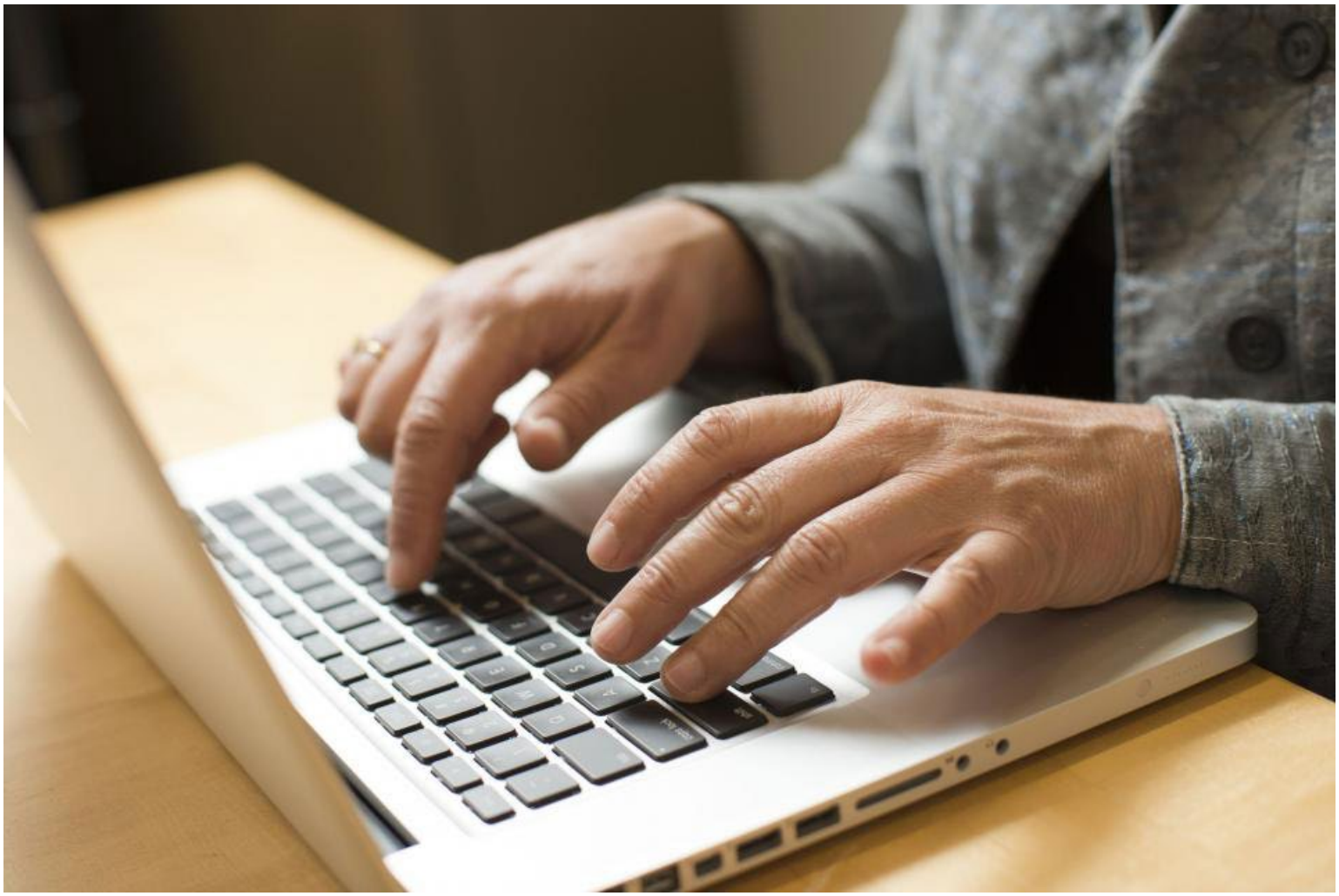
When AI Writes Your Email

KATHARINE MILLER May 6, 2020

Artificial intelligence is changing the way we communicate with each other, leading to questions of trust and bias.



LINDA A. CICERO / STANFORD NEWS SERVICE

Wording suggestions in email or text could alter the way we communicate and create bias.

Like Cyrano de Bergerac writing love letters on behalf of a friend, artificial intelligence (AI) tools are polishing our writing styles to make them more convincing, authoritative and trustworthy.

Spell check and autocorrect came first, fixing a few mistakes in our texts and emails. But now Gmail suggests entire sentences for us to use in emails or texts, while other AI tools polish our online profiles for Airbnb postings, job applications and dating websites. Down the line, AI systems might even send messages on our behalf with only minimal involvement on our part.

As the level of AI involvement in human-to-human communication grows, so too does the need for research into its impacts.

“There are interesting implications when AI starts playing a role in the most fundamental human business, which is communication,” says Jeff Hancock, the Harry and Norman Chandler Professor o Communication at Stanford University and founding director of the Stanford Social Media Lab.

In “AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations,” published in the *Journal of Computer-Mediated Communication* in January, Hancock and two Corne colleagues reflect on what happens when AI tools come between people and act on their behalf, fro how wording suggestions could alter our use of language and bake in bias, to the impact these communications could have on relationships and trust.

Language Change at Scale

For several years, Gmail users have had access to an AI tool called “Smart Reply,” which offers thre short email reply options for any email. For example, in response to an email proposing a meeting time, Gmail might suggest such replies as “Sounds good!”, “See you then!” or “Tuesday works for me!”

Recent research out of Cornell found that the language of Gmail Smart Reply tends to be overly positive rather than either neutral or negative (a result that was recently replicated by Stanford researchers). It’s even possible that the overly positive phrasing of Gmail Smart Reply primes recipients to respond in kind, with something like “Cool, it’s a plan!” even though they don’t know tha AI is involved. Since tens of millions of messages are sent by Gmail users every day, this tendency could lead to language change at an unforeseen scale: Our language might evolve toward Google’s optimistic tone.

“The simple bias of being positive has implications at Google scale,” Hancock says. “Maybe it’s no b deal and it’s what people would have said anyway, but we don’t know.”

During the current shutdown due to the Covid-19 pandemic, Hancock wonders whether the use of Gmail Smart Reply tools will decline because their positivity seems less appropriate. “It’s not as common to say ‘See you later,’” he notes. “Now the common signoff is more likely ‘Stay safe’ or ‘Be well.’ Will AI pick up on that?” Yet another item on the research agenda.

Built-In Bias

Natural language AI tools are typically built on a dataset consisting of a bucket of words and the various ways they have been assembled into sentences in the past. And these tools are designed to

optimize for a specific goal, such as trustworthiness or authoritativeness. Both of these aspects of AI can build bias into an AI-mediated communication system. First, the word bucket might not include a diversity of communication styles. Second, the optimization step might promote the communication style used by the dominant group in the culture. “If AI is optimizing for sounding authoritative, then everyone will be made to sound like older white males,” Hancock says.

For example, Hancock says, one can imagine that a young black woman might overcome the racial and gender biases she faces by writing a job application using an AI tool that optimizes for the authoritative communication style of an older white male. But this benefit might well come at some cost to her self-expression while also reinforcing the privileged status of the dominant group’s language usage.

Trust and Transparency

One recent study by Hancock and his colleagues hints at the issues of trust raised by AI-mediated communication. The researchers asked whether the belief that online Airbnb host profiles were written by AI affects how readers view them. Initially, readers trusted host profiles equally regardless of whether they were told they were written by humans or with AI assistance. But when readers were told that some were written by humans and others by AI, their level of trust fell for any profile that seemed formulaic or odd and therefore seemed more likely to have been written by AI.

“If people are uncertain and can’t tell if something is AI or human, then they are more suspicious,” Hancock says. This suggests that transparency (or lack of transparency) about the role that AI plays in our interpersonal communications may affect our relationships with one another.

Agency and Responsibility

Just as humans delegate agency to lawyers, accountants and business associates, humans are now delegating agency to AI communication assistants. But these agents’ responsibility for errors is unclear and their interests aren’t always fully aligned. “The agent is supposed to work on my behalf, but if it belongs to Google, does it have a separate interest adverse to my interest?” Hancock asks.

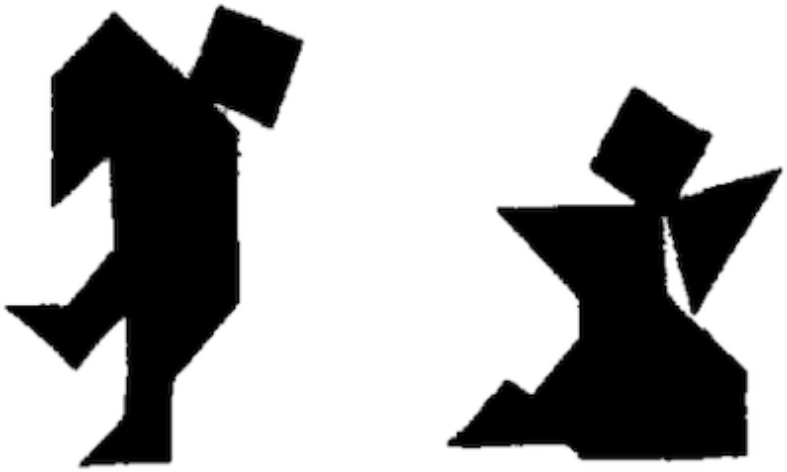
In *Cyrano de Bergerac,* the articulate Cyrano wants to win Roxane for himself while writing love letters to her on behalf of a man who can barely put a sentence together.

As AI systems become more sophisticated, will they step in as a personal Cyrano? And if so, will the responsible author be the human or AI? What happens if the human author behind an articulate

message proves to be a fool?

Sounds Good! Google Smart Reply's Positive Spin

To study how AI-assisted messaging affects language use, Hannah Mieczkowski, a PhD candidate i communication at Stanford University, recruited 35 pairs of strangers to perform a matching task using a Google messaging app. One member of each pair (the director) had to explain to the other (the matcher) how to arrange a set of 10 complex shapes in a particular order (see below). For half o the pairs, the text conversation involved no AI-assistance. But for the other half, the director was required to use Google Smart Reply, which suggests three responses.



Mieczkowski then used several standard language evaluation tools to analyze the conversation logs – including the triplet of smart replies that the directors were offered as well as the ones they actually chose. Preliminary analysis of the smart reply suggestions confirms prior work showing that the smart reply triplets ar overwhelmingly positive rather than negative. "We love to see replication," Mieczkowski says. But this work also goes further and compares the language of directors who were communicating with AI assistance to those who communicated without it. "That language was more positive as well," Mieczkowski says. That is, the smart replies actually changed the tone of the conversation.

Mieczkowski will also look at whether the positivity of the smart replies induced the matchers to respond in kind. "Can the AI actually help promote the rate at which language style matching happens?" she asks. "Is the matcher's language becoming more similar to their partner's?" Her findings will be presented in May 2020 at the International Communication Association's annual conference.