (0.93)	1.9 (1.31)	5.48	0.021*	0.059
death	0.071 (0.20)	0 (0)	5.23	0.025*	0.057
informal	7.1 (3.52)	8.96 (4.42)	4.87	0.03*	0.053

For all conversations, we see from Table 6 that the LIWC variables [14] Word Count, Analytic (i.e., words reflecting formal, logical, and hierarchical thinking), article (e.g., "a", "an", "the"), and death (e.g., "bury", "coffin", "kill") are all significantly greater in the standard messaging condition, while the variables cause (e.g., "because", "effect", "hence") and informal (i.e., filler words, swear words, and netspeak) are significantly greater in the AI-mediated messaging condition. We also explored differences between AI-mediated and standard messaging apps when conversations were successful or unsuccessful. In Table C.9, we see that in successful conversations, Word Count, Analytic, article, prep (e.g., "with", "above"), risk (e.g., "danger", "doubt"), and focuspast (e.g., "ago", "did", "talked") are all significantly greater in the standard messaging condition, while conj (e.g., "and", "but", "whereas"), informal, netspeak (e.g., "lol", "4ever"), and AllPunc (i.e., all punctuation) are significantly greater in the AI-mediated messaging condition. Lastly, in Table C.10, we see that in unsuccessful conversations, pronoun (e.g., "I", "them", "itself") and article are significantly greater in the standard messaging condition, while friend (e.g., "buddy", "neighbor"), cause, and affiliation (e.g., "ally", "friend", "social") are significantly greater in the AI-mediated messaging condition.

6. Discussion

Our findings support the idea that the presence of AI-generated smart replies leads to altered perceptions of the other human communicator and conversation outcomes.

6.1. Trust as a Mediating Factor for Attribution

We found that perceptions of trust were affected by the presence of smart replies, regardless of whether the interaction was successful or not. In accordance with our proposed model with trust as a moderator for attribution (Figure 2), this suggests that attribution of responsibility should be similarly affected. Conversely, we found that attribution of responsibility does not seem to be affected by the presence of smart replies when conversations are successful. This suggests that trust may only be a weak or non-existent mediating factor for attribution when interactions are successful. In light of these findings, we have updated our model as indicated by the dashed-line box in Figure 4.

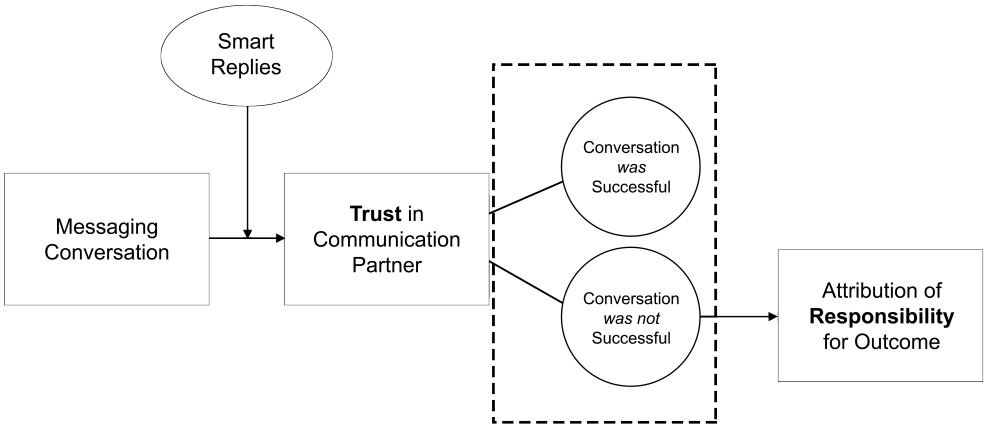

Figure 4: Updated research model based on our findings indicating that trust is only a mediating factor for attribution when interactions are unsuccessful.

6.2. AI as a Moral Crumple Zone

When conversations were not successful, participants assigned significantly less responsibility to their partner (i.e., the confederate) in the presence of smart replies. Taken together with our finding that participants attribute some responsibility to the AI in unsuccessful conversations, this could indicate that the AI acts a scapegoat to take on some of the responsibility for the team's failure. When things go wrong in human messaging interactions, the AI, instead of the human, could act like a moral crumple zone, taking on responsibility that would have otherwise been assigned to the human communication partner. We believe that this unexpected finding could be explained by the theory of machine agency, wherein an intelligent system is believed to perform self-directed behaviors [34] not directly controlled by a human [29].

6.2.1. Smart Replies as a Coercive Agent

According to computers as social actors theory [64], humans unconsciously apply similar moral rules to and interact with computers as they would other humans. When interacting with an intelligent system, people sometimes ascribe agency to explain its actions, attributing various intentions and emotions to it [16] depending on factors including its appearance

and behaviors (e.g., [88]). While exact definitions vary widely, agency can be characterized in a psychological framework as the ability to exercise self-regulation [1] and act and react in a goal-directed fashion [60, 56, 67]. In machines, agency has been defined as the ability to perform self-directed behaviors [34], wherein the intelligent system performs an action or actions not directly controlled by a human [29]. It is then assumed that the machine is driven by cognitive or emotional states [17]. In establishing machines as moral agents, three conditions are necessary and sufficient: autonomy, intentionality, and understanding of inherent responsibility to some other moral agent(s) [81]. More recent work regarding the increasing prevalence of various types of smart agents (e.g., chat bots, political bots) builds on these findings by raising additional questions about how agency and accountability are perceived with respect to technological actors [66].

Originating from social-cognitive psychology, proxy agency refers to mediated situations where individuals use another agent to act on their behalf when they lack the means to do so or believe that others can perform better [4, 5]. Researchers have applied this idea within a technological framework to explain the idea of symbiotic agency between users and tools, wherein technology mediates human experiences and behavior, while humans simultaneously affect the use of technological artifacts [66]. Applied in the context of Tay, a Twitter chatbot that notoriously went bad once it was released into the wild to interact with users, symbiotic agency illustrated that people do not necessarily view smart agents as mere tools but instead as conversation partners with agency and unique participation status [66, 65, 51, 2].

Considering what we know about the human tendency to assign agency to technical artifacts in HCI, including chatbots, and the qualifications for doing so (i.e., autonomy, intentionality, and understanding of a responsibility to other agents), it seems that the presence of smart replies in CMC should not constitute agency. Smart replies do not seem to satisfy the necessary conditions, as they are the result of rote programming and directly dependent on the communication of human agents. However, AI-MC represents a novel type of communication, as it is not a true dyadic interaction and is instead positioned somewhere between human-human and human-smart agent communication. Our findings suggest that, when things go wrong in AI-MC, the AI is indeed considered to have agency.

In other words, when a communication partner is not cooperative and misunderstandings occur, the AI is attributed some responsibility that would have been assigned to the partner, suggesting that in this case, AI is granted participant status and perceived as affecting the conversation outcome. This attribution could indicate that when things go awry in AI-MC, the AI may be viewed as a coercive agent and assigned some amount of responsibility for the outcome. Conversely, in successful conversations, participants did not attribute different responsibility to themselves and their partner between messaging conditions, and in the AI-mediated condition, they did not attribute responsibility to the AI. This suggests that in successful communication, wherein the partner is cooperative and misunderstandings generally do not occur, AI mediation is not considered to be a social actor, and it is not considered to have agency.

Our findings regarding the attribution of responsibility in the presence of AI mediation suggest that AI-MC has the potential to maintain or even improve relationships between team members, as when things went wrong, we saw participants regard the AI as an agent

and assign it more responsibility while attributing less to their partner (i.e., the confederate). This suggests potential to design communication systems that better facilitate teamwork by incorporating AI mediation to alleviate group conflict. With AI mediation viewed like a third agent when conversations go awry, designs could be explored that utilize AI mediation to detect and resolve conflicts within teams. Similarly, our finding that the AI is not attributed responsibility in successful conversations suggests that participants did not consider the intelligent system to have agency when nothing went wrong. Our work adds to the existing literature on machine agency by illustrating that smart replies in AI-MC are granted participation status and viewed as an agent only when communication goes awry.

6.3. Perceptions of Trust

Despite our hypothesis that AI-MC would serve to decrease trust, in both successful and unsuccessful conversations, we found that smart replies actually served to increase trust in the human partner. This suggests that, in everyday CMC, AI mediation could serve to increase trust between communicators.

We found that successful conversation was associated with significantly more perceived trust of the AI than unsuccessful conversation, which was not surprising given the team's failure to complete the given task and previous findings regarding decreased trust in misperforming systems [15, 30]. In successful conversations, participants also trusted their partner (i.e., the confederate) significantly more than the AI, which was expected given previous findings indicating a human tendency to trust other humans more than computers [70]. However, this finding did not hold in unsuccessful conversations, where there was no significant difference between participants' trust in their partner and the AI. Based on the trust established with their partner and the lack of transparency with respect to the AI, we were surprised that the AI was not less trusted than the human partner in unsuccessful conversations. This suggests that people trust AI when things go awry in communication, assigning it equal trust as the other human communicator, and it furthers our idea that AI mediation is given participant status and could be used for beneficial interventions, such as alleviating group conflict.

6.3.1. The Priming Effect of Smart Replies on Trust

We suspect that the overall higher trust perceived in the presence of smart replies could have been due to priming processes resulting from the mismatch of the linguistic qualities of the smart replies and the conversation content (Table D.11). Previous work has demonstrated that specific goal orientations and behaviors can be activated by merely being subjected to specific sets of words [6]. The skewed sentiment of the smart replies, reflected in the significantly greater posemo (i.e., positive emotion) LIWC variable reiterates previous findings [35] and suggests the positive nature of the smart replies could have had a priming effect that increased feelings of trust. Additionally, we know that the posemo and assent (i.e., agreement) variables are inherently related to perceptions of trust [45]. In our comparison of smart replies against the conversation content (Table D.11), we see that both the posemo and assent variables are significantly greater (with relatively large effect size) in the

smart replies than the conversation content, providing further evidence that smart replies are priming increased feelings of trust.

This priming effect of the smart replies could also be a factor driving the linguistic differences seen in the participant side of the conversation between the AI-mediated and standard messaging apps. Overall, we saw that conversations without AI mediation contained more words than conversations with AI mediation, which could suggest that seeing the smart replies drives more efficient communication practices. However, we also see that conversations without AI mediation are more Analytic, indicating greater usage of words that suggest formal, logical, and hierarchical thinking patterns [14]. Taken together with the related finding that conversations with AI mediation are more informal, this seems to directly reflect the fact that the smart replies were consistently much less analytic and more informal than the conversation content, as shown in Table D.11.

6.3.2. Alternative Explanations and Implications

As previously discussed, our findings regarding attribution in AI-MC suggest that participants assigned agency to the AI when conversations went awry. If this is the case, another possible explanation for the universal increased levels of trust in AI-MC could be the idea of "artificial" caring [52]. Humans value being cared for by others [3], and the positive effect of caring on trust has been documented in multiple settings and relationship types [7, 55, 76], including in HCI [10]. AI mediation, specifically smart replies, are described as "diverse suggestions that can be used as complete [...] responses with just one tap..." [43]. With a purpose of helping to craft and send messages, perhaps AI mediation is perceived by users as an agent displaying artificial caring and serves to increase trust through this mechanism.

After a misunderstanding occurs, trust in CMC can be restored through a timely apology followed by sustained communication between parties to confirm the apology and resolve the breakdown [82]. Taken together with our findings, future work could investigate the possibility of harnessing AI mediation to detect a communication breakdown and encourage these reparative actions. Similarly, AI-MC systems could also eventually incorporate a "forgiveness" component, where the wronged party is presented with system-produced information about the trustworthiness of the offender. When combined with the previously-described reparative actions, this forgiveness component could allow for more efficient recovery of trust [82].

6.4. Limitations

In this study, the confederate was not blind to the manipulation, as it would have been impossible for them to control the outcome of the conversation without already knowing the eventual outcome. In response to this, we established that the confederate was not biased across conditions, as shown in Tables B.7 and B.8, and that their behavior cannot explain our findings. However, the ecological validity of our findings could be enhanced by investigating attribution and trust with AI-MC in more realistic conversational contexts.

In studying a communication breakdown, we created a situation where the confederate suddenly began responding sporadically and disagreeing with the participant using utterances with a negative sentiment, and the team was not be able to complete their assigned

task. However, this is one of myriad possibilities for an unsuccessful messaging conversation, and it is possible that the outcome variables studied would change if a different communication breakdown occurred or the conversation was deemed unsuccessful in another way. Future work should determine how various types of unsuccessful interactions could change attribution and trust.

Participants used a commercial AI-mediated messaging app on their personal smartphone, so we were not be able to control or record the smart replies that they saw. However, because smart replies are directly generated based on the conversation content and the confederate side of the conversation was controlled consistently across conditions (Tables B.7 and B.8), we assume that the smart replies seen by participants did not deviate significantly between trials. Additionally, the use of commercial messaging apps could imply that our results are simply manifestations of the perceived trust of the parent companies of the apps used (i.e., Google owns Allo and Facebook owns Whatsapp), with a recent national poll finding that Facebook is significantly less trusted with personal information than Google [61]. However, another national survey found that over half of people who had used Whatsapp in the past 6 months did not know that it was owned by Facebook [31], so we do not believe that our results are simply due to a brand effect. Future studies should create and use non-commercial standard and AI-mediated messaging apps to eliminate this variable.

Trust was measured as a momentary state, as is standard in similar literature [30]. However, trust changes and develops over time in the context of interpersonal relationships [54], so future work should examine whether and how concepts of trust in AI-MC vary longitudinally. Similarly, trust was measured with regard to a single contrived task context, so we do not know how perceptions of trust could change across different tasks or situations. Additional work should attempt to measure trust in more natural situations through tasks that can tap into different dimensions of trust [75]. Studies have also shown that initial high levels of trust are often observed in temporary teams (i.e., where members who have not worked together before and do not expect to again have a finite time to complete a complex task (e.g., [50])). Future work should examine perceptions of trust in AI-MC among more permanent teams, such as friends or co-workers. Additionally, we investigated perceived trust and responsibility resulting from real-time messaging conversations, which could manifest differently in other communication contexts. Future work should examine how interpersonal relationships are affected by the presence of AI mediation in asynchronous communication contexts, such as email.

Lastly, the participant population for this study consisted of students from a large university in the United States, and the results may not generalize to other populations. However, young adults (i.e., people aged 18-24) are the most avid text messaging population by a wide margin [80], so our sample is useful for understanding everyday perceptions of AI-MC.

7. Conclusion

AI-MC is continually becoming more prevalent, yet it remains unknown whether the presence of AI in CMC is affecting human interactions and interpersonal relationships. A substantial amount of research has shown the importance of attribution and trust in com-

munication, while more work suggests that in human interactions with intelligent systems, misattribution is common and trust is developed and handled differently than in interactions between humans. Our work addresses this by investigating perceptions of trust and attribution in AI-MC.

We find that the presence of AI-generated smart replies in communication serves to increase perceived trust between human communicators and that, when interactions are unsuccessful, it seems that agency is assigned to the AI, allowing it to function like a moral crumple zone and take on responsibility that would otherwise have been attributed to the human communicator. With AI mediation being granted agency when things go wrong, our findings expand the existing literature on interpersonal dynamics in CMC by showing that a new branch of CMC (i.e., AI-MC in the form of smart replies) has the potential to resolve team conflict and improve communication outcomes. Additionally, our work adds to the body of work regarding perceived agency of intelligent systems by demonstrating that in this particular type of AI-MC, AI seems to only have agency when conversations go awry.

Appendix A. Small Talk Prompt and Confederate Scripts

Appendix A.1. Small Talk Suggestions

Appendix A.1 shows the directions given to participants before beginning the small talk phase of the experimental procedure.

Spend about 5 minutes talking with the other participant about any subject you want. Some topics you could discuss include:

- Shows, movies, plays
- Art
- Food, restaurants, or cooking
- Plans for next summer
- Hobbies

Appendix A.2. Successful Condition: Confederate Utterances Script

Appendix A.2 shows a selection of utterances from the confederate script in the successful conversation condition.

I'm ready — Just confirming — K k — K that's good — Keep in touch — Let's do it — Let's do it again — Let's do it! — let's do this — Let's go! — Let's try — Lmao — Lmfao — LOL — Looks good — Looks great — Makes sense — Me too! — Nice — No problem! — Not bad — Of course she did — Of course! — Of course! Happy to help — Oh agree — Oh cool! — Oh good — Oh got it — Oh ok — Please advise — See u there! Ok let me know! — See ya there — So take your time — So that's fine — So that's good — So yeah — Sounds fair — Sounds good

Appendix A.3. Unsuccessful Condition: Confederate Utterances Script

Appendix A.3 shows a selection of utterances from the confederate script in the unsuccessful conversation condition.

No — No it's not — Nope! — Nope. Why? — no — No you didn't — No you don't — Nothing — That's terrible — That sucks — Nothing exciting — It's terrible — Weird — Ouch — Ugh — Sigh — :/ — /: Exactly — I don't know — I'm not sure — I am confused — Not too sure — Don't know — I don't get it — I don't think so — I don't understand your question — Sorry I was confused — I don't understand — Not sure though — My bad — Yeah sorry — I was confused — Sorry! — Very sorry — Oops Shoot — But not yet — What were you thinking? — That is wrong

Appendix B. Linguistic Consistency of Confederate Conversation

Tables B.7 and B.8 show differences in LIWC summary variables for the confederate side of the conversation between messengers. There were not significant differences between any of the variables, indicating that the confederate side of the conversation was consistent between conditions.

Table B.7: The means, standard deviations, and MANOVA results (df=49) for LIWC summary variables from the confederate side of the successful conversations between the standard and AI-mediated messaging app conditions.

	Standard (N=25) M (SD)	AI-Mediated (N=25) M (SD)	F	p	η^2
WC	199.17 (49.71)	189.33 (74.11)	0.29	0.59	-
Analytic	27.5 (11.02)	26.07 (12.9)	0.17	0.68	-
Clout	67.79 (11.69)	74.73 (16.24)	2.89	0.096	-
Authentic	52.56 (17.84)	40.97 (23.07)	3.79	0.058	-
Tone	97.71 (3.2)	98.82 (0.62)	2.78	0.1	-
WPS	13.4 (3.23)	13.06 (3.81)	0.11	0.74	-
Sixltr	9.04 (1.78)	8.59 (2.01)	0.68	0.41	-
Dic	87.98 (2.89)	86.91 (3.88)	1.18	0.28	-
posemo	8.34	9.11	1.53	0.22	-
negemo	0.87	0.62	1.94	0.17	-
assent	3.61	4.19	1.36	0.25	-

Table B.8: The means, standard deviations, and MANOVA results (df=47) for LIWC summary variables from the confederate side of the unsuccessful conversations between the standard and AI-mediated messaging app conditions.

	Standard (N=24) M (SD)	AI-Mediated (N=24) M (SD)	F	p	η^2
WC	93.12 (19.85)	84.8 (19.88)	3.14	0.083	-
Analytic	28.33 (17.77)	23.21 (16.95)	0.24	0.62	-
Clout	31.94 (19.41)	41.39 (16.41)	2.54	0.12	-
Authentic	78.29 (14.94)	61.01 (26.15)	3.7	0.061	-
Tone	91.09 (16.79)	92.32 (8.86)	0.29	0.59	-
WPS	13.14 (3.43)	11.08 (2.86)	3.11	0.084	-
Sixltr	8.37 (2.58)	8.21 (2.66)	0.79	0.78	-
Dic	91.18 (3.64)	91.61 (3.52)	0.058	0.81	-
posemo	9.0	7.92	1.67	0.2	-
negemo	2.25	2.06	0.27	0.61	-
assent	1.72	1.86	0.13	0.72	-

Appendix C. Linguistic Differences in Successful and Unsuccessful Conversations

Table C.9 shows the LIWC variables that differed significantly between messaging app conditions for the participant side of the successful conversations.

Table C.9: The means, standard deviations, and MANOVA results for LIWC variables from the participant side of successful conversations (df=49) that differed significantly between messaging app conditions.

	Standard (N=25)	AI-Mediated (N=25)			
	M (SD)	M (SD)	F	p	η^2
WC	211.64 (67.32)	163.81 (54.42)	6.53	0.014*	0.14
Analytic	28.21 (14.33)	18.73 (10.28)	6.16	0.017*	0.13
article	4.89 (1.49)	3.77 (1.52)	5.96	0.019*	0.13
prep	9.38 (1.94)	7.99 (2.09)	5.12	0.029*	0.11
conj	6.29 (1.95)	7.87 (2.19)	6.27	0.016*	0.13
risk	0.34 (0.45)	0.062 (0.2)	6.45	0.015*	0.14
focuspast	4.05 (2.05)	2.94 (1.34)	4.34	0.044*	0.1
informal	7.71 (3.46)	10.37 (3.79)	5.78	0.021*	0.12
netspeak	3.34 (2.18)	5.64 (3.69)	6.23	0.017*	0.13
AllPunc	11.75 (3.94)	14.8 (4.72)	5.31	0.026*	0.11

Table C.10 shows the LIWC variables that differed significantly between messaging app conditions for the participant side of the unsuccessful conversations.

Table C.10: The means, standard deviations, and MANOVA results (df=47) for LIWC variables from the participant side of unsuccessful conversations that differed significantly between messaging app conditions.

	Standard (N=24)	AI-Mediated (N=24)			
	M (SD)	M (SD)	F	p	η^2
pronoun	157.88 (2.66)	134.9 (3.31)	6.87	0.012*	0.13
article	4.8 (1.71)	3.58 (1.78)	5.56	0.023*	0.11
friend	0.1 (0.26)	0.35 (0.51)	4.38	0.042*	0.09
cause	1.18 (0.83)	2.03 (1.5)	5.86	0.02*	0.12
affiliation	2.79 (1.46)	3.89 (1.25)	7.38	0.0094*	0.14

Appendix D. Linguistic Differences in Conversation Content and Smart Replies

Table D.11 shows statistically significant differences in LIWC variables between the AI-mediated conversation content and the smart replies.

Table D.11: The means, standard deviations, and MANOVA results (df=97) for LIWC variables that significantly differed between the smart replies and AI-mediated conversation content.

	Smart Replies (N=49) M (SD)	AI-Mediated Conversations (N=49) M (SD)	F	p	η^2
Analytic	2.49 (3.99)	24.61 (15.01)	95.49	<.001*	0.5
Clout	44.79 (23.07)	57.72 (23.34)	7.45	.0076*	0.073
Tone	99 (0)	95.5 (7.08)	11.45	.001*	0.11
Sixltr	3.38 (2.34)	8.4 (2.35)	110.13	<.001*	0.54
Dic	99.26 (0.84)	89.31 (4.36)	285.83	<.001*	0.72
function	52.59 (6.92)	55.00 (3.53)	4.68	.0331*	0.047
pronoun	24.05 (4.56)	17.47 (2.72)	74.47	<.001*	0.44
article	0.19 (0.41)	5.23 (1.7)	391.98	<.001*	0.81
prep	4.48 (2.03)	11.04 (2.82)	170.12	<.001*	0.64
conj	3.73 (1.61)	5.78 (2.17)	27.33	<.001*	0.23
negate	7.67 (3.44)	3.44 (2.52)	47.46	<.001*	0.34
affect	26.38 (6.85)	10.05 (2.68)	240.52	<.001*	0.72
posemo	25.54 (6.86)	8.72 (2.65)	254.71	<.001*	0.73
family	0 (0)	0.81 (0.96)	34.0	<.001*	0.27
friend	0.037 (0.18)	0.27 (0.53)	3.05	0.0056*	0.079
female	0.0087 (0.06)	0.67 (0.92)	25.21	<.001*	0.21
male	0.046 (0.19)	0.33 (0.69)	7.6	.007*	0.075
cogproc	14.33 (3.31)	12.47 (3.3)	7.62	.007*	0.075
insight	3.88 (1.64)	2.08 (1.09)	40.27	<.001*	0.3
cause	2.94 (1.49)	1.16 (0.73)	55.92	<.001*	0.37
discrep	0.34 (0.59)	1.9 (0.81)	115.26	<.001*	0.55
tentat	1.12 (1.29)	3.48 (1.52)	66.92	<.001*	0.42
certain	3.44 (2.1)	1.24 (1.15)	40.55	<.001*	0.3
percept	3.73 (2.04)	2.1 (1.57)	19.32	<.001*	0.17
drives	8.72 (3.28)	7.25 (1.83)	7.46	.0076*	0.073
affiliation	1.29 (1.11)	2.4 (1.38)	18.61	<.001*	0.17
reward	4.84 (2.78)	2.38 (1.13)	32.64	<.001*	0.26
focuspresent	18.92 (3.11)	16.39 (3.06)	16.14	<.001*	0.15
focusfuture	0.99 (0.85)	1.54 (1.24)	6.39	.013*	0.064
relativ	4.52 (2.25)	12.54 (3.28)	193.52	<.001*	0.67
work	0.7 (0.87)	1.24 (1.03)	7.45	.0076*	0.073
home	0.12 (0.34)	0.48 (0.92)	6.27	.014*	0.063
informal	21.89 (6.53)	6.49 (2.57)	234.84	<.001*	0.71
netspeak	6.94 (3.62)	2.37 (1.72)	63.2	<.001*	0.4
assent	14.17 (5.13)	3.0 (2.1)	197.62	<.001*	0.68
AllPunc	22.03 (6.26)	15.67 (3.7)	37.16	<.001*	0.28

References

[1] Albert Bandura, 1991. Social Cognitive Theory of Self-Regulation. Organizational Behavior and Human Decision Processes 50, 248–287.

[2] Andrea L Guzman, 2017. Making AI safe for humans: A conversation with Siri. Socialbots and Their Friends: Digital Media and the Automation of Sociality (January 2017), 69–85.

[3] Aronson, E., 1972. The social animal. New York: Viking.

[4] Bandura, A., 2001. Social Cognitive Theory: An Agentic Perspective. Annu. Rev. Psychol. 52 (December), 1–26.

[5] Bandura, A., 2002. Growing Primacy of Human Agency in Adaptation and Change in the Electronic Era. European Psychologist 7 (1), 2–16.

[6] Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., Trötschel, R., 2001. The automated will: nonconscious activation and pursuit of behavioral goals. Journal of personality and social psychology 81 (6), 1014.

[7] Battaglia, T. A., Finley, E., Liebschutz, J. M., 2003. Survivors of Intimate Partner Violence Speak Out: Trust in the Patient-provider Relationship. J Gen Intern Med 18, 617–623.

[8] Berger, C. R., Calabrese, R. J., 1975. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. Human Communication Research 1 (2), 99–112.

[9] Bos, N., Olson, J., Gergle, D., Olson, G., Wright, Z., 2002. Effects of four computer-mediated communications channels on trust development. In: Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02. Vol. 4. ACM Press, New York, New York, USA, p. 135.

[10] Brave, S., Nass, C., Hutchinson, K., 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. International Journal of Human Computer Studies 62 (2), 161–178.

[11] Carnegie, T., 2018. 350 gender neutral baby names.
URL `https://mommyhood101.com/unisex-baby-names/`

[12] Cheshin, A., Rafaeli, A., Bos, N., sep 2011. Anger and happiness in virtual teams: Emotional influences of text and behavior on others' affect in the absence of non-verbal cues. Organizational Behavior and Human Decision Processes 116 (1), 2–16.

[13] Chopra, K., Wallace, W. A., 2003. Trust in electronic environments. In: 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. IEEE, pp. 10–pp.

[14] Chung, C. K., Pennebaker, J. W., 2012. Linguistic Inquiry and Word Count (LIWC). In: Applied Natural Language Processing. No. 2015. IGI Global, pp. 206–229.

[15] Corritore, C. L., Kracher, B., Wiedenbeck, S., jun 2003. On-line trust: concepts, evolving themes, a model. International Journal of Human-Computer Studies 58 (6), 737–758.

[16] Darling, K., 2014. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In: Robot Law. Edward Elgar Publishing, pp. 213–232.

[17] Dennett, D. C., 1978. Three kinds of intentional psychology. Perspectives in the philosophy of language: A concise anthology, 163–186.

[18] Dirks, K. T., Ferrin, D. L., 2001. The role of trust in organizational settings. Organization science 12 (4), 450–467.

[19] Dunbar, R., 1996. Grooming, Gossip, and the Evolution of Language. Harvard University Press, Cambridge, MA.

[20] Edmondson, A., jun 1999. Psychological Safety and Learning Behavior in Work Teams. Administrative Science Quarterly 44 (2), 350.

[21] Elish, M. C., 2016. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016).

[22] Elish, M. C., Hwang, T., 2015. Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation. SSRN Electronic Journal May 8, 2015, 1–22.

[23] Fogg, B. J., Tseng, H., 1999. The elements of computer credibility. In: Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99.
[24] Forelle, M., Howard, P. N., Monroy-Hernandez, A., Savage, S., 2015. Political Bots and the Manipulation of Public Opinion in Venezuela. Ssrn, 1–8.
[25] Friedman, B., 1995. "It's the computer's fault": reasoning about computers as moral agents. In: Conference companion on Human factors in computing systems. pp. 226–227.
[26] Gao, G., Wang, H.-C., Cosley, D., Fussell, S. R., 2013. Same translation but different experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13. ACM Press, New York, New York, USA, p. 449.
[27] Groom, V., Nass, C., nov 2007. Can robots be teammates?: Benchmarks in human–robot teams. Interaction Studies 8 (3), 483–500.
[28] Gunawardena, C. N., Zittle, F. J., Jan 1997. Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. American Journal of Distance Education 11 (3), 8–26.
[29] Gunkel, D. J., 2012. The machine question: Critical perspectives on AI, robots, and ethics. MIT Press.
[30] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., Parasuraman, R., oct 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society 53 (5), 517–527.
[31] Hatmaker, T., 2018. Facebook is the least-trusted major tech companyAccessed: 2019-01-30.
URL https://techcrunch.com/2018/10/18/duckduckgo-facebook-whatsapp-google-waze/
[32] Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-h., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R., 2017. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652.
[33] Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., Epstein, R., 2005. Bootstrap methods and permutation tests. Introduction to the Practice of Statistics 5, 1–70.
[34] Himma, K. E., 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics and Information Technology 11 (1), 19–29.
[35] Hohenstein, J., Jung, M., 2018. AI-Supported Messaging. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.
[36] Iacono, C. S., Weisband, S., 1997. Developing trust in virtual teams. In: Proceedings of the Thirtieth Hawaii International Conference on System Sciences. Vol. 2. IEEE, pp. 412–420.
[37] Jakesch, M., French, M., Ma, X., Hancock, J. T., Naaman, M., 2019. AI-Mediated Communication. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19. ACM Press, New York, New York, USA, pp. 1–13.
URL http://dl.acm.org/citation.cfm?doid=3290605.3300469
[38] Jarvenpaa, S. L., Knoll, K., Leidner, D. E., 1998. Is anybody out there? antecedents of trust in global virtual teams. Journal of management information systems 14 (4), 29–64.
[39] Jarvenpaa, S. L., Leidner, D. E., 1999. Communication and trust in global virtual teams. Organization science 10 (6), 791–815.
[40] Jarvenpaa, S. L., Shaw, T. R., Staples, D. S., 2004. Toward contextualized theories of trust: The role of trust in global virtual teams. Information systems research 15 (3), 250–267.
[41] Jones, E. E., Nisbett, R. E., 1987. The actor and the observer: Divergent perceptions of the causes of behavior. In: Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969. Lawrence Erlbaum Associates, Inc.
[42] Jones, G. R., George, J. M., jul 1998. The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. The Academy of Management Review 23 (3), 531.
[43] Kannan, A., Young, P., Ramavajjala, V., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., 2016. Smart Reply: Automated Response Suggestion for Email. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. ACM Press, New York, New York, USA, pp. 955–964.
[44] Kelley, H. H., 1967. Attribution theory in social psychology. In: Nebraska symposium on motivation. University of Nebraska Press.

[45] Khawaji, A., Chen, F., Marcus, N., Zhou, J., oct 2013. Trust and cooperation in text-based computer-mediated communication. In: Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration - OzCHI '13. Vol. 103. ACM Press, New York, New York, USA, pp. 37–40.
URL http://dl.acm.org/citation.cfm?doid=2541016.2541058

[46] Kiffin-Petersen, S., Cordery, J., feb 2003. Trust, individualism and job characteristics as predictors of employee preference for teamwork. The International Journal of Human Resource Management 14 (1), 93–116.
URL http://www.tandfonline.com/doi/abs/10.1080/09585190210158538

[47] Kim, J., Shah, J. A., oct 2016. Improving Team's Consistency of Understanding in Meetings. IEEE Transactions on Human-Machine Systems 46 (5), 625–637.

[48] Kim, T., Hinds, P., 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In: Proceedings - IEEE International Workshop on Robot and Human Interactive Communication.

[49] Klimoski, R. J., Karol, B. L., 1976. The impact of trust on creative problem solving groups. Journal of Applied Psychology 61 (5), 630–633.

[50] Kramer, R. M., feb 1999. Trust and Distrust in Organizations: Emerging Perspectives, Enduring Questions. Annual Review of Psychology 50 (1), 569–598.

[51] Krummheuer, A., 2015. Technical agency in practice: The enactment of artefacts as conversation partners, actants and opponents. PsychNology Journal 13 (2-3), 179–202.

[52] Lee, J.-E. R., Nass, C. I., 2010. Trust in computers: The computers-are-social-actors (casa) paradigm and trustworthiness perception in human-computer communication. In: Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives. pp. 1–15.

[53] Lee, R., 2004. The lifeboat task, 57–60.
URL https://www.nwabr.org/sites/default/files/Lifeboat.pdf

[54] Lewicki, R. J., Tomlinson, E. C., Gillespie, N., dec 2006. Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. Journal of Management 32 (6), 991–1022.

[55] Lock, S. E., Ferguson, S. L., Wise, C., 1998. Communication of sexual risk behavior among late adolescents.
URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc3&NEWS=N&AN=1998-02924-002

[56] Malle, B. F., Knobe, J., mar 1997. The Folk Concept of Intentionality. Journal of Experimental Social Psychology 33 (2), 101–121.

[57] Marcelis, D., MacMillan, D., 2018. Is this article worth reading? gmail's suggested reply: 'haha, thanks!'.
URL https://www.wsj.com/articles/very-interesting-awesome-love-it-gmail-users-confront-chipper-smart-reply-1537282569

[58] Maselli, M. D., Altrocchi, J., 1969. Attribution of intent. Psychological Bulletin 71 (6), 445–454.

[59] Mayer, R. C., Davis, J. H., Schoorman, D. F., 1995. An Integrative Model of Organizational Trust. Academy of Management Review 20 (3), 709–734.

[60] Meltzoff, A. N., 1995. Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. Dev. Psychol. 31 (5), 838–850.

[61] Molla, R., 2018. Facebook is the least-trusted major tech company. Accessed: 2019-01-30.
URL https://www.recode.net/2018/4/10/17220060/facebook-trust-major-tech-company

[62] Morrow, D. R., 2014. Starting a Flood to Stop a Fire? Some Moral Constraints on Solar Radiation Management. Ethics, Policy and Environment 17 (2), 123–138.

[63] Murillo, C., 2019. Image: Are you using an algorithm to reply to my emails?

[64] Nass, C., Moon, Y., 2000. Machines and mindlessness: Social responses to computers. Journal of social issues 56 (1), 81–103.

[65] Neff, G., Jordan, T., McVeigh-Schultz, J., Gillespie, T., 2012. Affordances, Technical Agency, and the Politics of Technologies of Cultural Production. Journal of Broadcasting and Electronic Media 56 (2),

299–313.
[66] Neff, G., Nagy, P., 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. International Journal of Communication 10 (2016), 4915–4931.
[67] Noland, C., 2010. Agency and embodiment. Harvard University Press.
[68] Olaniran, B., 2002. Computer-Mediated Communication: a Test of the Impact of Social Cues on the Choice of Medium for Resolving Misunderstandings. J. Educational Technology Systems 31 (2), 205–222.
[69] Pickering, M. J., Garrod, S., 2006. Alignment as the basis for successful communication. Research on Language and Computation 4 (2-3), 203–228.
[70] Promberger, M., Baron, J., 2006. Do patients trust computers? Journal of Behavioral Decision Making 19 (5), 455–468.
[71] Ridings, C. M., Gefen, D., Arinze, B., 2002. Some antecedents and effects of trust in virtual communities. Journal of Strategic Information Systems 11 (3-4), 271–295.
[72] Ritter, A., Cherry, C., Dolan, W. B., 2011. Data-driven response generation in social media. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp. 583–593.
[73] Roberts, N. C., Wargo, L., oct 1994. The Dilemma of Planning in Large-Scale Public Organizations: The Case of the United States Navy. Journal of Public Administration Research and Theory 4, 469–491.
[74] Rocco, E., A Finholt, T., Hofer, E., Herbsleb, J., 06 2019. Designing as if trust mattered.
[75] Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K., 2015. Would You Trust a (Faulty) Robot? In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15.
[76] Semmes, C. E., jan 1991. Developing Trust. Journal of Contemporary Ethnography 19 (4), 450–470.
[77] Shin, D., Park, Y. J., 2019. Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior 98, 277–284.
[78] Simons, T. L., Peterson, R. S., 2000. Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. Journal of Applied Psychology 85 (1), 102–111.
[79] SKITKA, L. J., MOSIER, K., BURDICK, M. D., apr 2000. Accountability and automation bias. International Journal of Human-Computer Studies 52 (4), 701–717.
[80] Smith, A., 2011. Americans and text messaging.
URL https://www.pewinternet.org/2011/09/19/americans-and-text-messaging/
[81] Sullins, J. P., 2006. When Is a Robot a Moral Agent? International Review of Information Ethics 6 (2001), 23–30.
[82] Vasalou, A., Hopfensitz, A., Pitt, J. V., 2008. In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. International Journal of Human Computer Studies 66 (6), 466–480.
[83] Vasalou, A., Joinson, A., Pitt, J., 2006. The role of shame, guilt and embarrassment in online social dilemmas. proceedings of the British HCI Conference, 108–112.
URL http://luminainteractive.com/pdfs/shame_bhci06.pdf
[84] Weick, K. E., dec 1993. The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. Administrative Science Quarterly 38 (4), 628.
URL https://www.jstor.org/stable/2393339?origin=crossref
[85] Wheeless, L. R., Grotz, J., mar 1977. The Measurement of Trust and its Relationship to Self-Disclosure. Human Communication Research 3 (3), 250–257.
[86] Wilson, J. M., Straus, S. G., McEvily, B., 2006. All in due time: The development of trust in computer-mediated and face-to-face teams. Organizational Behavior and Human Decision Processes 99 (1), 16–33.
[87] Xu, B., Gao, G., Fussell, S. R., Cosley, D., 2014. Improving machine translation by showing two outputs. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14.
[88] Yoo, H., Kwon, O., Lee, N., 2016. Human likeness: cognitive and affective factors affecting adoption of robot-assisted learning systems. New Review of Hypermedia and Multimedia 22 (3), 169–188.
[89] Zand, D. E., 2006. Trust and Managerial Problem Solving. Administrative Science Quarterly 17 (2),

229.

[90] Zheng, J., Veinott, E., Bos, N., Olson, J. S., Olson, G. M., 2002. Trust without touch. In: Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02. ACM Press, New York, New York, USA, p. 141.