

# Human Heuristics for AI-Generated Language Are Flawed

Maurice Jakesch<sup>1,2,\*</sup>, Jeffrey T Hancock<sup>3</sup>, Mor Naaman<sup>1,2</sup>

<sup>1</sup>Cornell University, <sup>2</sup>Cornell Tech, <sup>3</sup>Stanford University

\* Corresponding author: [mpj32@cornell.edu](mailto:mpj32@cornell.edu)

Human communication is increasingly intermixed with language generated by AI. Across chat, email, and social media, AI systems produce smart replies, autocompletes, and translations. AI-generated language is often not identified as such but poses as human language, raising concerns about novel forms of deception and manipulation. Here, we study how humans discern whether one of the most personal and consequential forms of language – a self-presentation – was generated by AI. In six experiments, participants (N = 4,600) tried to detect self-presentations generated by state-of-the-art language models. Across professional, hospitality, and dating settings, we find that humans are unable to detect AI-generated self-presentations. Our findings show that human judgments of AI-generated language are handicapped by intuitive but flawed heuristics such as associating first-person pronouns, spontaneous wording, or family topics with humanity. We demonstrate that these heuristics make human judgment of generated language predictable *and* manipulable, allowing AI systems to produce language perceived as more human than human. We discuss solutions, such as AI accents, to reduce the deceptive potential of generated language, limiting the subversion of human intuition.

**Keywords:** human-AI interaction, language generation, cognitive heuristics, risks of AI

**Author Contributions:** M.J. conducted the study, analyzed the results, and wrote the initial draft of the manuscript. M.J. and M.N. developed the study design and research framework. All authors contributed to the final manuscript.

**Competing Interest Statement:** The authors declare that they have no conflict of interest.

**Funding information:** This material is based upon work supported by the National Science Foundation under Grant No. CHS 1901151/1901329 and the German National Academic Foundation.

**THIS WORKING PAPER HAS NOT YET BEEN PEER-REVIEWED**

First version: 6/14/2022

This version: 6/30/2022

## Introduction

Large language models like GPT-3 (1, 2) produce semantic artifacts closely resembling human language. Through applications like smart replies, writing auto-completion, grammatical assistance, and machine translation, AI systems powered by these models infuse human communication with generated language at a massive scale. AI-generated language enables novel interactions and reduces human effort but can facilitate novel forms of spam, plagiarism, manipulation, and deception (1, 3–8) when people mistake it for human language.

In a series of experiments, we analyzed how humans detect generated language in one of the most personal and consequential forms of speech – self-description. Previous research has extensively studied the importance of self-presentation (9–11), showing that impression formation based on self-descriptions is crucial for establishing the trust required for a wide variety of social interactions (12, 13). AI systems that generate self-presentations may invalidate signals that people rely on when assessing others (14), such as tone or compositional skill. Previous work has already shown that interpersonal trust declines when people suspect that others are using AI systems to generate or optimize their self-presentation (15).

Can humans detect self-presentation generated by the current generation of language models? Previous studies suggest that people struggle to discern AI-generated language in different settings (16–18). Here, we go beyond previous work by studying *how* people try to detect AI-generated language. Using qualitative and quantitative methods, we reconstruct heuristics that people rely on to detect generated language. We assess to what extent these heuristics successfully distinguish between human and AI-generated language and demonstrate that AI systems can predict and *manipulate* whether people perceive language as generated.

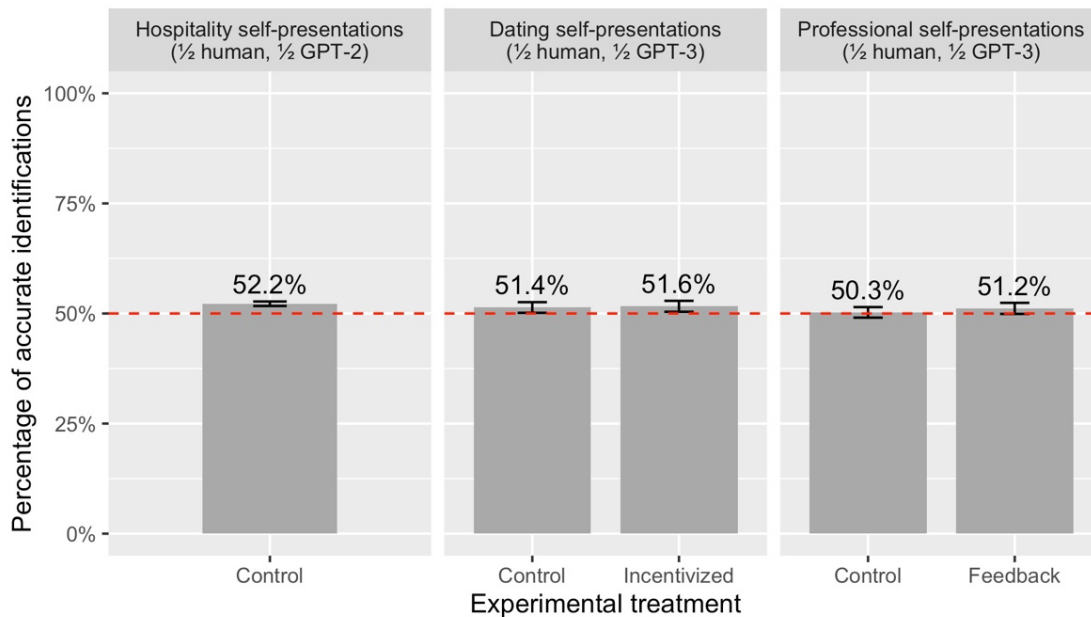
To examine how people detect AI-generated self-presentations, we performed six experiments patterned after the Turing test (19) where a human judge tries to identify a language-generating machine based on a conversation. We trained customized versions of state-of-the-art large language models (1, 2, 4) to generate self-presentations in three social contexts in which trust in a self-presentation is important for decision-making: professional (e.g., job applications) (20), romantic (e.g., online dating) (10), and hospitality (e.g., Airbnb host profiles) (13). Across six experiments, we asked 4,600 participants to read through a series of self-presentations—half AI-generated, half collected from real-world online platforms—and indicate which ones they thought were generated.

## Results

We start by computing the accuracy rates for participants' ability to distinguish between human and AI-generated self-presentations. Across scenarios, participants identified the source of a self-presentation with only 50% to 52% accuracy, as shown in Figure 1. Even when offered monetary incentives for accurate assessments (right bar in the central panel in Figure 1, 51.6%) or when receiving immediate feedback on their evaluations (right bar in the right panel, 51.2%), participants' accuracy remained close to chance. Further analyses revealed that no demographic group performed better than others.

Participants' evaluations were not random, however. For 23% of self-presentations, at least four out of five judges agreed on a human source. Such agreement would be expected only in 16% of cases if ratings were random. Overall, the observed agreement between participants' judgments was significantly higher than chance (Fleiss' kappa = 0.067,  $p < 0.0001$ ). Had this level of agreement been due to valid cues that differentiated human and AI-generated self-presentations, participants' accuracy would have

**Figure 1.** Participants were unable to detect self-presentations generated by the current generation of language models beyond chance. Error bars represent 95% confidence intervals for 6,000–16,000 judgements of 2,000–3,000 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.



been 62% to 66%. As the observed accuracy was close to chance, the agreement in participants' assessments must have been due to shared but flawed heuristics for AI-generated language.

To investigate participants' heuristics for AI-generated language, we asked them to explain their judgments. Two researchers independently coded a sample of responses into themes. 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 participants 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 human language. Grammatical errors were associated with a subpar AI by some participants but with fallible human authors by others. Some participants judged the self-presentation source by its tone (24%), associating warm and genuine language with humanity and impersonal, monotonous style with AI-generated language. Further detail on participants' explanations is included in the SI Appendix.

We quantitatively analyzed the extent to which language features in a self-presentation predicted participants' answers. We developed an initial set of language features motivated by participants' explanations of their judgments, to which we added various statistical metrics and commonly analyzed psychological language features such as LIWC (21). We also recruited 1,350 raters to label whether self-presentations were nonsensical, had grammatical issues, or seemed repetitive. After a feature selection process, we fit a regression model correlating selected features with participants' perception that a self-presentation was generated. The results are shown in Table 1.

**Table 1:** Logistic regression odds ratios predicting whether (1) participants suspected a self-presentation was AI-generated and (2) whether it actually was generated. Only nonsense, repetition, and conversation were functional cues, indicated by equal coefficient directions in models (1) and (2). Other heuristics were either inversely correlated or uncorrelated with the actual source of the self-presentation.

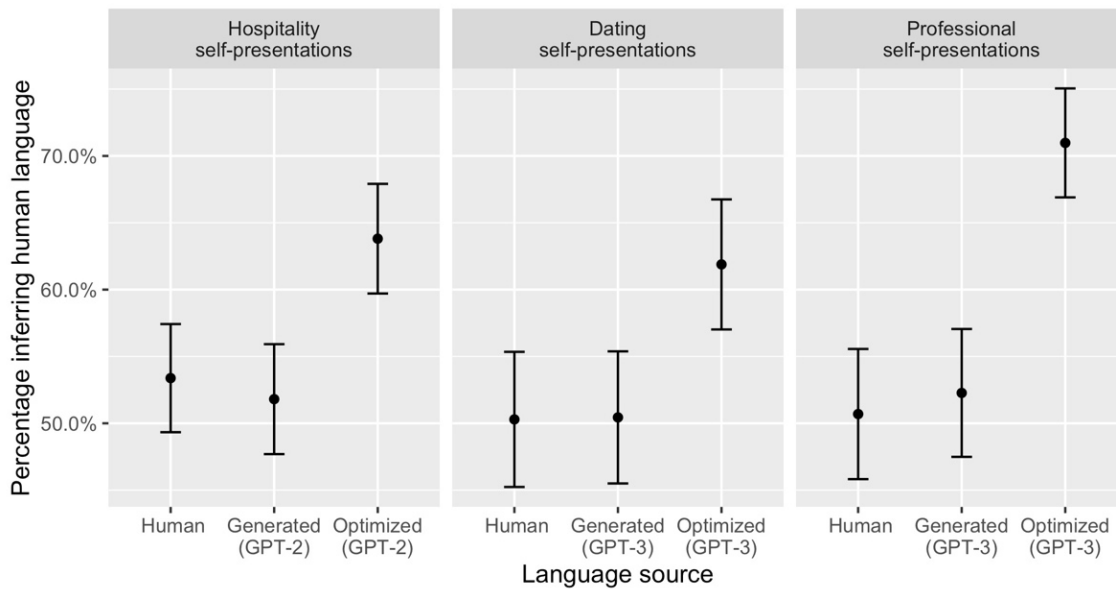
	Dependent variable:	
	(1) Perceived as AI-generated (95% CI)	(2) Actually AI-generated (95% CI)
Nonsensical content †	1.105*** (1.085, 1.126)	1.233*** (1.169, 1.296)
Repetitive content †	1.083*** (1.059, 1.106)	1.470*** (1.379, 1.561)
Second person pronouns	1.059*** (1.038, 1.079)	0.970 (0.908, 1.032)
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)
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)
Conversational words	0.947*** (0.925, 0.970)	0.898** (0.829, 0.967)
Contractions	0.947*** (0.924, 0.970)	1.134*** (1.065, 1.203)
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:

†manual labels, \* p\*\* p\*\*\* p<0.001

Participants relied on some valid cues. For example, self-presentations containing nonsensical content were 10.5% more likely to be seen as AI-generated (top row, left coefficient) and indeed were 23% more likely to be AI-generated (top row, right coefficient). Similarly, self-presentations with repetitive content were 8% more likely to be judged as generated and were 47% more likely to be generated. However, most heuristics participants relied on were *flawed*. They were 5% more likely to rate self-presentations with grammatical issues as generated, although grammatically flawed self-presentations were, in fact, 15% *less* likely to be generated. Participants often rated self-presentations with long words or rare bigrams as generated, while most self-presentations with long words or rare bigrams were actually written by humans. They also mistakenly judged first-person speech and family content as more human. Longer self-presentations that sounded authentic or spontaneous (22) or were focused on past events were more likely to be seen as human. These cues were not associated with actual generation.

**Figure 2.** Exploiting humans' flawed heuristics, AI systems can generate personal self-presentations perceived as more human than human. Error bars represent 95% confidence intervals for 350 to 450 judgments of 100 self-presentations per bar.



The regression model predicted participants' judgments of generated language with 57.6% accuracy when evaluated on a hold-out data set. We also tested whether language models can learn to predict human impressions of generated language without feature engineering input from the research team. A current language model (23) with a sequence classification head predicted participants' assessments of generated language with 58.1% accuracy when evaluated on hold-out validation data.

These results suggest that people rely on flawed heuristics to detect generated language and that AI systems can predict people's judgments of generated language. We conducted three independent replication experiments to validate these findings and test whether language models can exploit people's flawed heuristics to generate self-presentations perceived as more human than human. For the experiments, we reused the classifiers trained on participants' judgments in the previous experiments to generate self-presentations optimized for perceived humanity.

Figure 2 shows participants' evaluations of the generated self-presentations optimized for perceived humanity. Across domains, the optimized self-presentations (right bar in each panel) were more likely to be rated as human than generated self-presentations that were not optimized (middle bar, 65.7% vs. 51.6%,  $p < 0.001$ ). The optimized self-presentations were also more likely to be perceived as human than actual human self-presentation (left bar, 65.7% vs. 51.7%,  $p < 0.001$ ). The manipulation was strongest in the *professional* context (right panel), where we combined the regression and language model-based classifier to generate self-presentations perceived as human 71% of the time.

## Discussion

Our results affirm that humans are not able to detect self-presentations generated by current 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 generated news, recipes, and poetry (16–18).

Drawing on the extensive literature on deception detection (24, 25), we propose two explanations for human inability to detect generated self-presentation: First, the language generated by state-of-the-art AI systems may be so similar to human language that a lack of reliable cues limits accuracy. Second, participants' judgments may be inaccurate because they rely on flawed heuristics to detect generated language.

We found that AI-generated self-presentations have features that people may be able to detect. When we asked participants in a separate labeling task which self-presentations were nonsensical or repetitive, they successfully identified AI-generated language as more repetitive and nonsensical. Had participants used these heuristics only to detect AI-generated language, they would have achieved an accuracy of 59%. However, flawed intuitive heuristics reduced people's accuracy in detecting generated self-presentations to chance. These heuristics include associating humanity with first-person pronouns, spontaneous wording, or discussing family or past events. While creating an illusion of personhood, these features do not differentiate between human and generated language.

People's reliance on flawed intuitive heuristics to detect generated language demonstrates that the increased human-likeness of generated text is *not* necessarily indicative of increased machine intelligence. For example, emphasizing family topics increases the perceived humanity of generated self-presentations but does not require advances in machine intelligence. Recent work by Ippolito et al. (26) indeed suggests that developments in language model decoding methods have been optimized for fooling humans at the cost of introducing statistical anomalies that are easily detected by machines.

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. Whether people believe language is human or generated has consequences: Previous work has shown that not only are people more likely to disclose private information to and adhere to recommendations by non-human entities that they perceive as human (27), but they may start distrusting those they believe are using AI-generated language in their communication (15). People's flawed heuristics of AI-generated language also make their judgment predictable and manipulatable. In three replication studies, we have shown that AI systems can use people's flawed heuristics to produce language perceived as more *human than human*. From automated impersonation (8) to targeted disinformation campaigns (28), AI systems could be optimized to exploit human intuition, exacerbating concerns about novel automatized forms of deception, fraud, and identity theft (3–8).

Widespread AI education and technical tools that assist identification (29–31) might improve people's ability to detect generated language to some extent. However, the potential for improving human intuition for the detection of generated language is likely limited (16), and future generations of language models may invalidate learned heuristics. Rather than adapting humans to language models, we believe future language models should be designed *not* to undermine human intuition. Most AI applications do not genuinely require generating human-like language, and even natural-language interaction systems could use clearly non-human language without loss of functionality. Language models that announce their presence through dedicated AI accents or explicit in-built disclosures could support rather than undermine human intuition. AI systems that accommodate the limits and flaws of human judgment by design will more genuinely support human communication while reducing the risk of misuse.

## Materials and Methods

### *Experiment design*

The design of the six experiments combines elements of a simplified Turing test (19) with a classical data labeling task. After providing informed consent, participants were introduced to the *hospitality*, *dating*, or *professional* scenarios. They were told they were browsing a 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 language model. After rating six self-presentations, they were asked to explain their judgment in an open-ended response. 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 pre-registered the final two experiments prior to data collection ([https://aspredicted.org/blind.php?x=7DK\\_81P](https://aspredicted.org/blind.php?x=7DK_81P)).

To increase robustness and generalizability, experiments were performed in three social contexts. In addition, minor variations across experiments explored auxiliary hypotheses. Longer self-presentations were used in dating and professional contexts to test whether the length of self-presentations limited participants' accuracy. To keep the experiment duration comparable, we reduced the number of rated self-presentations to 12 in these experiments. Half of the participants in the *dating* context received a bonus payment if they rated three out of four self-presentations correctly to explore the effect of increased effort. There was no difference in performance between the two groups. Finally, to test whether participants could learn to detect generated self-presentations if they received feedback, half of the participants in the *professional* context were told whether their rating was correct after every rating, again with no difference in outcomes. An overview of the experimental designs is included in the SI Appendix.

### *Collecting and generating self-presentations*

For each context, we collected data from real-world platforms for the experiment. We used it to train a state-of-the-art language model to generate self-presentations. For the experiment in the *hospitality* context, we collected 28,890 self-presentation texts from host profiles on Airbnb.com that contained at least 30 and no more than 60 words. We used a random sample of 1,500 human-written self-presentations for the experiment. We also fine-tuned a 774M parameter version of GPT-2 (23) on the collected data 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 (32). We drew a random sample of 1,000 human self-presentations for each experiment. We used the full set of collected self-presentations in each context to fine-tune a 13B parameter version of GPT-3 (1) and created 1,000 AI-generated self-presentations for each experiment. For each validation experiment, we created a separate random sample of 100 human profiles from the collected data and generated a separate set of 100 profiles. To confirm that the language model had not plagiarized longer text sequences, we searched for identical 5-grams in the training data and generated text.

### *Predicting responses and optimizing self-presentations.*

We use multiple categories of language features to predict how participants evaluated the self-presentations. We developed an initial set of language features motivated by participants' explanations of their judgments. To this initial set, we added statistical metrics, readability scores, emotion classification, and psychological language features (21). We created three additional key features by recruiting raters (N

= 1,350) to label which self-presentations seemed nonsensical, repetitive, or had grammatical issues. The resulting set of about 180 is detailed in the SI Appendix. To reduce overfitting and increase interpretability, we reduced the set of relevant features to 15 in a feature selection process based on lasso regression. We reported the performance of a logistic regression model computed on a hold-out set not used for feature selection. In addition, to test whether modern language models can learn to predict human impressions of generated language, we trained a 117M parameter version of GPT-2 (23) with a sequence classification head to predict participants' judgments. The model's accuracy was evaluated on a separate hold-out data set.

For the three validation experiments, we ran these classifiers on a large new sample of self-presentations generated by the initial language models to select self-presentations more likely to be perceived as human. In the *dating* context, we used the regression classifier on the GPT-3 output to only keep those generated self-presentations likely to be perceived as human. In the *hospitality* context, we connected 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 both approaches using an ensemble classifier. Through this process, we produced 100 self-presentations optimized for perceived humanity for each of the three final experiments.

#### *Participant recruitment*

For the experiment in the *hospitality* context, we recruited a US-representative sample of 2,000 participants through Lucid (33). The experiment results indicated that demographics did not play a significant role in participants' answers and that a smaller sample size would be sufficient for follow-up experiments. In the *dating* and *professional* context, we recruited two gender-balanced samples of 1,000 US-based participants each from Prolific, a platform that enabled us to offer bonus payments. Participants from Prolific had a median age of 37 years, 67% had a college degree, and 27% were at least somewhat familiar with computer programming. 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 who correctly rated at least 9 out of 12 self-presentations received an additional \$3 bonus payment. We recruited a separate set of 1,350 participants to create the language features that could not be computed reliably. Finally, we recruited 200 participants for each validation experiment on the respective platforms. Tasks and payments were analogous to the initial experiments.

#### *Limitations and ethics statement*

Our results are limited to the current generation of language models and people's current heuristics for generated language. Developments in technology and culture may change the heuristics people rely on and the characteristics of generated language. However, it is unlikely that in other cultural settings or for future generations of language models, human intuition will per se coincide with the characteristics of 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 (34). We mitigate the risk of misuse by sharing training and model data with individual researchers after verification only.

#### **Acknowledgments**

We thank Benjamin Kim Carson for his assistance in collecting the self-presentation data, evaluating the qualitative data, and developing the language features.



## References

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** (2017).
3. S. Biderman, E. Raff, Neural Language Models are Effective Plagiarists. *ArXiv220107406 Cs* (2022) (May 2, 2022).
4. R. Bommasani, *et al.*, On the opportunities and risks of foundation models. *ArXiv Prepr. ArXiv210807258* (2021).
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 (2021).
8. L. Weidinger, *et al.*, Taxonomy of Risks posed by Language Models (2022).
9. 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*, (2017), pp. 740–754.
10. 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).
11. E. Schwämmlein, K. Wodzicki, What to Tell About Me? Self-Presentation in Online Communities. *J. Comput.-Mediat. Commun.* **17**, 387–407 (2012).
12. 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).
13. X. Ma, J. T. Hancock, K. L. Mingjie, M. Naaman, Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. in *CSCW*, (2017), pp. 2397–2409.
14. J. T. Hancock, M. Naaman, K. Levy, AI-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput.-Mediat. Commun.* **25**, 89–100 (2020).
15. 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*, (ACM, 2019), p. 239.
16. E. Clark, *et al.*, All That’s “Human” Is Not Gold: Evaluating Human Evaluation of Generated Text. *ArXiv210700061 Cs* (2021) (October 25, 2021).
17. 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).
18. S. E. Kreps, M. McCain, M. Brundage, All the News that’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Available SSRN 3525002* (2020).
19. A. Pinar Saygin, I. Cicekli, V. Akman, Turing test: 50 years later. *Minds Mach.* **10**, 463–518 (2000).
20. J. Guillory, J. T. Hancock, The effect of LinkedIn on deception in resumes. *Cyberpsychology Behav. Soc. Netw.* **15**, 135–140 (2012).
21. Y. R. Tausczik, J. W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
22. 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).
23. A. Radford, *et al.*, Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
24. C. F. Bond Jr, B. M. DePaulo, Accuracy of deception judgments. *Personal. Soc. Psychol. Rev.* **10**, 214–234 (2006).
25. 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).
26. D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled. *ArXiv Prepr. ArXiv191100650* (2019).
27. 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.
28. R. Zellers, *et al.*, Defending against neural fake news. *Adv. Neural Inf. Process. Syst.* **32** (2019).
29. S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: Statistical detection and visualization of generated

- text. *ArXiv Prepr. ArXiv190604043* (2019).
30. T. B. Hashimoto, H. Zhang, P. Liang, Unifying human and statistical evaluation for natural language generation. *ArXiv Prepr. ArXiv190402792* (2019).
  31. 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, 2022), pp. 7250–7274.
  32. A. Y. Kim, A. Escobedo-Land, OkCupid data for introductory statistics and data science courses. *J. Stat. Educ.* **23** (2015).
  33. 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).
  34. K. Macnish, J. van der Ham, Ethics in cybersecurity research and practice. *Technol. Soc.* **63**, 101382 (2020).

## Supplementary Information for

### Human Heuristics for AI-Generated Language Are Flawed

#### Supplementary Text

Below we provide additional information on several aspects of our experiments. Table S1 and S2 summarize the treatment, stimuli, and recruitment methods used across the six studies and three labeling tasks. Table S3 shows a sample of self-presentations for each study and treatment group.

Table S4 shows the results of an auxiliary analysis testing whether certain groups are better at detecting AI-generated language than others. Older participants were slightly more likely to detect generated self-presentations, with participants older than 50 years achieving an accuracy of 53% (compared to 51% for younger participants). No gender or ethnic group performed better than others. Participants with a university degree performed about 1% worse than those without, and self-reported technical knowledge was not correlated with more accurate ratings. Neither the time taken for the judgment nor the length of profiles predicted higher judgment accuracy. Across contexts, groups, and treatments, participants could not detect AI-generated self-presentations.

Table S5 and S6 provide further detail on the qualitative analysis of participants' explanations of why they thought certain self-presentations were generated or human. Two researchers independently coded a sample of responses into themes to provide an overview of participants' self-reported heuristics. Table S5 presents an overview of recurring themes. Participants most referred to the content of a self-presentation (blue-shaded regions in Table S5, representing 40% of responses). They associated specific content related to family and life experiences with human language and generic or nonsensical content with generated language. Participants also based their decisions on grammatical cues (gray, 28%), where first-person pronouns and the mastery of grammar were seen as indicative of human language. Some participants judged the self-presentation source by its tone (green, 24%), associating warm and genuine language with humanity and impersonal, monotonous style with AI-generated language. The codebook, along with theme frequencies and sample responses, is shown in Table S6.

Table S7 provides an overview of the developed language features and statistical summaries.

**Table S1: Overview of experiments**

#	Context	Stimuli	Treatment	Recruitment
1	Hospitality	1,500 self-presentations from Airbnb and 1,500 generated by GPT-2; 30-60 words each; 16 per subject	Within-subject variation of profile type	N = 2,000 US-representative sample via Lucid
2	Dating	1,000 self-presentations from OkCupid and 1,000 generated by GPT-3; 60-90 words; 12 per subject	Add. bonus payments for correct ratings	N = 1,000 gender-balanced sample via Prolific
3	Professional	1,000 self-presentations from Guru and 1,000 generated by GPT-3; 60-90 words each; 12 per subject	Add. feedback on answers	N = 1,000 gender-balanced sample via Prolific
4	Hospitality	100 self-presentations from Airbnb, 100 generated by GPT-2, and 100 optimized using the language model classifier; 16 per subject	Within-subject variation of profile type	N = 250 US-representative sample via Lucid
5	Dating	100 self-presentations from OkCupid, 100 generated by GPT-3, and 100 optimized by the regression classifier; 16 per subject	Within-subject variation of profile type	N = 200 gender-balanced sample via Prolific
6	Professional	100 self-presentations from Guru, 100 generated by GPT-3, and 100 optimized using an ensemble classifier; 16 per subject	Within-subject variation of profile type	N = 200 gender-balanced sample via Prolific

**Table S2: Overview of labeling tasks**

#	Context	Stimuli	Recruitment
L1	Hospitality	Same as in #1, 16 per participant	N = 600 US-representative sample via Lucid
L2	Dating	Same as in #2, 16 per participant	N = 350 gender-balanced sample via Prolific
L3	Professional	Same as in #3, 16 per participant	N = 350 gender-balanced sample via Prolific

**Table S3: Exemplary self-presentations**

Context	Source	Example
Hospitality	Human	My family has lived in DC for the past several years. Some of our favorite things about living on Capitol Hill are running through the neighborhood, exploring all the museums and exhibits that are walking distance from our home, and having a variety of great food offerings only steps away.

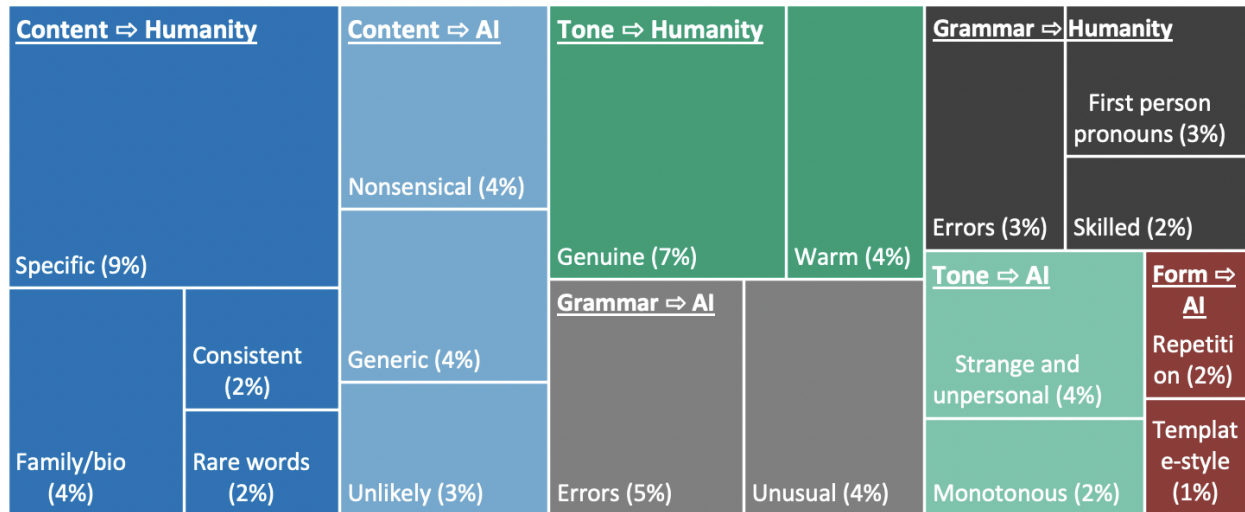
Hospitality	Generated (GPT-2)	A teacher and young entrepreneur, I love to ski and travel. My wife & I have lived in Vermont for the past 10 years and love the beauty and the snow that we get to ski during the summer.
Hospitality	Generated (GPT-2) & optimized (regression)	My husband and I have lived in Denver for 20 years. A few summers ago we visited my two brothers who live elsewhere so we decided to make our home available for others to enjoy as well. We love traveling in Europe, South America and anywhere new! Welcome to your home away from home.
Dating	Human	i'm an elementary school social worker and find my job both fulfilling and frustrating. an la native, i've also lived in the midwest and new england. i've been in sf for about 6 years now and love the people, politics, and food here. but, i do miss having seasons and look forward to my annual vacations back in the midwest, which generally involve lounging on a lake and drinking bell's beer. i enjoy being fit, active, and healthy, though i do eat ice cream for dinner on occasion.
Dating	Generated (GPT-3)	i just moved to the city last august and really don't know many people here yet. i'm interested in hanging out and maybe even finding someone special. i would love to be able to spend time together without any drama and want to get to know each other better. i'd love to find someone that i can share all of these exciting things in life with like art galleries, theatre, dinner, etc...
Dating	Generated (GPT-3) & optimized (GPT-2)	hey i moved to sf about 2 years ago, it's such a great city..i like to explore the city, always trying to find new hangouts and food... i've travelled a lot around the world and would love to travel more. i'm easy going and down to earth, i know what i want in life and am working towards my goals. message me if you want to know more :)
Professional	Human	I have 19 years of journalism experience. My work has appeared in daily and weekly newspapers, international trade magazines and textbooks. I also have worked in broadcast news, and my reporting has been picked up by the Associated Press. For six years, my interviews focused on C-level execs at Fortune 500 power companies, tech startups and government. In 2015, I became managing editor of a publication in the petroleum and fluid handling equipment industry.
Professional	Generated (GPT-3)	My name is Gary Stauch and I have been in the computer and electronics business for over 30 years. I have a A.S. in electronics, a B.S. in computer science and I am a registered professional engineer in Texas. In addition to my own company, I have worked for several others in the design and deployment of large scale network infrastructure in the data center and enterprise server market. I have designed and developed server platforms, workstations, servers, switches, routers and other devices that are part of large scale networks.
Professional	Generated (GPT-3) & optimized (regression and	I am a mother of three and a grandmother of two. I live in beautiful iowa and have been here all my life. I enjoy doing different things but I am a master at none. I love to tell stories and make people smile with laughter. I am very well at reading people and knowing what to

	GPT-2)	do to get the job done. I am very good at multi-tasking. I am very organized and very well at using my time.
--	--------	--

**Table S4:** Regression coefficients predicting the accuracy of a judgment based on treatment, social context, and participant demographics. No group performed much above chance level.

	<i>Dependent variable:</i>
	Likelihood of accurate identification OR (95% CIs)
Context: Dating profiles	0.974 (0.882, 1.065)
Context: Professional profiles	0.926 (0.845, 1.007)
Treatment: Feedback	1.038 (0.966, 1.110)
Treatment: Incentives	1.022 (0.944, 1.100)
Age	<b>1.002**</b> (1.001, 1.003)
Gender: Female	1.002 (0.967, 1.036)
Gender: Non-binary	1.010 (0.834, 1.186)
Race: African American	0.959 (0.895, 1.022)
Race: Asian	1.055 (0.976, 1.134)
Race: Hispanic	1.005 (0.940, 1.069)
Race: Other	0.973 (0.887, 1.059)
Level of education	<b>0.986**</b> (0.976, 0.996)
Technical knowledge	1.006 (0.982, 1.030)
Rating: Time taken	1.000 (1.000, 1.001)
Profile: Word count	1.000 (0.998, 1.002)
Constant	1.045 (0.925, 1.166)
Observations	53,411
Log Likelihood	-37,199.800
Akaike Inf. Crit.	74,435.610
<i>Note:</i>	*p**p***p<0.001

**Table S5.** Themes in participants' explanations of why they thought a self-presentation was human or generated language.  $N = 800$ , tile areas correspond to theme prevalence reported in brackets. Heuristics are classified by whether they refer to the content (blue), tone (green), grammar (gray), or form (red) of a self-presentation. Lighter tiles show cue that were associated with generated language.



**Table S6:** Examples themes and codes in participants' explanations of judgements

Category	Code	Freq.	Example
Content cues for AI	Nonsensical content	7%	"'travel here from around the world' in third sentence doesn't make sense"
Content cues for AI	Generic content	6%	"seems just a bit to generic and a bit random"
Content cues for AI	Unlikely content	4%	"A full time manager at a nuclear plant doesn't travel frequently enough to care about hotel amenities."
Content cues for Humanity	Specific content	14%	"How detailed descriptions were"
Content cues for Humanity	Family and biography	6%	"I determine this is a person because he says him and his wife and son travel and go places on there free time"
Content cues for Humanity	Consistent	3%	"Based primarily on the content, and whether each part of the statement made sense logically and thematically with the rest."
Form cues for AI	Repetitive	2%	"the repetition of the sentences make the whole thing sound lifeless and robotic."
Form cues for AI	Template-like	2%	"I looked for a stock template response"

			for AI, or for signs of a disjointed copy and paste from real user statements.”
Grammar cues for AI	Errors	7%	“If things are worded incorrectly.”
Grammar cues for AI	Unusual punctuation	7%	“There should be a comma after ‘I’m Kellie’”
Grammar cues for Humanity	Errors	5%	“Believe there was a grammar error where it should have been knowledgeable”
Grammar cues for Humanity	1st person speech	4%	“Using I, me, we language”
Grammar cues for Humanity	Good grammar	3%	“The English is good, but not great. It possibly is written by someone who is ESL.”
Grammar cues for Humanity	Rare words	3%	“Certain words that were unusual.”
Tone cues for AI	Strange and unpersonal	6%	“The personal touch is very unnatural sounding.”
Tone cues for AI	Monotonous	3%	“most people either put in little or more thought and AI just feels like a perfect monotone read”
Tone cues for Humanity	Genuinely personal	10%	“one can have a few replies per question and then have the AI Place together; but this isnt random.. it is Genuine”
Tone cues for Humanity	Warm and welcoming	6%	“Its how the phrase comes across, An AI Having Emotion...”

**Table S7:** Overview of the engineered language features and their correlations with participants’ judgements of generated speech.

Feature Name	Mean	SD.	Min	Max	Cor. with ratings	Cor. with source
Nonsensical (manual labels)	0.117	0.233	0	1	0.086	0.114
Repetitive (manual labels)	0.099	0.222	0	1	0.127	0.057
Grammatical issues (manual)	0.172	0.281	0	1	-0.086	0.057



LIWC Achieve	2.325	2.535	0	17.72	0.037	-0.009
LIWC Acquire	0.492	0.941	0	9.72	-0.012	0.01
LIWC Adjective	6.968	3.797	0	30.95	0.029	-0.007
LIWC Adverb	3.898	3.038	0	22.22	-0.094	0.023
LIWC Affect	7.368	4.983	0	34.48	0.028	0.024
LIWC Affiliation	3.15	4.238	0	25.81	0.033	0.032
LIWC Allnone	0.854	1.374	0	12.9	0.017	0.001
LIWC Allpunc	17.037	8.73	0	257.14	-0.009	-0.046
LIWC Allure	9.614	4.967	0	32.35	-0.056	0.09
LIWC Analytic	56.357	27.358	1	99	0.098	-0.051
LIWC Apostro	1.75	2.249	0	21.67	-0.109	0.03
LIWC Article	5.938	3.066	0	20.45	0.013	0.088
LIWC Assent	0.035	0.259	0	8.82	-0.047	-0.013
LIWC Attention	0.615	1.259	0	10.64	0.003	0.019
LIWC Auditory	0.336	0.976	0	11.54	-0.011	-0.036
LIWC Authentic	72.388	30.232	1	99	-0.197	0.031
LIWC Auxverb	7.599	3.585	0	25	-0.077	0.112

LIWC Bigwords	20.104	8.673	0	68.42	0.123	-0.128
LIWC Cause	0.94	1.367	0	9.52	0.033	-0.014
LIWC Certitude	0.35	0.89	0	9.3	-0.038	-0.01
LIWC Clout	33.423	35.196	1	99	0.162	0.01
LIWC Cognition	8.361	5.307	0	36.67	-0.012	-0.011
LIWC Cogproc	7.457	5.005	0	36.67	-0.017	-0.01
LIWC Comm	1.236	1.758	0	17.65	-0.035	-0.011
LIWC Comma	5.566	4.722	0	42.11	0.024	-0.075
LIWC Conflict	0.033	0.248	0	5	-0.02	-0.019
LIWC Conj	8.083	3.071	0	25.3	-0.033	0.055
LIWC Conversation	0.24	0.801	0	21.05	-0.089	-0.026
LIWC Culture	0.988	1.961	0	19.05	0.06	-0.02
LIWC Curiosity	0.983	1.601	0	12.5	-0.007	0.022
LIWC Death	0.02	0.19	0	3.61	-0.007	-0.012
LIWC Det	11.627	4.017	0	27.66	-0.021	0.06
LIWC Dic	88.99	6.611	36.84	100	-0.092	0.164
LIWC Differ	2.054	2.199	0	14.71	-0.04	0.013

LIWC Discrep	1.208	1.687	0	12.2	-0.005	0.01
LIWC Drives	6.244	4.802	0	29.41	0.069	0.008
LIWC Emo Anger	0.026	0.233	0	5.88	-0.022	-0.023
LIWC Emo Anx	0.033	0.277	0	8.22	-0.015	-0.023
LIWC Emo Neg	0.132	0.56	0	9.09	-0.032	-0.032
LIWC Emo Pos	2.502	2.662	0	17.65	-0.012	0.033
LIWC Emo Sad	0.016	0.173	0	5.08	0.023	-0.01
LIWC Emotion	2.679	2.747	0	20.59	-0.018	0.023
LIWC Ethnicity	0.122	0.675	0	16.39	0.002	-0.034
LIWC Exclam	0.76	1.68	0	26.58	-0.007	-0.024
LIWC Family	0.602	1.465	0	12.9	-0.083	0.011
LIWC Fatigue	0.014	0.164	0	4	-0.022	0.001
LIWC Feeling	0.267	0.738	0	6.67	0.018	-0.006
LIWC Female	0.426	1.197	0	19.35	-0.008	-0.015
LIWC Filler	0.005	0.098	0	4.11	-0.015	0.015
LIWC Focusfuture	0.919	1.624	0	16.67	0.022	-0.001
LIWC Focuspast	2.345	2.636	0	15.38	-0.111	0.008

LIWC Focuspresent	5.2	2.984	0	24.14	0.003	0.072
LIWC Food	0.737	1.657	0	19.05	-0.01	0.009
LIWC Friend	0.466	1.053	0	14.29	0.025	0.039
LIWC Fulfill	0.153	0.527	0	5.56	0.036	-0.017
LIWC Function	51.71	8.185	1.32	79.41	-0.129	0.162
LIWC Health	0.31	1.037	0	17.86	-0.006	-0.01
LIWC Home	0.721	1.531	0	22.86	0.021	-0.007
LIWC I me	7.962	4.525	0	24.39	-0.212	0.031
LIWC Illness	0.024	0.249	0	6.25	0.019	-0.026
LIWC Insight	1.674	1.994	0	15	0.022	-0.029
LIWC Ipron	2.301	2.483	0	22.06	-0.005	0.029
LIWC Lack	0.051	0.381	0	6.9	-0.003	-0.021
LIWC Leisure	1.975	2.788	0	19.35	-0.043	-0.004
LIWC Lifestyle	8.156	5.838	0	40	0.025	-0.013
LIWC Linguistic	66.286	9.102	6.58	91.18	-0.122	0.156
LIWC Male	0.572	1.223	0	15.69	-0.004	0.009
LIWC Memory	0.031	0.259	0	4.76	0.02	-0.021

LIWC Mental	0.022	0.246	0	8	-0.022	0.011
LIWC Money	1.089	2.075	0	20.51	0.065	-0.012
LIWC Moral	0.204	0.69	0	8.11	-0.016	-0.04
LIWC Motion	2.131	2.293	0	16.13	-0.032	0.018
LIWC Need	0.282	0.86	0	8.86	0.022	-0.009
LIWC Negate	0.521	1.082	0	12.5	-0.057	-0.019
LIWC Netspeak	0.184	0.686	0	21.05	-0.082	-0.028
LIWC Nonflu	0.022	0.196	0	4.35	-0.022	0.003
LIWC Number	1.364	1.965	0	27.27	-0.031	-0.026
LIWC Otherp	1.901	3.802	0	163.77	0.014	-0.043
LIWC Perception	11.3	5.608	0	43.24	-0.021	0.016
LIWC Period	7.001	4.89	0	245.71	0.003	0.019
LIWC Physical	1.785	2.432	0	23.81	-0.022	-0.006
LIWC Polite	0.38	0.995	0	10	0.043	-0.044
LIWC Politic	0.185	0.802	0	13.64	0.031	-0.005
LIWC Power	0.855	1.535	0	15.66	0.073	-0.054
LIWC Ppron	11.167	4.369	0	27.91	-0.133	0.064

LIWC Prep	13.841	4.042	0	29.51	-0.013	0.035
LIWC Pronoun	13.468	5.147	0	32.65	-0.115	0.068
LIWC Prosocial	0.887	1.507	0	13.33	0.089	-0.019
LIWC Qmark	0.061	0.431	0	13.24	-0.016	-0.002
LIWC Quantity	3.614	2.857	0	18.82	-0.067	-0.018
LIWC Relig	0.085	0.561	0	17.65	-0.001	-0.029
LIWC Reward	0.228	0.682	0	6.67	0.043	-0.012
LIWC Risk	0.094	0.432	0	7.69	0.028	-0.045
LIWC Sexual	0.026	0.246	0	7.81	-0.032	-0.041
LIWC Shehe	0.131	0.767	0	13.89	0.08	-0.005
LIWC Socbehav	4.371	3.262	0	23.33	0.032	-0.019
LIWC Social	11.563	6.541	0	48.72	0.074	0.028
LIWC Socrefs	6.542	5.332	0	36.17	0.07	0.045
LIWC Space	7.688	4.578	0	30.3	-0.016	0.02
LIWC Substances	0.084	0.465	0	10.2	0.019	0.018
LIWC Swear	0.025	0.213	0	4.23	-0.058	-0.02
LIWC Tech	0.682	1.653	0	19.05	0.055	-0.007

LIWC Tentat	1.583	2.181	0	15.79	-0.041	0.009
LIWC They	0.283	0.863	0	10	0.035	0.024
LIWC Time	3.959	2.977	0	24.39	-0.088	-0.01
LIWC Tone	79.83	26.516	1	99	0	0.03
LIWC Tone Neg	0.318	0.921	0	9.38	-0.043	-0.045
LIWC Tone Pos	6.986	4.917	0	31.03	0.039	0.034
LIWC Verb	15.177	5.054	0	36	-0.09	0.119
LIWC Visual	0.775	1.351	0	10.81	0.009	0.001
LIWC Want	0.321	0.829	0	8.99	-0.001	-0.007
LIWC Wordcount	60.942	17.212	28	97	-0.087	-0.006
LIWC We	1.479	3.23	0	22.58	0.04	0.029
LIWC Wellness	0.117	0.584	0	9.09	-0.001	-0.018
LIWC Work	4.9	5.389	0	40	0.039	-0.006
LIWC Words per sentence	15.624	6.985	3.47	97	0.014	-0.059
LIWC You	0.987	1.863	0	16.67	0.082	0.009
Part Of Speech CC	3.646	1.772	0	21	-0.042	0.056
Part Of Speech CD	0.668	0.998	0	16	-0.041	-0.039

Part Of Speech DT	4.359	2.361	0	17	-0.036	0.06
Part Of Speech EX	0.038	0.201	0	2	0.007	0.016
Part Of Speech FW	0.019	0.167	0	7	-0.012	-0.03
Part Of Speech IN	6.393	3.068	0	22	-0.074	-0.001
Part Of Speech JJ	6.444	3.134	0	23	-0.021	-0.059
Part Of Speech LS	0	0.012	0	1	-0.007	-0.012
Part Of Speech MD	0.523	0.833	0	7	-0.011	0.021
Part Of Speech NN	18.628	6.965	3	51	-0.024	-0.076
Part Of Speech PD	0.048	0.229	0	2	0.001	0.031
Part Of Speech PO	0.092	0.339	0	6	-0.007	-0.002
Part Of Speech PR	3.171	2.373	0	20	-0.014	0.022
Part Of Speech RB	3.026	2.459	0	20	-0.124	-0.01
Part Of Speech RP	0.254	0.538	0	4	-0.066	0.006
Part Of Speech SY	0.003	0.053	0	2	0.01	-0.005
Part Of Speech TO	1.992	1.508	0	13	-0.013	0.056
Part Of Speech UH	0.011	0.11	0	3	-0.019	0
Part Of Speech VB	11.998	4.558	0	29	-0.135	0.068



Part Of Speech WD	0.194	0.474	0	6	0.018	0.038
Part Of Speech WP	0.286	0.599	0	5	-0.018	0.016
Part Of Speech WR	0.218	0.503	0	4	-0.025	0.018
Contains List	2.25	2.125	0	26	0.024	-0.093
Number Negations	0.165	0.449	0	5	-0.055	-0.013
Number Of Addresses	0.003	0.053	0	1	0.005	0.021
Number Of Names	0	0.012	0	1	0.024	0.012
Number Of Numbers	0.783	1.559	0	30	-0.006	-0.034
Number Of Punctuation	8.255	5.18	0	174	-0.019	-0.043
Number Of Question Marks	0.04	0.277	0	9	-0.026	-0.005
Number Of Symbols	0.108	1.48	0	107	0.005	-0.005
URL Count	0.004	0.083	0	4	0.001	-0.021
Flesch Kincaid Grade Level	7.363	3.386	0	32.9	0.088	-0.113
Flesch Reading Ease Level	69.988	16.841	-23.45	111.78	-0.119	0.129
Sentiment AFINN	8.642	6.309	-16	44	0.014	0.047
Sentiment NRC Anger	0.009	0.019	0	0.22	-0.018	-0.032
Sentiment NRC Anticipation	0.077	0.054	0	0.316	-0.005	0.01

Sentiment NRC Disgust	0.007	0.017	0	0.22	0.002	-0.023
Sentiment NRC Fear	0.012	0.023	0	0.22	0.007	-0.049
Sentiment NRC Joy	0.107	0.076	0	0.5	-0.016	0.049
Sentiment NRC Negative	0.021	0.03	0	0.304	-0.012	-0.055
Sentiment NRC Positive	0.195	0.085	0	0.571	0.06	0.036
Sentiment NRC Sadness	0.017	0.026	0	0.222	-0.016	-0.014
Sentiment NRC Surprise	0.025	0.032	0	0.286	0.004	-0.012
Sentiment NRC Trust	0.093	0.062	0	0.429	0.061	-0.001
Sentiment Polarity	0.262	0.161	-0.443	1	0.025	0.018
Sentiment Subjectivity	0.51	0.148	0	1	-0.003	0.005
Sentiment Vader	0.812	0.265	-0.895	0.998	-0.021	0.015
Lexical Diversity	0.755	0.079	0.167	1	-0.016	-0.202
Character Count	341.203	107.151	126	705	-0.025	-0.058
Contractions Count	1.021	1.439	0	12	-0.152	0.02
Line Break Count	0.986	1.76	0	26	0.05	0.041
Longest Repetition Length	1.973	1.249	1	45	0.071	0.126
Mean Sentence Length	16	7.546	3.737	89	0.01	-0.066

Mean Word Length	4.565	0.559	3.265	7.933	0.142	-0.157
Number Of Exclamation Marks	0.39	0.843	0	21	-0.033	-0.03
Number Of Unique Words	46.039	11.648	15	77	-0.113	-0.089
Percentage Common 2-grams	0.048	0.055	0	0.385	-0.046	0.106
Percentage Common 3-grams	0.029	0.046	0	0.375	-0.025	0.113
Percentage Common 4-grams	0.011	0.043	0	1	-0.039	0.092
Percentage Common Words	0.156	0.096	0	0.688	-0.04	0.12
Percentage Rare 2-grams	0.691	0.153	0	1	0.082	-0.207
Percentage Rare Words	0.065	0.066	0	0.529	0.069	-0.223
Percentage Stop Words	0.476	0.075	0	0.733	-0.127	0.181
Word Density	0.183	0.018	0.112	0.241	-0.159	0.151
LDA Topic Vectors	Various techniques incl. structural topic models were explored but not used due to robustness and interpretability issues.					