AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness

Maurice Jakesch[1], Megan French[2], Xiao Ma[1], Jeffrey T. Hancock[2], Mor Naaman[1]
[1]Cornell Tech, Cornell University [2]Stanford University
mpj32@cornell.edu, mfrench2@stanford.edu, xm75@cornell.edu, jeff.hancock@stanford.edu, mor.naaman@cornell.edu

ABSTRACT

We are entering an era of AI-Mediated Communication (AI-MC) where interpersonal communication is not only mediated by technology, but is optimized, augmented, or generated by artificial intelligence. Our study takes a first look at the potential impact of AI-MC on online self-presentation. In three experiments we test whether people find Airbnb hosts less trustworthy if they believe their profiles have been written by AI. We observe a new phenomenon that we term the *Replicant Effect*: Only when participants thought they saw a *mixed* set of AI- and human-written profiles, they mistrusted hosts whose profiles were labeled as or suspected to be written by AI. Our findings have implications for the design of systems that involve AI technologies in online self-presentation and chart a direction for future work that may upend or augment key aspects of Computer-Mediated Communication theory.

CCS CONCEPTS

• **Human-centered computing → Interaction paradigms**; *HCI theory, concepts and models.*

KEYWORDS

AI-MC; Artificial Intelligence; CMC; Trust.

ACM Reference Format:
Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, Mor Naaman. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300469

CHI 2019, May 4–9, 2019, Glasgow, Scotland

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300469

1 INTRODUCTION

The introduction of artificial intelligence (AI) to interpersonal communication has the potential to transform how people communicate and present themselves in mediated environments. What used to be *Computer-Mediated Communication* (CMC) is turning into *AI-Mediated Communication* (AI-MC): interpersonal communication not simply transmitted by technology but *augmented*—or even generated—by *algorithms* to achieve *specific communicative or relational outcomes*. In AI-MC an AI system operates on behalf of the communicating person, e.g., by augmenting, generating or suggesting content. AI-MC is distinct from traditional CMC technologies that primarily transmit messages, and from typical machine-authored texts that do not represent a person. "Smart replies", for example, are AI-MC: The technology generates parts of the user's communication on the user's behalf.

Broadly defined, AI-MC can impact interactions from one-to-one exchanges such as messaging to one-to-many broadcasts like writing user profiles or appearing in a live YouTube video. In text-based communication—the focus of this work—we have already advanced from spell check and predictive auto-completion to early AI-MC instances, like the aforementioned auto-responses for chats and e-mails [31]. Natural language processing and generation technologies enhance people's writing style [23], write human-like texts [64], and produce online self-presentations [1]. As the power of language processing and generating technologies increases, AI-MC will become more pervasive and powerful. The use of AI in interpersonal communication challenges assumptions of agency and mediation in ways that potentially subvert existing social heuristics [15, 30, 56]. The emerging field presents a significant new research agenda with impact on core CMC and Human-Computer Interaction topics, from communication practices to relationship and interpersonal dynamics [51, p. 22].

This study examines the effect of AI-MC on online self-presentation dynamics, making theoretical contributions to a rich area of research [7, 12, 49, 54]. We ask a basic question, as appropriate for an initial study: Does the belief that AI may have written a profile affect evaluations by others? In

particular, to borrow the terminology of the Hyperpersonal Model [56, 57], will *receivers* evaluate *senders* differently if they believe AI is involved in authoring the *sender's* profile?

We study this question in the context of online lodging marketplaces like Airbnb [25, 46] with a focus on host trustworthiness. Trust and deception in online self-presentation have been studied extensively [24, 26–28, 52] and have been shown to play a critical role in online marketplaces [16, 35, 36]. The Airbnb scenario also allows us to build on previous work that investigated the influence of profile text on the trustworthiness of Airbnb hosts [38, 39].

In a series of three online experiments, we examine how the belief that a computer system has generated a host's profile changes whether the host is seen as trustworthy by others. Study 1 compares how hosts are evaluated in two hypothetical systems: one where profiles are supposedly written by AI, and one where hosts wrote their own profiles. In Study 2 and 3 participants evaluate hosts in an environment where they believe some profiles have been generated using AI, while others have been written by the hosts. In reality, all profiles shown were written by humans, and were selected from a publicly available dataset of Airbnb profiles [38] since we are interested in the future potential impact of AI-MC rather than its current capabilities.

In this first attempt to conceptualize AI-MC and to test one of its effects, we observe that (1) when people are presented with *all AI-generated* profiles they trust them just as they would trust *all human-written* profiles; (2) when people are presented with a *mixed* set of AI- and human-written profiles, they mistrust hosts whose profiles they believe were generated by AI. We term this phenomenon the *Replicant Effect*[1] – as in the movie Blade Runner, our (experimental) world was populated by non-human agents (the *replicants*) that imitated humans. Our results lend support to the Hyperpersonal Model of CMC [57] where receivers tend to exaggerate perceptions of the message sender, or, in this case, exaggerate textual hints that a profile was written by AI. Further, we show that different respondents reliably identify similar profiles to be written by AI, likely due to a folk theory [8, 17] of what AI-written text would look like. In the discussion, we draw on relevant theories that may explain our findings, offer directions for future AI-MC work, and project on how AI-MC may shift our use and understanding of CMC systems.

2 BACKGROUND

Our inquiry is motivated by the maturing ability of AI systems to *generate* natural language as well as the increasing use of AI in online self-presentation. Previous work in CMC has studied online self-presentation [13, 34] and the nature of human interactions with bots and agents [3, 5, 18, 44]. We build on these works to situate our discussion of AI-Mediated Communication. We also relate our work to previous studies on the perceived trustworthiness of user profiles.

Impression formation

CMC research has extensively studied how people present themselves online via technology [13, 34]. We expand on this research by analyzing how the introduction of AI into online self-presentation might shift impression formation. AI-mediation may influence how people interpret and scrutinize the content of profiles, as users interpret signals presented online to infer characteristics about other individuals [14, 58, 60]. The Hyperpersonal Model [56, 57], for example, argues that receivers may over-interpret cues from the sender's self-presentation because of the reduced cues in text-based CMC. When certain cues can be easily modified with the help of AI, receivers have to change how they evaluate them.

A number of theories touch on how information shared in online profiles becomes credible. Walther [59] introduced the principle of *warranting* to CMC, asserting that receivers rely more on information that is difficult for the sender to manipulate [6]. The warranting idea is highly related to *signaling theory*, used by Donath [9] to explain why online signals vary in their reliability as proxies of the sender's underlying qualities–from easily faked self-descriptions (e.g., "I go to Cornell") to difficult to fake signals (e.g., having a cornell.edu email address). The *Profile as Promise* framework explains how people assess signals in online profiles when they expect future interactions, like in online dating, or in lodging marketplaces. The framework asserts that people are expected to make minor–but not significant–misrepresentations in their profile.

Introducing AI to interpersonal communication may complicate these theories and models. Will self-descriptions generated by a computer be treated as "warranted", as earlier research suggested [40]? Can AI give credible promises on behalf of the sender? Will AI-MC change the assessment of online signals, and result in different behaviors by senders and receivers when people optimize their self-presentation algorithmically, as seen in recent work [7]? Studying AI-MC in the context of online self-presentation will test and extend these theories.

Interactions with bots and AI agents

Since Weizenbaum's early study [61], a large body of research adjacent to our work on AI-MC has explored natural language communication between man and machine. We know that people tend to apply social rules and attributes to computers [44, 45]. Technological advances now allow agents to produce more human-like dialogues. Studies of

[1] with apologies to Philip K. Dick and Ridley Scott

social bots [18] find that in these dialogues, people put more effort into establishing common ground when they perceive an agent as human [3] and that introducing anthropomorphism may generate strong negative user reactions [5].

Various researchers have explored how humans perceive machine generated content: In the context of automated journalism, scholars have observed differing levels of perceived credibility of computer-written articles: in some cases, computers were perceived as less credible, explained by the heuristic that machines are more artificial than humans [22, 55]. In other cases, there was no difference in the perceived credibility of human- and computer-written news [62], potentially because machines are seen as more objective than humans [50].

Unlike in interactions with bots, in AI-MC computers are *not* communicating on their own behalf, but on behalf of a person in interpersonal exchange. The findings and outcomes of past studies need to be re-evaluated in settings where bots and AI agents are used for interpersonal communication. Some early work suggests that the involvement of AI through "smart replies" can influence conversations, for example by offering primarily positive suggestions [31].

Trustworthiness, profiles, and Airbnb

We situate our work in the context of online lodging marketplaces, specifically Airbnb, where a range of prior work [16, 25, 38, 39] and publicly available data sets [38] allow us to ground our experiments in existing methods and discussions.

The trust that can be established based on user profiles [16, 20] is central to the functioning of social interactions and exchange, from online dating [19, 40, 53] to resumes [24] and lodging marketplaces like Airbnb [16, 35, 36, 38]. On Airbnb, *hosts* list properties that *guests* can book and rent. Hosts aim for their profiles to appear trustworthy, especially in situations where reputation signals are either unavailable or skew positively high for everyone [65].

The current work directly builds on a recent study of the trustworthiness of Airbnb hosts based on profile text [38]. The study revealed that the profile text impacts the perceived trustworthiness of hosts in a reliable way; in other words, the evaluations of host trustworthiness based on profile text are fairly consistent between raters [38]. Our experiments build on these established measurements of trustworthiness of Airbnb hosts to investigate whether the introduction of AI-MC affects perceived trustworthiness.

3 STUDY 1

Study 1 offers a first experimental attempt to understand the effect of AI-MC on perceived trustworthiness. It compares how hosts are evaluated in two hypothetical systems: one where their profiles are supposedly written by AI, and one where hosts wrote their own profiles. In reality, participants in both scenarios rate the same set of profiles.

Methods

Study 1 is a mixed-factorial-design online experiment where participants rate the trustworthiness of prospective hosts in an Airbnb-type scenario. Our procedure followed prior work on the trustworthiness of Airbnb host profiles [38]. We asked participants to imagine they were reviewing potential hosts in a lodging marketplace. We showed them a set of 10 Airbnb host profiles in randomized order. The profiles were selected from a publicly available dataset[2] of Airbnb profiles [38]. We only considered a set of profiles of comparable length (37 to 58 words) based on Ma et al.'s result showing the correlation between profile length and trustworthiness ratings. From this set, we chose five profiles that had received very high trustworthiness rankings (top 5%) in the prior study, and five profiles that had received very low trustworthiness ratings (bottom 5%). We defined an independent variable called the "profile baseline" based on this split. The variable allows us to observe whether the effect of AI-MC is different for high- and low-trustworthiness profiles. The profiles are listed in the appendix.

Experimental manipulation. Participants were randomly assigned to the control or treatment group. While all participants were reviewing the same profiles, subjects in the treatment group were led to believe that the profiles they rated have been generated using an AI system, similar to the "Wizard of Oz" approach used in other studies of interpersonal communication [4, 10, 37].

We developed our AI-MC illusion through multiple rounds of design iterations. We chose the wording of the task based on the results of a survey ($n = 100$) where we tested respondents' understanding of the terms *AI*, *algorithm* and *computer system*. We launched a pilot experiment ($n = 100$) to test the design and refined elements of the manipulation based on the feedback collected. In the final design, we explained the AI system as follows:

> To help hosts create profiles that are more attractive, this site provides a computer system using artificial intelligence that will write the description for hosts. The hosts simply enter some information and the artificial intelligence system generates the profile.

The participants in the treatment group then watched a 10-second demo video of a mock-up AI system (see Figure 1). In the video, an animation depicts a system automatically generating text for an Airbnb profile from a Facebook profile

[2]The dataset is available from https://github.com/sTechLab/AirbnbHosts

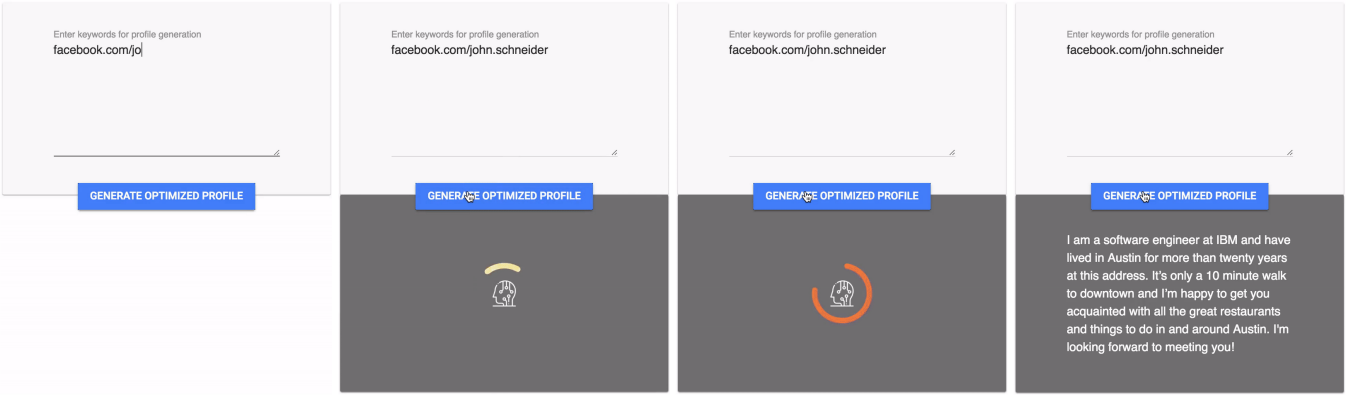

Figure 1: Screenshots of the "AI system" demo video participants in the treatment group watched before rating the profiles

URL provided by a user. We performed a manipulation check to verify that the participants in the treatment condition understood that a profile had been generated. To reinforce the manipulation, all profiles in the treatment group came with a label that reminded of the AI system.

Measured variables. We measured the perceived trustworthiness of the host as the outcome variable. We used perceived trustworthiness for several practical reasons: First, perceived trustworthiness is a variable that is conceivably affected by the introduction of AI-MC. Second, the variable has been shown to be a reliable measure in the context of Airbnb in previous studies [38, 39]. Finally, we had access to publicly available data on hosts and their perceived trustworthiness scores [38].

Perceived trustworthiness is defined as an attribute of a target individual [29, 33]– in our case, the host represented by the profile. We measured profile trustworthiness using a scale developed in [38] which builds on earlier measurements of Mayer et al. [41, 42]. As the six items in the original scale were highly correlated [38], to reduce respondent fatigue, we selected one item only from each dimension of trustworthiness (ability, benevolence, and integrity [42], Likert-style, 0–100). The items we used were:

(1) This person maintains a clean, safe, and comfortable household. *(ability)*
(2) This person will be concerned about satisfying my needs during the stay. *(benevolence)*
(3) This person will not intentionally harm, overcharge, or scam me. *(integrity)*

Following past studies [38], we combined the three items into a trust index by calculating their mean (Cronbach's $\alpha = .86$; $M = 66.6$, $SD = 18.5$, reliable and consistent with prior work).

After the main rating task, we asked participants to complete a generalized trust scale we adapted from Yamagishi's trust scale [63] and an AI attitude survey modeled after the well-established computer attitude scale [47]. We combined the multiple-item scales into mean generalized trust (Cronbach's $\alpha = .88$; $M = 64.1$, $SD = 16.7$) and AI attitude scores (Cronbach's $\alpha = .72$; $M = 70.1$, $SD = 18.7$).

Table 1: Overview of measurements

Name	Concept
Trustworthiness	The perceived trustworthiness of a host based on his or her Airbnb profile [38]
Generalized trust	A measure of how much the participant trusts other people in general [63]
AI attitude	An index of the participant's positive and negative attitudes toward AI [47]
Trust baseline	Whether the profile was rated as trustworthy or untrustworthy in a prior study [38]
AI score *(Study 2 and 3 only)*	A measure of how strongly the participant suspects a profile was written by AI

Participants also answered demographic questions (gender, age, education, and residential neighborhood type), as well as free-form questions explaining how they rated the profiles. We finally asked what they thought was the purpose of this study, and, in the treatment group, whether they had comments on the system. An overview of variables is shown in Table 1.

Participants. We recruited 527 participants via Amazon Mechanical Turk (AMT) [2, 32]. Participation was limited to adults in the US who had completed at least 500 tasks with an approval rate of ≥ 98%. Participants' mean age was 38, with 48% identifying as female. Participating workers received a $1.20 compensation based on an estimated work time of 6 minutes for a projected $12 hourly wage. The workers provided informed consent before completing the study and were debriefed after completion with an option to withdraw. The debrief is included in the appendix. The protocols

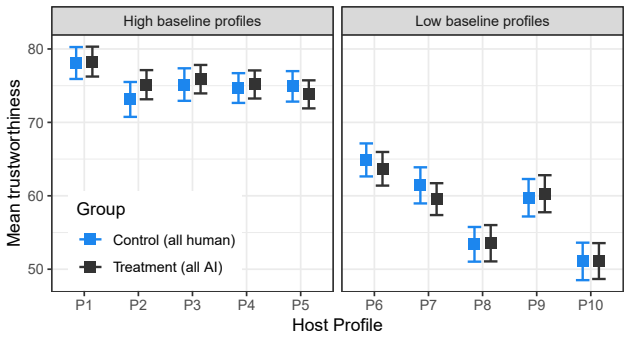

Figure 2: Study 1 host trustworthiness ratings by experimental condition, for profiles of high (left) and low (right) trustworthiness baseline

were approved by the Institutional Review Board at Cornell University (protocol #1712007684).

Data validation. We performed several integrity and attentiveness tests for our participants. We excluded responses that had failed the linguistic attentiveness check borrowed from Munro et al. [43] as well as participants who did not select the right scenario ("I am traveling and the person in the profile offers to host me.") in a second attentiveness test. We excluded workers whose median rating time per profile was less than five seconds and workers with mostly uniform responses ($SD < 5.0$). Furthermore, we removed participants whose average trust rating fell outside the $mean \pm 2SD$ statistic of participant rating averages, leaving us with 389 subjects. Finally, we examined the free-form responses participants in the treatment group gave after viewing the system demo. Almost all responses demonstrated a clear understanding that the system generated a profile, leading us to conclude that the manipulation was effective.

Open Science Repository. The full experimental data, analysis code and experiment preregistration are available from https://osf.io/qg3m2/ and https://github.com/sTechLab/aimc-chi19.

Results

When people are presented with either *all human-written* or *all AI-generated* profiles, do they assign different trustworthiness scores to hosts? The results of Study 1 provide a negative answer to this question.

Figure 2 illustrates our results: For each of the ten profiles (along the x-axis), we observed almost identical trustworthiness ratings (y-axis, along with confidence intervals) in the control (blue) and treatment (black) group. For example, profile 1 received average trust ratings of 78.3 by respondents who believed all profiles were written by humans, and 78.1 by respondents who thought an AI system generated all profiles. We conducted a 2x2 mixed factorial ANOVA to compare the main effects of perceived profile generation (human vs. AI), profile baseline (high vs. low), and their interaction effect on trust ratings. The ANOVA revealed significant differences between high baseline ($M = 75.4, SD = 14.6$) and low baseline ($M = 57.8, SD = 17.7$) profiles, $F(1, 387) = 2046, p < 0.001$. Since we selected the profiles to be of either high or low baseline trustworthiness based on a prior study this result was expected and validates the reliability of the trust measurement in the current study. The ANOVA results did *not* indicate a main effect of perceived profile generation (human vs. AI); in other words, we did not find significant differences in trustworthiness ratings when we told participants that all profiles were written by the hosts ($M = 66.64, SD = 18.1$) and when we told them all profiles were AI-generated ($M = 66.64, SD = 18.8$). We briefly note that consistent with previous work, respondents' generalized trust levels were predictive of the trustworthiness ratings they assigned ($\beta = 0.30, p < .001$) and AI attitude ($\beta = 0.08, p < .001$) was predictive of the trust ratings as well.

4 STUDY 2

Study 2 explores whether people perceive the trustworthiness of profiles differently when they encounter a mixed-source environment that includes both AI- and human-written profiles without knowing how each profile was written.

Methods

We ran Study 2 with an almost identical setup to Study 1, but we told participants this time that "*some* of the profiles [they see] have been generated by a computer system using artificial intelligence, while others have been written by the host." Participants were not told which or how many profiles were generated by AI. We showed them the same demo video of the AI system and checked the efficacy of the manipulation as described in Study 1. Participants rated the 10 host profiles from Study 1. The full experimental data, analysis code and experiment preregistration are publicly available on OSF.

Measured variables. We measured the same variables as in Study 1. In addition, respondents indicated whether they thought each profile was (1) "Definitely Human-written" to (6) "Definitely AI-generated" on a 6-point Likert-style scale. We refer to this measurement as the "AI score" of a profile. In follow-up questions after the rating task, we asked participants how they decided whether a profile had been generated by AI. We aggregated indices for the trust ratings (Cronbach's $\alpha = .86$; $M = 65.5$, $SD = 17.9$) as in Study 1.

Participants. We recruited 286 participants using the same procedure, parameters, and payments we used in Study 1. Participants who had participated in Study 1 were not eligible

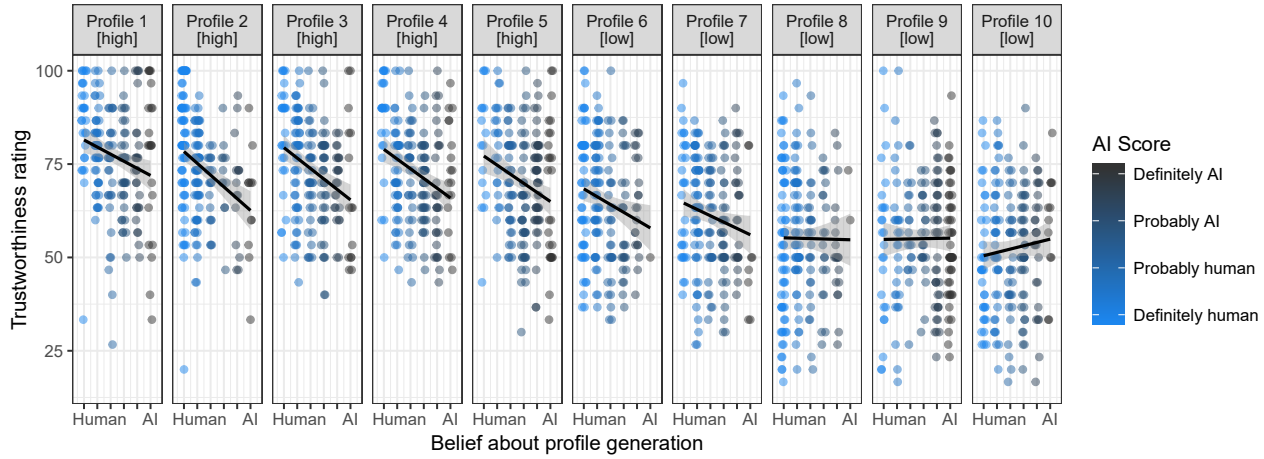

Figure 3: Study 2 host trustworthiness (y-axis) versus the participant's belief whether a profile was AI-generated (x-axis), for profiles of high (left) and low (right) trustworthiness baseline

for Study 2. Participants' mean age was 37; 56% of them identified as female. We performed manipulation checks and attentiveness tasks to control for low-quality responses using the same procedure as in Study 1, excluding 89 out of 285 participants.

Results

In a mixed-source environment, where participants do not know whether a profile was written by the host or AI, do they evaluate hosts with profiles they suspect to be AI-generated differently? The results show a clear trend: the more participants believed a profile was AI-generated, the less they tended to trust the host.

Our observations are visualized in Figure 3 showing an overview of the raw trustworthiness scores participants gave (y-axis), grouped by host profiles 1–10 (x-axis), and further plotted over the AI score assigned. The AI score is also represented by color, from "more human" (blue, left) to "more AI" (grey, right). For example, the top-most, left-most point on the figure shows a participant that gave Profile 1 a perfect trustworthiness score (100), and a low AI score (1), corresponding to the belief that the profile was "definitely human-written". Just like Study 1, the five profiles on the left are high baseline trustworthiness profiles. The figure suggests that participants who believed that a profile was written by the host assigned higher trust ratings to the host than participants who suspected the *same profile* was AI-generated. We visualize this trend by fitting a basic linear model to the data. The slope of the fitted line indicates that there may be an interaction: while for the high baseline trustworthiness profiles the slope is strongly and consistently negative, the slope of low baseline trustworthiness profiles is less pronounced.

To test how the particular characteristics of an observation affected ratings, we calculated a multiple linear regression predicting trustworthiness based on AI score, profile baseline, and their interaction (R^2=.231, $F(3, 1966) = 196.8, p < .001$). As expected, a low baseline is predictive of lower trust ratings ($B = -21.7, SE = 1.55, p < .001$). More interestingly, the AI score participants assigned to a profile significantly predicted lower trustworthiness ratings ($B = -2.51, SE = 0.31, p < .001$): the more a participant believed a profile to be AI-generated, the less trustworthy the participant judged the host. We also find a significant interaction between baseline trustworthiness and AI score, predicting that the negative effect of AI score will be weaker for low baseline trustworthiness profiles ($B = 1.68, SE = 0.45, p < .001$). We repeated the analysis with a multilevel model with a random effect per subject and computed two additional models including fixed effects for generalized trust and AI attitude. All models showed similar coefficients and significance of baseline trustworthiness, AI score, and their interactions. We thus omit the model details for brevity.

Taken together, Studies 1 and 2 demonstrate that AI-MC has an effect on trustworthiness. Study 3 replicates the effect and extends the results by investigating what factors contributed to the lower evaluations that hosts with profiles perceived as AI-generated received in Study 2, but not Study 1.

5 STUDY 3

Study 3 investigates key questions raised by the previous studies. While Study 1 exposed no differences in trust, Study 2 provided initial evidence that perceived AI-generation affects trustworthiness in mixed-source environments. We designed

Study 3 to clarify the conditions under which AI-Mediated Communication is distrusted.

Specifically, Study 3 asked whether the uncertainty in the mixed-source environment led to distrust. Did hosts receive lower trust ratings due to source uncertainty–as in Study 2 participants did not know what type of profiles they rated–or due to the heightened salience of the type of profile in a mixed-source environment? We tested the role of uncertainty in one experimental group where profiles were labeled, disclosing their supposed generation type. In addition, Study 2 forced participants to assign an AI score to a profile before they provided trust ratings, perhaps priming their responses. Study 3 explored the impact of asking participants to assign AI scores before rating a profile. Furthermore, we replicated the trend observed in Study 2 on a wider set of profiles. Both Study 1 and Study 2 used the same set of 10 profiles. We conducted Study 3 with a different and larger set of profiles to show that the effect observed in the earlier studies was not due to specific characteristics of the chosen profiles. Finally, we designed Study 3 as a randomized controlled trial by showing participants profiles that we pretested to be more AI-like or more human-like. Conjointly, Study 3 has been designed to offer strong experimental evidence for the existence of an AI-MC effect.

We hypothesized that *"AI" profiles in the treatment conditions will be rated as less trustworthy than the same profiles are rated in the control condition.* In other words, we predicted that when we tell participants they are rating a mixed set of profiles, regardless of whether the AI-like profiles are labeled as such, these "AI" profiles will be rated as less trustworthy compared to the ratings they receive in a control group that assumed all profiles to be written by humans. We preregistered our hypotheses and the full experimental design prior to the collection of data. The full experimental data, analysis code and preregistration are publicly available on OSF.

Methods

Study 3 used the procedures and techniques from Study 1 and 2, introducing new experimental conditions and a new and larger set of 30 profiles that we pretested to be either human- or AI-like.

Selection of profiles. In a preliminary study, we collected a set of profiles that were generally seen as either human-like or AI-like. To identify such profiles, we tested 100 random profiles from the same public dataset we used in the first two studies [38] on AI score. To keep the studies comparable, we only selected profiles of 37-58 words length. While the profiles in Study 1 and 2 were selected to explore the difference between high or low trustworthiness profiles, in Study 3 we selected profiles of average trustworthiness ($mean \pm 0.5SD$ statistic) to minimize potential confounds due to differences in trustworthiness.

Table 2: Overview of Study 3 conditions

Name	Manipulation
Control	Subjects believed they were rating regular profiles
Unlabeled	Subjects believed that some of the profiles were AI-generated, while others were written by the host
Labeled	In addition, "AI" profiles were labeled as AI-generated
Primed	Instead of labels, subjects assigned AI scores to profiles

We recruited 80 workers on Amazon Mechanical Turk to each rate 16 of the 100 profiles, indicating whether they thought a profile was (1) "Definitely Human-written" to (6) "Definitely AI-generated" on a 6-point Likert-style scale. After excluding uniform or incomplete answers, we analyzed the 945 AI scores received. We selected the 15 profiles that received the highest mean AI scores for the "AI " profile group and the 15 profiles receiving the lowest mean AI scores for the "human" profile group. The selected profiles are available on OSF.

Study design and procedure. Participants rated 10 profiles in randomized order: five "AI" profiles (out of the 15 profiles rated as AI-like in the preliminary selection) and five "human" profiles (out of the 15 profiles rated human-like). We randomly assigned participants to one of four groups: The control group participants were told they were rating regular profiles written by the host (akin to the "host-written" group in Study 1). In the treatment groups, participants were told that "some of the profiles [they] see have been generated by a computer system using artificial intelligence, while others have been written by the host." Treatment group participants also viewed the system demo used in Studies 1 and 2.

The three treatments, different versions of the "mixed-source" environment, were designed to test under which conditions "AI" profiles are distrusted. Participants in the *labeled* condition saw a 'generated profile' label above the "AI" profiles and a 'regular profile' label above the "human" profiles. Participants in the *unlabeled* condition did not see any label identifying the profiles. Subjects in the *primed* condition were not shown any labels, but we asked them, as we had done in Study 2, to rate the AI score of a profile before they rated the host's trustworthiness. An overview of conditions is shown in Table 2. We measured the same variables as in Study 1 and 2 and computed an index for the trust ratings (Cronbach's $\alpha = .87$; $M = 69.8$, $SD = 16.7$).

Participants. We recruited 323 participants that had not participated in Studies 1 or 2 for the experiment using the procedure, parameters, and payments of Study 1. Participants' mean age was 35.6; 44% of them identified as female. We

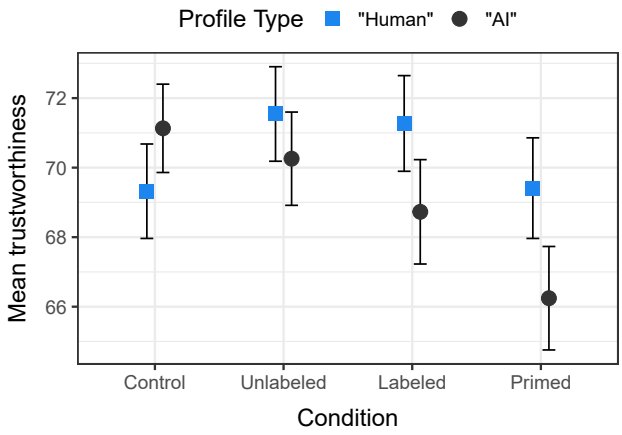

Figure 4: Study 3 trustworthiness ratings for hosts in the "AI" profile set versus hosts in the "human" profile set, across all experimental conditions

Table 3: Regression table predicting trust ratings based on profile type and treatment

	Model 1		*Model 2*	
	B	*SE*	*B*	*SE*
(Intercept)	69.321***	0.99	69.321***	1.709
"AI" type profile	1.810	1.414	1.810	1.016
Unlabeled condition	2.222	1.401	2.222	2.407
Labeled condition	1.950	1.510	1.950	2.594
Primed condition	0.089	1.434	0.089	2.463
"AI" x Unlabeled condition	-3.096	1.981	-3.096*	1.430
"AI" x Labeled condition	-4.352*	2.135	-4.352**	1.542
"AI" x Primed condition	-4.976*	2.028	-4.976***	1.464
Random effects:				*SD*
1 | Subject				11.61
N	*2,080*		*2,080*	
R^2	*0.0095*		*0.4873*	

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

performed manipulation checks and filtering tasks to exclude low-quality responses using the same procedure as in Study 1 and 2, excluding 115 participants. In addition to the checks of the prior studies, we performed a multiple-choice manipulation check after the rating task to make sure participants remembered the AI-generation. Only four participants failed the additional manipulation check, confirming that our former procedure was effective at removing low-quality responses and that the manipulation had been understood and remembered. We decided not to exclude these four participants due to their small number and the risks associated with post-randomization exclusions.

Results

Figure 4 shows the trust ratings that the different profile types received in the different treatment groups. Black circles show the mean trust ratings (and confidence intervals) of AI-like profiles, blue squares represent human-like profiles. The different experimental conditions are shown on the x-axis. We see that "AI" profiles received slightly higher ratings ($M = 71.13, SD = 10.8$) than "human" profiles ($M = 69.32, SD = 11.54$) in the control group, where participants believed all profiles were written by the hosts. However, in the treatment groups, where respondents believed some profiles were written by the host, while others were generated using an AI system, the ratings of AI-like profiles dropped considerably to their lowest observed mean of 66.24 in the *primed* condition.

We calculated a multiple linear regression of our 4x2 mixed design to estimate how the different treatments and profile types affected the trust ratings. Model 1, shown in Table 3, predicts respondents' trust ratings based on treatment (*control, unlabeled, labeled or primed*), profile type ("AI" or "human"), and their interaction. The baseline is "human" profiles in the control group. None of the main effects were significant predictors; as expected, the treatment did not have a significant effect on the evaluation of "human" profiles. However, in the *labeled* and *primed* conditions hosts with AI-like profiles received significantly lower trust ratings. Model 2 includes a random effect per subject, in order to control for participants' differing trust baselines. In the multilevel model, AI-like profile characteristics predicted significantly lower trust ratings in all treatment groups. Following our preregistration, we also conducted a 4x2 mixed ANOVA on the influence of profile type, experimental treatment, and their interaction on the trust ratings. Similar to the regression, the ANOVA reveals a significant interaction of treatment and profile type ($F(1, 1966) = 4.534, p < 0.001$).

We separately analyzed the data collected in the *primed* treatment where participants rated the profiles' AI scores. We wanted to confirm that our selection of "AI" and "human" profiles based on the pre-study aligned with the AI scores profiles received in the experiment. We find that indeed, profiles in the "AI" group received significantly higher AI scores ($M = 3.56, SD = 1.70$) than profiles in the "human" group ($M = 2.77, SD = 1.51, t(512) = -5.60, p < 0.001$), demonstrating that AI score is a reliable measure.

The *primed* condition furthermore allows us to expand on the analysis of Study 2 (Figure 3), directly re-evaluating the relationship between AI score and trustworthiness. Figure 5 shows the means and confidence intervals of ratings in the *primed* condition plotted over the associated AI scores. For example, when participants rated profiles as "definitely

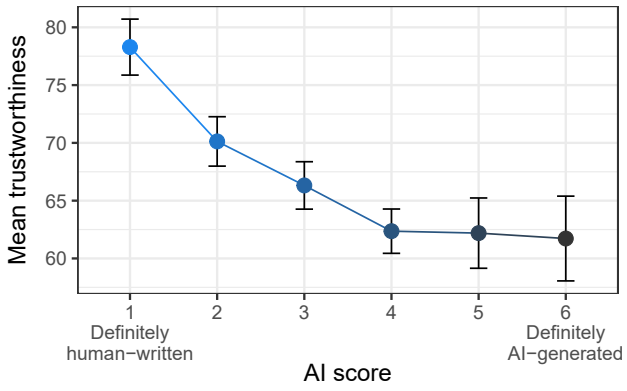

Figure 5: Trustworthiness ratings in the *primed* experimental condition by AI score assigned

human-written" they gave these profiles the highest trustworthiness rating ($M = 78.29, SD = 12.75$) – an average of 16.6 points higher than ratings they gave when they had evaluated a profile as "definitely AI-generated" ($M = 61.72, SD = 13.94$). Interestingly, we observe a floor effect: once a participant suspected a profile to be AI-generated (corresponding to an AI score of 4) the trustworthiness ratings dropped to the lowest level.

6 DISCUSSION

Taken together, the results of the three studies show a robust effect of AI-Mediated Communication on the perceived trustworthiness of hosts and give an early indication of how online self-presentation may be affected by AI-MC. To our knowledge, this is the first study to demonstrate a significant effect of perceived AI-mediation on a communication outcome, namely loss of trustworthiness.

Study 1 participants were willing to accept and trust the AI-mediation, possibly due to the uniform application of the technology. Recall that in Study 1, treatment group participants were told that *all profiles* were written by AI. This result aligns with other studies where researchers have found that people accept contributions of automated agents: In Wölker and Powell's study [62], readers rated automated and human-written news as equally credible; similarly, Edwards et al. [11] found no differences in source credibility between an otherwise identical human and bot account on Twitter.

In contrast, in Study 2 and 3, where participants encountered a *mix* of supposedly AI- and human-written profiles, respondents consistently rated profiles that were labeled or suspected to be AI-generated as less trustworthy. We term this phenomenon the *Replicant Effect*. As in the movie Blade Runner, our (experimental) world was populated by both humans and non-human agents that imitated humans–the *replicants*. Our results show a robust trend: in such a mixed-source world, the knowledge, or even suspicion, that a profile is a replicant (i.e., AI-generated) results in distrust.

While we observed this phenomenon the first time in Study 2, the results of Study 3 replicated the effect on a wider set of profiles in a randomized controlled trial. Study 3 clarified under which conditions the effect occurs. We hypothesized that the effect may be due to the additional uncertainty in the mixed-source environment: In Study 1, participants knew all profiles were AI-generated, whereas, in Study 2, they could not be sure of the source. In Study 3, however, hosts of profiles that were disclosed as 'AI-generated' still were trusted less, suggesting uncertainty did not drive the lower trust ratings. We also examined whether the distrust in AI-generation in Study 2 may have been a result of priming by forcing participants to assign AI-scores. The results of the *unlabeled* condition of Study 3 show that participants distrusted hosts with profiles that they suspected to be AI-generated even when they were not explicitly asked about AI scores, demonstrating that priming was not necessary for a *Replicant Effect*.

This result is consistent with the Hyperpersonal Model of CMC, where receivers tend to make over-attributions based on minimal cues [56, 57]. In our mixed-source environments (Studies 2 and 3), participants scrutinized profiles that were deemed AI-like and made strong negative attributions of the host based on minimal cues (Figure 5). The Hyperpersonal Model may explain why there were no such effects in Study 1: when participants encountered only one kind of source, there was no reason to use the source of the profile as a cue in their trustworthiness attributions. Further support for the Hyperpersonal Model and over-attribution is provided by the fact that highlighting differences by labeling profiles, or by making participants assign AI scores, made the *Replicant Effect* stronger and increased distrust.

The Elaboration Likelihood Model (ELM) [48] further formalizes this explanation by differentiating two major routes to processing stimuli: the *Central Route* and the *Peripheral Route*. Under the *Peripheral Route*, information is processed mindlessly, relying on basic cues and rules of thumb. The results from Study 1, where participants encountered profiles from the same source only, could be due to peripheral processing. Because the source of the profile was not salient, respondents relied on the same social cues they used to judge human-written profiles for all the profiles. In contrast, the *Central Route* involves more careful and thoughtful consideration of the information presented. The mixed-source environment with both AI and human-generated profiles may have made the source of the profile more salient, leading participants to engage in more careful processing of the profiles. Under *Central Route* processing the source of the

profile became part of the evaluation, leading to the *Replicant Effect* observed in Studies 2 and 3.

The current work clarified the conditions under which a *Replicant Effect* occurs and provided evidence that it depends on the salience of the source. The results raise the questions about why participants mistrusted profiles that they suspected were AI-generated. While the quantitative findings from Study 3 suggest that higher uncertainty or priming did not cause the effect, an examination of the open-ended responses in the studies provides some insight: Participants rarely criticized accuracy of generated profiles, maybe due to a *promise of algorithmic objectivity* [21] of AI systems. However, they often noted that AI-generated profiles lacked emotion or authenticity. Multiple participants also expressed resentment toward the host for using an AI-generated profile ("They can be handy, but also a bit lazy. Which makes me question what else they'll be lazy about."). Further studies are needed to clarify *why* AI-generated profiles are seen as less trustworthy in mixed-source environments.

A further takeaway from our results is that people have folk theories [17] about what AI-written profile text looks like. For our experimental design of Study 3, we pretested a range of profiles on their AI-scores. The ratings that profiles received in Study 3 are largely consistent with the pretest. This not only shows that AI-score is a reliable measure, but that people consistently evaluate some profiles as more human while others are consistently rated as more AI-like. Again, an informal analysis of the open-ended responses to our studies offers hints into what factors made profiles AI-like. For instance, a colloquial style of writing and personal details are taken as evidence of human writing, while bullet-style "lists of irrelevant information", as one participant put it, make profiles more "AI". Future research can expand on understanding when people suspect interpersonal communications to be written by AI. Of course, we expect that such perceptions might shift and change as people are exposed to future AI-MC technologies.

A further aspect to be explored is who owns and controls the AI technology. In this work, participants' assumptions about the nature or characteristics of the "AI" were not considered. Future studies will need to explore what kind of control and assurances users need from a system to develop trust in AI-MC.

Limitations

Our work has several important limitations. First, the context of our study was limited, as our experimental setup only explored the specific scenario of a lodging marketplace. It is not immediately clear that such findings will generalize to other online environments. Second, our studies offered strong experimental evidence of the manipulation's effect, but did not assess behavioral consequences (e.g., renting from the host). Evidence that AI-MC might cause changes in behavior is still needed. Future studies in different contexts such as dating or e-commerce will help to provide a better understanding of the *Replicant Effect*.

In addition, while we pre-tested, checked, and re-checked the manipulation, it is still possible that our manipulation is not ecologically valid. Given the novelty of AI-MC, it is not clear how, or if, AI-generated profiles will be identified or described by real systems. Furthermore, since we used human-written text and only *manipulated* the perception it was created by AI, our results are limited to understanding the perception of AI's involvement–and not based on reactions to actual AI-generated text.

Lastly, we note that while this initial examination of AI-MC exposed a robust effect of introducing AI into mediated-communication, we did not directly test theories that can explain the mechanisms behind the findings. Such investigations will be needed to help advance the *conceptual* groundwork for AI-MC and are an exciting avenue for future work.

7 CONCLUSION

We have shown a first example of the impact of AI-Mediated Communication (AI-MC) and developed an initial understanding of how people's perceptions might shift through the introduction of AI to interpersonal communication. In the context of evaluating the trustworthiness of potential Airbnb hosts based on their text profile, we found that when people face a mixed environment of AI- and human-generated profiles, hosts with profiles suspected to be AI-generated were trusted less. We termed this phenomenon the *Replicant Effect*. As technology developments bring us closer to an AI-mediated world, it is important to understand the theoretical and practical implications of this trend. Future studies can investigate AI-MC's impact on other key outcomes such as relationship formation or personal attraction, and in other contexts and media.

Issues related to AI-MC are not only important to the field of HCI but have far-reaching societal implications. The current findings suggest that there are ways to design AI-MC technologies that users find acceptable, while others lead to loss of trust. AI-MC is in its early stages today, but it may become so pervasive and sophisticated that people have to second-guess what parts of their interpersonal communications are optimized or generated by AI. To maintain trust in digital communications, we need to establish appropriate design guidelines and fair use policies before AI-MC is widespread. We hope our study and future investigation in this area will contribute to such informed policies.

REFERENCES

[1] Microsoft Blogs. 2017. Bringing AI to job seekers with Resume Assistant in Word, powered by LinkedIn. https://bit.ly/2Di34QB.

[2] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[3] Kevin Corti and Alex Gillespie. 2016. Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior* 58 (2016), 431–442.

[4] Niels Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies — Why and How. *Knowledge-Based Systems* 6, 4 (Dec. 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[5] Antonella De Angeli, Graham I Johnson, and Lynne Coventry. 2001. The unfriendly user: exploring social reactions to chatterbots. In *Proceedings of The International Conference on Affective Human Factors Design, London*. 467–474.

[6] David C DeAndrea. 2014. Advancing warranting theory. *Communication Theory* 24, 2 (2014), 186–204.

[7] Michael A. DeVito, Jeremy Birnholtz, and Jeffery T. Hancock. 2017. Platforms, People, and Perception: Using Affordances to Understand Self-Presentation on Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 740–754. https://doi.org/10.1145/2998181.2998192

[8] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. Algorithms ruin everything:# RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3163–3174.

[9] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251. https://doi.org/10.1111/j.1083-6101.2007.00394.x

[10] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech communication* 50, 8-9 (2008), 630–645.

[11] Chad Edwards, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton. 2014. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior* 33 (2014), 372–376.

[12] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication* 11, 2 (2006), 415–441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

[13] Nicole B Ellison and Danah M Boyd. 2013. Sociality through social network sites. In *The Oxford handbook of internet studies*.

[14] Nicole B Ellison and Jeffrey T Hancock. 2013. Profile as promise: Honest and deceptive signals in online dating. *IEEE Security and Privacy* 11, 5 (Sept. 2013), 84–88.

[15] Nicole B Ellison, Jeffrey T Hancock, and Catalina L Toma. 2012. Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *New Media & Society* 14, 1 (2012), 45–62.

[16] Eyal Ert, Aliza Fleischer, and Nathan Magen. 2016. Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management* 55 (2016), 62–73. https://doi.org/10.1016/j.tourman.2016.01.013

[17] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I like it, then I hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. ACM, 2371–2382.

[18] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of The ACM* 59, 7 (2016), 96–104.

[19] Jennifer L Gibbs, Nicole B Ellison, and Chih-Hui Lai. 2010. First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research* (2010), 0093650210377091.

[20] Jennifer L Gibbs, Nicole B Ellison, and Chih-Hui Lai. 2011. First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research* 38, 1 (2011), 70–100. https://doi.org/10.1177/0093650210377091

[21] Tarleton Gillespie, Pablo J Boczkowski, and Kirsten A Foot. 2014. *Media technologies: Essays on communication, materiality, and society*. MIT Press.

[22] Andreas Graefe, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19, 5 (2018), 595–610.

[23] Grammarly 2017. Free Grammar Checker - Grammarly. https://www.grammarly.com/.

[24] Jamie Guillory and Jeffrey T Hancock. 2012. The effect of Linkedin on deception in resumes. *Cyberpsychology, Behavior, and Social Networking* 15, 3 (2012), 135–140.

[25] Daniel Guttentag. 2015. Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism* 18, 12 (2015), 1192–1217.

[26] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45, 1 (2007), 1–23.

[27] Jeffrey T. Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. 2004. Deception and Design: The Impact of Communication Technology on Lying Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 129–134. https://doi.org/10.1145/985692.985709

[28] Jeffrey T. Hancock, Catalina Toma, and Nicole Ellison. 2007. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 449–452. https://doi.org/10.1145/1240624.1240697

[29] Russell Hardin. 2002. *Trust and trustworthiness*. Russell Sage Foundation.

[30] Susan C. Herring. 2002. Computer-mediated communication on the internet. *Annual Review of Information Science and Technology* 36, 1 (2002), 109–168. https://doi.org/10.1002/aris.1440360104

[31] 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*. ACM, LBW089.

[32] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14, 3 (2011), 399–425.

[33] Toko Kiyonari, Toshio Yamagishi, Karen S Cook, and Coye Cheshire. 2006. Does trust beget trustworthiness? Trust and trustworthiness in two games and two cultures: A research note. *Social Psychology Quarterly* 69, 3 (2006), 270–283.

[34] Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar Face(book): Profile elements as signals in an online social network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 435–444. https://doi.org/10.1145/1240624.1240695

[35] Airi Lampinen and Coye Cheshire. 2016. Hosting via Airbnb: Motivations and financial assurances in monetized network hospitality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 1669–1680. https://doi.org/10.1145/2858036.2858092

[36] Debra Lauterbach, Hung Truong, Tanuj Shah, and Lada Adamic. 2009. Surfing a web of trust: Reputation and reciprocity on couchsurfing. com. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 346–353.

[37] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37, Supplement C (2014), 94 – 100. https://doi.org/10.1016/j.chb.2014.04.043

[38] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 2397–2409. https://doi.org/10.1145/2998181.2998269

[39] Xiao Ma, Trishala Neeraj, and Mor Naaman. 2017. A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI.

[40] Xiao Ma, Emily Sun, and Mor Naaman. 2017. What Happens in Happn: The Warranting Powers of Location History in Online Dating. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 41–50. https://doi.org/10.1145/2998181.2998241

[41] Roger C Mayer and James H Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology* 84, 1 (1999), 123.

[42] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[43] Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 122–130.

[44] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[45] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[46] Riley Newman and Judd Antin. 2016. Building for trust: Insights from our efforts to distill the fuel for the sharing economy. http://nerds.airbnb.com/building-for-trust.

[47] Gary S Nickell and John N Pinto. 1986. The computer attitude scale. *Computers in human behavior* 2, 4 (1986), 301–306.

[48] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[49] Eva Schwämmlein and Katrin Wodzicki. 2012. What to tell about me? Self-presentation in online communities. *Journal of Computer-Mediated Communication* 17, 4 (2012), 387–407.

[50] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).

[51] Crispin Thurlow, Laura Lengel, and Alice Tomic. 2004. *Computer Mediated Communication*. Sage Publishing.

[52] Catalina L. Toma and Jeffrey T. Hancock. 2012. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication* 62, 1 (2012), 78–97. https://doi.org/10.1111/j.1460-2466.2011.01619.x

[53] Catalina L Toma, Jeffrey T Hancock, and Nicole B Ellison. 2008. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin* 34, 8 (2008), 1023–1036.

[54] Suvi Uski and Airi Lampinen. 2014. Social norms and self-presentation on social network sites: Profile work in action. *New Media & Society* (2014).

[55] T Franklin Waddell. 2018. A Robot Wrote This? How perceived machine authorship affects news credibility. *Digital Journalism* 6, 2 (2018), 236–255.

[56] Joseph Walther. 2011. Theories of computer-mediated communication and interpersonal relations. In *The Handbook of Interpersonal Communication*. 443–479.

[57] Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43.

[58] Joseph B. Walther, Tracy Loh, and Laura Granka. 2005. Let Me Count the Ways: The Interchange of Verbal and Nonverbal Cues in Computer-Mediated and Face-to-Face Affinity. *Journal of Language and Social Psychology* 24, 1 (2005), 36–65.

[59] Joseph B Walther and Malcolm R Parks. 2002. Cues filtered out, cues filtered in. *Handbook of interpersonal communication* (2002), 529–563.

[60] Joseph B Walther, Brandon Van Der Heide, Lauren M Hamel, and Hillary C Shulman. 2009. Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication research* 36, 2 (2009), 229–253.

[61] Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[62] Anja Wölker and Thomas E Powell. 2018. Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism* (2018), 1464884918757072.

[63] Toshio Yamagishi. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology* 51, 1 (1986), 110.

[64] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. 2017. Automated Crowdturfing Attacks and Defenses in Online Review Systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. ACM, New York, NY, USA, 1143–1158. https://doi.org/10.1145/3133956.3133990

[65] Georgios Zervas, Davide Proserpio, and John Byers. 2015. A first look at online reputation on Airbnb, where every stay is above average. *Social Science Research Network* (2015).

A OPEN SCIENCE REPOSITORY

The full experimental data, analysis code and experiment preregistration, as well as profiles and measures used are available from https://osf.io/qg3m2/ and https://github.com/sTechLab/aimc-chi19.

B PROFILES (STUDY 1 AND 2)

High baseline trustworthiness profiles:

(1) We own and manage several vacation rental properties. We're always readily available if you need anything at all during your stay. The homes are fully furnished with everything that you would need for a short term or long term stay. If you have any questions about Austin or the area, we'd love to help!

(2) I'm creative, fun-loving and sociable whilst respecting the space and peace of others. I'm told my home is a reflection of that with a homely feel surrounded with art, music beautiful plants, fruit trees and good vibes! Please note my reviews mention my London home (my home in Los Angeles is just as beautiful!) ;)
(3) I'm a professional working full time as an IT consultant in a Healthcare company. I like hosting. I enjoy exchanging conversations and sharing the experiences of life with good people who are all around the world. I enjoy cooking Indian food and feeding my guests is my favorite hobby :)
(4) We have lived in Seattle for over 20 years! We enjoy traveling with our family in search of what makes a place unique and special. We know many great hidden spots all over Seattle to make your stay special no matter what your age and interests!
(5) I'm a mother of two, a writer and a drummer. I am a native of the Boston area. I worked in the Hospitality Industry for 20 years and can help you find whatever you need in Boston!

Low baseline trustworthiness profiles:

(6) Hey There! I'm Denise. Just a laid back professional making my way in the city. Raised a military brat, I'm a nomad at heart. Lived in Japan, Germany, and Greece and have traveled all over the world.
(7) My stepfather, Raul, and I are from the Dominican Republic. We are friendly and accommodating. I came to NYC when I was 7 years old. I've lived in Manhattan and the Bronx. When I'm not at work I like to relax at home or go out with my girlfriend.
(8) I live here in Hancock park with my older brother. We are songwriters and producers developing our careers here. On the side I go to school for Envl Sci, and my bro does graphic design. We are very laid-back dudes.
(9) Organic gardener with purpose of raising my own food. Owner of two cats. My 17 year old dog left me a few years ago. Music fan and know my way around when the city is crowded. Kayaking.
(10) I am Sharma, a simple and single guy. Monday to Friday I like to stay busy at my work and Saturdays and Sundays are party night, love going out for fun food, music, bars and lounges. Most of my time goes in market and working. In summer time like going beaches and water parks and in winters on Mountains.

C PARTICIPANT DEBRIEF

"Thank you for participating in this study. In order to get the information we were looking for, we provided you with incorrect information about some aspects of this study. Now that the experiment is over, we will describe the deception to you, answer any of your questions, and provide you with the opportunity to make a decision on whether you would like to have your data included in this study.

What the study really is about is how the knowledge that text profiles have been enhanced by artificial intelligence may impact their perceived trustworthiness. To test this, we have given you the impression that the profiles you saw were generated by AI. However, the profiles you have seen were written by real hosts and have not been generated.

The findings of this study will help to better understand the potential benefits and risks that algorithmic mediation brings to our societies. For example, will people lose trust in each other's words as their communication becomes more AI-improved? We will analyze the answers given to provide some first insight. For any questions, please contact <mpj32@cornell.edu>.

Although you have already completed the survey, your involvement is still voluntary, and you may choose to withdraw the data you provided prior to debriefing without penalty. Withdrawing your submission will not adversely affect your relationship with Cornell University, the researchers, or any of our affiliates."