PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES

Human heuristics for AI-generated language are flawed

Maurice Jakesch[a,b,1], Jeffrey T. Hancock[c], and Mor Naaman[a,b]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received June 29, 2022; accepted December 27, 2022

Human communication is increasingly intermixed with language generated by AI. Across chat, email, and social media, AI systems suggest words, complete sentences, or produce entire conversations. AI-generated language is often not identified as such but presented as language written by humans, raising concerns about novel forms of deception and manipulation. Here, we study how humans discern whether verbal self-presentations, one of the most personal and consequential forms of language, were generated by AI. In six experiments, participants (N = 4,600) were unable to detect self-presentations generated by state-of-the-art AI language models in professional, hospitality, and dating contexts. A computational analysis of language features shows that human judgments of AI-generated language are hindered by intuitive but flawed heuristics such as associating first-person pronouns, use of contractions, or family topics with human-written language. We experimentally demonstrate that these heuristics make human judgment of AI-generated language predictable and manipulable, allowing AI systems to produce text perceived as "more human than human." We discuss solutions, such as AI accents, to reduce the deceptive potential of language generated by AI, limiting the subversion of human intuition.

human–AI interaction | language generation | cognitive heuristics | risks of AI

Significance

Human communication is now rife with language generated by AI. Every day, across the web, chat, email, and social media, AI systems produce billions of messages that could be perceived as created by humans. In this work, we analyze human judgments of self-presentations written by humans and generated by AI systems. We find that people cannot detect AI-generated self-presentations as their judgment is misguided by intuitive but flawed heuristics for AI-generated language. We demonstrate that AI systems can exploit these heuristics to produce text perceived as "more human than human." Our results raise the question of how humanity will adapt to AI-generated text, illustrating the need to reorient the development of AI language systems to ensure that they support rather than undermine human cognition.

Large generative language models (1, 2) produce semantic artifacts closely resembling language created by humans. Through applications like smart replies, writing autocompletion, grammatical assistance, and machine translation, AI-enabled systems infuse human communication with generated language at a massive scale. Large language models like OpenAI's GPT-3 and AI language applications like ChatGPT (1, 2) produce coherent writing pieces and generate entire conversations. AI-generated language enables novel interactions that reduce human effort but can facilitate novel forms of plagiarism, manipulation, and deception (1, 3–8) when people mistake AI-generated language for language created by humans.

In a series of experiments, we analyzed how humans detect AI-generated language in one of the most personal and consequential forms of speech—verbal self-presentation. Self-presentation refers to behaviors designed to control impressions of the self by others (9), while verbal self-presentation focuses on the words used to accomplish impression management. In this work, we operationalize self-presentation as self-descriptions of the type prevalent in online profiles (10), e.g., on professional or dating platforms. Researchers have extensively studied the importance of online self-presentation (11–13), showing that impression formation based on self-descriptions is crucial for establishing the trust required for various social interactions (14, 15). AI systems that generate human-like self-presentations may invalidate signals that people rely on when assessing others (16), such as tone or compositional skill. Earlier work on AI-mediated communication (16) has shown that interpersonal trust declines when people suspect that others are using AI systems to generate or optimize their self-presentation (17).

Previous studies suggest that people struggle to discern AI-generated language in different settings (18–20). Here, we go beyond prior work by providing strong evidence that people use flawed heuristics to detect AI-generated language. Using qualitative, quantitative, and computational methods, we reconstruct a set of potential heuristics that people may rely on to detect AI-generated language, expanding on related analyses in previous work (18). We then measure the extent to which people actually use these heuristics and whether the heuristics help or hinder their attempts to distinguish between human- and AI-generated language. Finally, we demonstrate that AI systems can predict and manipulate whether people perceive AI-generated language as human.

Results

To examine how people detect AI-generated self-presentations, we performed six experiments broadly patterned after the Turing test (21). While participants in the original test were asked to identify a language-generating machine through a text-based conversation, participants in our studies were asked to judge whether a personal self-presentation was written by a person

Author affiliations: [a]Department of Information Science, Cornell University, Ithaca, NY 14850; [b]Jacobs Institute, Cornell Tech, New York, NY 10044; and [c]Department of Communication, Stanford University, Stanford, CA 94305

Preprint Servers: arXiv.org perpetual, nonexclusive license (https://arxiv.org/abs/2206.07271).

Author contributions: M.J. and M.N. designed research; M.J. performed research; M.J. analyzed data; M.N. consulted the analysis; and M.J., J.T.H., and M.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mpj32@cornell.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2208839120/-/DCSupplemental.

Published March 7, 2023.

or generated by an AI system. We trained multiple customized versions of state-of-the-art AI language models (1, 2, 4) to generate self-presentations in three social contexts where trust in a self-presentation is important for decision-making: professional (e.g., job applications) (22), romantic (e.g., online dating) (12), and hospitality services (e.g., Airbnb host profiles) (15). Across three main and three validation experiments, we asked 4,600 participants to read through a total of 7,600 self-presentations—some AI-generated and some collected from real-world online platforms—and indicate which ones they thought were generated by AI.

We start by computing the accuracy rates for participants' ability to distinguish between human and AI-generated self-presentations. In our three main experiments, using two different language models to generate verbal self-presentations across three social contexts, participants identified the source of a self-presentation with only 50 to 52% accuracy. These results, with a breakdown by experiments and treatments, are shown in Fig. 1. In the hospitality context (shown in the *Left* panel), participants correctly identified the source of a self-presentation 52.2% of the time. In the dating context, we introduced experimental treatments testing whether incentivizing participants to increase their efforts (23) would increase their accuracy. In the professional context, we tested whether providing training (18) in the form of feedback would improve participants' judgments. However, participants' accuracy remained close to chance even when offered monetary incentives for accurate assessments (right bar in the second panel in Fig. 1, 51.6%) and when receiving immediate feedback on their evaluations (right bar in the third panel, 51.2%). Further analyses (included in *SI Appendix*) revealed that no demographic group performed better than others.

Participants' evaluations were not random, however. The observed agreement between participants' judgments was significantly higher than chance (Fleiss' kappa = 0.07, $P < 0.0001$). As the observed accuracy was close to chance, the agreement in participants' assessments must have been due to shared but flawed heuristics that participants relied on to identify AI-generated language. To investigate participants' heuristics for AI-generated language, we next conducted a qualitative analysis of the heuristics participants thought they relied on.

After completing half of the ratings, we asked participants to explain one of their judgments. Two researchers independently coded a sample of their responses and grouped them into themes: content, grammar, tone, and form. These themes are extending categories identified in previous research (18). Participants commonly referred to the content of a self-presentation (40% of responses): Self-presentations with specific content related to family and life experiences led many to infer a human author. Participants also referred to grammatical cues (28%), where first-person pronouns and the mastery of grammar were seen as indicative of language created by humans. Replicating findings from earlier research (18), grammatical errors were associated with a subpar AI by some participants but with fallible human authors by others. Participants also judged the self-presentation source by its tone (24%), associating warm and genuine language with humanity and impersonal, monotonous style with AI-generated language. Details on participants' self-reported explanations of their judgments are included in *SI Appendix*.

As self-reports on mental processes can be unreliable and even misleading (24), we conducted additional analyses to evaluate participants' judgments independently of their self-reported explanations. While participants may not always know why they did something (25), a multiparadigm approach (26) based on a statistical analysis of their judgments combined with a computational analysis of language features present in the self-descriptions allows us to independently reconstruct heuristics they rely on (27). Rather than drawing conclusions from participants' self-reported heuristics like previous research (18), we used their self-reports as a starting point for extracting potentially relevant language features from the self-presentation texts. We computationally created a range of language features present in the self-presentations, including measurements for personality, sentiment, and perspective (28, 29). We also conducted an additional labeling task to create language features that could not be reliably computed.

For the feature labeling task, we recruited a separate sample of 1,300 crowdworkers. We asked them to read 12 self-presentations and indicate whether they were nonsensical, had grammatical issues, or seemed repetitive. Two to three crowdworkers (M = 2.3) evaluated each of the 7,000 human-written and AI-generated

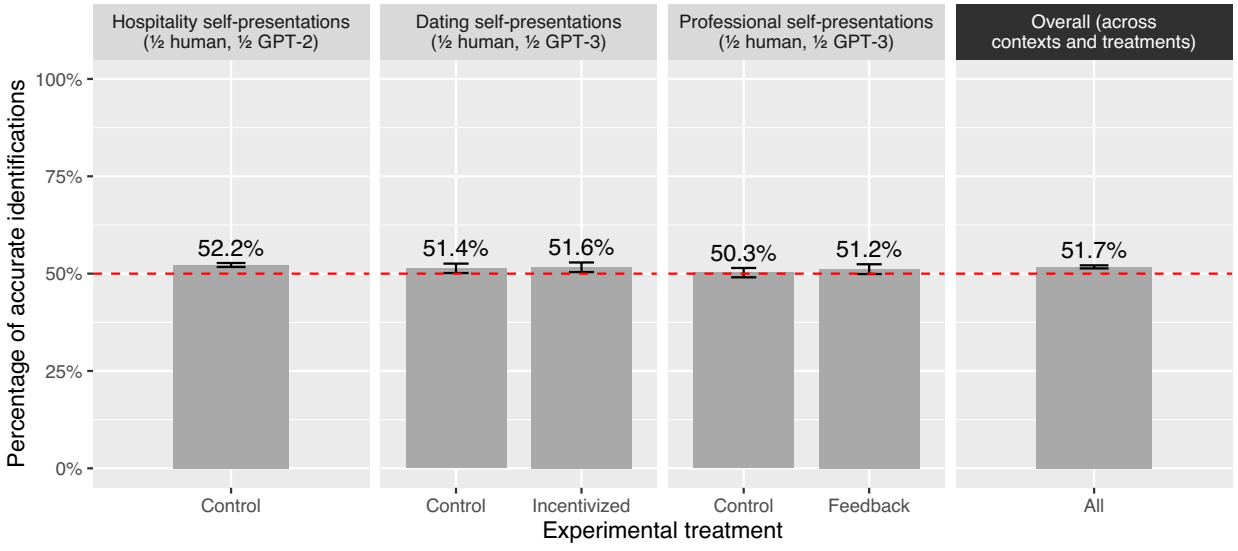

Fig. 1. Participants could not detect self-presentations generated by the current AI language models beyond chance in the three main experiments. Error bars represent 95% CIs for 6,000 to 16,000 judgments of 2,000 to 3,000 self-presentations per bar. Across three social contexts, discernment remained close to chance. Providing monetary incentives for accurate answers or telling participants whether their answers were correct did not increase accuracy.

Table 1. Logistic regression models predicting (1) whether participants in the three main experiments rated a self-presentation as AI-generated and (2) whether a self-presentation was actually generated by AI. Only nonsense, repetition, and conversational words were functional cues (top section), indicated by aligned odds ratios in models (1) and (2). The remaining features indicative of participants' heuristics were either inversely related (center) or unrelated (bottom) to features indicative of the actual source of the self-presentation.

	Dependent variable	
	(1) Perceived as AI-generated (odds ratios with 95% CI)	(2) Actually AI-generated (odds ratios with 95% CI)
Aligned features		
Nonsensical content †	1.105*** (1.085, 1.126)	1.233*** (1.169, 1.296)
Repetitive content †	1.083*** (1.059, 1.106)	1.470*** (1.379, 1.561)
Conversational words	0.947*** (0.925, 0.970)	0.898** (0.829, 0.967)
Misaligned features		
Grammatical issues †	1.048*** (1.028, 1.069)	0.851*** (0.788, 0.913)
Rare bigrams	1.042*** (1.019, 1.065)	0.666*** (0.596, 0.736)
Long words	1.034** (1.009, 1.059)	0.783*** (0.706, 0.861)
Contractions	0.947*** (0.924, 0.970)	1.134*** (1.065, 1.203)
Nonindicative		
Second-person pronouns	1.059*** (1.038, 1.079)	0.970 (0.908, 1.032)
Filler words	1.009 (0.990, 1.027)	1.119* (1.021, 1.218)
Swear words	0.969** (0.948, 0.989)	0.965 (0.905, 1.024)
Authentic words	0.946*** (0.921, 0.971)	0.945 (0.870, 1.021)
Focus on past	0.938*** (0.917, 0.959)	1.002 (0.940, 1.064)
First-person pronouns	0.925*** (0.886, 0.963)	0.992 (0.868, 1.117)
Family words	0.910*** (0.889, 0.932)	1.014 (0.950, 1.077)
Word count	0.904*** (0.874, 0.935)	1.076 (0.986, 1.165)
Constant	0.850*** (0.830, 0.870)	1.007 (0.947, 1.068)
Observations	38,866	4,690
Log likelihood	−26,318.460	−3,029.542
Akaike Inf. Crit.	52,670.930	6,093.085

Note: †manually labeled feature, *P**P***P < 0.001.

self-presentations used in the main experiments. The results indicate that crowdworkers' ratings in the labeling task, to some extent, differentiated between human-written and AI-generated self-presentations. Crowdworkers rated AI-generated self-presentations as nonsensical more often than human-written self-presentations (13.6% vs. 9.6%, $P < 0.0001$). They also rated AI-generated self-presentations as more repetitive (12.7% vs. 7.1%, $P < 0.0001$) and found fewer grammatical issues with AI-generated self-presentations than with human-written self-presentations (14.8% vs. 19.6%, $P < 0.0001$). These rates differed somewhat between contexts (*SI Appendix*). We explored the potential of the rating task labels to distinguish AI- and human-written self-presentations. We created a classifier that predicted that a profile was generated when at least one in three raters in the labeling task marked it as nonsensical or repetitive. The classifier predicted the source of a self-presentation with 58.8% accuracy, compared to the 51.7% accuracy participants achieved in the main experiment when directly asked about the source of the self-presentations.

With the language features we created—both computationally and through the labeling task—we quantitatively tested whether the presence of these features was associated with participants' judgments in the main experiments. After a feature selection process, we fit a regression model correlating selected features with participants' perception that a self-presentation was generated by AI. We fit a second model to understand whether the same features are indeed predictive of AI-generated self-presentations. The results suggest that participants relied on several cues in their ratings, some valid and others flawed. Table 1 shows which features were predictive of self-presentations being perceived as AI-generated (on the left) and which features were actually predictive of AI-generated self-presentations (on the right) in the three main experiments.

Some heuristics participants relied on to identify AI-generated self-presentations were indeed indicative of such language. For example, the odds ratios in the top row in Table 1 indicate that self-presentations containing nonsensical content were 10.5% more likely to be seen as AI-generated (left) and, indeed, were 23% more likely to be generated by AI (right). Similarly, self-presentations with repetitive content were 8% more likely to be rated as AI-generated and 47% more likely to be AI-generated in our experiments. However, most heuristics participants relied on were flawed: Participants were 5% more likely to rate self-presentations with grammatical issues as AI-generated, although grammatically flawed self-presentations were, in fact, 15% less likely to be AI-generated. Participants often rated self-presentations with long words or rare bigrams as generated by AI, while most self-presentations with long words or rare bigrams had been written by humans. Participants also judged first-person speech and family content as more human. However, these cues were not significantly associated with either AI or human-written language. Similarly, self-presentations that were longer, that included authentic or spontaneous words (30), or were focused on past events were more likely to be rated as human by participants. However, these features were not significantly associated with human-written or AI-generated self-presentations in our data.

Following the correlation analysis, we tested whether the presence of language features in a self-presentation could predict participants' judgments. A regression model based on the features above predicted participants' judgments with 57.6% accuracy when evaluated on a hold-out data set. We also tested whether AI language models can learn to predict human impressions of AI-generated language without feature engineering input from the research team. A current language model (31) with a sequence classification head predicted participants' assessments of AI-generated language with 58.1% accuracy when evaluated on hold-out validation data. These results suggest that the flawed heuristics people rely on to detect AI-generated language allow AI systems to predict their judgments, at least to some extent.

We conducted three additional experiments to validate and extend these findings: If the three main experiments correctly identified features people associate with self-descriptions that are written by humans, self-presentations selected based on the presence of these features would be more likely to be perceived as human-written in independent validation experiments. The validation studies thus tested whether language models can exploit people's flawed heuristics to produce self-presentations perceived as "more human than human." For these validation experiments, we created an additional sample of human-written and AI-generated self-presentations; and used the classifiers trained on participants' judgments in the main studies to create a set of AI-generated self-presentations optimized for perceived humanity.

The Fig. 2. shows that participants evaluated the AI-generated self-presentations optimized for perceived humanity as more human than the human-written and the nonoptimized AI-generated self-presentations. Across all three validation experiments (aggregated in the panel on the *Right*), optimized self-presentations were rated as human more often than regular generated self-presentations (65.7% vs. 51.6%, $P < 0.0001$). The optimized self-presentations were also more likely to be seen as human than self-presentations that were actually written by humans (65.7% vs. 51.7%, $P < 0.0001$). When creating the optimized self-presentations, we used different classifiers in each context to increase generalizability and to independently validate both the regression and language- model-based classifiers. The increase in perceived humanity of optimized self-presentations was strongest in the professional context, where a combination of the regression- and language-model-based classifiers produced self-presentations that were perceived as human 71% of the time.

Discussion

Our results reaffirm that humans are not able to detect verbal self-presentations generated by current AI language models. Across contexts and demographics, and independent of effort and expertise, human discernment of AI-generated self-presentation remained close to chance. These results align with recent work showing that humans struggle to detect AI-generated news, recipes, and poetry (18–20), suggesting that the era of the Turing test may be coming to an end. Our results go beyond earlier efforts by providing an empirically grounded explanation of why people fail to identify AI-generated language. Drawing on the extensive literature on deception detection (32–34), we consider two explanations for people's inability to detect AI-generated self-presentation: First, the language generated by state-of-the-art AI systems may be so similar to human-written language that a lack of reliable cues limits accuracy. Second, people's judgments may be inaccurate because they rely on flawed heuristics to detect AI-generated language.

The results of a separate labeling task we conducted suggest that the AI-generated self-presentations in our studies had certain features that people, in principle, may be able to detect. The participants in the labeling task rated AI-generated self-presentations as nonsensical and repetitive significantly more often than human-written self-presentations. This finding contradicts the idea that AI-generated language has become entirely indistinguishable from human-written language: While future generations of AI language technologies may change this, the language generated by AI technologies available at the time of the study had some human-detectable features. Yet, when we directly asked participants whether self-presentations were AI-generated in the main experiments—rather than asking them whether self-presentations were nonsensical or repetitive—the accuracy of their judgments remained close to chance.

Our analysis of the heuristics people used to identify AI-generated language provides a more nuanced picture than previous research: While people can sometimes identify certain

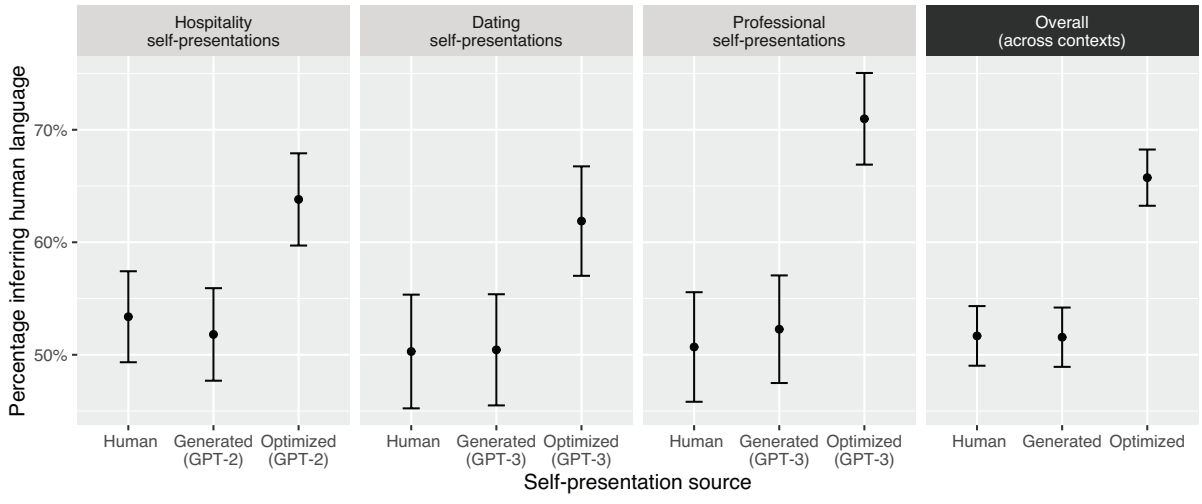

Fig. 2. The three validation experiments show that AI systems can exploit humans' flawed heuristics to generate optimized self-presentations (rightmost in each pane) more likely to be perceived as human than human-written self-presentations (leftmost) and regular AI-generated profiles (center). Error bars represent 95% CIs for 350 to 450 judgments of 100 self-presentations per bar.

characteristics of AI-generated language, they rely on other flawed cues that simultaneously impair their judgment. Participants in our studies relied to some extent on functional cues, such as nonsensical and repetitive text, to identify AI-generated self-presentations. Had participants relied on those cues only, they could have achieved a detection accuracy of 58.8%. However, participants also relied on cues like grammatical issues, rare bigrams, or long words to identify AI-generated language, although those cues were more indicative of human-written language in our data. Most other language features that participants relied on to identify human-written language, such as family words or first-person pronouns, were equally present in human-written and AI-generated self-presentations. These misleading heuristics reduced people's accuracy in detecting AI-generated self-presentations to chance, partially explaining why people in our research and in previous work failed to identify AI-generated language (18–20).

People's reliance on flawed intuitive heuristics to detect AI-generated language demonstrates that the increased human-likeness of AI-generated text is not necessarily indicative of increased machine intelligence. For example, emphasizing family topics does not require advances in machine intelligence but does increase the perceived humanity of AI-generated self-presentations. Recent work by Ippolito et al. (35) suggests that language model–decoding methods have been optimized for fooling humans at the cost of introducing statistical anomalies easily detected by machines. Previous research also suggests that domain expertise may be somewhat more effective than personal intuition in identifying AI-generated content (23). Rather than interpreting human inability to detect AI-generated language as an indication of machine intelligence, we propose to view it as a sign of human vulnerability. People are unprepared for their encounters with language-generating AI technologies, and the heuristics developed through media exposure and other social contexts are dysfunctional when applied to state-of-the-art AI language systems.

People's inability to detect AI-generated language has important consequences: As demonstrated in the three validation experiments, AI systems can use people's flawed heuristics to manipulate their judgments and produce language perceived as "more human than human." Previous work has shown that not only are people more likely to disclose private information to and adhere to recommendations by nonhuman entities that they perceive as human (36) but they may start distrusting those they believe are using AI-generated language in their communication (17). People's heuristics also can be exploited by malevolent actors. From automated impersonation (8) to targeted disinformation campaigns (37, 38), AI systems could be optimized to undermine human intuition, exacerbating concerns about novel automatized forms of deception, fraud, and identity theft (3–8). Further, new widely available applications like ChatGPT allow anyone to generate human-like text tailored to certain tasks in any requested style (e.g., informal), lowering the barrier to automatically creating language that is deceptively human.

Widespread AI education and technical tools that assist identification (39–41) might improve people's ability to detect AI-generated language to some extent. However, the potential for improving human intuition for the detection of AI-generated language is likely limited (18), and future adaptations of language models may invalidate learned heuristics (35). At the same time, how to transparently identify the use of AI systems in communication is an open and challenging problem. A recent blueprint for an AI Bill of Rights from the US White House calls for "Notice and Explanation" when "an automated system is being used" (42). Similarly, a regulation proposal issued by the EU states that "if an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means" (43). However, such policies can be difficult to apply in AI-mediated communication (16) where AI technologies modify, augment, or generate communication between people. For example, it hardly seems necessary to add notice to every message people write with AI-enabled autocorrections, smart replies, or translations. Research also shows that typical notice and consent disclosures are largely ignored by users (44).

Identifying context-appropriate and effective disclosure mechanisms for the use of AI in communication is an urgent question that requires further research (45). Our results suggest that one could develop AI language technologies that are self-disclosing by design: Rather than training AI language systems to imitate human language, AI systems could be optimized to fulfill their specific communicative function while preserving the validity of human intuitive judgment (35). Many AI applications could use language that is clearly not written by humans without loss of functionality. AI language systems could reduce the risks of false identifications by producing language that humans intuitively connect to AI sources and avoiding language that people wrongly associate with humanity, such as informal and colloquial speech. Disclosures that preserve the fluidity of communication and support human goals in communication could also be achieved through dedicated AI accents: Requiring AI systems to generate language with a dedicated dialect or accent could facilitate people's intuitive judgments without interrupting the flow of communication. Rather than undermining human cognition, AI language technologies that, by design, accommodate the limits and flaws of human judgment may genuinely support human communication and reduce the risk of misuse.

Materials and Methods

Experiment Design. The six experiments combined elements of a simplified Turing test (21) with a data labeling task. After providing informed consent, participants were introduced to the hospitality, dating, or professional scenario. They were told that they were browsing an online platform where some users had written their self-presentations while an AI system generated other self-presentations. Participants completed two comprehension checks and rated 16 self-presentations, half generated by a state-of-the-art AI language model. They were asked to evaluate whether each self-presentation was generated by AI on a five-point Likert scale from "definitely AI-generated" to "definitely human-written." Mirroring truth default behaviors observed in deception research (34), participants marked the self-presentations as "likely human-written" or "definitely human-written" in 53.8% of cases. In the remaining 46.2% of cases, they showed suspicion and selected either "not sure," "likely AI-generated," or "definitely AI-generated." To allow for concise analysis, we used these two roughly balanced groups to create a binary signal corresponding to participants' suspicion that a self-presentation may not be human-written. A robustness check using the full scale as the primary outcome metric showed similar results. Halfway through the rating task, participants in the three main experiments were asked to explain their judgment in an open-ended response. Asking participants to explain their reasoning did not change the accuracy of their subsequent ratings (see *SI Appendix* for details). Following the rating task, participants provided demographic information and indicated their experience with computer programming and AI technologies. Participants were debriefed about their performance and the purpose of the study. The Cornell University Institutional Review Board approved the study protocols. We preregistered the final two validation experiments prior to data collection (https://aspredicted.org/bz7x7.pdf).

We performed the experiments in three social contexts to increase robustness and generalizability. In addition, minor variations across experiments explored auxiliary hypotheses. We used longer self-presentations in the dating- and professional-context experiments to test whether the length of self-presentations limited participants' accuracy. To keep the three main experiments' duration comparable, we reduced the number of rated self-presentations to 12 in these two experiments. To explore the effect of increased effort (23), we offered half

of the participants in the dating context a bonus payment if they rated at least 75% of the self-presentations correctly. There was no difference in performance between the bonus and no-bonus groups. Finally, to test whether participants could learn to detect generated self-presentations if they received feedback (18), half of the participants in the professional context were told whether their choice was correct after every rating, again with no difference in outcomes. An overview of the experimental designs is included in *SI Appendix*.

Collecting and Generating Self-Presentations. We collected data from real-world platforms in each of the three social contexts for the experiments. The data collected were used in two ways: A subset was shown to participants in the experiments, and the full data were used to train state-of-the-art large language models to generate self-presentations. We employed different AI models for generating self-presentations as new and more powerful models became available over the course of this research, providing further generalizability of our findings. An overview of the models used and the setup of each experiment is included in *SI Appendix*.

For the main experiment in the hospitality context, we collected 28,890 verbal self-presentations that contained at least 30 and no more than 60 words from host profiles on Airbnb.com. We drew a random sample of 1,500 human-written self-presentations for the experiment. We fine-tuned a 774M-parameter version of GPT-2 (31) for four epochs with a learning rate of 0.00002 on the collected data. We used the fine-tuned model and nucleus sampling (46) at $P = 0.95$ to produce 1,500 AI-generated hospitality self-presentations. In the professional context, we collected 37,450 profile self-presentations with at least 60 and no more than 90 words from Guru.com, a platform where companies find freelance workers for commissioned work. In the dating context, we used a publicly available dataset of 59,940 OkCupid.com self-presentation essays collected with the platform operators' permission (47). We drew a random sample of 1,000 human-written self-presentations for the professional and dating main experiments. We used the full set of collected self-presentations in each of these contexts to fine-tune a 13B-parameter version of GPT-3 (1) for four epochs with a learning rate multiplier of 0.1. We used these fine-tuned models to produce 1,000 AI-generated self-presentations for each experiment with temperature sampling at $t = 0.9$.

We confirmed that there were no duplicate self-presentations and used multiple techniques to check that the models did not plagiarize the training data. For example, we searched for identical sentences in the training data and AI-generated text and found that 95% of sentences in the AI-generated texts were not present in the training data. As we found no signs of substantial plagiarism, we used the AI-generated self-presentations without further preprocessing.

Predicting Responses and Optimizing Self-Presentations. We developed a set of text-based language features for the quantitative language analysis of participants' judgments in the three main experiments. The full set of about 180 features is included in *SI Appendix*. We used two approaches to create these features: One set of language features were computational features that could be automatically extracted from the text. For the computational features, we manually developed measures motivated by participants' explanations of their judgments. To this initial set, we added readability scores, emotion language classification, and other psychological language features (29). We relied on a labeling task for features that could not be reliably computed. We created three additional key features by recruiting crowdworkers ($N = 1,300$) to label which self-presentations seemed nonsensical, contained repetitive text, or had grammatical issues.

For the prediction task, to reduce overfitting and increase interpretability, we reduced the set of relevant features to 15 in a feature selection process based on lasso regression performed on 20% of the self-presentations. Table 1 reports the coefficients of a logistic regression model fitted to 4,900 self-presentations (70%) that were not used for feature selection. In addition, to test whether modern language models can learn to predict human perceptions of AI-generated language without predeveloped features, we trained a large language model with a sequence classification head on 4,900 self-presentations to predict participants' judgments. We trained the 117M parameter version of GPT-2 (31) with a learning rate of 0.00005 on 70% of the data and stopped training when performance on the validation data set (20%) decreased. The predictive accuracy of the regression and sequence classification models was evaluated on a separate hold-out data set consisting of the 700 remaining self-presentations (10%).

Generating Language Optimized for Perceived Humanity. For the three validation experiments, we drew a separate sample of 100 human-written self-presentations from the collected data. We created an additional set of 100 AI-generated self-presentations using the methods described in the main studies. We then produced an additional set of 100 self-presentations optimized for perceived humanity. To create these optimized self-presentations, we first generated a large number of self-presentations in each context using the same models as in the initial experiments. We then used the classifiers developed above to select self-presentations that the model predicted would be perceived as written by humans.

We employed different classifiers to select self-presentations in each context to increase generalizability and to validate both the regression and the language-model-based classifier. In the dating context, we used the regression-based classifier on the GPT-3 output to select those generated self-presentations that were more likely to be perceived as human-written. In the hospitality context, we used a classifier based on language models to perform the same task, connecting the GPT-2 generation model with the GPT-2 sequence classifier trained to predict participants' evaluation of self-presentations. In the professional context, we combined the regression and language-model classifiers using an ensemble approach. In each context, we selected the top 20% percentile of self-presentations that the classifier predicted were likely to be perceived as human-written. We drew a random sample of 100 self-presentations optimized for perceived humanity from these sets for each of the three validation experiments.

Participant Recruitment. For the main experiment in the hospitality context, we recruited a US-representative sample of 2,000 participants through Lucid (48). The experiment's results indicated that participants' answers did not vary significantly across demographics and that a smaller sample size would be sufficient for follow-up experiments. In the main dating and professional experiments, we recruited two gender-balanced samples of 1,000 US-based participants each from Prolific (49), a platform that enabled us to process bonus payments. Participants from Prolific had a median age of 37 y, 67% had a college degree, and 27% were at least somewhat familiar with computer programming. The median time participants spent on evaluating each self-presentation was 14.3 s (mean = 23.1, SD = 39.6). In return for their time, participants received compensation of $1.40 at a rate of about $12.5 per hour. Participants in the bonus condition in the dating context received an additional $3 bonus payment if they correctly rated at least 9 out of 12 self-presentations. We recruited a separate set of 1,300 crowdworkers to create the language features that could not be reliably computed for the 7,000 self-presentations in the main experiments. These crowdworkers were recruited from the same platforms as the participants in the main experiments and rated 12 self-presentations each, receiving compensation of $1.10. We recruited 200 participants for each of the three validation experiments on the respective platforms. Tasks and payments were analogous to the main experiments.

Limitations and Ethics Statement. Our results are limited to the current generation of language models and people's current heuristics for AI-generated language. Developments in technology and culture may change both the heuristics people rely on and the characteristics of AI-generated language. However, it is unlikely that in other cultural settings or for future generations of language models, human intuition will naturally coincide with the characteristics of AI-generated language. Our findings show that humans' flawed heuristics leave them vulnerable to large-scale automated deception. In disclosing this vulnerability, we face ethical tensions similar to cybersecurity researchers: On the one hand, publicizing a vulnerability increases the chance that someone will exploit it; on the other, only through public awareness and discourse effective preventive measures can be taken at the policy and development level. While risky, decisions to share vulnerabilities have led to positive developments in computer safety (50).

Data, Materials, and Software Availability. The data and code for the analyses performed across the three main studies and three validation experiments are publicly available through an Open Science Foundation repository (https://osf.io/284yv/). Previously published data were used for this work (47).

ACKNOWLEDGMENTS. We thank Benjamin Kim Carson for his assistance in collecting the self-presentation data, evaluating the qualitative data, and developing the language features. This material is based on work supported by the NSF under grant no. CHS 1901151/1901329 and the German National Academic Foundation.

1. T. Brown *et al.*, Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
2. A. Vaswani *et al.*, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
3. S. Biderman, E. Raff, Neural language models are effective plagiarists. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2201.07406 (Accessed 2 May 2022).
4. R. Bommasani *et al.*, On the opportunities and risks of foundation models. arXiv [Preprint] (2021). https://doi.org/10.48550/arXiv.2108.07258 (Accessed 12 August 2022).
5. N. A. Cooke, *Fake News and Alternative Facts: Information Literacy in a Post-Truth Era* (American Library Association, 2018).
6. L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **30**, 681–694 (2020).
7. B. Buchanan, A. Lohn, M. Musser, K. Sedova, Truth, lies, and automation. *Cent. Secur. Emerg. Technol.* **1**, 2 (2021).
8. L. Weidinger *et al.*, "Taxonomy of risks posed by language models" in *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (Association for Computing Machinery, New York, NY, USA, 2022), pp. 214–229.
9. B. R. Schlenker, "Self-presentation" in *Handbook of Self and Identity* (The Guilford Press, ed. 2, 2012), pp. 542–570.
10. B. Van Der Heide, J. D. D'Angelo, E. M. Schumaker, The effects of verbal versus photographic self-presentation on impression formation in Facebook. *J. Commun.* **62**, 98–116 (2012).
11. M. A. DeVito, J. Birnholtz, J. T. Hancock, "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 (Association for Computing Machinery, New York, NY, 2017), pp. 740–754.
12. N. Ellison, R. Heino, J. Gibbs, Managing impressions online: Self-presentation processes in the online dating environment. *J. Comput. Mediat. Commun.* **11**, 415–441 (2006).
13. E. Schwämmlein, K. Wodzicki, What to tell about me? Self-presentation in online communities. *J. Comput. Mediat. Commun.* **17**, 387–407 (2012).
14. E. Ert, A. Fleischer, N. Magen, Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tour. Manag.* **55**, 62–73 (2016).
15. X. Ma, J. T. Hancock, K. Lim Mingjie, M. Naaman, "Self-disclosure and perceived trustworthiness of Airbnb host profiles" in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (Association for Computing Machinery, New York, NY, 2017), pp. 2397–2409.
16. J. T. Hancock, M. Naaman, K. Levy, AI-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput. Mediat. Commun.* **25**, 89–100 (2020).
17. M. Jakesch, M. French, X. Ma, J. T. Hancock, M. Naaman, "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* (Association for Computing Machinery, New York, NY, 2019), p. 239.
18. E. Clark *et al.*, "All that's 'human' is not gold: Evaluating human evaluation of generated text" in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Association for Computational Linguistics, Cedarville, OH, 2021), pp. 7282–7296.
19. N. Köbis, L. D. Mossink, Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav.* **114**, 106553 (2021).
20. S. E. Kreps, M. McCain, M. Brundage, All the news that's fit to fabricate: AI-generated text as a tool of media misinformation (2020). Available at SSRN: https://ssrn.com/abstract=3525002.
21. A. Pinar Saygin, I. Cicekli, V. Akman, Turing test: 50 years later. *Minds Mach.* **10**, 463–518 (2000).
22. J. Guillory, J. T. Hancock, The effect of Linkedin on deception in resumes. *Cyberpsychol. Behav. Soc. Netw.* **15**, 135–140 (2012).
23. M. Karpinska, N. Akoury, M. Iyyer, "The perils of using mechanical turk to evaluate open-ended text generation" in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, Cedarville, OH, 2021), pp. 1265–1285.
24. M. T. Cox, Metacognition in computation: A selected research review. *Artif. Intell.* **169**, 104–141 (2005).
25. J. W. Pennebaker, The secret life of pronouns. *New Sci.* **211**, 42–45 (2011).
26. P. Slovic, S. Lichtenstein, Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organ. Behav. Hum. Perform.* **6**, 649–744 (1971).
27. J. Berger *et al.*, Uniting the tribes: Using text for marketing insight. *J. Mark.* **84**, 1–25 (2020).
28. C. Hutto, E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text" in Proceedings of the International AAAI Conference on Web and Social Media (Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 2014), pp. 216–225.
29. Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods. *J. Language Soc. Psychol.* **29**, 24–54 (2010).
30. M. L. Newman, J. W. Pennebaker, D. S. Berry, J. M. Richards, Lying words: Predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.* **29**, 665–675 (2003).
31. A. Radford *et al.*, Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
32. C. F. Bond Jr., B. M. DePaulo, Accuracy of deception judgments. *Pers. Soc. Psychol. Rev.* **10**, 214–234 (2006).
33. M. Hartwig, C. F. Bond Jr., Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol. Bull.* **137**, 643 (2011).
34. T. R. Levine, *Duped: Truth-Default Theory and the Social Science of Lying and Deception* (University Alabama Press, 2019).
35. D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, "Automatic detection of generated text is easiest when humans are fooled" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Cedarville, OH, 2020), pp. 1808–1822.
36. C. Ischen, T. Araujo, H. Voorveld, G. van Noort, E. Smit, "Privacy concerns in chatbot interactions" in *International Workshop on Chatbot Research and Design*, (Springer, 2019), pp. 34–48.
37. R. Zellers *et al.*, Defending against neural fake news. *Adv. Neural Inf. Process. Syst.* **32**, 9054–9065 (2019).
38. J. A. Goldstein *et al.*, Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2301.04246 (Accessed 22 February 2022).
39. S. Gehrmann, H. Strobelt, A. M. Rush, "GLTR: Statistical detection and visualization of generated text" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Association for Computational Linguistics, Cedarville, OH, 2019), pp. 111–116.
40. T. B. Hashimoto, H. Zhang, P. Liang, "Unifying human and statistical evaluation for natural language generation" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Cedarville, OH, 2019), pp. 1689–1701.
41. Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. Smith, Y. Choi, "Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Cedarville, OH, 2022), pp. 7250–7274.
42. A. Nelson, S. Friedler, F. Fields-Meyer, Blueprint for an AI bill of rights: A vision for protecting our civil rights in the algorithmic age. *White House Office of Science and Technology Policy* (2022). (18 October 2022).
43. European Commission, Proposal for a regulation laying down harmonised rules on artificial intelligence. *Shaping Europe's. Digital Future* (2021). (18 October 2022).
44. A. Acquisti, L. Brandimarte, J. Hancock, How privacy's past may shape its future. *Science* **375**, 270–272 (2022).
45. J. Williams, Should AI always identify itself? It's more complicated than you might think. *Electronic Frontier Foundation* (2018). (18 October 2022).
46. A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration. arXiv [Preprint] (2019). https://doi.org/10.48550/arXiv.1904.09751 (Accessed 12 August 2022).
47. A. Y. Kim, A. Escobedo-Land, OkCupid data for introductory statistics and data science courses. *J. Stat. Educ.* **23** (2015).
48. A. Coppock, O. A. McClellan, Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Polit.* **6**, 2053168018822174 (2019).
49. S. Palan, C. Schitter, Prolific. ac–A subject pool for online experiments. *J. Behav. Exp. Finance* **17**, 22–27 (2018).
50. K. Macnish, J. van der Ham, Ethics in cybersecurity research and practice. *Technol. Soc.* **63**, 101382 (2020).