AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task

HANNAH MIECZKOWSKI and JEFFREY T. HANCOCK, Stanford University, USA
MOR NAAMAN, Cornell Tech, Cornell University, USA
MALTE JUNG and JESS HOHENSTEIN, Cornell University, USA

AI-Mediated Communication (AI-MC) is interpersonal communication that involves an artificially intelligent system that can modify, augment, or even generate content to achieve communicative and relational goals. AI-MC is increasingly involved in human communication and has the potential to impact core aspects of human communication, such as language production, interpersonal perception and task performance. Through a between-subjects experimental design we examine how these processes are influenced when integrating AI-generated language in the form of suggested text responses (Google's smart replies) into a text-based referential communication task. Our study replicates and extends the impacts of a positivity bias in AI-generated language and introduces the adjacency pair framework into the study of AI-MC. We also find preliminary yet mixed evidence to suggest that AI-generated language has the potential to undermine some dimensions of interpersonal perception, such as social attraction. This study contributes important concepts for future work in AI-MC and offers findings with implications for the design of AI systems in human-to-human communication.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: AI-mediated communication; linguistic alignment; impression formation; tasks; sentiment

ACM Reference Format:
Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 17 (April 2021), 14 pages. https://doi.org/10.1145/3449091

1 INTRODUCTION

Over the past few years, artificial intelligence, defined in this context as "computational systems that involve algorithms, machine learning methods, natural language processing, and other techniques that operate on behalf of an individual" [15, p.90], has played an increasingly large role in how people communicate with each other. Although predictive text technology has been available for several decades [10], AI-infused systems are already becoming more prominent in a host of text-based interpersonal domains. For example, using Google's "smart replies," a message sender can select one of several responses produced by AI, purportedly based on the content of the message

Authors' addresses: Hannah Mieczkowski, hnmiecz@stanford.edu; Jeffrey T. Hancock, hancockj@stanford.edu, Stanford University, 450 Jane Stanford Way, Stanford, California, USA, 94305; Mor Naaman, mor.naaman@cornell.edu, Cornell Tech, Cornell University, 2 West Loop Road, New York, New York, USA; Malte Jung, mfj28@cornell.edu; Jess Hohenstein, jch378@cornell.edu, Cornell University, 236 Gates Hall, Ithaca, New York, USA.

2573-0142/2021/4-ART17 $15.00
https://doi.org/10.1145/3449091

received. The uptake of these systems has been widespread, with billions of smart replies used daily in human communication [5].

In comparison to previous technology that assumed a more passive, mediating role, AI is an active and dynamic entity that has the potential to alter the norms and dynamics of human communication. Previous research refers to the use of these technologies in human interaction as AI-Mediated Communication (AI-MC), in which messages are not simply transmitted by technology, but modified, augmented, or even generated by a computational agent to achieve communication goals [15]. In this short paper, we examine how AI-MC technology in the form of AI-generated language (Google's smart replies) are integrated into a text-based communication task, and how this form of AI-MC may influence 1) the language patterns used by conversation partners in both AI-generated and human-generated language, 2) perceptions of the conversation partner and 3) the performance of the conversation partners in a text-based referential communication task. The primary contributions of the short paper are: a replication and extension of the impacts of a positivity bias in AI-generated language, the introduction of an adjacency pair framework into the study of AI-MC, and preliminary yet mixed evidence to suggest that AI-generated language has the potential to undermine some dimensions of interpersonal perception, such as social attraction.

2 LANGUAGE PATTERNS IN AI-MEDIATED COMMUNICATION

Our first question of interest is how the use of AI-generated language (herein referred to as AI language) influences human-generated language (herein referred to as human language) production and comprehension. Pickering and Garrod's [27] interactive alignment model argues that production and comprehension are strongly linked, and claims that over the course of a conversation, partners will become more linguistically aligned with one another. Linguistic alignment has been observed in both face-to-face and technologically mediated settings [9, 13]. The alignment process is driven by the linguistic features of the conversation participants. Thus, when AI language is inserted into human-human dialogue, it has potential to modify not only the sender's linguistic features but also the receiver's. We examine two aspects of language use: how the presence of potentially positivity-biased smart replies affect the emotional expression in human language, and what kind of language structures correspond to use of smart replies in messages.

2.1 Positivity Bias

AI language suggestions in messaging platforms typically provide three options for a response, and these responses may prime human senders to use language differently than if the AI language was not available. For example, recent work on AI language revealed that Google's smart reply options tend to be more emotionally positive (e.g. "Great!" or "Sounds good!") than negative or neutral [17]. If smart replies have a positivity bias relative to human language, does this positivity bias influence human senders? To examine this question we first examined the sentiment of the smart replies, including both the options that were used and the options that were suggested but not used. Based on the initial findings suggesting that smart replies are biased towards positive sentiment [17], our first hypothesis predicted a positivity bias in the AI language. We used both human coding and computational analysis to compare the sentiment of the AI language to the human language:

> *H1a) The smart reply options will include more positive sentiment than negative or neutral sentiment, and b) the smart reply options will be more positive than human messages sent without the presence of these options.*

If we find that the AI language has a positivity bias, then will senders who use the smart replies have more positive messages than senders who do not have access to smart replies? Recent work suggests that the positivity of AI language can influence the language of human participants in

non-interpersonal settings [2]. We examine this question in two ways. First, when a sender is exposed to smart replies, are their messages more positive than senders who are not because they have incorporated the AI's positivity bias into their messages? Second, does the AI's positivity bias prime the sender such that they use more positive language in their own text? We hypothesize that:

> *H2) Senders using smart replies will a) produce messages that overall are more positive than senders that do not have access to smart replies and b) include human language that is more positive.*

Finally, psychological models of language use, such as Pickering and Garrod's [27] linguistic alignment model, predict that a receiver of messages with smart replies is also likely to become increasingly positive as the conversation progresses in order to align with the sender's more positive messages. Thus, we predict:

> *H3) The receiver's messages will be more positive when the sender has access to the smart replies than when the sender does not have access to smart replies.*

2.2 Adjacency Pairs and AI-MC

How do senders incorporate AI language into their conversations? The AI system behind smart replies uses information from prior utterances to offer suggested text responses that are probabilistically relevant to the previous message [23]. One way to understand the flow of turns in a conversation is through the concept of adjacency pairs [22, 28]. Adjacency pairs are consecutive utterances by two people in which the "second-pair part" depends on the "first-pair part". For example, one type of adjacency pair is "question/answer," in which one participant proposes a question (first-pair part) and the other provides an answer (second-pair part). Although human communication is deeply complex, many conversations only include a few types of adjacency pairs. One adjacency pair typology includes nine types of pairs [8], such as question/answer, assessment/agreement or request/acceptance.

Given that current smart replies in messaging contexts are designed to be relevant and meaningful to the prior utterance, rather than all of the prior turns in the conversation, and the creators of AI systems often refer to smart replies as part of a message "pair" [23, 30], the adjacency pair concept applies well to this form of AI-MC. In this study we examine what kinds of adjacency pairs the smart replies are used for in our task. We also examine how the human sender incorporated the smart reply into their messages. When senders use a smart reply option, how often do they choose to use the smart reply as a complete conversation turn without adding any of their own words (e.g. ["Yes"]), as a modification to their conversation turn through strategies such as conjunction words (e.g. ["Yes], but I think…"), or add a new conversation turn afterwards via a new sentence or text message (e.g. ["Yes]. We should try…")? Therefore, our research question is:

> *RQ1 a) What types of adjacency pairs are smart replies used for in our task and b) how do senders incorporate the smart replies into the second-pair part of the adjacency pair?*

3 INTERPERSONAL PERCEPTION

Language is tightly linked with another key aspect of human communication – interpersonal perception. People perceive a variety of traits in their communication partners, but the dimensions of warmth and competence have surfaced repeatedly as two of the most important factors in impression formation [24].

Linguistic features are regularly associated with various aspects of impression formation, including warmth [4, 12, 19]. If we find that senders using smart replies have more positive language in their messages, we hypothesize:

H4) Senders using smart replies will be perceived by receivers as more warm than senders that do not use them.

However, it is unclear whether or not an increase in positive language would affect other dimensions of interpersonal perception. Holoein and Fiske [19] suggest that an increase in the perceived warmth of another person may lead to a decrease in perceived competence, regardless of positivity in the language. Another possibility is that the smart replies will come across as unusual or strange messages, given the potential positivity bias, which may undermine perceptions of social or task attractiveness, or how personable or skilled the person is in a specific task environment [25]. We formulate our second research question as:

RQ2) How will perceptions of competence, social attraction and task attraction be affected by the use of AI-MC?

4 TASK PERFORMANCE

Interpersonal communication activities often can be conceptualized through a model of group tasks, where most goal-oriented interaction falls into one of four categories: "generating," "choosing," "negotiating," and "executing" [25]. AI is regularly involved in a range of these interpersonal communication tasks. For example, AI-assistance in online profile creation would be considered a "generating" task and including AI language in a work email about a group project could be a "choosing" task.

Claims about efficiency in completing certain tasks, as well as "state-of-the-art, large-scale deep learning" across multiple platforms have been made by the creators of AI systems [16, 23, 30], and there is recent evidence to suggest AI-assistance can prompt shorter, more simplistic descriptions of photos [3]. As such, how might the introduction of AI-MC into a task affect the human's ability to perform the task? Do changes such as a positivity bias or differences in interpersonal perception affect participants' abilities to complete a task? To address these questions, we ask the following:

RQ3) How will task accuracy, length of conversation, and word count be affected by AI-MC?

5 THE PRESENT STUDY

In this work, we use a lab-based experiment to measure the effect of smart replies on human communication, as hypothesized above. For our communication task we adopt a classic referential communication paradigm, the Tangram task, in which two partners collaborate to identify targets in an array of abstract images called tangrams [7]. This paradigm is well-established in both face-to-face and computer-mediated settings [14, 28], and can be considered a "choosing" task, in which partners must collaborate to determine correct answers [25]. We employed a between-subjects experimental design, in which dyads were randomly assigned to a condition with smart replies (AI-MC condition) or without smart replies (Control condition). One participant in each dyad was randomly assigned the Director role and the other the Matcher role. For dyads in the AI-MC condition, the Director was instructed to use one of the three available smart replies every time that they were available when texting the Matcher, and could write anything they wanted afterward. These Directors were asked not to alter the text of the smart reply itself. We required these Directors to choose smart replies, as we were not interested in how often people engaged with the AI language, but instead in how they incorporated this language into their own messages. The smart reply functionality was disabled for Directors in the control condition and for Matchers in all conditions.

6 METHOD

6.1 Participants

For this study, we conducted an *a priori* power analysis anticipating small to medium effect sizes in our interpersonal perception findings, which indicated a minimum sample size of approximately 150 participants. During recruitment and data collection in March 2020, however, the COVID-19 pandemic prompted a full shutdown of in-person behavioral research at our university, resulting in our ability to only collect data from 70 participants, or 35 dyads. The data from one dyad was excluded because they arrived simultaneously and noted that they already knew each other. Our final sample had 34 dyads and a total of 68 participants (62% female, 38% male; 43% Asian/Asian American, 13% Black/African American/African, 1% Hispanic/Latinx, 34% White/Caucasian/European American, 9% Mixed). The average age of participants was 20 years old (SD = 2.00). We recruited students from a large US research university through a participant pool allocated for the authors' department. No students were working with any of the authors or were familiar with the experiment design beforehand. All participants spoke English. The study received IRB approval in August 2019.

6.2 Procedure

Upon arrival, participants were led into separate rooms where they provided their informed consent and received instructions. Participants randomly assigned to the Matcher role received a sheet with 15 unnumbered tangrams, or abstract images, in a random order. All Matchers were asked to initiate the conversation with their respective Directors, to provide consistency across dyads. As a result, Directors in the AI-MC condition were always able to see smart replies before sending their first message. Participants randomly assigned to the Director role received a sheet with ten numbered tangrams. The Matchers had to determine which tangrams were matched with which number by texting with the Director through Google Hangouts Chat, a popular text-based communication application offering AI language suggestions. Participants were given as much time as needed to complete the task. All participants completed surveys after the task and were then debriefed about the purpose of the experiment.

All conversations were screen-recorded with the ScreenCam app, and smart replies were transcribed from these recordings by research assistants. Participants were given one research credit, part of a course requirement, for their participation.

6.3 Measures

After performing the task, participants received questions about the following measures:

Warmth and Competence. We asked participants to rate how warm or competent [11] their partner was on a 5-point scale anchored by 1 (not at all) to 5 (extremely). On average, participants rated their partners as very warm (M = 4.08, SD = .58) and very competent (M = 3.90, SD = .52).

Social and Task Attraction. Participants completed McCroskey and McCain's [24] social and task attraction scales. Response options were anchored by 1 (not at all) and 5 (extremely). On average, participants felt the social attraction statements represented their attitudes towards their partner moderately well (M = 3.40, SD = .90), and the task attraction statements extremely well (M = 4.50, SD = .43).

Familiarity with AI and Smart Replies. We asked participants how often they interacted with AI and/or smart replies, as well as "What are your thoughts about AI?" and "What are your thoughts about suggested text responses (e.g. Gmail's "smart reply" feature)?" On average, participants interacted with AI a little to a moderate amount (M = 2.69, SD = 1.18) and used smart replies a little (M = 1.68, SD = .82).

Funneled Debriefing. At the end of the survey, we were interested in whether or not any Matchers detected AI language in the Directors' messages. Matchers were asked the following questions sequentially and in an open-ended format: "Was there anything strange about the interaction with your partner?" "Was there anything strange about the language your partner used?" and "How 'machine-like' or 'automated' did your partner's language sound?"

A full list of survey questions, the raw data, and participant instructions are available at: https://bit.ly/2X0Domz.

6.4 Data Analysis Approach

The texts from the conversations were collected through Google Takeout, which allowed us to export and download logs of the conversations. In order to delineate AI from human language, smart replies were transcribed from screen recordings of the conversations. A total of 3,272 messages were sent (M = 96.24, SD =34.49 messages per dyad) and Directors in the AI-MC condition saw an average of 41.57 (SD = 18.72) smart replies in their conversation. We employed a content analysis and a lexicon/dictionary-based approach for data analysis. For the content analysis, two research assistants classified the sentiment of messages from Directors in both conditions, as well as the 873 smart replies that were available to Directors, including the smart replies that Directors did not select for their conversation. The smart reply options were coded as expressing positive, negative or neutral sentiment. Despite our instructions, sometimes the Directors neglected to choose a smart reply. Research assistants classified how the smart reply was or was not incorporated into their messages. Any discrepancies between classifications were resolved by the experimenter.

For the lexicon/dictionary-based approach, we used Valence Aware Dictionary and sEntiment Reasoner (VADER; [20]) and Linguistic Inquiry and Word Count (LIWC; [26]), which are two common dictionaries that have been well-validated for conversations. VADER's compound emotionality score is a "normalized, weighted composite score" that sums and normalizes valence scores of each word (less than or equal to -.05 = negative; greater than -0.05 and less than 0.05 = neutral; greater than or equal to .05 = positive). It is particularly appropriate for short text, and incorporates slang words and emojis as indicators of emotion. LIWC is one of the most common linguistic tools in social scientific analyses and follows a frequency-based approach, counting the presence of affective words to measure emotion.

6.4.1 Human-AI Composite Messages. Messages that include both AI language and human language are Human-AI Composite Messages (see Figure 1). We refer to the smart reply component as AI language and the component written by the Director as human language. For example, the composite message "Sounds great! Let's move on to the next figure" includes AI language ("Sounds great!") and the Director's human language ("Let's move on to the next figure"). We analyzed these messages in three ways by considering the AI language, the Human-AI composite message, the human language in separate analyses.

6.4.2 Statistical Tests and COVID-19 Restrictions. In this study, we investigate multiple types of interpersonal behavior, including language patterns, person perception and task performance. As noted in the Participants section, the COVID-19 pandemic resulted in a stoppage of in-person behavioral studies at the university where the study was conducted, so we were not able to collect data from the number of participants recommended by an *a priori* power analysis. Because our study is therefore underpowered, we report our statistical tests with and without Bonferroni post-hoc corrections for multiple tests. We believe that reporting both the more liberal and post hoc corrected statistics under these circumstances provides the reader with more information for interpreting the statistical conclusions.

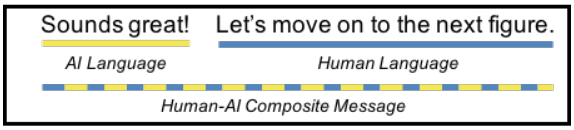

Fig. 1. Example Human-AI Composite Message.

7 RESULTS

In our results, we focus on language patterns in AI-MC through the lens of positivity bias and adjacency pairs, as well as emphasizing potential differences in interpersonal perception and task performance between conditions.

7.1 Positivity Bias in Smart Replies

Consistent with prior work [17] and with our first hypothesis, our human-coded content analysis found that 93% of smart replies suggestions included positive sentiment, while only 6% included a negative sentiment (all following αs are for inter-rater reliability; α = .75). We also confirmed the positivity bias in our computational analyses. The VADER analysis revealed that the emotionality of the AI language was significantly above the neutral cutoff (.05) with an average score of .23 (95% CI [.22, .25], $t(869) = 21.4$, $p < .001$, $d = .73$; see Figure 1). Similarly, a paired t-test using LIWC showed that the AI language included a higher percentage of positive emotion words per smart reply ($M = 24.62$, 95% CI [21.68, 26.83]) than negative emotion words ($M = .36$, $t(888.7) = 18.48$, $p < .001$, $d = .89$). As a proportion, for every negative emotion word in the AI's suggested language there were 94 positive emotion words.

Hypothesis 1b asked how the positivity bias compared with human language. The human language of the Directors and Matchers in the control condition, who were never exposed to smart replies, had an average emotionality score of .20 ($SD = .27$), which was significantly less positive than the AI language ($t(1661.6) = -3.61$, $p < .001$, $d = .14$). Human language also had three times fewer positive emotion words compared to the AI language ($M = 9.47$, $SD = 23.4$, $t(1132.3) = -10.84$, $p < .001$, $d = .53$). Our results suggest that the sentiment in the AI language has a strong positivity bias, with overwhelmingly positive sentiment compared to human language.

7.2 Human-AI Composite Messages and Human Language

Did the positivity bias in the smart replies influence the language of the conversation participants in the AI-MC condition?

We first examined the language of the Directors' messages in the AI-MC condition to address H2a: that senders with access to smart replies will produce messages that overall are more positive than senders that do not have access to smart replies. We test our hypotheses using a linear mixed model, with messages nested within participants. The condition (AI-MC or Control) was included as a fixed variable. We used VADER's compound emotionality score as the dependent variable. We found that Directors in the AI-MC condition sent composite messages that had higher positive sentiment ($M = .27$, $SE = .01$) compared to the Directors' messages in the control condition, which only included human language ($M = .20$, $SE = .01$, $t(29.02) = 2.99$, $p < .01$, $d = .25$; see Figure 2). Our LIWC analysis revealed that the Directors' Human-AI composite messages had marginally more positive emotion words ($M = 11.70$, $SE = .87$) compared to the control condition ($M = 8.28$, $SE = .78$, $t(25.42) = 2.00$, $p = .06$, $d = .13$). Our human coding ($\alpha = .92$) also corroborated these findings, with more positive sentiment in the AI-MC condition ($M = .29$, $SE = .02$) than in the control condition ($M = .18$, $SE = .02$, $p < .05$, $t(28.90) = 2.70$, $d = .22$).

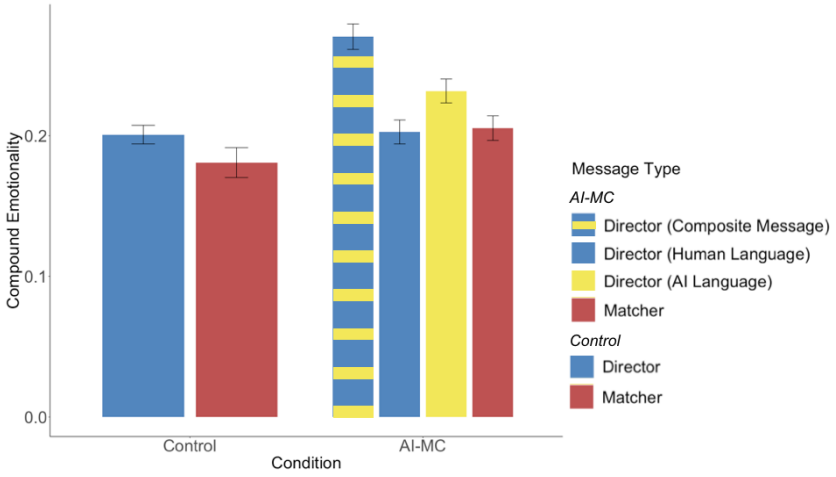

Fig. 2. Effect of AI-MC on Mean Compound Emotionality Score (SE).

Note: VADER's compound emotionality score is a "normalized, weighted composite score" that sums and normalizes valence scores of each word (less than or equal to -.05 = negative; greater than -0.05 and less than 0.05 = neutral; greater than or equal to .05 = positive).

Did the Director's own language in the AI-MC condition change, even when not using the AI suggestions, given that the AI language they were using had a strong positivity bias? Surprisingly, there were no significant differences in compound emotionality score between conditions for the human language (i.e., when the AI language was removed from the Human-AI composite messages; p = .76 and p = .47 for VADER and LIWC analyses, respectively; see Figure 2), suggesting that Directors did not adapt their own language to imitate or reduce the positivity bias of the smart replies.

Recall that Matchers never used any smart replies and were unaware of the Director's potential use in the AI-MC condition. Only two out of 34 Matchers reported that their partner sounded "a bit" or "slightly" machine-like, indicating that the AI language was not obvious to the Matchers. Although Matchers in the AI-MC condition were exposed to the positivity bias, their language was not more positive than Matchers in the control condition (p = .19 and p = .93 for VADER and LIWC analyses, respectively; see Figure 2), which did not support H3.

7.3 Pragmatics of AI-MC

How did the Directors incorporate the smart replies into their Human-AI composite messages (RQ1)? Recall that Directors in the AI-MC condition were instructed to choose one smart reply option every time it was available in the conversation, and then they could add their own language afterwards. The Directors were largely compliant, and chose a smart reply 77.8% of the time they were available. Our content analysis found evidence of six types of adjacency pairs and three types of utterances, or how the participant did or did not incorporate the AI language into their message (see Table 1).

The content analysis of adjacency pairs (α = .92) revealed that the most common type of adjacency pair for using AI language was for answering a question the Matcher had asked (34% of the time). Consider the following example (Dyad #006):

> Matcher: Okay so does the right side "shoulder" jutt out more than the rest
> Director: Yes, it does [from smart reply options: "Yes, that's it" "Yes, it does" "Yes, it is"]

Table 1. Adjacency Pair Types

Type	Example	Utterance Type	% of Type	Freq.
assessment/ agreement	-Yeah they all look like people except the one that looks like a bunny -[Right!] I'm thinking that...	Complete	18.2	8
		Modified	25.0	11
		Additional Turn	52.3	23
		Did Not Use	4.5	2
assessment/ disagreement	-Not a diamond. He is sitting and... -[Okay], hmmm I don't think I have that...	Complete	0.0	0
		Modified	50.0	1
		Additional Turn	50.0	1
		Did Not Use	0.0	0
compliment/ accept	-You were a great partner!! Thanks for the instructions haha -[haha, no problem!] I'm glad...	Complete	30.8	4
		Modified	0.0	0
		Additional Turn	69.2	9
		Did Not Use	0.0	0
inform/ acknowledge	-Got it!! -[Great!] It will be funny...	Complete	22.7	22
		Modified	21.6	21
		Additional Turn	55.7	54
		Did Not Use	0.0	0
question/ answer	-Both hands forward on the right? -[Correct], well not really correct...	Complete	37.9	39
		Modified	31.1	32
		Additional Turn	30.1	31
		Did Not Use	0.1	1
request/ accept	-Hmm ya sure let's come back to it lol -[Haha okay!] Figure 3...	Complete	8.6	3
		Modified	28.6	10
		Additional Turn	62.9	22
		Did Not Use	0.0	0

Type refers to Adjacency Pair Type. % is within Adjacency Pair Type. Utterance Type refers to how the participant did or did not incorporate the smart reply into their message. Freq. refers to the total N of messages. Brackets indicate the smart reply text.

In this case, the smart reply was incorporated as a complete turn; the Director did not make any modifications to the smart reply, and did not add any turns afterwards. Across all types of adjacency pairs, Directors incorporated them as a complete turn 25% of the time (utterance type α = .76).

As noted above, 22.2% of the time Directors did not use one of the smart reply options. Our analysis for these cases found that the majority of the time Directors chose not to use the smart reply because the options were not relevant second parts to the first part of the adjacency pair. For example (Dyad #005):

> Matcher: Ok cool how about 1?
> Director: Small Square (rotated 45 deg) atop a black rectangle that has some cutout in right and bottom right [from smart reply options: "That should work" "Yes" "Deal"]

This suggests that when the smart replies were not used it was because they could not be incorporated as an appropriate adjacency pair.

7.4 Interpersonal Perception and Task Performance

We were also interested in how interpersonal perception and task performance might be impacted by AI-MC. Despite the positivity bias in AI language, there were no significant differences between

conditions for Directors (AI-MC: M=4.10, SD= .68; Control: M=3.79, SD= .48, p= .16) or Matchers (AI-MC: M=4.24, SD= .58; Control: M=4.08, SD= .47, p= .41) on the variable of warmth, which did not support H4. In answer to RQ2, we found no differences between conditions for Directors or Matchers on the variables of competence or task attraction with and without controlling for multiple corrections. However, Matchers in the AI-MC condition (M = 3.13, SD =.20) reported less social attraction toward their partner than Matchers in the control condition (M = 3.76, SD = .21, $t(29.5) = 2.15$, $p < .05$, $d = .72$; see Figure 3), although this effect did not reach significance when using a Bonferroni correction.

To answer RQ3, we examined how task performance in terms of accuracy, conversation length and word count per message was influenced by the introduction of AI. Accuracy reflects the percentage of correctly numbered tangrams (i.e. the target of the task) and is calculated per dyad. Conversation length is the total number of messages a dyad sent to each other. There were no significant differences in accuracy (p = .75) and only marginal differences in conversation length (p = .09) between conditions. Word count was significantly higher for Directors in the AI-MC condition (M = 9.94, SD = 13.1) as compared to Directors in the control condition (M =7.68, SD = 7.01, $t(30.1) = 2.12$, $p < .05$, $d = .21$) and there were no significant differences in word count between conditions for Matchers (p = .78). Further, there were no significant differences between the AI-MC condition (M = 37.33, SD = 11.30) and control condition (M = 33.75, SD = 9.02) in the time in minutes that it took for dyads to complete the task (p = .33). Our findings suggest that the positivity provided by AI may undermine one type of interpersonal perception between partners, but it may not be detrimental to the interaction overall.

8 DISCUSSION

In this short paper, we replicate prior research on AI-MC in the form of smart replies, and develop a novel set of analyses to investigate how the introduction of AI language in human conversation affects language use, interpersonal perception, and task performance. We confirmed earlier, more informal findings [17] showing that AI language in the form of smart replies has a strong positivity bias. In our study, the smart replies were also more positive than human language, similar to past work in this area. Directors' Human-AI composite messages, which included both AI and human language, were more positive than the language of Directors in the control condition. We did not find evidence, however, to suggest that Directors exposed to the smart replies adapted their own language to match the increased level of positivity, as the interactive alignment model would predict [27]. Since the smart replies were overwhelmingly positive, Directors may have chosen not to incorporate this positivity bias into their own language to avoid sounding like the AI system.

From the Matcher's perspective in the AI-MC condition, there was no indication that they knew that AI was involved in their partner's messages, yet they still did not incorporate the positivity bias of the smart replies in their own language, despite the predictions of the interactive alignment model [27]. It is possible that the excess positivity they saw through the Director's Human-AI composite messages violated their expectations of a typical conversation. Even positive violations, or violations that exceed an individual's expectations, can introduce unintended negative consequences in an interaction [6].

We found that the concept of adjacency pairs was a useful framework for analyzing smart replies. Smart replies were used as the second-pair part in six types of adjacency pairs, and Directors incorporated AI language into their conversation turns in three different ways. Unlike past typologies of adjacency pairs in human conversation, in this task we did not find evidence that smart replies were used in pairs such as "blame/denial" or "clarification/puzzlement" [8]. Given the positivity bias of smart replies, they may be less useful in adjacency pairs expressing some kind of negative sentiment, which we see in the rarity of "assessment/disagreement" in our sample. We also found

that approximately 25% of messages with a smart reply included only the smart reply and no human language, providing evidence for the appropriateness of the application of adjacency pairs to AI-MC. When Directors resisted smart replies, it was typically because the AI language options were not relevant for completing an adjacency pair.

The positivity bias of AI language may have had a detrimental effect on social attraction, although this effect needs to be replicated in studies with additional power. Matchers interacting with Directors using AI language rated their conversation partners as less socially attractive than Matchers in the control condition, although this negative effect was not observed in perceptions of warmth. Since our definition of warmth involved constructs such as sincerity and tolerance, it is possible that participants did not associate a positivity bias with these qualities.

Lastly, the introduction of AI language did not have a deleterious effect on task performance in terms of accuracy or conversation length. While Directors with access to smart replies used more words per message on average, this did not seem to undermine performance on the task. In our sample, we do not see evidence to suggest that the positivity bias increased confusion regarding the task or made the conversation more difficult to understand.

8.1 Limitations

Our study has several important limitations. Although we were able to conduct our language analyses with sufficient power, our interpersonal perception analyses were underpowered because we had to reduce our intended sample size due to restrictions of conducting in-person behavioral studies in the lab as a result of COVID-19. As such, there may be differences between conditions that we were unable to detect with our current sample size. Relatedly, we only recruited participants from a university in the United States who all spoke English. Understanding the intersection between demographic characteristics and experiences with AI is important in this research area. As such, we aim to collect data from a larger and more diverse sample to build on our current findings and believe that replications of this work with additional power is required to shed new light on these questions.

This study focused on a particular type of social task on a specific platform: a "choosing" task [25] on Google Hangouts Chat. A Tangram task, where two partners collaborate to identify target images, is obviously not representative of the full spectrum of tasks that people complete in their daily conversations. While these design considerations focused on experimentally uncovering the dynamics of smart replies on interpersonal communication, the design of the study places limitations on the contexts these findings generalize to. Given that there may be more than 36 billion emails or other messages sent per day that contain AI language [1, 5], there are a wide variety of contexts in everyday communication AI messages are deployed in. Future work must build on current research investigating the role of AI-mediated communication in other kinds of tasks and conversational settings [2, 18, 21].

For our language analyses, we used both computational tools and human coding and found congruent results throughout, especially in regard to the positivity bias. Nonetheless, it is possible that the positivity bias results from LIWC and VADER are inflated to some extent due to errors regarding negation (e.g. LIWC does not recognize "not good" as negative sentiment). As such, replications of this finding with additional computational and human coding methods would advance our understanding of the nature of this bias.

Finally, in most current implementations, people are not *required* to use smart replies in their conversations. Our instructions to the Directors to use the smart replies were designed to increase usage and maximize any effects of introducing AI language into human communication. While this requirement allowed us to examine how Directors used and incorporated the smart replies, future work will need to examine AI-MC under more naturalistic and ecologically valid conditions.

8.2 Implications

The findings from our short paper have notable implications for future work in this area. This study corroborates past research that suggests the a positivity bias of AI language in the form of smart replies. Although in we did not find that AI-MC influenced warmth and competence in our constrained task setting, the positivity bias may have undermined social attraction between the partners. Changes in language patterns and interpersonal perception at scale have a number of ethical and social implications [15], but very little work has explored this potential consequence of AI-MC.

One important ethical question surrounding AI-MC is in regard to attributions of blame for when things go wrong. Consider, for example, an individual who chooses a suggested response option (e.g., "Hope you have a great weekend!") because it is readily available, even if it does not fit the interpersonal context well. If this message negatively impacts the receiver's perception of the sender, who is perceived at fault - the sender? The AI system? Or maybe even the creator of the AI system? Although there is evidence to suggest AI can shoulder some blame in unsuccessful interactions [18], recent events concerning accidents and self-driving cars indicate that humans may be considered to be at fault as well [29]. Understanding patterns of blame attribution will continue to be important as AI is further integrated into interpersonal communication.

Another ethical dimension concerns whose discourse the AI messages are modeled after. If text-based AI in English is trained on linguistic data from users whose demographics are strongly skewed towards older white males with higher socioeconomic status, the system is likely to produce the discourse of that demographic. The ubiquity of this kind of AI language may lead to the homogenizing of linguistic norms, and the removal of crucial cultural variations in language. It could also prioritize certain racial and economic types of discourses, reinforcing systemic inequalities. Although the biases in algorithmic decision-making have received increased attention in recent years, it is unclear how these biases will affect social dynamics of human communication at scale and over time, as well as in non-English contexts.

However, designers of AI systems could play a large role in the future of AI-MC, and their work stands to impact interpersonal dynamics beyond the level of conversations. The concept of a positivity bias in AI language and the notion of adjacency pairs could provide useful guidance for design options. What are the advantages and disadvantages of smart replies expanding to include a larger range of emotional expression and types of discourse, or continuing to optimize for efficiency and positivity? How might smart replies be tailored to the entire conversation, or histories of conversation, rather than just the prior utterance? These questions are important for the those behind the design of AI systems, as these choices could introduce new interpersonal dynamics into our everyday communication.

8.3 Conclusion

Our results indicate that although AI language tends to have a positivity bias, senders in our study did not incorporate this bias into their own language, though this bias may have undermined the sender's social attractiveness. Our analysis suggests that the adjacency pair framework may be a useful way to conceptualize AI messages within conversations. We found that senders were often successful in integrating AI language as part of adjacency pairs in their conversations, producing a novel form of message we refer to as Human-AI composite messages. We encourage researchers to consider potential differences between AI language and human language, as well as impression formation and the notion of efficiency in their work on AI-MC. Future work can examine how language may change over the course of conversation, as well as over time as people are increasingly exposed to AI in their interpersonal communication experiences.

ACKNOWLEDGMENTS

The research team would like to thank the following research assistants for their work on this project: Eleni Aneziris, Caroline Ghisolfi, Alyssa Horeczko, Daniella McMahon, Maia Rocklin, and Mark York. This material is based upon work supported by the National Science Foundation under Grant No. CHS 1901151/1901329.

REFERENCES

[1] 2015. Email Statistics Report, 2015-2019. https://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf
[2] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2018. Sentiment bias in predictive text recommendations results in biased writing. In *Graphics Interface*. 8–11.
[3] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.
[4] Diane S Berry, James W Pennebaker, Jennifer S Mueller, and Wendy S Hiller. 1997. Linguistic bases of social perception. *Personality and Social Psychology Bulletin* 23, 5 (1997), 526–537.
[5] Greg Bullock. 2017. Save time with Smart Reply in Gmail. https://www.blog.google/products/gmail/save-time-with-smart-reply-in-gmail/
[6] Judee K Burgoon. 1993. Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology* 12, 1-2 (1993), 30–48.
[7] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.
[8] John H Connolly, Roel M Vismans, Christopher S Butler, and Richard A Gatward. 2011. *Discourse and pragmatics in Functional Grammar*. Vol. 18. Walter de Gruyter.
[9] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*. 745–754.
[10] Roy W Feinson. 1988. Interpretive tone telecommunication method and apparatus. US Patent 4,754,474.
[11] Susan T Fiske, Juan Xu, Amy C Cuddy, and Peter Glick. 1999. (Dis) respecting versus (dis) liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of social issues* 55, 3 (1999), 473–489.
[12] Pamela Gibbons, Jon Busch, and James J Bradac. 1991. Powerful versus powerless language: Consequences for persuasion, impression formation, and cognitive response. *Journal of Language and Social Psychology* 10, 2 (1991), 115–133.
[13] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.
[14] Jeffrey T Hancock and Philip J Dunham. 2001. Language use in computer-mediated communication: The role of coordination devices. *Discourse Processes* 31, 1 (2001), 91–110.
[15] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.
[16] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652* (2017).
[17] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[18] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.
[19] Deborah Son Holoien and Susan T Fiske. 2013. Downplaying positive impressions: Compensation between warmth and competence in impression management. *Journal of Experimental Social Psychology* 49, 1 (2013), 33–41.
[20] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
[21] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 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.
[22] Emily Jamison and Iryna Gurevych. 2014. Adjacency Pair Recognition in Wikipedia Discussions using Lexical Pairs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. 479–488.
[23] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 955–964.

[24] James C McCroskey and Thomas A McCain. 1974. The measurement of interpersonal attraction. (1974).
[25] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.
[26] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[27] Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences* 36, 4 (2013), 329–347.
[28] Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica* 8, 4 (1973), 289–327.
[29] Tom Simonite. 2020. As Machines Get Smarter, How Will We Relate to Them? https://www.wired.com/story/as-machines-get-smarter-how-will-we-relate-to-them/
[30] Yue Weng, Huaixiu Zheng, Franziska Bell, and Gokhan Tur. 2019. OCC: A Smart Reply System for Efficient In-App Communications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2596–2603.

Received June 2020; revised October 2020; accepted December 2020