ASSESSING THE EFFECTS AND RISKS OF LARGE LANGUAGE MODELS IN AI-MEDIATED COMMUNICATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Maurice Jakesch

December 2022

© 2022 Maurice Jakesch
ALL RIGHTS RESERVED

ASSESSING THE EFFECTS AND RISKS OF LARGE LANGUAGE MODELS IN AI-MEDIATED COMMUNICATION

Maurice Jakesch, Ph.D.

Cornell University 2022

Large language models like GPT-3 are increasingly becoming part of human communication. Through writing suggestions, grammatical assistance, and machine translation, the models enable people to communicate more efficiently. Yet, we have a limited understanding of how integrating them into communication will change culture and society. For example, a language model that preferably generates a particular view may influence people's opinions when integrated into widely used applications. This dissertation empirically demonstrates that embedding large language models into human communication poses systemic societal risks. In a series of experiments, I show that humans cannot detect language produced by GPT-3, that using large language models in communication may undermine interpersonal trust, and that interactions with opinionated language models change users' attitudes. I introduce the concept of AI-Mediated Communication–where AI technologies modify, augment, or generate what people say–to theorize how the use of large language models in communication presents a paradigm shift from previous forms of computer-mediated communication. I conclude by discussing how my findings highlight the need to manage the risks of AI technologies like large language models in ways that are more systematic, democratic, and empirically grounded.

BIOGRAPHICAL SKETCH

Maurice grew up on a little island between Germany and Switzerland. He was first trained as an electrical engineer (B.Sc.) at ETH Zurich. Through electives at the ETH Department of Humanities and an exchange to University of Toronto, he developed an interest in understanding how technologies shape people's lives. He continued his studies with an M.Sc. in Information Technology at the Hong Kong University of Science and Technology where he gained first research experience in social data analytics. To deepen his understanding of questions between technology and society, he completed an M.A. in Philosophy of Science and Technology at the TU Munich. In addition to exchange semesters in Singapore and Tanzania, he receiveed an honors degree in Technology Management at the Center for Digital Technologies and Management during this time. Looking for an interdisciplinary and dynamic environment for researching technology's societal impacts, Maurice joined the Cornell University Department of Information Science in 2017. He developed his first research program assessing the risks of AI technologies in human communication advised by Mor Naaman, Michael Macy, and Nathan Mathias. Maurice was an early student and Digital Life Initiative fellow at the new Cornell Tech campus in New York City. He was a visiting researcher at the MIT Institute for Data, Systems, and Society, as well as the Leibniz Institute for the Social Sciences (GESIS). Beyond his academic work, he interned at Microsoft Research, Facebook Core Data Science, and GE's Digital Industrial Foundries. The German National Academic Foundation, the Cornell Graduate School, and the US National Science Foundation have generously supported him.

This thesis is dedicated to my parents and grandparents

who have paved the way for me to pursue what I like.

ACKNOWLEDGEMENTS

This dissertation has been in the making for a while. If there is any clarity in its results, it belies the uncertainty of the way there. I want to thank those who have helped me along the way through their guidance, support, and company.

First, I am indebted to my advisor Mor Naaman, the most thoughtful and gentle giant you will meet around Central Park Summer Stage. He taught me to think like a water bender when working on research (and real-life) problems without an apparent angle of attack. He showed me the importance of asking good questions and helped me hone my intuition for perspectives that matter. He taught me the tools of the trade, made sure that I always had what I needed, and edited countless irreferential pronouns out of my drafts.

Michael Macy set an example to me of the value of scientific clarity and precision in times when the future of truth is uncertain. If I ever write a clean introduction, cogent paper title, or meaningful figure caption, I will have learned from him. Nathan Matias, the third member of my dissertation committee, has been a role model to me for taking an utopian research vision and turning it into reality step by step. Our conversations have helped me to reflect on the broader context of my research.

I want to thank all those that I was lucky to work with and learn from: Jeff Hancock, our favorite collaborationist at Stanford, for lending us his storytelling talent and the adventurous energy he brings into research. Xiao Ma, my predecessor in the Cornell Tech Social Technologies Lab, who developed ideas in her research on networked trust that I build on in this dissertation. I thank Arpita Ghosh for helping me think deeply about conceptual questions early in my Ph.D., and Sarah Kreps for showing me that research doesn't have to be all complicated later in my Ph.D.

I am grateful to my hosts at the different stations of a varied journey: Dean Eckles for my stay at MIT, Claudia Wagner at GESIS, Alexandra Olteanu and Saleema Amershi at Microsoft Research, Thomas Leeper at Facebook Core Data Science, and Hycham Basta at the GE Digital Foundries. I would also like to thank our superb admin teams at Cornell and Cornell Tech, particularly Barbara Woske, for their incessant support.

There are people without whom this journey would have been the Ph.D. grind that Philip Guo describes in his memoir: Palashi and Anthony, who welcomed me into the program have given me the best company between the firepit in Farm Street, board games on cabin weekends, and trips to the Yucatan Peninsula and Gujarat. Yiqing, my favored companion for off-off-broadway shows, Greek temple ruins, and research trips during pandemic outbreaks. Marianne and Natalie, who brought their wit and positive presence to Roosevelt Island every day. Ziv and Mattia, who made my stays in Boston and Cologne memorable. And Dana, Ru, Martin, Minsu, Matt, and Maggie, talking to whom has always been the best use of my time. Finally, I thank Katja and my family for covering my back and for bearing the distance and occasional absent-mindedness that the Ph.D. brought about. This dissertation would not have been possible without you.

CONTENTS

Chapter 1
Introduction

Large language models like GPT-3 (Brown et al., 2020; Vaswani et al., 2017) produce semantic artifacts closely resembling language created by humans. Enabled by developments in computer hardware and software architecture (Vaswani et al., 2017), large language models passably continue dialogue, narrative, technical and poetic writing. The models have also attracted attention for their ability to respond to a range of natural language tasks that they were not explicitly trained for (Brown et al., 2020). Figure 1.1 shows an example of natural language produced by GPT-3: Prompted to extend a fragment of the opening scene of Goethe's Faust (Goethe, 1808), the model creates a coherent and original continuation as it imitates the author's style and responds to the topic and mood of the prompt.

Large language models are increasingly becoming part of human communication. Applications such as grammatical assistants (Koltovskaia, 2020), writing suggestions (Dang et al., 2022), and machine translators (Gaspari et al., 2014) inject models' output in what people say, write and read. Large language models are currently at the research and development stage (Brown et al., 2020; Radford et al., 2019), but their capabilities have been advancing rapidly (Bommasani et al., 2021). When we conducted the initial study for this dissertation (Jakesch et al., 2019), we had to work with hypothetical scenarios as even comparatively weaker

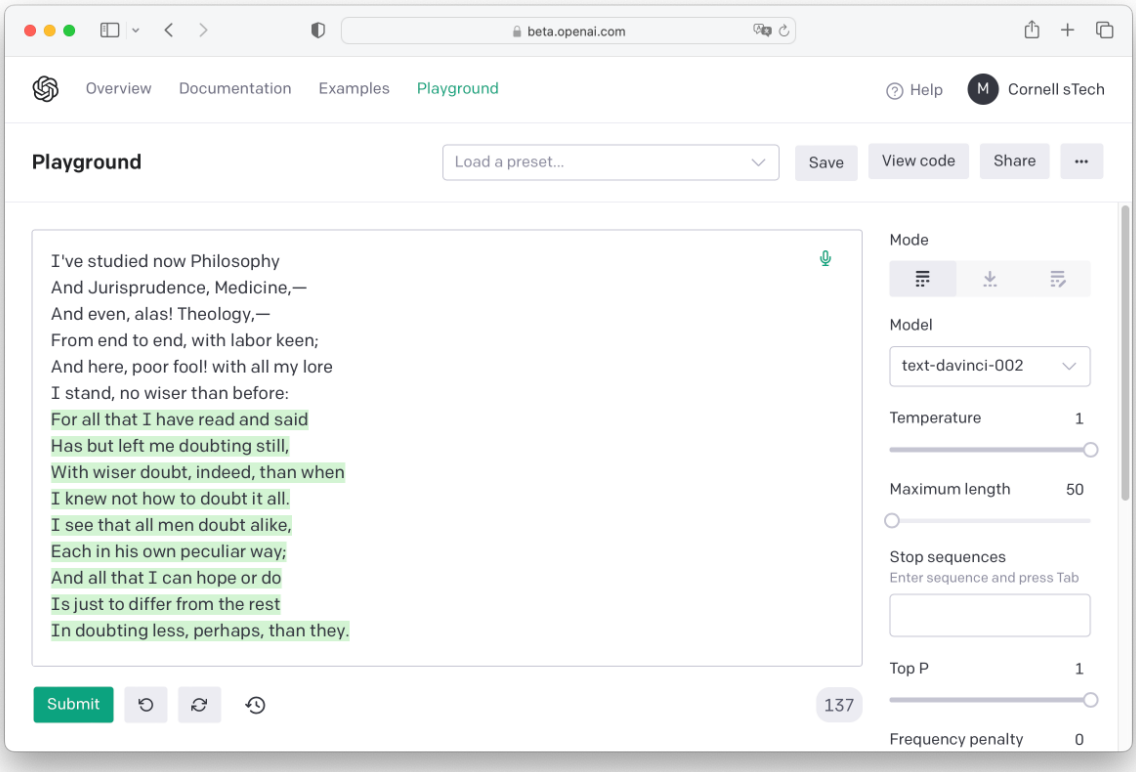

Figure 1.1: OpenAI's GPT-3 model continues the opening monologue from Goethe's Faust. The initial prompt is shown in black; text generated by the model is highlighted in green. The model imitates the authors' writing style and meter. It picks up on the speakers' frustration with the limits of human knowledge and completes the monologue coherently.

models like BERT or GPT-2 (Devlin et al., 2018; Radford et al., 2019) were not available yet. Three years later, we were able to experiment with powerful beta-level models offered for product integrations through commercial APIs[1]. As of writing, more than 36 billion messages sent daily (Mieczkowski et al., 2021b) are generated by language models through widely used *smart reply* (Kannan et al., 2016) features. As the models proliferate, we expect that a substantial part of what people read and say will become modified, augmented, or even generated by AI language technologies.

[1]See OpenAI's public API for its GPT-3 models (`https://beta.openai.com/`)

Using large language models in communication may enable people to respond to correspondences efficiently (Kannan et al., 2016), write mistake-free (Koltovskaia, 2020), translate between languages (Gaspari et al., 2014), and generate ideas (Stevenson et al., 2022). However, we have a limited understanding of how integrating such technology into human communication will change our society and culture (Bommasani et al., 2021). For example, if language models make it easier to express certain views, other views may be expressed less often. If models enable everyone to write convincingly, judging someone's skill or knowledge based on their writing may become difficult. If sending a thoughtful and elaborate reply becomes as simple as a mouse click, relationships may change. Far from minor side-effects of technological innovation, technology-induced changes in relationships, judgment, or discourse have far-reaching political and cultural consequences (McLuhan, 1994).

Previous research on the risks of large language models has looked at the risk of adversarial use and at harms caused by offensive or discriminating output. A commonly voiced concern is that large language models may enable highly automated forms of fake news, disinformation, and propaganda (Buchanan et al., 2021; Zellers et al., 2019; Evans et al., 2021). Initial studies have demonstrated that current models generate credible news stories (Kreps et al., 2022a) and compelling extremist text (McGuffie and Newhouse, 2020) that individuals cannot identify as generated (Clark et al., 2021; Ippolito et al., 2019). Language models also can degenerate into producing toxic or offensive content from even innocuous prompts (McGuffie and Newhouse, 2020; Askell et al., 2021; Rae et al., 2021). They learn stereotypes and biases from their training data (Fortuna and Nunes, 2018; Garrido-Muñoz et al., 2021) that they may amplify (Caliskan et al., 2017; Blodgett et al., 2020). Researchers have further been attempting to map out the ethical and so-

cietal risks of large language models more generally (Bender, 2019; Bommasani et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021). However, most previous work on the risks of large language models in communication has been speculative or policy-oriented. There is little principled empirical work investigating how deploying large language models into communication may change our society and culture.

Empirically assessing the effects and risks of large language models in communication has its unique challenges. Since the models are new and not widely deployed yet, what can be learned through observation or analogy is limited. Due to the amount of resources and infrastructure required to operate large language models only a few private companies can train state-of-the-art models. Some companies have made model functions available through public APIs, but the lack of access to technology and usage data limits impact assessments by third parties (Bommasani et al., 2021). Assessments are further complicated as the effects of a language model in a *single* interaction with *individual* users will likely be minimal, and only repeated interactions with a large group of users will surface the model's full effects.

Yet, early impact assessments are necessary as once the technology is widely deployed, it becomes costly or impossible to change (Genus and Stirling, 2018). Even if early assessments involve an amount of guesswork, they can reduce the uncertainty about how large language models will affect communication. Interviews and surveys tell us how people expect to use the models. Qualitative field studies and data analyses can explore how large language models may affect the people and communities. And quantitative lab experiments allow us to robustly evaluate the effect of large language models in a controlled environment. In this thesis,

we have chosen the latter approach. We hypothesize how large language models might be used in future applications and create speculative prototypes of what such applications may look like. We then recruit a large number of participants and observe their interactions with the prototypes to understand how the model affects their thoughts and behavior.

We empirically demonstrate that embedding large language models in human communication poses systemic societal risks that require careful evaluation and management. Our contribution is threefold:

1. We conduct **four quantitative studies** providing strong evidence that embedding large language models in communication has consequences for people's ability to make judgments, form opinions, and trust each other.
2. We introduce the concept of **AI-Mediated Communication** (AI-MC) to theorize how communicating through AI technologies like large language models presents a paradigm shift from previous forms of Computer-Mediated Communication.
3. We develop **methods for assessing the effects** of AI technologies like large language models that provide empirically grounded insights into relevant risks even without direct access to the technology.

We next provide detail on the different parts of this thesis: Chapter 2 tests whether humans can tell whether one of the most personal and consequential forms of language – a self-presentation – was generated by AI. If people cannot identify self-presentations generated by AI systems, they will be vulnerable to novel automatized forms of deception, fraud, and identity theft (Biderman and Raff, 2022; Bommasani et al., 2021; Cooke, 2018; Floridi and Chiriatti, 2020; Buchanan et al.,

2021; Weidinger et al., 2022). Across six experiments, we show that humans cannot identify self-presentation produced by large language models like GTP-2 or GPT-3. We also demonstrate that human judgment of AI-generated language is handicapped by intuitive but flawed heuristics such as associating first-person pronouns, spontaneous wording, or family topics with humanity. These heuristics make people's impressions of generated language predictable and *and* manipulable, allowing AI systems to produce language perceived as more human than human. We discuss solutions to reduce the deceptive potential of generated language.

Chapter 3 shows that people's inability to detect generated language has consequences for how people trust each other and mediated communication in general. Trust established through impression formation based on self-descriptions is crucial for various social interactions (Ert et al., 2016; Ma et al., 2017a). AI systems that generate human-like self-presentations may invalidate signals that people rely on when assessing others (Hancock et al., 2020), such as tone or compositional skill. In three experiments, we test whether people consider others less trustworthy if they believe their self-presentation was generated by AI. We find that people try to identify who uses AI technologies in communication and mistrust those who they think use them. However, since people cannot identify generated language (as shown in Chapter 2) they will often mistrust others who do not use AI in communication but sound artificial for other reasons.

In Chapter 4, we extend our investigation of the effects and risks of large language models to the realm of politics and opinions. Here, large language models may not only perpetuate biases and stereotypes (Huang et al., 2019; Brown et al., 2020; Nozza et al., 2021), but may change people's opinions at an unseen scale. We developed a custom experimental platform to empirically test how language

models that produce certain views more often than others influence what their users write and believe. In an online experiment, we asked participants to articulate their opinion in a discussion. Some participants saw suggestions from a large language model configured to support a specific side of the debate. We measured participants' post-task opinions in a survey and asked a separate set of judges to evaluate participants' written opinions. Our results show that interacting with an opinionated language model affects written opinions and reported attitudes in a subsequent survey.

Chapter 5 proceeds from assessing the effects and risks of using large language models to the question of AI risk management. Given that large language models likely change our cultural and political landscapes, we ask how risks could be assessed in more democratic and inclusive ways. Currently, engineering teams and expert groups decide on the development and deployment of communicative AI systems, often without consulting wider and more diverse populations that may be affected by the technology. We develop opinion research methods to ask a wider public about their priorities for AI risk management. Drawing on prior work on empirical ethics, value elicitation, and responsible AI, we create an AI value survey that compares people's valuations of AI risks different demographic groups. We field the survey with AI practitioners, crowdworkers, and a representative sample. Our results show that the risk priorities of AI practitioners significantly differ from those of the general public. AI practitioners appear to consider responsible AI values as less important, while self-identified women and black respondents found a responsible approach to AI risks more important than other groups. Our findings demonstrate how important it is to pay attention to who makes decisions about large language models' risks.

In the concluding chapters, we argue that embedding large language models in human communication presents a paradigm shift from previous forms of Computer-Mediated Communication with risks that need to be managed more carefully. We introduce the concept of AI-Mediated Communication–where language models augment, optimize, or generate what people say–and describe how the work by us and others has contributed to an initial understanding of its effects. We discuss how interdisciplinary research can reduce the risk of adverse and unwanted consequences. Finally, we argue that our findings highlight the need to manage the risks of AI technologies like large language models in ways that are more systematic, inclusive, and empirically grounded.

Chapter 2

Human Heuristics for AI-Generated Language are Flawed

In this chapter, we describe six experiments, participants (N = 4,600) examining whether humans can discern whether one of the most personal and consequential forms of language – a self-presentation – was generated by a large language model. Across professional, hospitality, and dating settings, we find that humans are unable to detect AI-generated self-presentations. The results also indicate that 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.

Introduction

Large language models like GPT-3 (Brown et al., 2020; Vaswani et al., 2017) produce semantic artifacts closely resembling language created by humans. 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 plagiarism, manipulation, and deception (Brown et al., 2020; Biderman and Raff, 2022; Bommasani et al., 2021; Cooke, 2018; Floridi and Chiriatti, 2020; Buchanan et al., 2021; Weidinger et al., 2022) when people mistake it 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 (Schlenker, 2012), 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 (Van Der Heide et al., 2012), e.g., on professional or dating platforms. Researchers have extensively studied the importance of online self-presentation (DeVito et al., 2017; Ellison et al., 2006; Schwämmlein and Wodzicki, 2012), showing that impression formation based on self-descriptions is crucial for establishing the trust required for various social interactions (Ert et al., 2016; Ma et al., 2017a). AI systems that generate human-like self-presentations may invalidate signals that people rely on when assessing others (Hancock et al., 2020), such as tone or compositional skill. Earlier work has shown that interpersonal trust declines when people suspect that others are using AI systems to generate or optimize their self-presentation (Jakesch et al., 2019).

Previous studies suggest that people struggle to discern AI-generated language

in different settings (Clark et al., 2021; Köbis and Mossink, 2021; Kreps et al., 2022a). Here, we go beyond previous work by providing strong evidence of the flawed heuristics people use to detect AI-generated language. Using qualitative, quantitative, and computational methods, we first reconstruct a set of potential heuristics that people may rely on to detect AI-generated language, expanding on related analyses in previous work (Clark et al., 2018). We then measure the extent to which people actually use these heuristics and to what extent 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.

To examine how people detect AI-generated self-presentations, we performed six experiments broadly patterned after the Turing test (Pinar Saygin et al., 2000). While participants in the Turing test try to identify a language-generating machine through a text-based conversation, participants in our studies were asked to evaluate whether a personal self-presentation was written by a person or generated by an AI system. We trained multiple customized versions of state-of-the-art AI-based large language models (Brown et al., 2020; Vaswani et al., 2017; Bommasani et al., 2021) to generate self-presentations in three social contexts where trust in a self-presentation is important for decision-making: professional (e.g., job applications) (Guillory and Hancock, 2012), romantic (e.g., online dating) (Schwämmlein and Wodzicki, 2012), and hospitality online services (e.g., Airbnb host profiles) (Ma et al., 2017a). 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, some collected from real-world online platforms–and indicate which ones they thought were AI-generated.

Results

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 in three different social contexts, participants identified the source of a self-presentation with only 50% to 52% accuracy. These results, including a breakdown by experiments and treatments, are shown in Figure 2.1. In the hospitality context (shown in the left panel), participants rated human-written self-presentations as human or AI-generated self-presentations as generated 52.2% of the time. In the dating context, we introduced experimental treatments testing whether incentivizing participants to increase their efforts (Karpinska et al., 2021) would increase their accuracy. In the professional context, we tested whether providing training (Clark et al., 2021) 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 Figure 2.1, 51.6%) and when receiving immediate feedback on their evaluations (right bar in the third panel, 51.2%). Further analyses (included in the extended materials section) 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 $= 0.067$, $p < 0.0001$). Had this level of agreement been due to valid cues that differentiated human and AI-generated self-presentations, participants' accuracy would have 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 that participants relied on to identify AI-generated language.

As a first step to investigate participants' heuristics for AI-generated language, we conducted a qualitative analysis of the heuristics participants thought they applied. 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, overlapping and extending categories identified in previous research (Clark et al., 2021). 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 language created by humans. Replicating findings from earlier research (Clark et al., 2021), 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. Details on participants' self-reported explanations of their judgments are included in the extended materials section.

As self-reports on mental processes can be unreliable and even misleading (Cox, 2005), we conducted additional analyses to evaluate participants' judgments independently of their self-reported explanations. While participants may not always know why they did something (Pennebaker, 2011), a multi-paradigm approach (Slovic and Lichtenstein, 1971) based on a statistical analysis of their judgments combined with a computational analysis of language features in the self-descriptions allows us to independently reconstruct heuristics they rely on (Berger

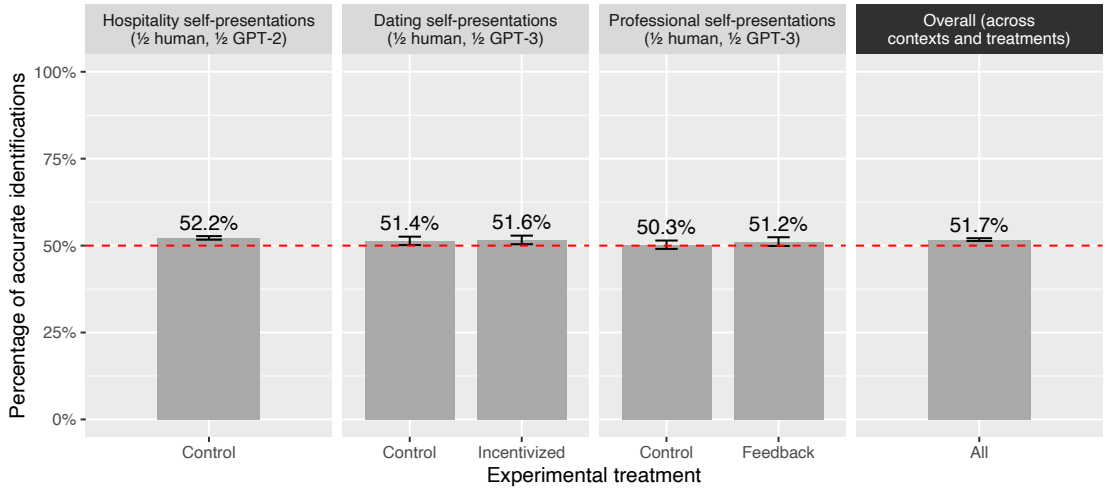

Figure 2.1: Participants could not detect self-presentations generated by the current generation of language models beyond chance in our three main experiments. Error bars represent 95% confidence intervals for 6,000–16,000 judgments of 2,000–3,000 self-presentations per bar. Across three social contexts, discernment remained close to chance. Providing monetary incentives for accurate answers or telling participants whether their answers were correct did not increase accuracy.

et al., 2020). Rather than drawing conclusions from participants' self-reported heuristics like previous research (Clark et al., 2021), 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 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 through 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 self-presentations used in the main experiments. The results show that crowdworkers' ratings in the labeling task, to some extent, differentiate 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 (see extended materials section).

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 AI-generated. Our 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 self-presentations being AI-generated (on the right) in the three main experiments.

Table 1: *Logistic regression odds ratios predicting whether (1) participants in the three main experiments rated a self-presentation as AI-generated and (2) whether it actually was generated. Only nonsense, repetition, and conversation were functional cues, indicated by equal coefficient directions (same text color) in models (1) and (2). Other heuristics were either inversely related (different color) or unrelated (black) to features in the actual source of the self-presentation.*

	Dependent variable:	
	(1) Perceived as AI-generated	**(2) Actually AI-generated**
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

†manual labels, *p**p***p<0.001

For example, the top row in Table 1 shows that self-presentations containing nonsensical content were 10.5% more likely to be seen as AI-generated and, indeed, were 23% more likely to be generated by AI. 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 AI-generated, 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, longer self-presentations that sounded authentic or spontaneous (Newman et al., 2003) or were focused on past events were more likely to be rated as human by participants but were equally likely to be human-written and AI-generated.

Following the correlation analysis, we tested whether the presence of language features in a self-presentation would predict participants' judgments. The regression model predicted participants' judgments of AI-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 AI-generated language without feature engineering input from the research team. A current language model (Radford et al., 2019) 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 people not only rely on flawed heuristics to detect AI-generated language but that AI systems can predict people's

judgments.

We conducted three additional experiments to validate and extend these findings: If the three main experiments correctly identified features associated with the perception that self-descriptions are human-written, new 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 also tested whether language models can exploit people's flawed heuristics to produce self-presentations perceived as "more human than human." For the validation experiments, we created a new sample of human-written and AI-generated self-presentations. We used the classifiers trained on participants' judgments in the main studies to create a set of AI-generated self-presentations optimized for perceived humanity.

Figure 2.2 shows that participants evaluated the AI-generated self-presentations optimized for perceived humanity as more human than the human-written and the non-optimized 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 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 self-presentations was strongest in the professional context, where we combined the regression- and language-model-based classifiers to select self-presentations that were perceived as human 71% of the time.

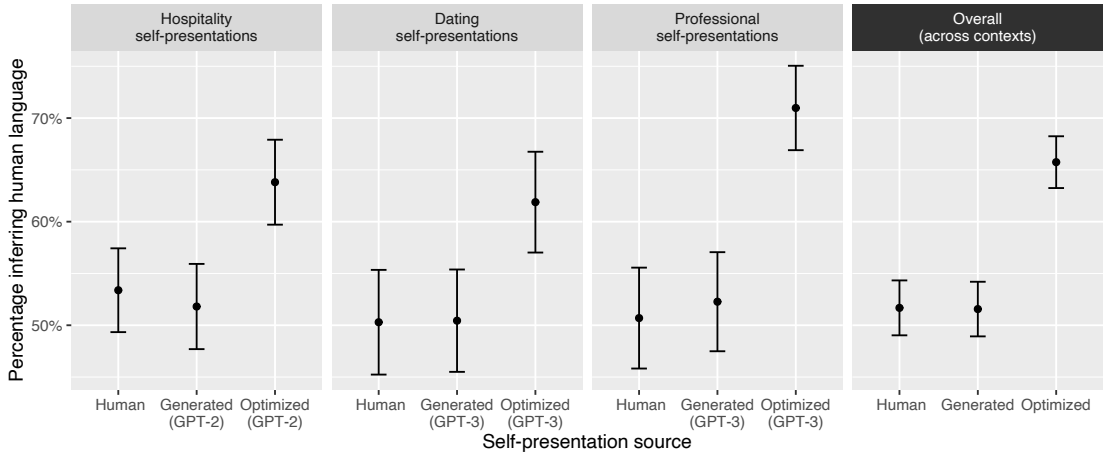

Figure 2.2: Exploiting humans' flawed heuristics, the three validation experiments show that AI systems can generate verbal self-presentations perceived as more human than human-written ones. Error bars represent 95% confidence intervals for 350 to 450 judgments of 100 self-presentations per bar.

Discussion

Our results affirm 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 (Clark et al., 2018; Kreps et al., 2022a; Köbis and Mossink, 2021). 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 (Bond Jr and DePaulo, 2006; Hartwig and Bond Jr, 2011), 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 may be able to detect, lending support to the latter explanation. When explicitly asked about the presence of nonsensical text, repetitiveness, or grammar issues in the self-presentations, crowdworkers evaluated AI-generated text as nonsensical or repetitive more often than human-written self-presentations. However, when we directly asked participants whether self-presentations were AI-generated in the main experiments, their accuracy in identifying generated self-presentations remained close to chance.

Our analysis of the heuristics people use to identify AI-generated language provides a more nuanced picture than previous research: While people can sometimes identify a few characteristics of AI-generated language, they also rely on other flawed cues that invalidate their judgment. Participants in our studies relied on some functional cues, such as nonsensical and repetitive text, to identify AI-generated verbal self-presentations. Had participants relied on those cues only, they would have achieved a detection accuracy of about 59%. 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. Many other cues 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 flawed 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 (Clark et al., 2018; Kreps et al., 2022a; Köbis and Mossink, 2021).

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 increases the perceived humanity of AI-generated self-presentations. Recent work by Ippolito et al. Ippolito et al. (2019) suggests that language model decoding methods have been optimized for "fooling" humans at the cost of introducing statistical anomalies that are easily detected by machines. Previous research also suggests that domain expertise can be somewhat more effective than personal intuition in identifying AI-generated content (Karpinska et al., 2021). 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. As demonstrated in our three validation experiments, AI systems can use people's flawed heuristics to manipulate their judgments and produce language perceived as more human than human.

People's inability to detect AI-generated language has important 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 (Ischen et al., 2019), but they may start distrusting those they believe are using AI-generated language in their communication (Jakesch et al., 2019). People's flawed heuristics also can be exploited by malevolent actors. From automated impersonation (Weidinger et al., 2022) to targeted disinformation cam-

paigns (Zellers et al., 2019), AI systems could be optimized to undermine human intuition, exacerbating concerns about novel automatized forms of deception, fraud, and identity theft (Biderman and Raff, 2022; Bommasani et al., 2021; Cooke, 2018; Floridi and Chiriatti, 2020; Buchanan et al., 2021; Weidinger et al., 2022). Widespread AI education and technical tools that assist identification (Gehrmann et al., 2019; Hashimoto et al., 2019; Dou et al., 2022) 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 (Clark et al., 2018), especially given the possibility of future adaptations of language models that may invalidate learned heuristics (Ippolito et al., 2019).

At the same time, when and 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 U.S. White House calls for "Notice and Explanation" when "an automated system is being used" (Nelson et al., 2022a). 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" (Commission, 2021). However, such policies can be difficult to apply when AI technologies modify, augment or generate communication between people. For example, it hardly seems necessary to add notice and explanation to every message when people communicate with AI-enabled auto-corrections or translations. Prior research also shows that typical notice and consent disclosures are often ignored by users (Acquisti et al., 2022). Identifying context-appropriate and effective disclosure mechanisms for the use of AI in communication is an urgent question that requires further research (Williams, 2018).

In some cases, though, our results suggest that it may be possible to create language models that are self-disclosing by design: Rather than training AI language systems that imitate human language, AI systems could be optimized to fulfill their specific communicative function while preserving the validity of human intuitive judgment (Ippolito et al., 2019). Many AI applications could use language that is clearly not written by humans without loss of functionality and avoid generating language that people wrongly associate with humanity, such as first-person speech or family words. Explicit disclosures that preserve the fluidity of communication might also be achieved through dedicated AI accents: AI language standards could require systems to generate language with a dedicated dialect that would enable intuitive identification without interrupting the flow of communication. Rather than undermining human intuition, AI systems that accommodate the limits and flaws of human judgment by design will genuinely support human communication and reduce the risk of misuse.

Materials and methods

Experiment design

The design of the six experiments combined elements of a simplified Turing test (Pinar Saygin et al., 2000) 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 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. 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 their accuracy for their subsequent ratings (see extended materials section 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/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 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 (Karpinska et al., 2021), half of the participants in the dating context received 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 (Clark et al., 2021), 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 extended materials section.

Collecting and generating self-presentations

We collected data from real-world platforms in each of the three contexts for the experiments. We used the data collected to train state-of-the-art large language models to generate self-presentations. We used different AI models for generating self-presentations as new and more powerful models were made 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 the extended materials section.

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 (Radford et al., 2019) for four epochs with a learning rate of 0.00002 on the collected data. We used the fine-tuned model and nucleus sampling (Holtzman et al., 2019) 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 (Kim and Escobedo-Land, 2015). We drew a random sample of 1,000 human self-presentations for the professional and dating main experiments. We used the full set of collected self-presentations in each of these two contexts to fine-tune a 13B parameter version of GPT-3 (Brown et al., 2020) for four epochs with a learning rate multiplier of 0.1. We produced 1,000 AI-generated self-presentations for each experiment using the fine-tuned models

with temperature sampling at t=0.9. We used multiple techniques to confirm 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.

To perform the quantitative language analysis of participants' judgments in the three main experiments, we developed a set of text-based language features. The full set of about 180 features is included in the extended materials section. We used two different 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 statistical metrics, readability scores, emotion classification, and psychological language features (Tausczik and Pennebaker, 2010). 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 (1) seemed nonsensical, (2) contained repetitive text, or (3) had grammatical issues.

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 the use of pre-developed 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 (Radford et al., 2019) 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 new sample of 100 human-written self-presentations from the collected data. We produced a new 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 self-presentations, we first generated a large set of self-presentations in each context using the same models as in the initial experiments. We then used the classifiers developed above to remove self-presentations that would likely be perceived as AI-generated. We used different classifiers to select self-presentations in each context to increase generalizability and to independently validate both the regression and language-model-based classifiers. In the dating context, we used the regression-based classifier on the GPT-3 output to remove those generated self-presentations that were more likely to be perceived as AI-generated. 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 removed the top 80% percentile of self-presentations that the classifier predicted were likely to be perceived as generated by AI. From the remaining 20%, we drew a random sample of 100 self-presentations optimized for perceived humanity 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 (Coppock and McClellan, 2019). The experiment 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 (Palan and Schitter, 2018), 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. The median time participants spent on evaluating each self-presentation was 14.3 seconds (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. Finally, 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 (Macnish and van der Ham, 2020).

Extended Materials

Below we provide additional information on several aspects of our experiments. Table S2.1 and S2 summarize the treatment, stimuli, and recruitment methods used across the six studies and three labeling tasks. Table S2.3 shows a sample of

self-presentations for each study and treatment group.

Table S2.4 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 AI-generated self-presentations, with participants older than 50 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.

Figure 2.3 provides further detail on the qualitative analysis of participants' explanations of why they thought certain self-presentations were AI-generated or human-written. Two researchers independently coded a sample of responses into themes to provide an overview of participants' self-reported heuristics. Figure 2.3 presents an overview of recurring themes. Participants most commonly referred to the content of a self-presentation (blue-shaded regions in Figure 2.3 representing 40% of responses). The participants reported associating specific content related to family and life experiences with language written by humans and generic or nonsensical content with AI-generated language. Participants also reported basing their decisions on grammatical cues (gray, 28%), where first-person pronouns and the mastery of grammar were mentioned as indicative of human-generated language. Some participants saw grammatical errors as associated with a subpar AI, but others claimed they associated them with fallible human authors. Another category of cues mentioned by participants was the tone (green, 24%). Participants

reported associating warm and genuine language with humanity and impersonal, monotonous style with AI-generated language.

Prior research suggests that asking participants to explain their responses could have changed their subsequent evaluations or degraded performance (Wilson and Schooler, 1991). We thus conducted an analysis testing whether participants' performance had changed after being asked to explain their judgment. The results are shown in Figure 2.4. There was no evidence for such change in our data as participants' accuracy before and after the open-ended response did not change across any of the three contexts. Note that open-ended responses were only solicited for the three main experiments. The validation experiments did not include open-ended responses, showing similar outcomes and providing further evidence that participants' ratings — and our findings — were not affected by the explanations.

Figure 2.5 shows how crowdworkers evaluated human-written and AI-generated self-presentations in a separate labeling task when asked whether the text was nonsensical, seemed repetitive, or had grammatical issues. Crowdworkers were significantly more likely to rate AI-generated self-presentations as nonsensical (13.6% vs. 9.6%, $p<0.0001$). This was the case in the hospitality context in particular where we had used the older GPT-2 model to generate self-presentations. Crowdworkers also rated generated self-presentations as more repetitive (12.7% vs. 7.1%%, $p<0.0001$), particularly in the professional context. Finally, crowdworkers labeled generated self-presentations as having fewer grammatical issues than human-written text (14.8% vs. 19.6%, $p<0.0001$). This difference was most pronounced in the dating and professional contexts where we had used the more advanced GPT-3 model to generate self-presentations.

Table S2.1: *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 S2.2: *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 S2.3: *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 SS. 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 GPT-2)	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 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 S2.4: *Regression coefficients predicting the accuracy of a judgment based on treatment, social context, and participant demographics. No group performed much above chance level.*

	Dependent variable:
	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	**1.002**** (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	**0.986**** (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
Note:	*p**p***p<0.001

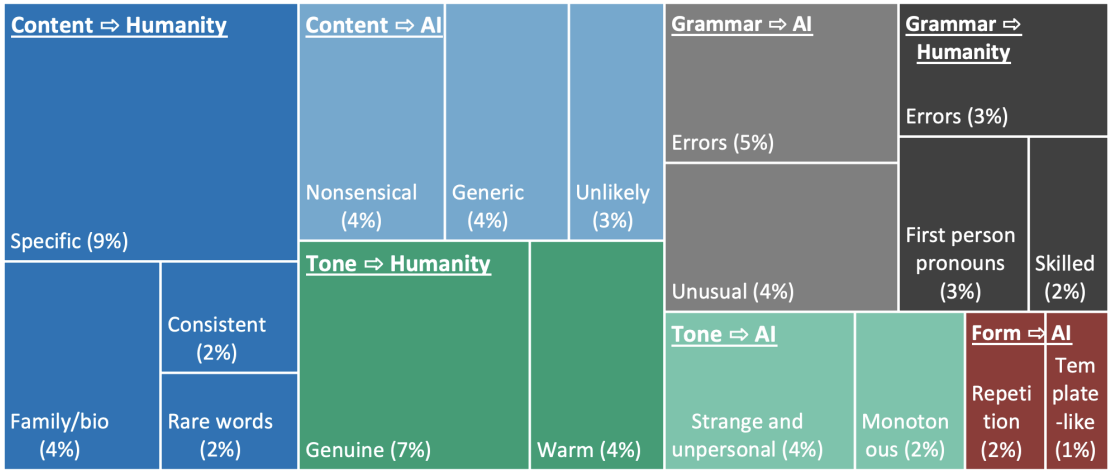

Figure 2.3: Themes in participants' explanations of why they thought a self-presentation was human or generated language. N = 800, tile areas correspond to theme prevalence reported in brackets. Heuristics are classified by whether they refer to the content (blue), tone (green), grammar (gray), or form (red) of a self-presentation. Lighter tiles show cue that were associated with generated language.

Figure 2.4: Participants' performance in identifying generated self-presentations did not change throughout the experiment. Error bars represent 95% confidence intervals for 6,000–16,000 judgments of 2,000–3,000 self-presentations per bar.

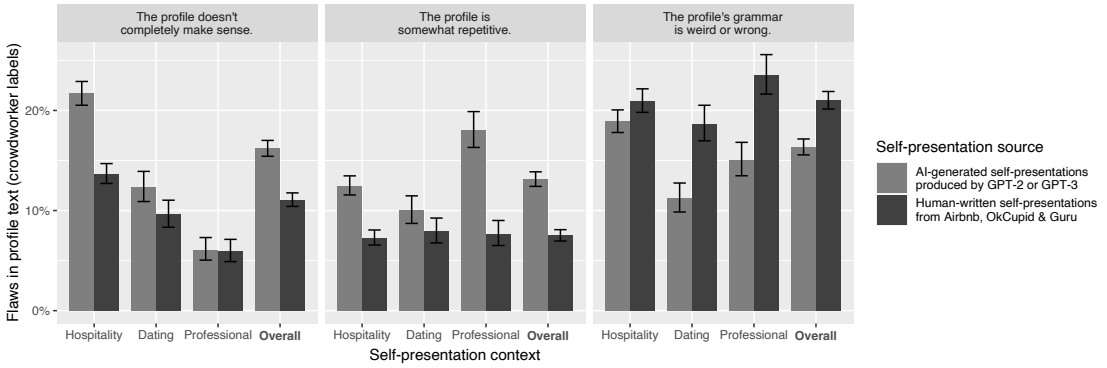

Figure 2.5: Participants in a separate labeling task rated AI-generated self-presentations as nonsensical and repetitive more often than human-written self-presentations. Error bars represent 95% confidence intervals for 1,898–4,704 judgments of 1,000–1,500 self-presentations per bar.

Chapter 3

The Suspicion That Text was Generated Reduces Trustworthiness

This chapter shows that people's inability to identify generated language has far-reaching implications and may undermine trust in other and mediated communication more generally. In three experiments we test whether people find Airbnb hosts less trustworthy if they believe their profiles have been written by AI. We observe a new phenomenon that we term the *Replicant Effect*: Only when participants thought they saw a *mixed* set of AI- and human-written profiles, they mistrusted hosts whose profiles were labeled as or suspected to be written by AI. Our findings have implications for the design of systems that involve AI technologies in online self-presentation and chart a direction for future work that may upend or augment key aspects of Computer-Mediated Communication theory.

Introduction

Using large language models in communication can impact interactions from one-to-one exchanges such as messaging to one-to-many broadcasts like writing user profiles or appearing in a live YouTube video. In text-based communication—the focus of this work—we have already advanced from spell check and predictive auto-

completion to using large language models to generate our communication, like the aforementioned auto-responses for chats and e-mails (Hohenstein and Jung, 2018). AsThis use of large language models in interpersonal communication challenges assumptions of agency and mediation in ways that potentially subvert existing social heuristics (Ellison et al., 2012; Walther, 2011; Herring, 2002).

This series of studies examines the effect of large language models' use on online self-presentation dynamics, making theoretical contributions to a rich area of research (Ellison et al., 2006; Schwämmlein and Wodzicki, 2012; DeVito et al., 2017; Uski and Lampinen, 2014). We ask a basic question, as appropriate for an early study: Does the belief that AI may have written a profile affect evaluations by others? In particular, to borrow the terminology of the Hyperpersonal Model (Walther, 2011, 1996), will *receivers* evaluate *senders* differently if they believe AI is involved in authoring the *sender's* profile?

We study this question in the context of online lodging marketplaces like Airbnb (Newman and Antin, 2016; Guttentag, 2015) with a focus on host trustworthiness. Trust and deception in online self-presentation have been studied extensively (Guillory and Hancock, 2012; Toma and Hancock, 2012; Hancock et al., 2007a,b, 2004) and have been shown to play a critical role in online marketplaces (Ert et al., 2016; Lauterbach et al., 2009; Lampinen and Cheshire, 2016). The Airbnb scenario also allows us to build on previous work that investigated the influence of profile text on the trustworthiness of Airbnb hosts (Ma et al., 2017a,b).

In a series of three online experiments, we examine how the belief that a language model has generated a host's profile changes whether the host is seen as trustworthy by others. Study 1 compares how hosts are evaluated in two hypothetical systems: one where profiles are supposedly written by AI, and one where

hosts wrote their own profiles. In Study 2 and 3 participants evaluate hosts in an environment where they believe some profiles have been generated using AI, while others have been written by the hosts. In reality, all profiles shown were written by humans, and were selected from a publicly available dataset of Airbnb profiles (Ma et al., 2017a) since we have already shown that humans are unable to differentiate between generated and human self-presentations.

Our results show that that (1) when people are presented with *all AI-generated* profiles they trust them just as they would trust *all human-written* profiles; (2) when people are presented with a *mixed* set of AI- and human-written profiles, they mistrust hosts whose profiles they believe were generated by large language models. Our results lend support to the Hyperpersonal Model of CMC (Walther, 1996) where receivers tend to exaggerate perceptions of the message sender, or, in this case, exaggerate textual hints that a profile was written by language models. In the discussion, we draw on relevant theories that may explain our findings.

Background

Our inquiry is motivated by the maturing ability of AI systems to *generate* natural language as well as the increasing use of AI in online self-presentation. Previous work in CMC has studied online self-presentation (Ellison and Boyd, 2013; Lampe et al., 2007) and the nature of human interactions with bots and agents (Nass and Moon, 2000; Ferrara et al., 2016; Corti and Gillespie, 2016; De Angeli et al., 2001). We also relate our work to previous studies on the perceived trustworthiness of user profiles.

Impression formation

CMC research has extensively studied how people present themselves online via technology (Ellison and Boyd, 2013; Lampe et al., 2007). We expand on this research by analyzing how the introduction of AI into online self-presentation might shift impression formation. Using large language models in communication may influence how people interpret and scrutinize the content of profiles, as users interpret signals presented online to infer characteristics about other individuals (Walther et al., 2009, 2005; Ellison and Hancock, 2013). The Hyperpersonal Model (Walther, 2011, 1996), for example, argues that receivers may over-interpret cues from the sender's self-presentation because of the reduced cues in text-based CMC. When certain cues can be easily modified with the help of AI, receivers have to change how they evaluate them.

A number of theories touch on how information shared in online profiles becomes credible. Walther (Walther and Parks, 2002) introduced the principle of *warranting* to CMC, asserting that receivers rely more on information that is difficult for the sender to manipulate (DeAndrea, 2014). The warranting idea is highly related to *signaling theory*, used by Donath (Donath, 2007) to explain why online signals vary in their reliability as proxies of the sender's underlying qualities–from easily faked self-descriptions (e.g., "I go to Cornell") to difficult to fake signals (e.g., having a cornell.edu email address). The *Profile as Promise* framework explains how people assess signals in online profiles when they expect future interactions, like in online dating, or in lodging marketplaces. The framework asserts that people are expected to make minor–but not significant–misrepresentations in their profile.

Introducing large language models to interpersonal communication may complicate these theories and models. Will self-descriptions generated by a language model be treated as "warranted", as earlier research suggested (Ma et al., 2017c)? Can language models give credible promises on behalf of the sender? Will large language models change the assessment of online signals, and result in different behaviors by senders and receivers when people optimize their self-presentation algorithmically, as seen in recent work (DeVito et al., 2017)? Studying large language models in the context of online self-presentation will test and extend these theories.

Interactions with bots and AI agents

Since Weizenbaum's early study (Weizenbaum, 1966), a large body of research adjacent to our work on large language models has explored natural language communication between man and machine. We know that people tend to apply social rules and attributes to computers (Nass et al., 1994; Nass and Moon, 2000). Technological advances now allow agents to produce more human-like dialogues. Studies of social bots (Ferrara et al., 2016) find that in these dialogues, people put more effort into establishing common ground when they perceive an agent as human (Corti and Gillespie, 2016) and that introducing anthropomorphism may generate strong negative user reactions (De Angeli et al., 2001).

Various researchers have explored how humans perceive machine generated content: In the context of automated journalism, scholars have observed differing levels of perceived credibility of computer-written articles: in some cases, computers were perceived as less credible, explained by the heuristic that machines are more artificial than humans (Graefe et al., 2018; Waddell, 2018). In other cases, there

was no difference in the perceived credibility of human- and computer-written news (Wölker and Powell, 2018), potentially because machines are seen as more objective than humans (Sundar, 2008).

Unlike in interactions with bots, when large language models are used in communication they are *not* communicating on their own behalf, but on behalf of a person in interpersonal exchange. The findings and outcomes of past studies need to be re-evaluated in settings where bots and AI agents are used for interpersonal communication. Some early work suggests that the involvement of AI through "smart replies" can influence conversations, for example by offering primarily positive suggestions (Hohenstein and Jung, 2018).

Trustworthiness, profiles, and Airbnb

We situate our work in the context of online lodging marketplaces, specifically Airbnb, where a range of prior work (Ert et al., 2016; Guttentag, 2015; Ma et al., 2017a,b) and publicly available data sets (Ma et al., 2017a) allow us to ground our experiments in existing methods and discussions.

The trust that can be established based on user profiles (Gibbs et al., 2011; Ert et al., 2016) is central to the functioning of social interactions and exchange, from online dating (Gibbs et al., 2006; Toma et al., 2008; Ma et al., 2017c) to resumes (Guillory and Hancock, 2012) and lodging marketplaces like Airbnb (Ert et al., 2016; Lauterbach et al., 2009; Lampinen and Cheshire, 2016; Ma et al., 2017a). On Airbnb, *hosts* list properties that *guests* can book and rent. Hosts aim for their profiles to appear trustworthy, especially in situations where reputation signals are either unavailable or skew positively high for everyone (Zervas et al.,

2015).

The current work directly builds on a recent study of the trustworthiness of Airbnb hosts based on profile text (Ma et al., 2017a). The study revealed that the profile text impacts the perceived trustworthiness of hosts in a reliable way; in other words, the evaluations of host trustworthiness based on profile text are fairly consistent between raters (Ma et al., 2017a). Our experiments build on these established measurements of trustworthiness of Airbnb hosts to investigate whether the introduction of large language models to online self-presentation affects perceived trustworthiness.

Study 1: Transparent AI involvement

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

Methods

Study 1 is a mixed-factorial-design online experiment where participants rate the trustworthiness of prospective hosts in an Airbnb-type scenario. Our procedure followed prior work on the trustworthiness of Airbnb host profiles (Ma et al., 2017a). We asked participants to imagine they were reviewing potential hosts in a lodging marketplace. We showed them a set of 10 Airbnb host profiles in randomized order.

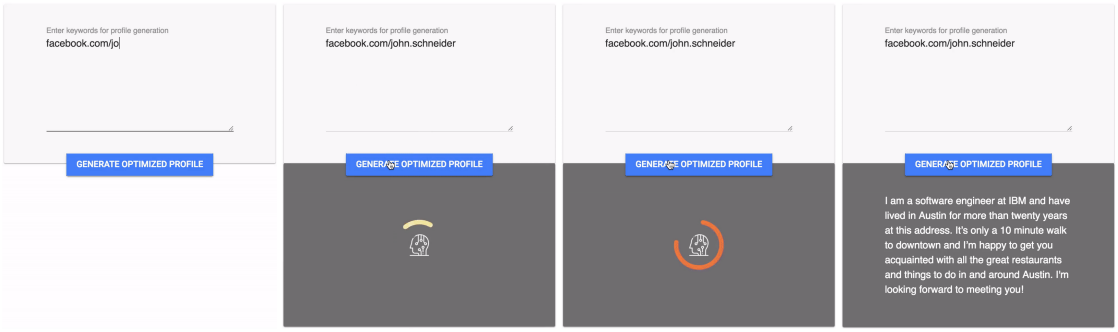

Figure 3.1: Screenshots of the "AI system" demo video participants in the treatment group watched before rating the profiles

The profiles were selected from a publicly available dataset[1] of Airbnb profiles (Ma et al., 2017a). We only considered a set of profiles of comparable length (37 to 58 words) based on Ma et al.'s result showing the correlation between profile length and trustworthiness ratings. From this set, we chose five profiles that had received very high trustworthiness rankings (top 5%) in the prior study, and five profiles that had received very low trustworthiness ratings (bottom 5%). We defined an independent variable called the "profile baseline" based on this split. The variable allows us to observe whether the effect of AI-MC is different for high- and low-trustworthiness profiles. The profiles are listed in the appendix.

Experimental manipulation

Participants were randomly assigned to the control or treatment group. While all participants were reviewing the same profiles, subjects in the treatment group were led to believe that the profiles they rated have been generated using an AI system, similar to the "Wizard of Oz" approach used in other studies of interpersonal communication (Dahlbäck et al., 1993; Lucas et al., 2014; Edlund et al., 2008).

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

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

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

The participants in the treatment group then watched a 10-second demo video of a mock-up AI system (see Figure 3.1). In the video, an animation depicts a system automatically generating text for an Airbnb profile from a Facebook profile URL provided by a user. We performed a manipulation check to verify that the participants in the treatment condition understood that a profile had been generated. To reinforce the manipulation, all profiles in the treatment group came with a label that reminded of the AI system.

Measured variables

We measured the perceived trustworthiness of the host as the outcome variable. We used perceived trustworthiness for several practical reasons: First, perceived trustworthiness is a variable that is conceivably affected by the introduction of large language models. Second, the variable has been shown to be a reliable measure in the context of Airbnb in previous studies (Ma et al., 2017a,b). Finally, we

had access to publicly available data on hosts and their perceived trustworthiness scores (Ma et al., 2017a).

Perceived trustworthiness is defined as an attribute of a target individual (Hardin, 2002; Kiyonari et al., 2006)– in our case, the host represented by the profile. We measured profile trustworthiness using a scale developed in (Ma et al., 2017a) which builds on earlier measurements of Mayer et al. (Mayer and Davis, 1999; Mayer et al., 1995). As the six items in the original scale were highly correlated (Ma et al., 2017a), to reduce respondent fatigue, we selected one item only from each dimension of trustworthiness (ability, benevolence, and integrity (Mayer et al., 1995), Likert-style, 0–100). The items we used were:

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

Following past studies (Ma et al., 2017a), we combined the three items into a trust index by calculating their mean (Cronbach's $\alpha = .86$; $M = 66.6$, $SD = 18.5$, reliable and consistent with prior work).

After the main rating task, we asked participants to complete a generalized trust scale we adapted from Yamagishi's trust scale (Yamagishi, 1986) and an AI attitude survey modeled after the well-established computer attitude scale (Nickell and Pinto, 1986). We combined the multiple-item scales into mean generalized trust (Cronbach's $\alpha = .88$; $M = 64.1$, $SD = 16.7$) and AI attitude scores (Cronbach's $\alpha = .72$; $M = 70.1$, $SD = 18.7$).

Table 3.1: Overview of measurements

Name	**Concept**
Trustworthiness	The perceived trustworthiness of a host based on his or her Airbnb profile (Ma et al., 2017a)
Generalized trust	A measure of how much the participant trusts other people in general (Yamagishi, 1986)
AI attitude	An index of the participant's positive and negative attitudes toward AI (Nickell and Pinto, 1986)
Trust baseline	Whether the profile was rated as trustworthy or untrustworthy in a prior study (Ma et al., 2017a)
AI score *(Study 2 and 3 only)*	A measure of how strongly the participant suspects a profile was written by AI

Participants also answered demographic questions (gender, age, education, and residential neighborhood type), as well as free-form questions explaining how they rated the profiles. We finally asked what they thought was the purpose of this study, and, in the treatment group, whether they had comments on the system. An overview of variables is shown in Table 3.1.

Participants

We recruited 527 participants via Amazon Mechanical Turk (AMT) (Buhrmester et al., 2011; Horton et al., 2011). Participation was limited to adults in the US who had completed at least 500 tasks with an approval rate of $\geq 98\%$. Participants' mean age was 38, with 48% identifying as female. Participating workers received a \$1.20 compensation based on an estimated work time of 6 minutes for a projected \$12 hourly wage. The workers provided informed consent before completing the study and were debriefed after completion with an option to withdraw. The debrief

is included in the appendix. The protocols were approved by the Institutional Review Board at Cornell University (protocol #1712007684).

Data validation

We performed several integrity and attentiveness tests for our participants. We excluded responses that had failed the linguistic attentiveness check borrowed from Munro et al. (Munro et al., 2010) as well as participants who did not select the right scenario ("I am traveling and the person in the profile offers to host me.") in a second attentiveness test. We excluded workers whose median rating time per profile was less than five seconds and workers with mostly uniform responses ($SD < 5.0$). Furthermore, we removed participants whose average trust rating fell outside the $mean \pm 2SD$ statistic of participant rating averages, leaving us with 389 subjects. Finally, we examined the free-form responses participants in the treatment group gave after viewing the system demo. Almost all responses demonstrated a clear understanding that the system generated a profile, leading us to conclude that the manipulation was effective.

Open Science Repository

The full experimental data, analysis code and experiment preregistration are available from `https://osf.io/qg3m2/`.

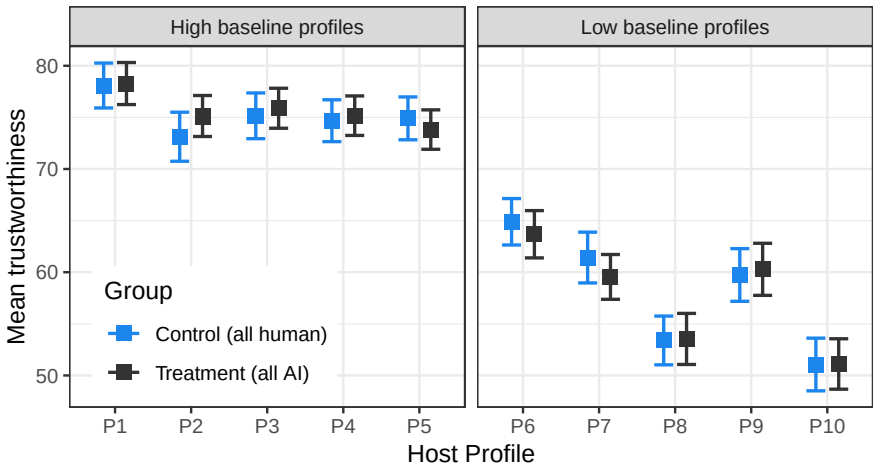

Figure 3.2: Study 1 host trustworthiness ratings by experimental condition, for profiles of high (left) and low (right) trustworthiness baseline

Results

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

Figure 3.2 illustrates our results: For each of the ten profiles (along the x-axis), we observed almost identical trustworthiness ratings (y-axis, along with confidence intervals) in the control (blue) and treatment (black) group. For example, profile 1 received average trust ratings of 78.3 by respondents who believed all profiles were written by humans, and 78.1 by respondents who thought an AI system generated all profiles. We conducted a 2x2 mixed factorial ANOVA to compare the main effects of perceived profile generation (human vs. AI), profile baseline (high vs. low), and their interaction effect on trust ratings. The ANOVA revealed significant differences between high baseline ($M = 75.4, SD = 14.6$) and low baseline ($M = 57.8, SD = 17.7$) profiles, $F(1, 387) = 2046, p < 0.001$. Since we selected

the profiles to be of either high or low baseline trustworthiness based on a prior study this result was expected and validates the reliability of the trust measurement in the current study. The ANOVA results did *not* indicate a main effect of perceived profile generation (human vs. AI); in other words, we did not find significant differences in trustworthiness ratings when we told participants that all profiles were written by the hosts ($M = 66.64, SD = 18.1$) and when we told them all profiles were AI-generated ($M = 66.64, SD = 18.8$). We briefly note that consistent with previous work, respondents' generalized trust levels were predictive of the trustworthiness ratings they assigned ($\beta = 0.30, p < .001$) and AI attitude ($\beta = 0.08, p < .001$) was predictive of the trust ratings as well.

Study 2: Uncertain AI involvement

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

Methods

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

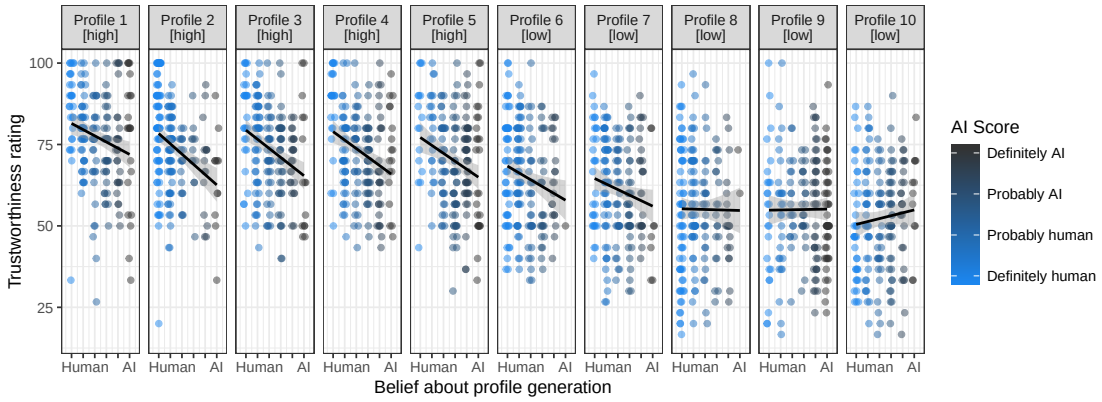

Figure 3.3: Study 2 host trustworthiness (y-axis) versus the participant's belief whether a profile was AI-generated (x-axis), for profiles of high (left) and low (right) trustworthiness baseline

Study 1. The full experimental data, analysis code and experiment preregistration are `https://osf.io/qg3m2/`publicly available on OSF.

Measured variables

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

Participants

We recruited 286 participants using the same procedure, parameters, and payments we used in Study 1. Participants who had participated in Study 1 were not eligible for Study 2. Participants' mean age was 37; 56% of them identified as female. We performed manipulation checks and attentiveness tasks to control for low-quality responses using the same procedure as in Study 1, excluding 89 out of 285 participants.

Results

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

Our observations are visualized in Figure 3.3 showing an overview of the raw trustworthiness scores participants gave (y-axis), grouped by host profiles 1–10 (x-axis), and further plotted over the AI score assigned. The AI score is also represented by color, from "more human" (blue, left) to "more AI" (grey, right). For example, the top-most, left-most point on the figure shows a participant that gave Profile 1 a perfect trustworthiness score (100), and a low AI score (1), corresponding to the belief that the profile was "definitely human-written". Just like Study 1, the five profiles on the left are high baseline trustworthiness profiles. The figure suggests that participants who believed that a profile was written by the host assigned higher trust ratings to the host than participants who suspected the

same profile was AI-generated. We visualize this trend by fitting a basic linear model to the data. The slope of the fitted line indicates that there may be an interaction: while for the high baseline trustworthiness profiles the slope is strongly and consistently negative, the slope of low baseline trustworthiness profiles is less pronounced.

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

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

Study 3: Validation and extensions

Study 3 investigates key questions raised by the previous studies. While Study 1 exposed no differences in trust, Study 2 provided initial evidence that perceived AI-generation affects trustworthiness in mixed-source environments. We designed Study 3 to clarify the conditions under which AI-generated self-presentations are distrusted.

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

We hypothesized that *"AI" profiles in the treatment conditions will be rated as less trustworthy than the same profiles are rated in the control condition.* In other

words, we predicted that when we tell participants they are rating a mixed set of profiles, regardless of whether the AI-like profiles are labeled as such, these "AI" profiles will be rated as less trustworthy compared to the ratings they receive in a control group that assumed all profiles to be written by humans. We preregistered our hypotheses and the full experimental design prior to the collection of data. The full experimental data, analysis code and preregistration are `https://osf.io/qg3m2/`publicly available on OSF.

Methods

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

Selection of profiles

In a preliminary study, we collected a set of profiles that were generally seen as either human-like or AI-like. To identify such profiles, we tested 100 random profiles from the same public dataset we used in the first two studies (Ma et al., 2017a) on AI score. To keep the studies comparable, we only selected profiles of 37-58 words length. While the profiles in Study 1 and 2 were selected to explore the difference between high or low trustworthiness profiles, in Study 3 we selected profiles of average trustworthiness ($mean \pm 0.5SD$ statistic) to minimize potential confounds due to differences in trustworthiness.

We recruited 80 workers on Amazon Mechanical Turk to each rate 16 of the 100 profiles, indicating whether they thought a profile was (1) "Definitely Human-

Table 3.2: Overview of Study 3 conditions

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

written" to (6) "Definitely AI-generated" on a 6-point Likert-style scale. After excluding uniform or incomplete answers, we analyzed the 945 AI scores received. We selected the 15 profiles that received the highest mean AI scores for the "AI " profile group and the 15 profiles receiving the lowest mean AI scores for the "human" profile group. The selected profiles are `https://osf.io/qg3m2/`available on OSF.

Study design and procedure

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

The three treatments, different versions of the "mixed-source" environment, were designed to test under which conditions "AI" profiles are distrusted. Participants in the *labeled* condition saw a 'generated profile' label above the "AI" profiles and a 'regular profile' label above the "human" profiles. Participants in the *unlabeled* condition did not see any label identifying the profiles. Subjects in the *primed* condition were not shown any labels, but we asked them, as we had done in Study 2, to rate the AI score of a profile before they rated the host's trustworthiness. An overview of conditions is shown in Table 3.2. We measured the same variables as in Study 1 and 2 and computed an index for the trust ratings (Cronbach's $\alpha = .87$; $M = 69.8$, $SD = 16.7$).

Participants

We recruited 323 participants that had not participated in Studies 1 or 2 for the experiment using the procedure, parameters, and payments of Study 1. Participants' mean age was 35.6; 44% of them identified as female. We performed manipulation checks and filtering tasks to exclude low-quality responses using the same procedure as in Study 1 and 2, excluding 115 participants. In addition to the checks of the prior studies, we performed a multiple-choice manipulation check after the rating task to make sure participants remembered the AI-generation. Only four participants failed the additional manipulation check, confirming that our former procedure was effective at removing low-quality responses and that the manipulation had been understood and remembered. We decided not to exclude these four participants due to their small number and the risks associated with post-randomization exclusions.

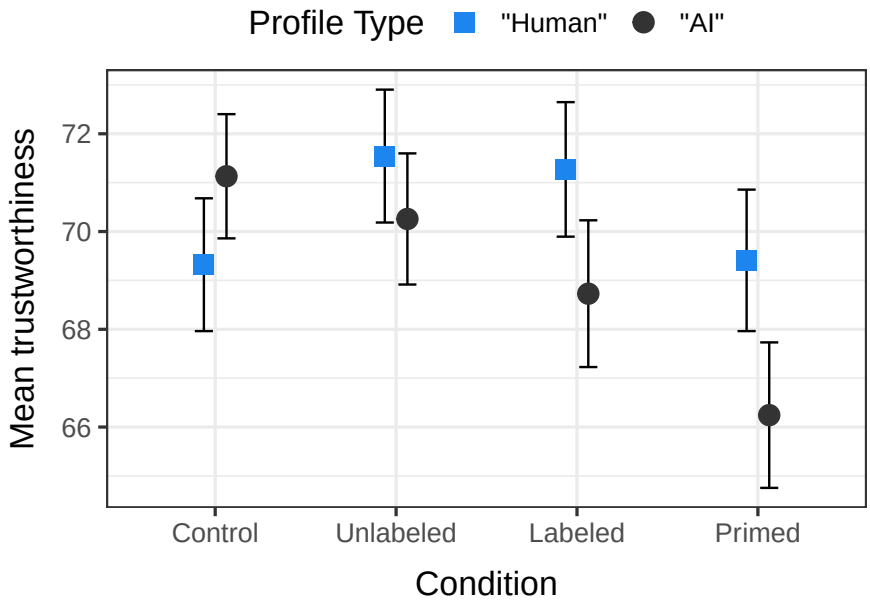

Figure 3.4: Study 3 trustworthiness ratings for hosts in the "AI" profile set versus hosts in the "human" profile set, across all experimental conditions

Results

Figure 3.4 shows the trust ratings that the different profile types received in the different treatment groups. Black circles show the mean trust ratings (and confidence intervals) of AI-like profiles, blue squares represent human-like profiles. The different experimental conditions are shown on the x-axis. We see that "AI" profiles received slightly higher ratings ($M = 71.13, SD = 10.8$) than "human" profiles ($M = 69.32, SD = 11.54$) in the control group, where participants believed all profiles were written by the hosts. However, in the treatment groups, where respondents believed some profiles were written by the host, while others were generated using an AI system, the ratings of AI-like profiles dropped considerably to their lowest observed mean of 66.24 in the *primed* condition.

Table 3.3: Regression table predicting trust ratings based on profile type and treatment

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

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

We calculated a multiple linear regression of our 4x2 mixed design to estimate how the different treatments and profile types affected the trust ratings. Model 1, shown in Table 3.3, predicts respondents' trust ratings based on treatment (*control, unlabeled, labeled or primed*), profile type ("AI" or "human"), and their interaction. The baseline is "human" profiles in the control group. None of the main effects were significant predictors; as expected, the treatment did not have a significant effect on the evaluation of "human" profiles. However, in the *labeled* and *primed* conditions hosts with AI-like profiles received significantly lower trust ratings. Model 2 includes a random effect per subject, in order to control for participants' differing

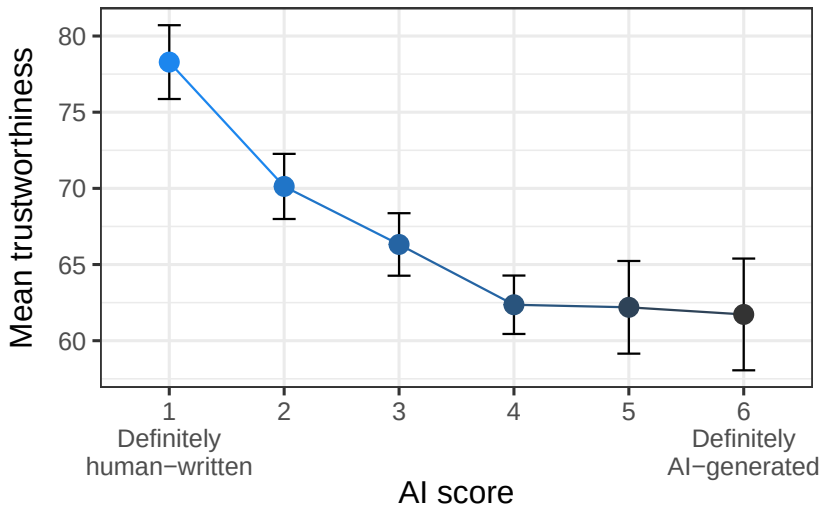

Figure 3.5: Trustworthiness ratings in the *primed* experimental condition by AI score assigned

trust baselines. In the multilevel model, AI-like profile characteristics predicted significantly lower trust ratings in all treatment groups. Following our preregistration, we also conducted a 4x2 mixed ANOVA on the influence of profile type, experimental treatment, and their interaction on the trust ratings. Similar to the regression, the ANOVA reveals a significant interaction of treatment and profile type ($F(1, 1966) = 4.534, p < 0.001$).

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

The *primed* condition furthermore allows us to expand on the analysis of

Study 2 (Figure 3.3), directly re-evaluating the relationship between AI score and trustworthiness. Figure 3.5 shows the means and confidence intervals of ratings in the *primed* condition plotted over the associated AI scores. For example, when participants rated profiles as "definitely human-written" they gave these profiles the highest trustworthiness rating ($M = 78.29, SD = 12.75$) – an average of 16.6 points higher than ratings they gave when they had evaluated a profile as "definitely AI-generated" ($M = 61.72, SD = 13.94$). Interestingly, we observe a floor effect: once a participant suspected a profile to be AI-generated (corresponding to an AI score of 4) the trustworthiness ratings dropped to the lowest level.

Discussion

Taken together, the results of the three studies show a robust effect of using large language models in self-presentation on the perceived trustworthiness of hosts and give an early indication of how online self-presentation may be affected by large language models.

Study 1 participants were willing to accept and trust self-presentation generated by large language models, possibly due to the uniform application of the technology. Recall that in Study 1, treatment group participants were told that *all profiles* were written by AI. This result aligns with other studies where researchers have found that people accept contributions of automated agents: In Wölker and Powell's study (Wölker and Powell, 2018), readers rated automated and human-written news as equally credible; similarly, Edwards et al. (Edwards et al., 2014) found no differences in source credibility between an otherwise identical human and bot account on Twitter.

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

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

This result is consistent with the Hyperpersonal Model of CMC, where receivers tend to make over-attributions based on minimal cues (Walther, 2011, 1996). In our mixed-source environments (Studies 2 and 3), participants scrutinized profiles that were deemed AI-like and made strong negative attributions of the host based

on minimal cues (Figure 3.5). The Hyperpersonal Model may explain why there were no such effects in Study 1: when participants encountered only one kind of source, there was no reason to use the source of the profile as a cue in their trustworthiness attributions. Further support for the Hyperpersonal Model and over-attribution is provided by the fact that highlighting differences by labeling profiles, or by making participants assign AI scores, made the *Replicant Effect* stronger and increased distrust.

The Elaboration Likelihood Model (ELM) (Petty and Cacioppo, 1986) further formalizes this explanation by differentiating two major routes to processing stimuli: the *Central Route* and the *Peripheral Route*. Under the *Peripheral Route*, information is processed mindlessly, relying on basic cues and rules of thumb. The results from Study 1, where participants encountered profiles from the same source only, could be due to peripheral processing. Because the source of the profile was not salient, respondents relied on the same social cues they used to judge human-written profiles for all the profiles. In contrast, the *Central Route* involves more careful and thoughtful consideration of the information presented. The mixed-source environment with both AI and human-generated profiles may have made the source of the profile more salient, leading participants to engage in more careful processing of the profiles. Under *Central Route* processing the source of the profile became part of the evaluation, leading to the *Replicant Effect* observed in Studies 2 and 3.

The current work clarified the conditions under which a *Replicant Effect* occurs and provided evidence that it depends on the salience of the source. The results raise the questions about why participants mistrusted profiles that they suspected were AI-generated. While the quantitative findings from Study 3 suggest that

higher uncertainty or priming did not cause the effect, an examination of the open-ended responses in the studies provides some insight: Participants rarely criticized accuracy of generated profiles, maybe due to a *promise of algorithmic objectivity* (Gillespie et al., 2014) of AI systems. However, they often noted that AI-generated profiles lacked emotion or authenticity. Multiple participants also expressed resentment toward the host for using an AI-generated profile ("They can be handy, but also a bit lazy. Which makes me question what else they'll be lazy about."). Further studies are needed to clarify *why* AI-generated profiles are seen as less trustworthy in mixed-source environments.

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

Limitations

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

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

Lastly, we note that while this initial examination of large language models' use in online self-presentation exposed a robust effect on trust, we did not directly test theories that can explain the mechanisms behind the findings. Such investigations will be needed to help advance the *conceptual* groundwork for understanding large language models' effects and are an exciting avenue for future work.

Chapter 4

Interacting with Opinionated Language Models Changes Users' Views

In this chapter, we extend our investigation of large language model's societal impact to the realm of politics, opinion dynamics, and democratic institutions. If large language models like GPT-3 produce certain views more often than others, they may influence people's opinions on an unknown scale. We explore whether language models that preferably generate a particular opinion change what their users write and believe. In an online experiment, participants (N=1,500) replied to a social media post discussing whether social media is good for society. Some participants received suggestions from GPT-3 configured to support a specific side of the debate. We asked a separate set of judges (N=500) to evaluate participants' written opinions and measured participants' post-task opinions in a survey. Our results show that interacting with an opinionated language model affects both written opinions and reported attitudes in a subsequent survey considerably. We conclude that the opinion preferences built into large language models need to be monitored and engineered more carefully.

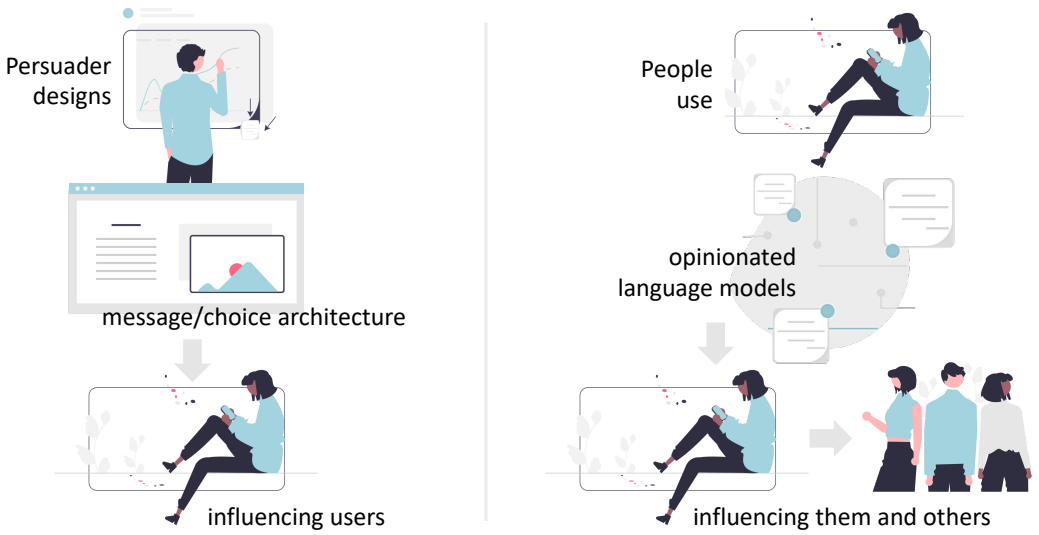

Figure 4.1: **Conventional technology-mediated persuasion (left) compared to *latent persuasion* by language models (right).** In conventional influence campaigns, a central persuader designs an influential message or choice architecture distributed to recipients. In *latent persuasion*, language models produce some opinions more often than others, influencing what their users write which is, in turn, read by others.

Introduction

Large generative language models like GPT-3 (Winata et al., 2021; Bommasani et al., 2021; Vaswani et al., 2017) may change how we form opinions and influence each other. In conventional forms of persuasion, a persuader crafts a compelling message and delivers it to recipients – either face-to-face or mediated through contemporary technology (Simons, 2011). More recently, user researchers and behavioral economists have shown that choice architectures in technical designs can influence people's behavior as well (Leonard, 2008; Fogg, 2002). With the emergence of generative language models that produce human-like language (Jakesch et al., 2022c; Buchanan et al., 2021), interactions with technology may change not

only behavior but also opinions: when language models produce some views more often than others, they may persuade their users. We call this new paradigm of influence *latent persuasion by language models*, as illustrated in Figure 4.1.

Latent persuasion by language models extends the critical insight of *nudge theory* (Leonard, 2008; Fogg, 2002) – that choice defaults change people's behavior – to the field of language and persuasion. Where *nudges* change behavior by making certain behaviors easier than others, language models may shift opinions by making it easier to say certain things than others. However, while choice architectures are intentional and visible, opinion preferences built into language models may be opaque to users, policymakers, and even system developers. Furthermore, while *nudges* are single interventions created by a designer to target a specific audience, a language model adapts its output to users' input and can be deployed across products and contexts.

The research on the risks of generative language models to date has focused on the conventional paradigm of persuasion, where language models may automate and optimize the production of content for advertising (Karinshak et al., 2022; Duerr and Gloor, 2021) or misinformation (Kreps et al., 2022a; Buchanan et al., 2021; Zellers et al., 2019). Studies have also shown that language models reproduce stereotypes and biases (Huang et al., 2019; Brown et al., 2020; Nozza et al., 2021) and support certain cultural values more than others (Johnson et al., 2022). While emerging research on co-writing with large language models suggests that models become increasingly active partners in people's writing (Lee et al., 2022; Yang et al., 2022; Yuan et al., 2022), little is known about how the opinions produced by language models may affect users' views. Initial work by Arnold et al. (2018) and Bhat et al. (2021, 2022) has suggested that a biased writing assistant may shift

movie or restaurant reviews, but whether models affect users' opinions remains an open and urgent question.

This study investigates whether large language models that generate certain opinions more often than others may change what their users write and think. In an online experiment (N=1,500), participants wrote a short statement discussing the benefits and risks of social media. Treatment group participants were offered a writing assistant that suggested text generated by a large language model. The model, GPT-3 (Winata et al., 2021), was configured to either generate text that argues that social media is good for society or text that argues the opposite. Following the writing task, we asked participants about their assessment of social media's societal impact in a survey. A separate sample of human judges (N=500) evaluated the opinions participants had expressed in their writing.

Our quantitative analysis tests whether the interactions with the opinionated language model changed participants' writing and survey opinions. We also explore how this opinion change may have occurred. To preview our results – we find that both the statements participants had written, *and* their attitude towards social media reported in the later survey were considerably influenced by the model's preferred opinion. We conclude by discussing how the possibility of *latent persuasion* by language models requires audits of the opinions built into models and broader conversations about what kind of opinions models should (not) produce.

Background

We draw on prior research on social influence and persuasion, interactions with writing assistants, and the societal risks of large language models.

Social influence and persuasion

Social influence is defined as a shift in an individual's thoughts, feelings, attitudes, or behaviors as a result of interaction with others (Rashotte, 2007). While social influence is integral to human collaboration, landmark studies have shown that it can also lead to unreasonable or unethical behavior. On a personal level, people may conform to majority views against their better judgement (Asch, 1951) and obey an authority figure even if it means harming others (Milgram, 1963). On a societal level, researchers have shown that social influence drives speculative markets (Shiller, 2015), affects voting patterns (Lazarsfeld et al., 1968) and contributes to the spread of unhealthy behaviors such as smoking and obesity (Christakis and Fowler, 2007, 2008).

Following the rise of social media, how online interactions affect people's opinions and decisions has been studied extensively. Research has shown that a variety of sources influences users' attitudes and behaviors, including friends, family, experts, and internet celebrities (Goel et al., 2012; Marwick and Boyd, 2011); the latter group was labeled *influencers* due to their influence on a large group of 'followers' (Bakshy et al., 2011). Research has also found that in online settings, users can be influenced by non-human entities such as brand pages, bots, and algorithms (Ferrara et al., 2016). Studies have evaluated the influence that technical artifacts such as personalized recommendations, chatbots, and choice architectures have on users' decision-making (Berkovsky et al., 2012; Leonard, 2008; Cosley et al., 2003; Gunaratne et al., 2018).

The influence that algorithmic entities have on people depends on how people perceive the algorithm, for example, whether they attribute trustworthiness to its

recommendations (Logg et al., 2019; Gunaratne et al., 2018). With the public's growing awareness of developments in artificial intelligence, people have also begun to regard *smart* algorithms as a source of authority (Kapania et al., 2022; Logg et al., 2019; Araujo et al., 2020). The influence of algorithms on individuals tends to increase as the environment becomes more uncertain and decisions become more difficult (Bogert et al., 2021). However, there is recent evidence that people may accept algorithmic advice even in simple cases when it is clearly wrong (Liel and Zalmanson, 2020). In the related field of automation, such over-reliance on machine output has been referred to as *automation bias* (Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010; Wickens et al., 2015).

Interaction with writing assistants

Historically, HCI research for text entry has predominantly focused on efficiency (Kristensson and Vertanen, 2014). Typical text entry systems attend to language context at the word (Vertanen et al., 2015; Bi et al., 2014) or sentence level (Arnold et al., 2016; Buschek et al., 2021). They suggest one to three subsequent words based on underlying likelihood distributions (Dunlop and Levine, 2012; Fowler et al., 2015; Gordon et al., 2016; Quinn and Zhai, 2016). More recent systems also provide multiple short reply suggestions (Kannan et al., 2016) or a single long phrase suggestion (Chen et al., 2019). More extensive suggestions are usually avoided because the time taken to read and select them might exceed the time required to enter that text manually. Several studies indicate that features such as auto-correction and word suggestions can negatively impact typing performance and user experience (Banovic et al., 2019; Dalvi et al., 2016; Buschek et al., 2018; Palin et al., 2019).

However, with the emergence of larger and more powerful language models (Winata et al., 2021; Bommasani et al., 2021; Vaswani et al., 2017), there has been a growing interest in design goals beyond efficiency. Studies have investigated interface design factors and interactions with writing assistants that directly or indirectly support inspiration (Lee et al., 2022; Singh et al., 2022; Yuan et al., 2022; Bhat et al., 2022), language proficiency (Buschek et al., 2021), story writing (Singh et al., 2022; Yuan et al., 2022), text revision (Cui et al., 2020; Zhang et al., 2019) or creative writing (Clark et al., 2018; Gero and Chilton, 2019). Here, language models are framed as *active writing partners* or *co-authors* (Lee et al., 2022; Yang et al., 2022; Yuan et al., 2022), rather than tools for prediction or correction. There is also evidence that a system that suggests phrases rather than words (Arnold et al., 2016) is more likely to be perceived as a collaborator and content contributor by users.

The more writing assistants become *active writing partners* rather than mere tools for text entry, the more the writing process and output may be affected by their "co-authorship". Bhat et al. (2022) discuss how writers evaluate the suggestions provided by the model and integrate them into different cognitive writing processes. Work by Singh et al. (2022) suggests that writers make active efforts or 'leaps' to integrate system-generated content into their writing. Buschek et al. (2021) conceptualized nine behavior patterns that indicate varying degrees of engagement with suggestions, from ignoring them to chaining multiple ones in a row. Writing with suggestions correlates with shorter and more predictable texts being written (Arnold et al., 2020), along with increased use of standard phrases when writing with a language model (Buschek et al., 2021; Bhat et al., 2022). Furthermore, the sentiment of the suggested text may influence the sentiment of the written text (Arnold et al., 2018; Hohenstein and Jung, 2020).

Societal risks of large language models

Technical advances have given rise to a generation of language models (Bommasani et al., 2021) that produces language so natural that humans can barely distinguish it from real human language (Jakesch et al., 2022c). Enabled by improvements in computer hardware and the transformer architecture (Vaswani et al., 2017), models like GPT-3 (Brown et al., 2020; Radford et al., 2019) have attracted attention for their potential to impact a range of beneficial real-world applications (Bommasani et al., 2021). However, more cautious voices have also warned about the ethical and social risks of harm from large language models (Weidinger et al., 2021, 2022), ranging from discrimination and exclusion (Huang et al., 2019; Brown et al., 2020; Nozza et al., 2021) to misinformation (Kreps et al., 2022a; Lin et al., 2021; Rae et al., 2021; Zellers et al., 2019) and environmental (Strubell et al., 2019) and socioeconomic harms (Bender et al., 2021).

Comparatively little research has considered widespread shifts in opinion, attitude, and culture that may result from a comprehensive deployment of generative language models. It is known that language models work and perform better for the languages and contexts they are trained in (primarily English or Mandarin Chinese) (Brown et al., 2020; Rae et al., 2021; Winata et al., 2021). Small-n audits have also suggested that the values embedded in models like GPT-3 were more aligned with reported dominant US values than those upheld in other cultures (Johnson et al., 2022). Work by Jakesch et al. (2022b) has highlighted that the values held by those developing AI systems differ from those of the broader populations interacting with the systems. The adjacent question of AI alignment – how to build AI systems that act in line with their operators' goals and values – has received comparatively more attention (Askell et al., 2021), but primarily

from a control and safety angle.

A related topic, the political repercussions of social media and recommender systems (Zhuravskaya et al., 2020), has received extensive research attention, however. After initial excitement about social media's democratic potential (Khondker, 2011), it became evident that technologies that affect public opinion will be subject to powerful political and commercial interests (Bradshaw and Howard, 2017). Rather than mere technical platforms, algorithms become constitutive features of public life (Gillespie, 2014) that may undermine the democratic institutions (Aral and Eckles, 2019). Even without being designed to change opinions, it has been found that algorithms may contribute to political polarization by reinforcing divisive opinions (Bruns, 2019; Cinelli et al., 2021; Bail et al., 2018).

Methods

To investigate whether interacting with opinionated language models shifts people's writing and affects people's views, we conducted an online experiment asking participants (N=1,500) to respond to a social media post in a simulated online discussion using a writing assistant. The language model powering this writing assistant was configured to generate text supporting one or the other side of the argument. We compared the essays and opinions of these participants to a control group that wrote their social media posts without any writing assistance.

Experiment design

To study interactions between model opinion and participants' opinion in a possibly realistic and relevant setting, we created the scenario of an opinionated discussion on social media platforms like Reddit. Such discussions have a large readership (Medvedev et al., 2017), pertain to political controversies, and are plausible application settings for writing assistants and language models. To identify a discussion topic, we searched ProCon.org[1], an online resource for research on controversial issues. We selected "Is Social Media Good for Society?" as a discussion topic. We chose this topic because it is an easily accessible discussion topic that is politically relevant but not considered so controversial that entrenched views may limit constructive debate.

To run the experiment, we created a custom experimental platform combining a mock-up of a social media discussion page, a rich-text editor, and a writing assistant. The assistant was powered by a language generation server and included comprehensive logging tools. To provide a realistic-looking social media mock-up, we copied the design of a Reddit discussion page and drafted a question based on the ProCon.org discussion topic. Figure 4.2 shows a screenshot of the experiment. We asked participants to write at least five sentences expressing their take on social media's societal impact. We randomly assigned participants to three different groups:

1. *Control group:* participants wrote their answers without a writing assistant
2. *Techno-optimist language model treatment:* participants were shown suggestions from a language model configured to argue that social media is good

[1] `https://www.procon.org/`

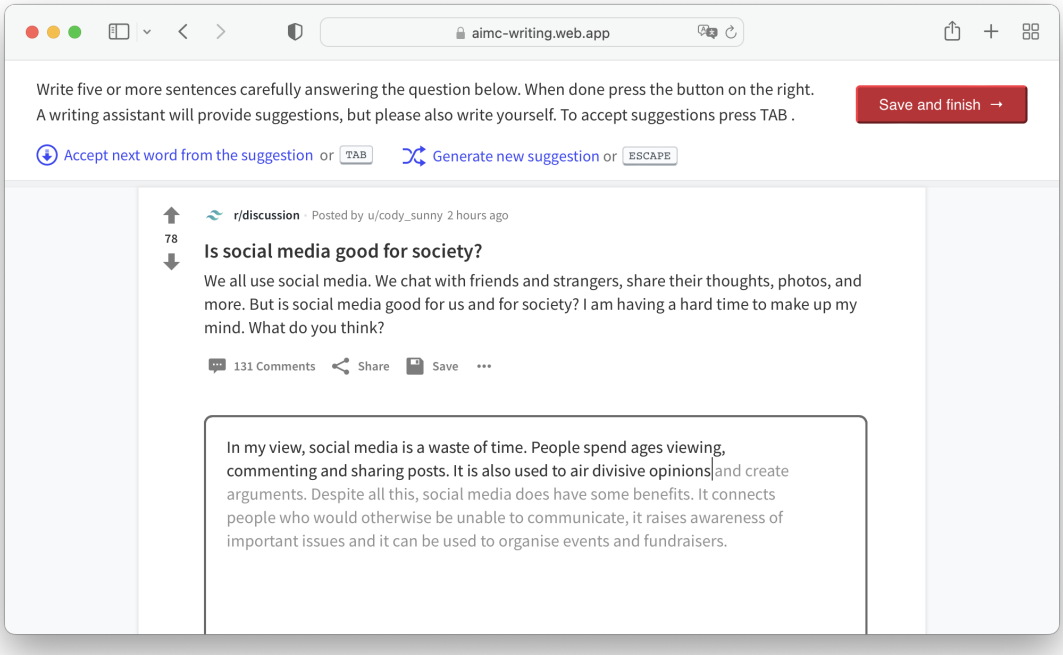

Figure 4.2: **Screenshot of the writing task.** Participants read a Reddit-style discussion post to which they were asked to reply. During their writing process, a writing assistant displayed writing suggestions (shown in grey). The participant in the screenshot wrote an argument critical of social media, but the model was configured to argue that social media is *good* for society.

for society.

3. *Techno-pessimist language model treatment:* participants received suggestions from a language model configured to argue that social media is bad for society.

Building the writing assistant

Similar to Google's *Smart Compose* (Chen et al., 2019), and Microsoft's predictive text in Outlook, the writing assistant in the treatment groups suggested possible

continuations (sometimes called "completions") to text that participants had entered. We integrated the suggestions into a customized version of the rich-text editor Quill.js[2]. The client sent a generation request to the server whenever a participant paused their writing for a certain amount of time (750ms). Including round-trip and generation time, a suggestion appeared on participants' screens about 1.5 seconds after they paused their writing.

When the editor client received a text suggestion from the server, it revealed the suggestion letter by letter with random delays calibrated to resemble a co-writing process (cf. (Lehmann et al., 2022)). Once the end of a suggested sentence was reached, the editor would pause and request from the server an extended generation until at least two sentences had been suggested. Participants could accept each suggested word by pressing the tab key or clicking an accept button on the interface. In addition, they could reset the generation, requesting a new suggestion by pressing a button or key.

We hosted the required cloud functions, files, and interaction logs on Google's Firebase platform.

Configuring an opinionated language model

In this study, we experimented with language models that *strongly* favored one view over another. We chose a strong manipulation as we wanted to explore the *potential* of language models to affect users' opinions and understand whether they could be used or abused to change people's views (Bagdasaryan and Shmatikov, 2021).

[2] `https://quilljs.com/`

We used GPT-3 (Brown et al., 2020) with carefully designed prompts to generate text suggestions for the experiment in real-time. Specifically, we accessed OpenAI's most potent 175B parameter model ("text-davinci-002"), setting the randomness parameter (sampling temperature) to 0.85. We designed prompts (Brown et al., 2020) to configure the model to produce suggestions that support one side or the other. To steer the model behavior, we inserted *"Is social media good for society? Explain why social media is good/bad for society:"* before participants' written texts when generating continuation suggestions. This approach produced consistent opinions to start with. Yet, when participants argued strongly against the opinion we had prompted the model, they would make the model follow their opinion. To make the model opinion persist even in the face of disagreements with the participant, we inserted an instruction affirming the model opinion (*"One sentence continuing the essay explaining why social media is good/bad:"*) after the last sentence of participants' entered text when generating new suggestions. These prompts were not visible to participants in their editor UI; they were inserted before generation and removed before sending the output to the client. Validation of the model opinion manipulation is provided in the results section. We also experimented with creating an opinionated version of GPT-3 using fine-tuning (Howard and Ruder, 2018), but the resulting model did not consistently reproduce the intended opinion.

Outcome measures and covariates

We collected different types of outcome measures to investigate interactions between participants' opinions and the model opinion:

Opinion expressed in the post: To evaluate expressed opinion, we split par-

ticipants' written texts into sentences and asked crowd workers to evaluate the opinion expressed in each sentence. Each crowd worker assessed 25 sentences, indicating whether each argued that social media is good for society, bad, or both good and bad. A fourth label was offered for sentences that argued neither or were unrelated. For example, *"Social media also promotes cyber bullying which has led to an increase in suicides" (P#421)* was labeled as arguing that social media is bad for society, while *"Social media also helps to create a sense of community" (P#1169)* was labeled as *social media is good for society.* We collected one to two labels for each sentence participants wrote and collected labels for a sample of the writing assistant's suggestions. On sentences that we collected multiple labels for, the labels provided by different raters agreed 84.1% of the time (Cohen's $\kappa = 0.76$).

Real-time writing interaction data: We gathered comprehensive interaction logs at the key-stroke level of how participants interacted with the model's suggestions. We recorded which text the participant had written, what text the model had suggested, and what suggestions participants had accepted from the writing assistant. We measured how long they paused to consider suggestions and how many suggestions they accepted.

Opinion survey (post-task): After finishing the writing task, participants completed an opinion survey. The central question, "Overall, would you say social media is good for society?" was designed to assess changes in participants' attitude. This question was not shown immediately after the writing task to reduce demand effects. The following secondary questions were asked to understand participants' opinions in more detail: "How does social media affect your relationships with friends and family?", "Does social media usage lead to mental health problems or addiction?", "Does social media contribute to the spread of false information and

hate?", "Do you support or oppose government regulation of social media companies?" The questions were partially adapted from Morning Consults' National Tracking Poll (Consult, 2016); answers were given on typical 3- to 5-point Likert scales.

User experience survey (post-task): Participants in the treatment groups completed a survey about their experience with the writing assistant following the opinion surveys. They were asked, "How useful was the writing assistant to you?", whether "The writing assistant understood what you wanted to say" and whether "The writing assistant was knowledgeable and had expertise." To explore participants' awareness of the writing assistants' opinions and their own opinion changes, we asked them whether "The writing assistant's suggestions were reasonable and balanced" and whether "The writing assistant inspired or changed my thinking and argument." Answers were given on a 5-point Likert scale from "strongly agree" to "strongly disagree." An open-ended question asked participants what they found most useful or frustrating about the writing assistant.

Covariates: We asked participants to self-report their age, gender, political leaning, and their highest level of education at the end of the study. We also constructed a "model alignment" covariate estimating whether the opinion the model supported was aligned with the participant's opinion. We did not ask participants to report their overall judgment before the writing task to avoid commitment effects. Instead, we asked them at the end of the study whether they believed social media was good for society before participating in the discussion. While imperfect, this provides a proxy for participants' pre-task opinions. It is biased by the treatment effect observed on this covariate, which amounts to 14% of its standard deviation.

Participant recruitment

We recruited 1,506 participants (post-exclusion) for the writing task, corresponding to 507, 508, and 491 individuals in the control, techno-optimist, and techno-pessimist treatment groups, respectively. The sample size was calculated based on effect sizes observed in the pilot studies' post-task question, "Overall, would you say social media is good for society?" at a power of 80%. The sample was recruited through Prolific (Palan and Schitter, 2018). The sample included US-based participants at least 18 years old (M= 37.7, SD= 14.2); 48.5% self-identified as female, and 48.6% identified as male. Six out of ten indicated liberal leanings; 57.1% had received at least a Bachelor's degree. Participants who failed the pre-task attention check (8%) were excluded, and six percent of participants admitted to the task did not finish it. We paid participants $1.50 for an average task time of 5.9 minutes based on an hourly compensation rate of $15. For the labeling task, we recruited a similar sample of 500 participants through Prolific. The experimental protocols were approved by the [anonymized] Institutional Review Board.

Results

We first analyze the opinions participants expressed in their social media posts. We then examine whether participants may have accepted the models' suggestions out of mere convenience and whether the model influenced participants' opinions in a later survey. Finally, we present data on participants' perceptions of the model's opinion and influence.

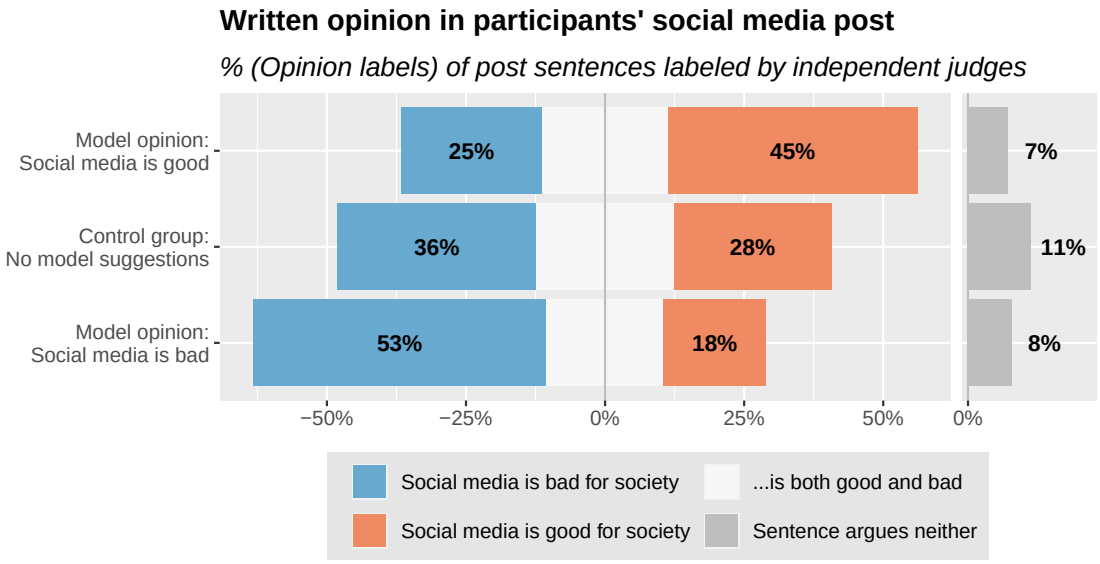

Figure 4.3: **Participants assisted by a model supportive of social media were more likely to argue that social media is good for society in their posts (and vice versa).** N_s=9,223 sentences written by N_p=1,500 participants evaluated by N_j=500 judges. The y-axis indicates whether participants wrote their social media posts with assistance from an opinionated language model that was supportive (top) or critical of social media (bottom). The x-axis shows how often participants argued that social media is bad for society (blue), good for society (orange), or both good and bad (white) in their writing. Sentences that argued neither or were unrelated to the topic are shown in grey.

Did the language model affect participants' writing?

Figure 4.3 shows how often participants in each of the treatment conditions (y-axis) argued that social media is good or bad for society (x-axis) in their writing. The social media posts written by participants in the control group (middle row) were slightly critical of social media: They argued that social media is bad for society in 38% and that social media is good in 28% of their sentences. In about 28% of their sentences, control group participants argued that social media is both

good and bad, and 11% of their sentences argued neither or were unrelated.

Participants who received suggestions from a language model supportive of social media (top row of Figure 4.3) were 2.04 more likely than control group participants ($p<0.0001$, 95% CI [1.83, 2.30]) to argue that social media is good. In contrast, participants who received suggestions from a language model that criticized social media (bottom row) were 2.0 times more likely ($p<0.0001$, 95% CI [1.79, 2.24] to argue that social media is bad than control group participants. We conclude that using an opinionated language model in their writing changed participants' writing such that the text they wrote was more likely to support the model's preferred view.

Did participants accept suggestions out of mere convenience?

Participants may have accepted the models' suggestions out of convenience, even though the suggestions did not match what they would have wanted to say. In particular, paid participants in online studies may be motivated to accept suggestions to swiftly complete the task.

Our data shows that, across conditions and treatments, most participants did not blindly accept the model's suggestions but interacted with the model to co-write their social media posts. Figure 4.4 shows that, on average, participants wrote 63% of their sentences themselves without accepting suggestions from the model. About 25% of participants' sentences were written by both the participant and the model, which typically meant that the participant wrote some words and accepted the model's remaining sentence suggestion. Only 11.5% of sentences were fully accepted from the model. Participants whose personal views were likely

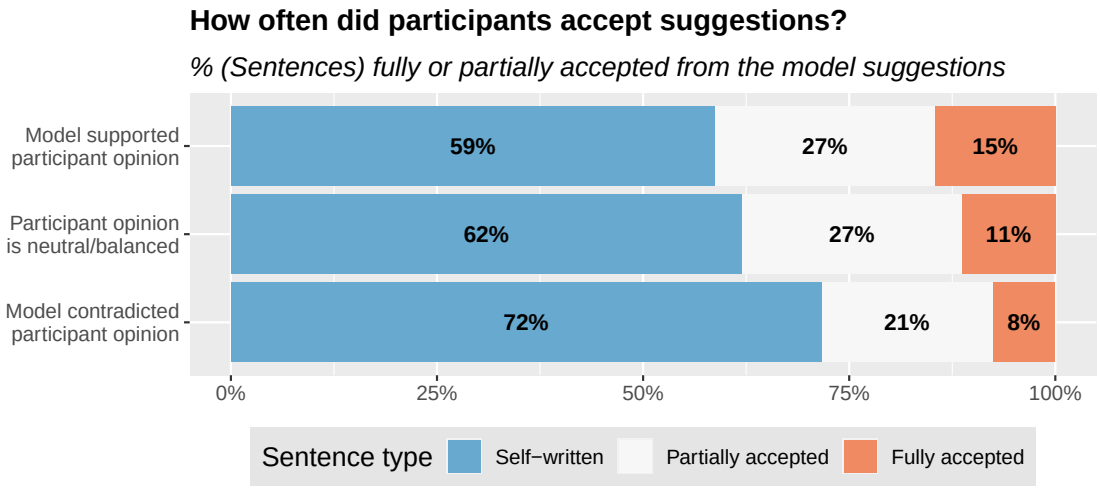

Figure 4.4: **Participants were more likely to accept suggestions if the model's opinion aligned with their own views** N_s=6,14 sentences by N_p=1,000 participants. The x-axis shows how many of the sentences participants had written themselves (blue), together with the model (white), or fully accepted from the model's suggestions (orange). The y-axis splits results based on whether the model suggestions were in line with participants' likely pre-task opinion.

aligned with the model were more likely to accept suggestions, while participants with opposing views accepted fewer suggestions. About one in four participants did not accept any model suggestion, and one in ten participants had more than 75% of their post written by the model.

Did conveniently accepted suggestions increase the observed differences in written opinion? However, we do find that the writing of participants who spent little time to write their post was more influenced by the model's opinion. We use the time participants took to write their posts to estimate to what extent they may have accepted suggestions without due consideration. For a concise statistical analysis, we treat the ordinal opinion scales as an interval scale. Since the opinion scale has comparable-size intervals and a zero point, continuous analysis is meaningful

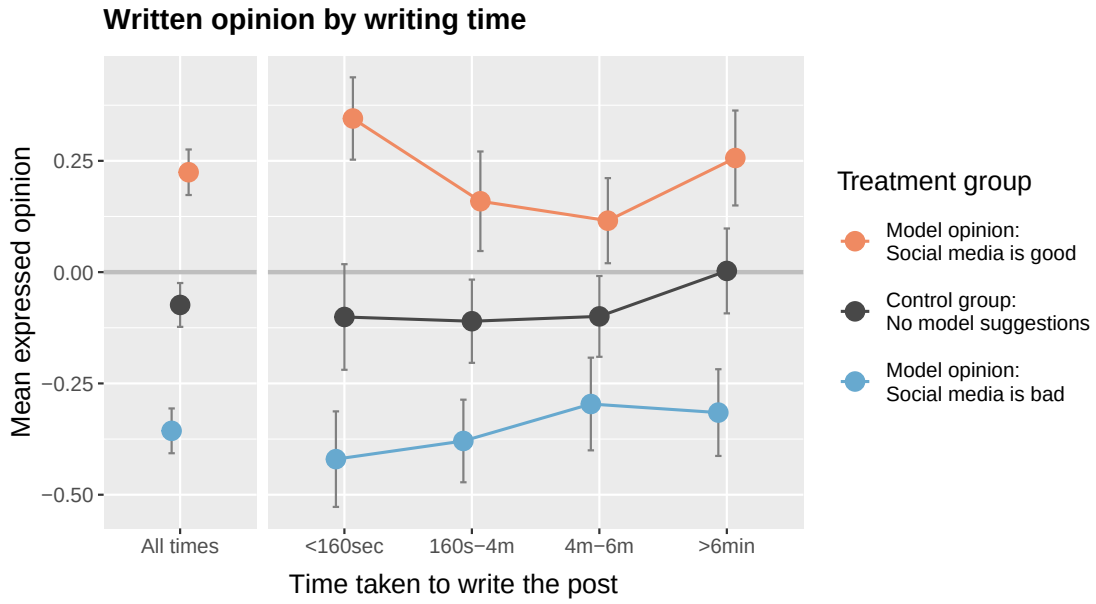

Figure 4.5: **The opinion change in participants' writing was larger when they finished the task quickly.** N=1,500. The y-axis shows the mean opinion expressed in participants' social media posts based on aggregated sentence labels ranging from -1 for "social media is bad for society" to 1 for "social media is good for society". The x-axis indicates how much time participants took to write their posts. For reference, the left panel shows expressed opinions aggregated across writing times.

and justifiable (Knapp, 1990). We treat "social media is bad for society" as -1 and "social media is good for society" as 1. Sentences that argue both or neither are treated as zeros.

Our analysis shows that accepting suggestions out of convenience has contributed to the differences in the written opinion but was not the primary factor causing the difference. Figure 4.5 shows the mean opinion expressed in participants' social media posts depending on treatment group and writing time. The left panel shows participants' expressed opinions across times for reference, with a mean opinion difference of about 0.29 ($p<0.001$, 95% CI [0.25, 0.33], SD=0.58)

between each treatment group and the control group (corresponding to a large effect size of d=0.5). Participants who took little time to write them (less than 160 seconds, left-most data in right panel) were more affected by the opinion of the language model (0.38, $p<0.001$, 95% CI [0.31, 0.45]). However, even for participants who took four to six minutes to write their posts, we observed significant differences in opinions across treatment groups (0.20, $p<0.001$, 95% CI [0.13, 0.27], corresponding to a treatment effect of d=0.34).

Did the language model affect participants' attitudes?

The opinion differences in participants' writing may be due to actual opinion change caused by interacting with the opinionated model. Here, we estimate how interactions with the language model led to a change in participants' attitudes based on a post-task survey asking participants whether they thought social media was good for society. An overview of participants' answers is shown in Figure 4.6.

The figure shows the frequency of different survey answers (x-axis) for the participants in each condition (y-axis). Participants who did not interact with the opinionated models (middle row in Figure 4.6) were balanced in their evaluations of social media: 33% answered that social media is not good for society (middle, blue); 35% said social media is good for society. In comparison, 45% of participants who interacted with a language model supportive of social media (top row) answered that social media is good for society. Converting participants' answers to an interval scale, this change in opinion corresponds to an effect size of d=0.22 ($p<0.001$). Similarly, participants that had interacted with the language model critical of social media (bottom row) were more likely to say that social media was bad for society afterward (d=0.19, $p<0.005$).

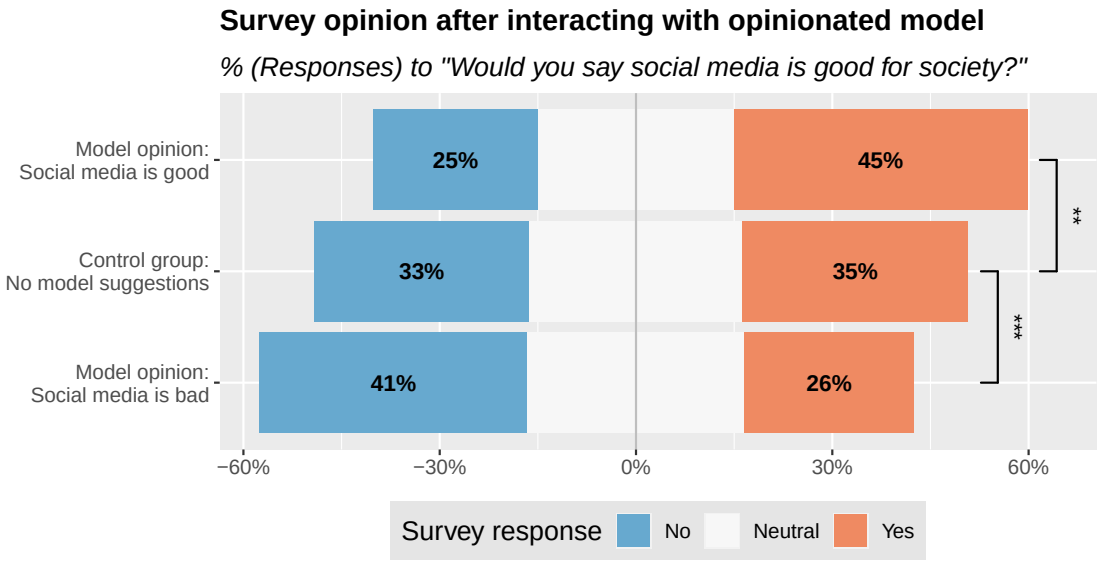

Figure 4.6: **Participants interacting with a model supportive of social media were more likely to say that social media is good for society in a later survey (and vice versa).** N_r=1,500 survey responses by N_r=1,500 participants. The y-axis indicates whether participants received suggestions from a model supportive or critical of social media during the writing task. The x-axis shows how often they said that social media was good for society (orange) or not (blue) in a subsequent attitude survey. Undecided participants are shown in white. Brackets indicate statistically significant differences in mean opinion at the **p<0.005 and ***p<0.001 level.

Did the opinionated model gradually convince the participant? While we cannot ascertain the exact mechanism of persuasion, our results provide further insight into how this process might have occurred.

Figure 4.7 shows how participants' written opinions changed throughout their writing process. In the control group (shown in black), participants started their posts with two positive statements, followed by two more critical statements and an overall critical conclusion. Participants interacting with a model that evaluated social media positively (orange) consistently evaluated social media more favorably

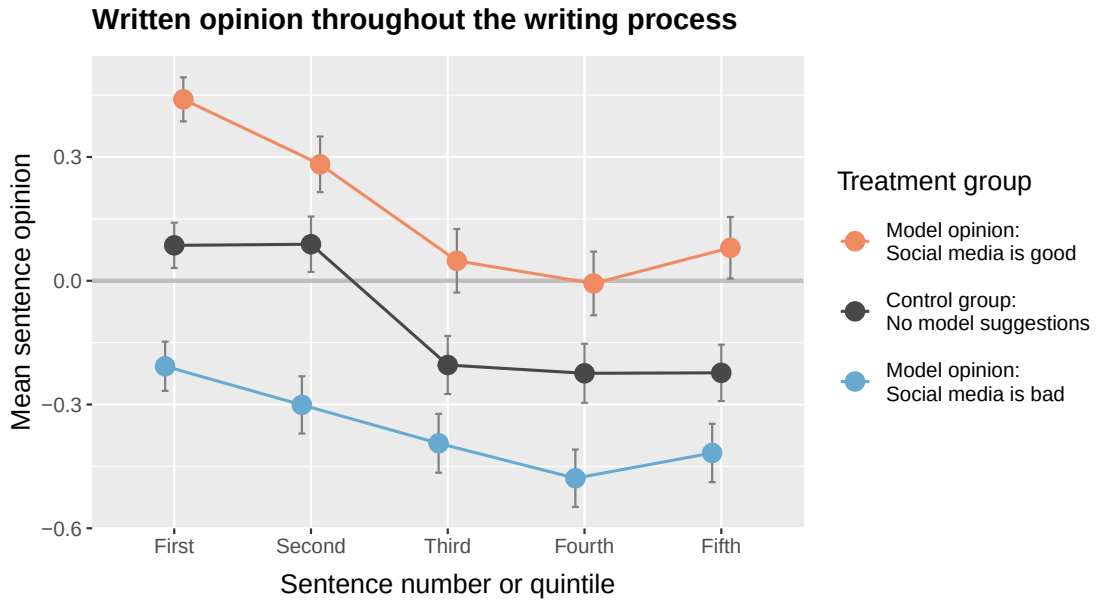

Figure 4.7: **Participants' writing was affected by the model equally throughout the writing process.** N_s=9,223 sentences by N_p=1,500 participants. The y-axis shows the mean opinion expressed in participants' sentences. The x-axis indicates whether the sentence was positioned earlier or later in participants' social media posts. Since most participants wrote five sentences as requested, the quintiles roughly correspond to sentence numbers.

throughout their entire statement. Participants interacting with a model critical of social media (blue) also wrote sentences that were more critical of social media, starting with their first sentence. Similar to the control group, they were more positive at the beginning and more critical towards the end of their post, showing that the writing assistant augmented rather than replaced their narrative.

Were participants aware of the model's opinion and influence?

After the writing task, we asked treatment group participants about their experience with the writing assistant. We use their answers to estimate to what extent

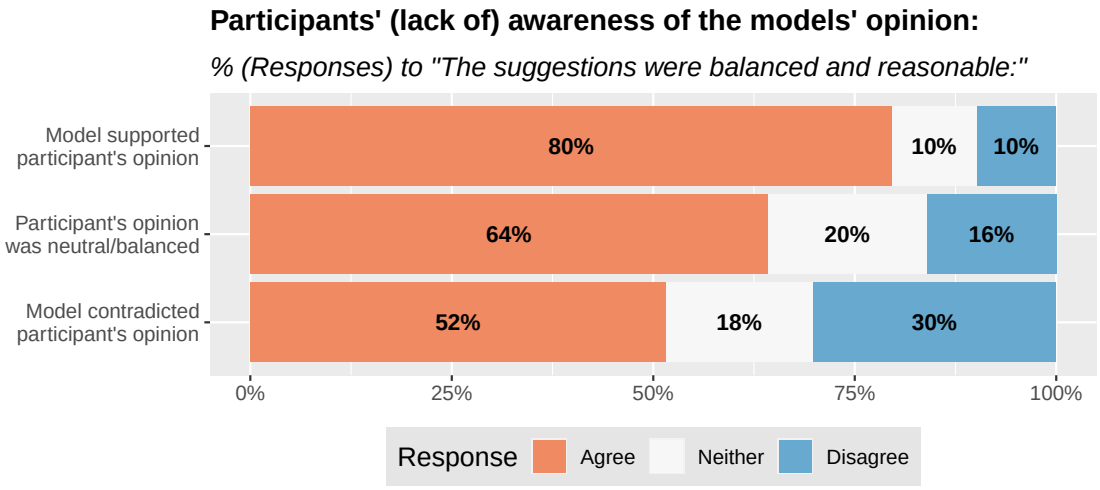

Figure 4.8: **Participants were often unaware of the model's opinion.** N_p=1,000 treatment group participants. The x-axis indicates whether participants found the model's suggestions balanced and reasonable. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

they were aware of the model's opinion and influence.

While the language model was configured to support one specific side of the debate, the majority of participants said that the model's suggestions were balanced and reasonable. Figure 4.8 shows that, in the group of participants whose opinion was supported by the model, only 10% noticed that its suggestions were imbalanced (top row in blue). When the model contradicted participants' opinions, they were more likely (30%) to notice its skew, but still, more than half agreed that the model's suggestions were balanced and reasonable (bottom row in orange).

Figure 4.9 shows that the majority of participants were not aware of the model's influence on their writing. Participants using a model aligned with their view – and accepting suggestions more frequently – were slightly more aware of the model's influence (34%, top row in orange). In comparison, only about 20% of

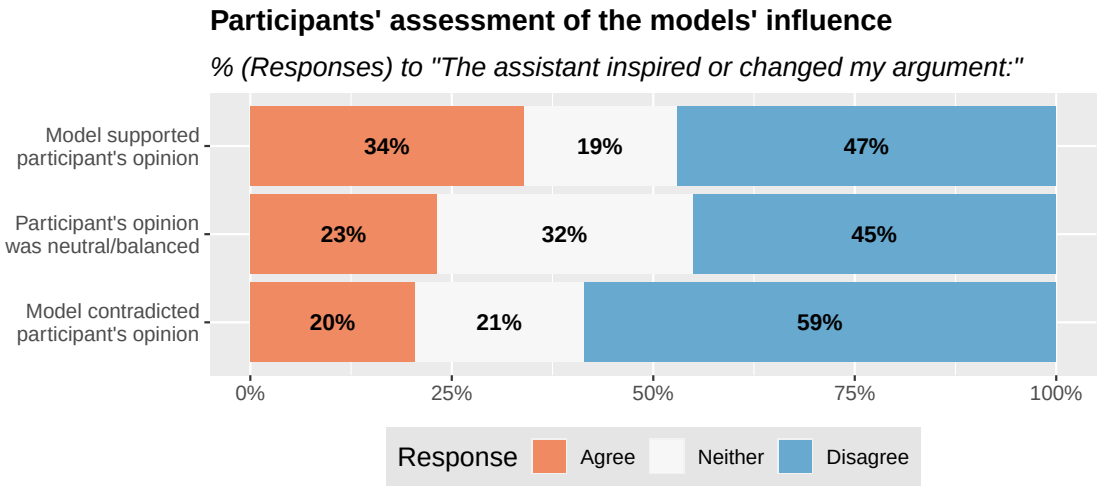

Figure 4.9: **Participants interacting with a model that supported their opinion were more likely to indicate that the model changed their argument.** N_p=1,000 treatment group participants. The x-axis indicates whether participants thought that the model changed their argument. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

the participants who did not share the model's opinion believed that the model influenced them. Overall, we conclude that participants were often unaware of the model's opinion and influence.

Did participants perceive the writing assistant as useful?

We also observed that participants who used a model sharing their opinions found the model more useful. Participants' evaluation of the model's usefulness is shown in Figure 4.10. While 67% of participants assigned to a model that likely shared their opinion said the assistant was useful or very useful (top right), only 39% of participants using a model contradicting their personal view found the model at

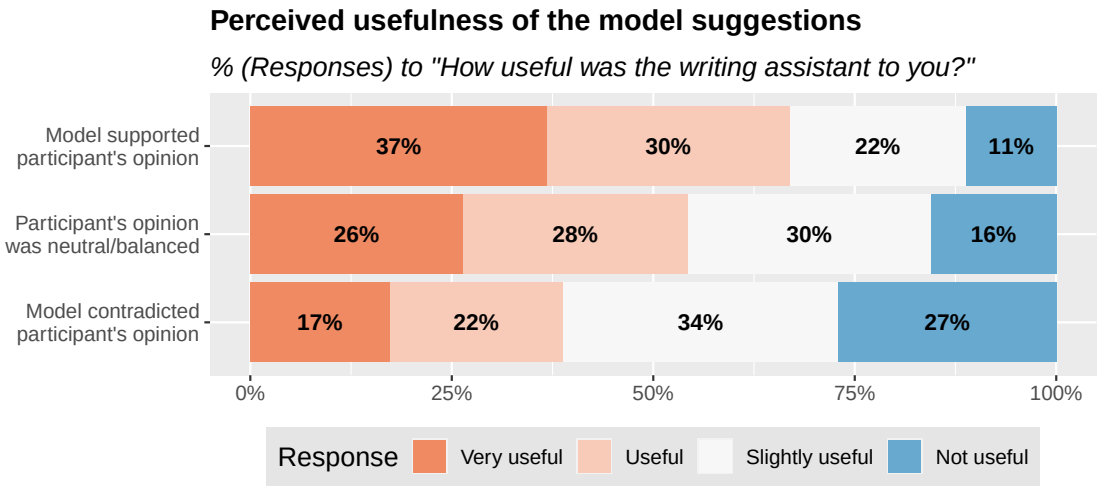

Figure 4.10: **Participants interacting with a model that supported their opinion found the assistant more useful than those writing with a model contradicting their view.** N_p=1,000 treatment group participants. The x-axis indicates how useful participants found the model's suggestions for their writing. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

least useful (bottom right).

The vast majority of participants thought the language model had expertise and was knowledgeable – even if it contradicted their personal views. As shown in Figure 4.11, 84% of participants said that the assistant was knowledgeable and had expertise when the language model supported their opinion. When the model contradicted their opinion, only 15% of participants said that it was not knowledgeable or lacked expertise.

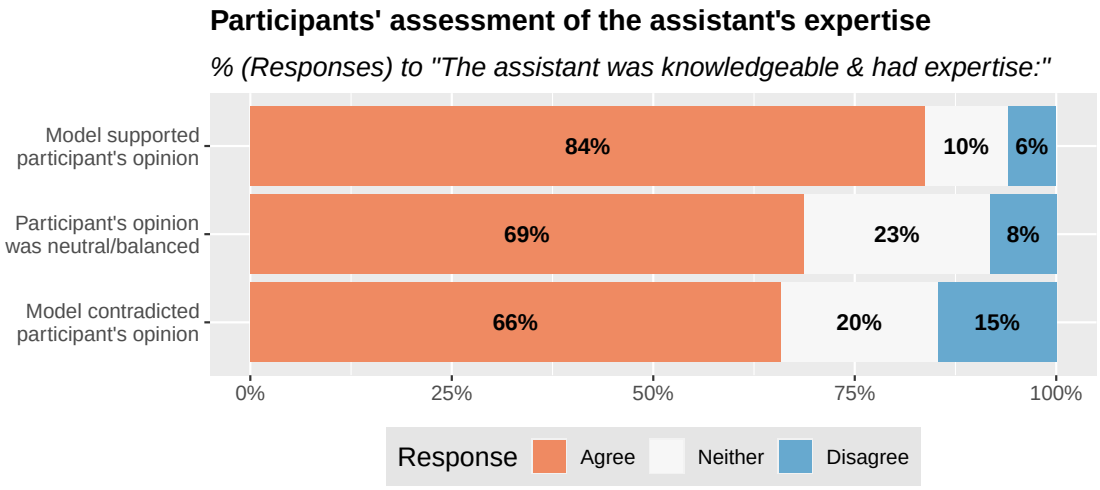

Figure 4.11: **Participants agreed that the language model was knowledgeable – even if it did not share their opinion.** N_p=1,000 treatment group participants. The x-axis indicates whether participants believed the language model had expertise and was knowledgeable. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

Robustness and validation

We finally validate that the experimental manipulation worked as intended and address potential concerns about experimenter demand effects.

Did manipulating the models' opinion work as intended? To validate that the prompting technique led to model output opinionated as intended, we sampled a subset of all suggestions shown to participants and asked a separate set of raters to indicate the opinion expressed in each. We found that of the full sentences suggested by the model, 86% were labeled as supporting the intended view, and 8% were labeled as balanced. For partially suggested sentences, that is, suggestions where the participants had already begun a sentence and the model completed it,

the ratio of suggestions that were opinionated as intended dropped to 62%, with another 19% arguing that social media is both good and bad. Overall, this indicates that the prompting technique guided the model to generate the target opinion with a high likelihood.

Could participants have accepted the model suggestion and changed their opinion to satisfy the experimenters? As in all subject-based research, there is a chance that participants change their behavior to fit their interpretation of the study's purpose. However, we have reason to believe that demand effects do not threaten the validity of our results. When participants were asked what they perceived as the purpose of the study, most thought we were studying what people think about social media or how they use writing assistants. Only about 14% mentioned that we might be studying the assistants' influence on people's opinions. Further, based on our post-task survey, most participants were not aware of the model's opinion and believed that the model did not change their argument. These results suggest that participants did not change their views as they felt the research team expected them to.

Discussion

Our results show that language models that produce some views more often than others can change what their users write. Participants assisted by an opinionated language model were significantly more likely to support the model's opinion in their social media posts than control group participants who did not interact with a language model. Even participants who took five minutes to write their post – ample time to write the five required sentences – were significantly affected by

the model's preferred view, showing that the model's effect on participants writing cannot be explained by people swiftly accepting suggestions. Most importantly, the interactions with the opinionated model also caused an opinion change in a later attitude survey, suggesting that the change in written opinion may be associated with a shift in actual attitudes.

Our results allow us to ascertain *that* interacting with opinionated language models changes the opinions in users' writing and in a subsequent attitude survey. We cannot ascertain *how* exactly the language model changed users' views. However, our secondary results do suggest that some vectors of influence are more likely than others.

First, the language model shifted participants' written opinions consistently throughout the writing process (as opposed to a gradual or incremental change in opinion). This consistent shift is contrasted by research showing that among human co-writers, opinions converge in a gradual process where co-writers introduce their positions and restructure their text to develop a shared position (Kimmerle et al., 2012). Had the language models changed participants' views through convincing arguments, we would expect to observe a larger shift in opinion towards the end of writing than at the beginning. These observations suggest that the model did not change participants' views through *informational influence* (Myers, 2008), that is, due to new information or convincing arguments.

Second, the language model may have shifted participants' views through *normative influence* (Myers, 2008), where opinion change happens out of reciprocity or obedience. Participants in our experiment attributed a high degree of expertise to the assistant (see Figure 4.11). The wider literature similarly suggests that people may regard AI systems as authoritative sources (Kapania et al., 2022; Logg et al.,

2019; Araujo et al., 2020). Further, our participants were often unaware of the language model's skewed opinion and influence. A lack of awareness of the models' influence similarly supports the idea that the model's influence was through the subconscious peripheral route (Petty and Cacioppo, 1986) and intuitive "System 1" processing (Kahneman, 2011).

Third, the interactions with the language models may have changed participants' views by changing their behavior through *latent persuasion*. The suggestions provided by the language model repeatedly prompted participants to consider the model's opinion and decide whether to accept them into their writing. Similar to *nudges*, the suggestions changed participants' behavior and changed what they spent their attention on. According to self-perception theory (Bem, 1972), such changes in behavior may lead to changes in opinion: when people do not have strongly formed attitudes, they may infer their opinion from their own behavior. The findings that participants who accepted the models' suggestions more frequently were more influenced by the model's view corroborates that some of the opinion change has been through behavioral routes.

Implications

Our results caution that interactions with opinionated language models may change users' opinions systematically, even if unintended. While we used a "strongly opinionated" model in the experiment, our results likely underestimate the opinion changes a widespread deployment of weakly opinionated models could cause: In the experiment, participants only interacted with the model once, while people could interact with a widely deployed model regularly over an extended period of time. Further, in real-world settings, people will not interact with models individ-

ually, but millions will interact with the same model, and what they write with the model will be read by others, leading to further reinforcement of the model's opinion. Finally, language models that insert their preferred views into people's writing increase the prevalence of their opinion in future training data, leading to even more opinionated future models.

We conclude that we have to be more careful about the opinions that are built into deployed language models. As of this writing, we are not aware of systematic efforts seeking to understand and map the opinions built into large language models. Critical audits of language models have primarily focused on detecting and reducing discriminating (Huang et al., 2019; Brown et al., 2020; Nozza et al., 2021) or otherwise offensive (Askell et al., 2021) output. A first exploration of the opinions built into GTP-3 by Johnson et al. (2022) suggests that the model's preferred views seem to align with dominant US public opinion. In addition, a version of GPT trained on 4chan data has led to controversy about the ideologies and types of speech that should be avoided in training data. But for the most widely used language models like GPT-3, we still do not completely understand the type of opinions, attitudes, and ideologies they may reinforce.

As we get a better understanding of what opinions are currently built into language models, more research is required on how to create models that have balanced opinions or support a desired opinion. Further, a broad public debate will be needed to decide what opinion models should perpetuate in the first place. While there is widespread agreement that large language models should not be offensive or perpetuate stereotypes (Huang et al., 2019; Brown et al., 2020; Nozza et al., 2021), it is less clear what a language model should have to say about issues such as immigration, social inequality, and vegetarianism. Given that the opinions

of the model will affect the writing and views of millions of users, what opinions are built into models is a political decision that should not be limited to system developers alone.

Chapter 5

People have Different Priorities for Managing Risk in AI

The previous chapters have demonstrated that humans are unable to detect language produced by GPT-3, that using large language models in self-presentation may damage interpersonal trust, and that interactions with large language models changes expressions and attitudes. We now proceed from providing empirical evidence that using large language models in human communication has far-reaching consequences to the to the question of how these risks can be managed. In Chapter 6, we will argue that the there is a need for more democratic forms of governing large language models' risks. In the current chapter, we produce empirical data to substantiate this argument as well as a tool that can facilitate a participatory management of language models' risk. We develop and AI value survey and field it across three groups: a representative sample of the US population (N=743), a sample of crowdworkers (N=755), and a sample of AI practitioners (N=175). Our results empirically show that AI practitioners' value priorities differ from those of the general public. Compared to the US-representative sample, AI practitioners appear to consider responsible AI values as less important and emphasize a different set of values. In contrast, self-identified women and black respondents found responsible AI values more important than other groups. Surprisingly, more

liberal-leaning participants, rather than participants reporting experiences with discrimination, were more likely to prioritize fairness than other groups. Our findings highlight the importance of paying attention to who gets to set the priorities in managing the risks of large language models. Since our considerations on risk management for large language models' apply to the risk management of AI systems more generally, we refer to AI systems more generally in this chapter.

Introduction

The advances in language models and artificial intelligence discussed in this dissertation have the potential to benefit people and society, but they also raise ethical challenges and concerns about possible adverse impacts (Montreal, 2017). Being prone to errors and biases, AI systems may harm people (Awad et al., 2018) for instance by reinforcing stereotypes (Blodgett et al., 2020) or by increasing social inequality (Eubanks, 2018). While the larger consequences of AI can be difficult to anticipate (Boyarskaya et al., 2020), systems developed with broader human and societal values in mind stand a better chance of preserving these values (Raji et al., 2020; Friedman, 1996; Agre and Agre, 1997). To support the development of socially beneficial AI technologies, several private companies, public sector organizations, and academic groups have published ethics guidelines with values they consider important for responsible AI (Jobin et al., 2019).

These AI ethics guidelines have been found to largely converge on five central values (Jobin et al., 2019): transparency, fairness, safety, accountability, and privacy. But these values may differ from what a broader and more representative population would consider important for the AI technologies they interact

with. While prior work has shown that value preferences depend on peoples' backgrounds and personal experiences (Davis and Steinbock, 2021; Intemann, 2010), AI technologies are often developed by relatively homogeneous and demographically skewed subsets of the population (Landivar, 2013; Crawford, 2016; House, 2016). Given the lack of reliable data on other groups' priorities for responsible AI, practitioners may unknowingly encode their own biases and assumptions into their concept and operationalization of responsible AI (Martin, 2019; Raji et al., 2020).

In this work, we present the results of a survey we developed, validated, and fielded to elicit peoples' value priorities for responsible AI. Drawing on the traditions of empirical ethics (Musschenga, 2005; Dunn et al., 2012) and value elicitation research (Fischhoff, 1991; Schwartz, 2007), our survey asks participants about the perceived importance of a set of 12 responsible AI values both in general and in specific deployment scenarios. To increase robustness, respondents assessed values from three perspectives: value selection, contextual assessment of values, and comparative prioritization of values (detailed in §5).

We administered this survey to three different populations. We analyzed how value priorities of a US census-representative sample (N=743), a crowdworker sample (N=755), and an AI practitioner sample (N=175) vary by deployment scenario and individuals' backgrounds and experiences. We surveyed the value priorities of AI practitioners as they are often the ones making decisions about the AI technologies that are being developed, and compared their preferences to those of a more representative sample. We also consulted crowdworkers as they are already involved in producing data that AI systems are evaluated on to explore the feasibility of involving them in the ethical assessment of AI systems as well.

Our results provide evidence that responsible AI values are perceived and prioritized differently by different groups. AI practitioners, on average, rated responsible AI values less important than other groups. At the same time, AI practitioners prioritized fairness more often than participants from the US-census representative sample who emphasized safety, privacy, and performance. We also find differences in value priorities along demographic lines. For example, women and black respondents evaluated responsible AI values as more important than other groups. We observed the most disagreement in how people traded-off fairness with performance. Surprisingly, participants reporting past experiences of discrimination did not prioritize fairness more than others, but liberal-leaning participants prioritized fairness more than conservative-leaning participants.

Our results highlight the need for AI practitioners to contextualize and probe their ethical intuitions and assumptions. The empirical approach to AI ethics explored in this study can help to increase the context sensitivity of the responsible AI development process. However, as we elaborate in the discussion, opinion research can inform ethical decision-making, but cannot replace sound ethical reasoning.

Background

Our study draws on prior work on responsible AI, value sensitive design (Friedman, 1996), empirical ethics (Musschenga, 2005), value elicitation (Fischhoff, 1991; Schwartz, 2007), and standpoint theory (Intemann, 2010).

AI ethics guidelines and value-sensitive design

Science and technology studies theorize that computing technologies incorporate a tacit understanding of human nature (Winograd et al., 1986). Algorithms are described as value-laden artifacts (Martin, 2019) that encode developer assumptions, including ethical and political values (Raji et al., 2020). From this perspective, a product team that decides to maximize the chance that a disease detection system will recognize a disease at the cost of increasing false alarms prioritizes certain values over others. Past work has shown that machine learning development and research often narrowly focus on technical values such as accuracy, efficiency, and generalization (Birhane et al., 2021; Nanayakkara et al., 2021). In contrast, proponents of value-sensitive design (Friedman, 1996; Friedman and Nissenbaum, 1996), reflective design (Sengers et al., 2005), and critical technical practice (Agre and Agre, 1997) advocate that AI systems should be designed with broader human and societal values in mind.

What values developers of responsible AI systems should emphasize remains a key question. Some argue these values should be naturally embedded in an organization's culture (Raji et al., 2020). Several organizations have also published guidelines describing what values they believe AI systems should embody. Jobin et al. (2019) found these guidelines to converge around central values, but differ in how they construe these values and concepts. Critics note that reliable methods to translate values into practice are often missing (Raji et al., 2020; Mittelstadt, 2019). Some also argue that statements of high-level values and principles are too ambiguous and may gain consensus simply by masking the complexity and contending interpretations of ethical concepts (Whittlestone et al., 2019). For example, people may agree on the importance of fairness, but "fairness" in and by

itself has little to say about what is fair and why (Binns, 2018).

Our study validates and contextualizes value priorities outlined in AI ethics guidelines. To date, there is little empirical data on values a broader and more representative public finds important for the AI technologies they interact with. Our empirical approach to AI ethics probes for possible blind spots in AI practitioners' and researchers' assumptions.

Empirical studies of human values and AI ethics

Eliciting people's values is a central pursuit in the social sciences (Fischhoff, 1991). Economists explain choices in the marketplace based on value theory, sociologists seek to understand which values are held by a community and how they change. Psychologists use value elicitation for therapy and counsel, and empirical ethicists enhance the context-sensitivity of their arguments by combining social scientific methods with ethical reasoning (Musschenga, 2005). While drawing normative conclusions from empirical results is difficult, empirical data on ethical preferences can inform decision making (Musschenga, 2005).

Several studies have examined people's ethical intuitions concerning AI technologies. In the "moral machine" experiment, Awad et al. (2018) generated a variety of moral dilemmas a self-driving car might find itself in and ask participants which course of action they recommend. They report significant cross-cultural differences in ethical preferences correlated with modern institutions and cultural traits. Hidalgo et al. (2021) explored how people judge humans and machines differently when they make mistakes. They found that people tend to forgive machines more in scenarios with high intentionality. Similarly, Malle et al. (2015) compared how

people apply moral norms to humans versus robots. Most related to the empirical study of responsible AI values, Saxena et al. (2019) have compared public perceptions of different fairness paradigms. Similarly, Grgic-Hlaca et al. (2018) and Pierson (2017) have studied which features people find fair to include in a prediction algorithm. They found substantial disagreement among participants (Grgic-Hlaca et al., 2018), with e.g., women being less likely to include gender as a feature in a course recommendation algorithm if this might result in female students seeing fewer recommendations for science courses (Pierson, 2017).

Going beyond previous work, we develop a responsible AI value survey to explore what values people find most important for responsible AI. Where previous studies have elicited preferences concerning specific technical implementations with convenience samples, we provide a first high-level perspective on a representative public's priorities for the AI system they interact with and might be affected by.

The impact of background and context on value priorities

Feminist empiricists and standpoint theorists argue that knowledge is achieved from a particular standpoint (Wylie et al., 2003) and that social location systematically influences our experiences and decisions (Intemann, 2010). They hold that homogeneous communities are prone to false consensus effects (Ross et al., 1977) where individuals believe that the collective opinion of their own group matches that of the larger population. In homogeneous communities, inaccurate assumptions or biases can be hard to recognize and correct (Intemann, 2010; Boyarskaya et al., 2020). In communities comprised of individuals with diverse values and experiences, however, how assumptions influence reasoning becomes more visible (Intemann, 2010; Longino, 2020; Rolin, 2006). Including historically under-

represented groups, in particular, may lead to rigorous critical reflection as their experiences may facilitate the identification of problematic background assumptions (Intemann, 2010).

Demographics and experiences not only affect background assumptions (Dobbe et al., 2018), but also shape people's values and ethical preferences (Fumagalli et al., 2010; Graham et al., 2016). Rather than stemming from overarching belief systems, values often arise through particular social practices in a specific context (MacIntyre, 1981). As such, ethical intuition is contextual and socially situated (Davis and Steinbock, 2021). For instance, what's fair to some people may seem unfair to others (Lee and Baykal, 2017), and some people value privacy and autonomy more than others (Whittlestone et al., 2019). The population of AI practitioners is demographically skewed (Landivar, 2013; Crawford, 2016; House, 2016) with e.g., women and black people being underrepresented (Dillon Jr et al., 2015). With their specific demographics and experiences, AI practitioners may bring their own preferences to what it means for AI to be "responsible" or "ethical", such as a bias towards deployment (Kaur et al., 2020). Responsible AI technologies developed within homogeneous communities may fail to account for the experiences and needs of various groups, so it remains crucial to scrutinize who gets to define AI ethics (Jobin et al., 2021).

By surveying representative population samples about their priorities for responsible AI, we seek to validate the value prioritization in AI ethics frameworks. We explore the social relativity of responsible AI values to provide grounds for more critical reflection about possibly inaccurate assumptions and false consensus effects.

Methods

To study how people perceive and prioritize responsible AI values, we combine instruments from value elicitation research (Fischhoff, 1991) with the concepts and principles found in AI ethics guidelines (Jobin et al., 2019). We fielded an iteratively developed online survey with 743 census-representative participants, 755 crowd workers, and 175 AI practitioners.

Survey development

We adapted the Schwartz Value Survey (Schwartz, 1992, 1994) to apply it to responsible AI values. The Schwartz Value Survey has been used to study individual and intercultural differences in general human values in over 60 countries (Schwartz, 2007). Based on an inventory of human values, the Schwartz Value Survey asks respondents to self-report which values are most important to them. Respondents rate the importance of each value on a Likert scale while explanations for each value are shown.

Selecting and explaining responsible AI values. To adapt the Schwartz Value Survey to the study of AI ethics, we constructed an inventory of responsible AI values. The responsible AI values we chose for our survey are based on a review of published AI ethics guidelines. We drew on work by Jobin et al. (2019) finding that AI ethics guidelines commonly refer to transparency, justice & fairness, non-maleficence, accountability, privacy, beneficence, freedom & autonomy, trust, and dignity. To this list, we added system performance, as it is a central value in AI research and development (Birhane et al., 2021) that is often used to compare AI

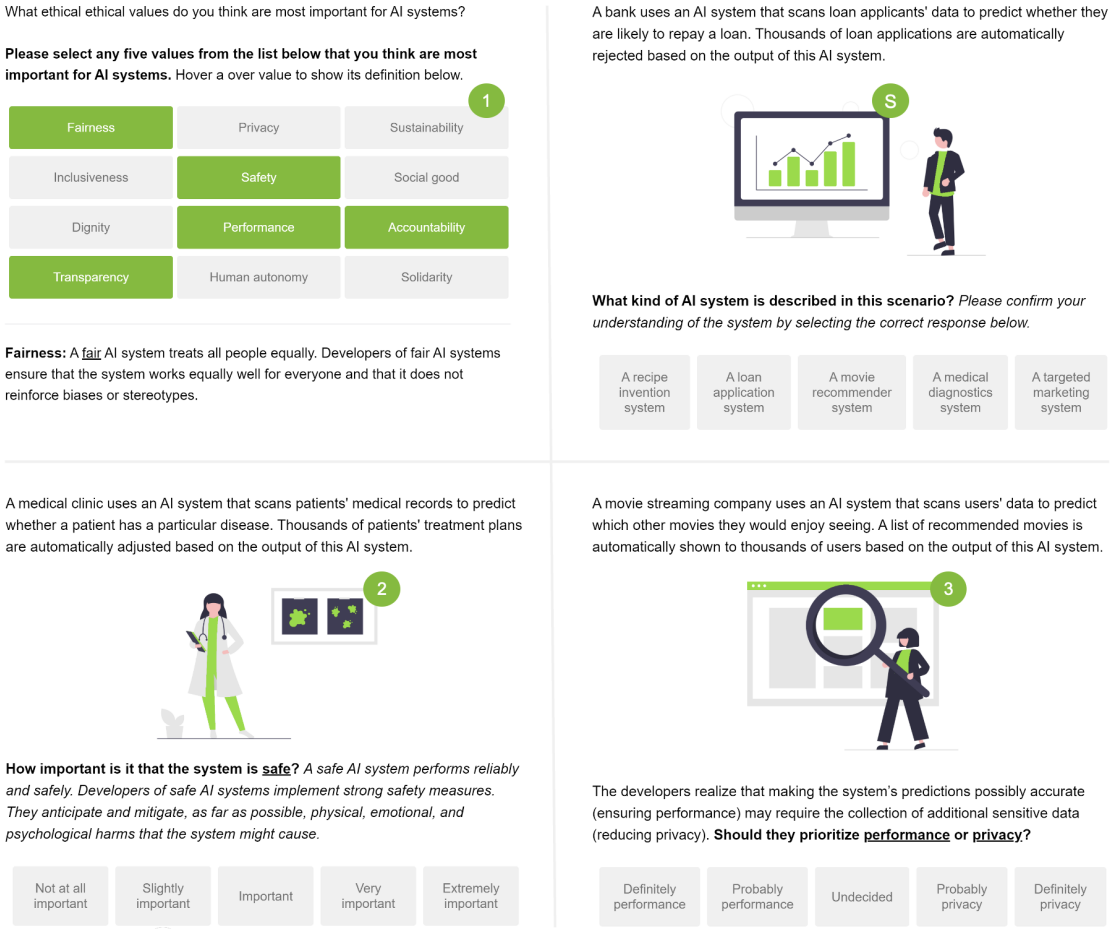

Figure 5.1: Overview of the main survey components. Participants first completed a value selection task **(1)**. After confirming the understanding of the respective deployment scenario **(S)**, they evaluated how the importance of values in context **(2)**. Finally, participants indicated how they would prioritize values when they are in conflict **(3)**.

models and to make deployment decisions.

As responsible AI values are abstract and participants may not easily understand how they apply in the context of AI technologies (Cave et al., 2018), we provided additional explanations. To formulate explanations for each value, we drew again on existing AI ethics guidelines, including Microsoft's responsible AI principles (Microsoft, 2020), Google's AI Principles (Google, 2020), the Montreal

Declaration for the Responsible Development of Artificial Intelligence (Montreal, 2017), the Deloitte AI ethics guide (Deloitte, 2020), IBM's Principles for Trust and Transparency (IBM, 2020), and the EU's Ethics guidelines for trustworthy AI (Union, 2020).

We tested and iterated on different explanations of responsible AI values in four crowdsourcing pilot studies (N_1=40, N_2=80, N_3=40, N_4=160). Each pilot asked participants whether they understood an explanation through both Likert scales and open-ended responses. Based on the pilot results, we substituted "non-maleficence" with "safety" and "beneficence" with "social good," as the former were not well-understood by participants. We also explicitly referred to "*human* autonomy" to avoid confusion with autonomous cars and robots. Finally, we did not include "trust" as it appeared overly general and overlapped with other values such as transparency and accountability.

We phrased the explanations in simple, non-technical language, all following the same structure. Each explanation starts with a sentence describing what a system embodying the value would do, followed by an example of steps developers might take to realize a value, e.g.: "*An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives.*" By complementing a general definition with specific operationalizations of a value, the framing provides a tangible understanding of the value while maintaining a degree of generality.

Identifying pairs of possibly conflicting responsible AI values. In addition to assessments of values themselves, we asked participants about their preferences in cases of conflicting values (Barocas and Boyd, 2017). For example, ensuring

fairness might require collecting additional sensitive data, potentially diminishing privacy. To identify value conflicts, we searched for mentions of conflicts in the literature for each pair of values in the responsible AI value inventory. We found prior discussions of trade-offs between privacy & performance (Bagdasaryan et al., 2019; Shokri and Shmatikov, 2015), fairness & privacy (Bagdasaryan et al., 2019; Ekstrand et al., 2018), fairness & performance (Corbett-Davies et al., 2017; Pleiss et al., 2017), safety & transparency (Hua et al., 2021; Meijer et al., 2014; Cappelli et al., 2010), and autonomy & safety (Livingstone et al., 2011). We combined the value explanations developed above to introduce the conflicts to participants, e.g. "*The developers realize that minimizing the collection of sensitive data (ensuring privacy) may make the system's predictions less accurate (reducing performance). Should they prioritize privacy or performance?*"

Constructing hypothetical AI deployment scenarios. We used hypothetical scenarios to make value assessments more tangible and to elicit judgments in specific contexts. We produced four hypothetical deployment settings validated through two pilot studies (N_1=180, N_2=160). To design these scenarios, we selected 25 AI systems people may have encountered in everyday settings starting with a list of general AI use cases (Dilmegani, 2018). We developed short explanations of these use cases and asked pilot participants whether they found them understandable and relatable. Based on the pilot results, we further refined the scenarios and kept only the 10 scenarios that were most easily understood by pilot participants. The second pilot then asked participants which scenarios they understood best and whether the AI system's decisions were highly consequential. Based on the responses, we selected two well-understood high-stake and low-stake scenarios for the study:

(a) Medical: An AI system used by a medical clinic to predict whether a patient has a disease (high-stake)

(b) Banking: An AI system used by a bank to predict whether an applicant will repay a loan (high-stake)

(c) Marketing: An AI system used by a marketing company to match ads to viewers (low-stake)

(d) Streaming: An AI system used by a streaming company to recommend movies to users (low-stake)

Each scenario states the entity controlling the AI system and the type of data the system is using. It then elaborates what predictions are being made and what actions are being taken based on the prediction, e.g.: "A medical clinic uses an AI system that scans patients' medical records to predict whether a patient has a particular disease. Thousands of patients' treatment plans are automatically adjusted based on the output of this AI system."

Survey procedure

After providing informed consent, participants received a high-level introduction both covering the general goals of AI and noting the complex decision-making involved in the AI system development beyond technical challenges.Figure 5.1 illustrates the subsequent survey steps which combined three value elicitation tasks: (1) *value selection*—select five responsible AI values (out of the 12) that are deemed most important in general, (2) *contextual assessment*—evaluate the perceived importance of seven central responsible AI values (transparency, fairness, safety, accountability, privacy, autonomy, and performance) in a specific deployment setting,

and (3) *comparative assessment*—recommend what product teams should do when values are in conflict.

Participants selected the five most important values for AI systems in general, with explanations displayed when a value was hovered over. They then read the first scenario and confirmed their understanding of the deployment setting. Overall, participants encountered four scenarios. In scenarios 1 and 2, participants indicated how important they thought three responsible AI values were in the given situation on a 5-point Likert scale. In scenario 3, participants evaluated one more value and then two value conflicts by indicating which value they thought should be prioritized in the given situation. Finally, they evaluated three value conflicts in the fourth and last scenario. For every rating, participants were given the option to explain their choices.

After completing the rating tasks, participants indicated their familiarity with machine learning, user research, and their personal experiences with discrimination. We selected these experiential correlates based on the hypothesis that personal experience might inform ethical preferences (Davis and Steinbock, 2021). For example, user researchers may have learned to empathize with users, whereas respondents trained in ML may have better insight into the technical constraints of responsible AI. We also asked participants to report their gender identity, age, ethnicity, political views, sector of work, and highest level of education. Again, these demographic correlates were selected to explore to what extent social location influences the perceived importance of responsible AI values (Intemann, 2010). For all experiential and demographic questions, participants could choose not answer.

Participant recruitment

To examine how different groups assess responsible AI values, we surveyed three populations:

A US census-representative sample ($N=743_1$) was recruited by Qualtrics to gain insights into how the general population assesses the importance of responsible AI values. The recruitment process combined a variety of methods to minimize biases and performed stratified random sampling to match the US census along gender, age, race, region, and household income. Participant compensation was handled by Qualtrics.

A convenience US-based crowdworker sample ($N_2=755$) was recruited via the Clickworker crowdsourcing platform. Participants were US-based and likely previously contributed to the training of AI models by e.g., providing data labels. Each participant received USD 2.8 for a median participation time of 8 minutes. While crowdworkers are not directly involved in the AI development process, their judgments are often a key ingredient to machine learning systems. We explored whether their assessments could serve as proxies for the ethical intuition of a more representative population.

A sample of AI practitioners ($N_3=175$) was recruited through an open call on Twitter (N=156) and internal mailing lists (N=19) at a large tech company. Our call for participation targeted US-based participants whose work is related to AI/ML. We confirmed their background in the survey, but ultimately rely on self-reported expertise. For the internal mailing lists, we specifically targeted teams doing AI/ML related work. Participants could choose to enter a raffle to win

one of five $50 gift vouchers after study completion. AI practitioners are a relevant population that makes key decisions throughout the AI development process. We explore whether their value judgments differ from those of the more general population.

We had to work with different types of compensation due to differences in respondent type and recruitment method across samples. However, we aimed to provide roughly commensurate compensation across recruitment methods. The study was IRB approved, and we obtained informed consent from all our participants.

Data quality control

To counterbalance ordering effects, the arrangement of scenarios, values, and conflict questions was randomized. In addition, the order of response options was randomly flipped per participant. For the conflict questions, we also randomized the internal order of the conflict, e.g. fairness vs. performance was inverted to performance vs. fairness. A pop-up window asked participants to slow down whenever they attempted to submit responses in under 3 seconds per survey page to deter spammers and inattentive participants. The four scenario introductions throughout the survey served as attention and comprehension checks for our participants. We removed all participants that had failed more than one attention check from our analysis to increase response quality, reducing the relevant samples to N_1=516, N_2=607, N_3=140 respectively.

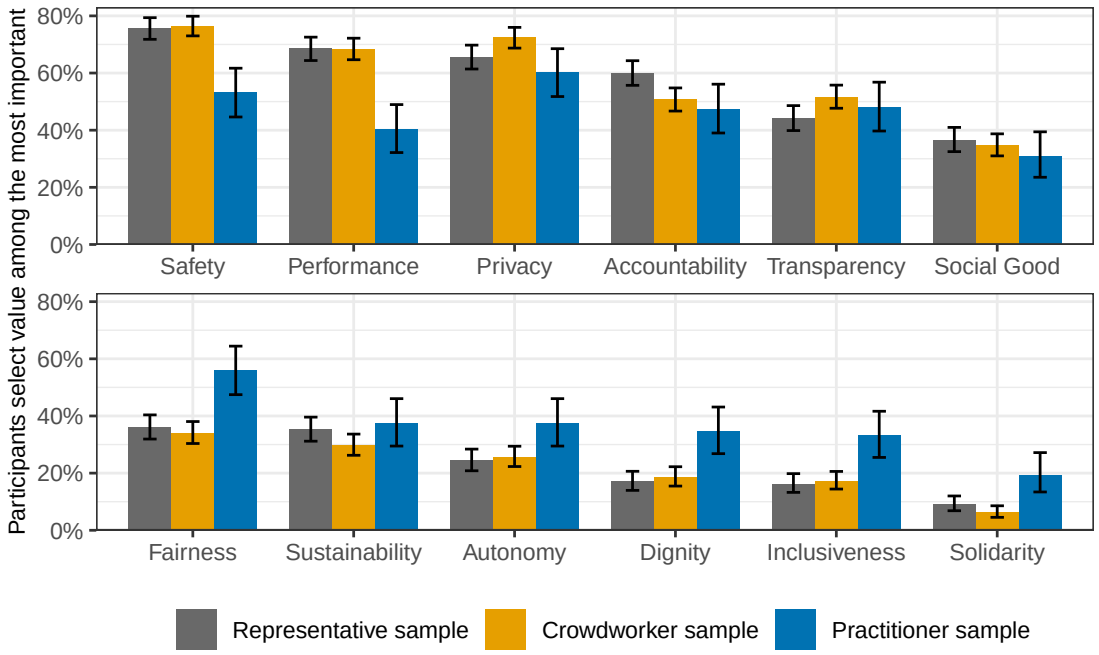

Figure 5.2: **AI practitioners' value priorities differ from those of the general public.** N_1=516, N_2=607, N_3=140. The x-axis shows the 12 responsible AI values respondents chose from, while the y-axis indicates how often respondents selected a value among the five most important. Participants from the US-census representative sample and the crowdworker sample selected safety, performance, and privacy most often among their five most important values, while practitioners selected fairness more often.

Results

What values are deemed as most important in general?

In Task 1, participants selected five values they deemed most important for AI systems out of an inventory of 12 responsible AI values (Figure 5.2). 76% of respondents from the US-census representative sample selected safety among the top 5 responsible AI values. Over 60% of participants in this representative panel

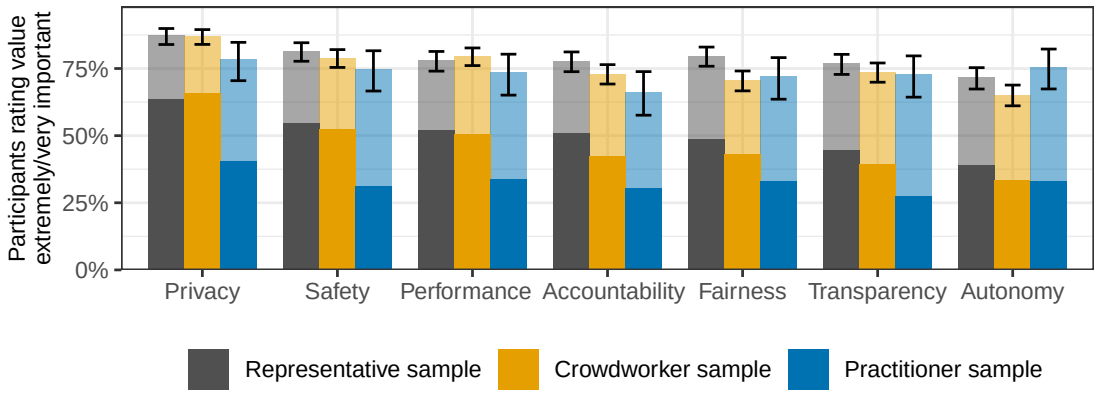

Figure 5.3: **Representative participants rated responsible AI values as more important than AI practitioners did.** N=140 to 607 ratings per bar. The x-axis shows the assessed responsible AI values and the y-axis indicates how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark).

also selected performance, privacy, and accountability among the most important values. Respondents from the crowdworker sample selected accountability less often, but their preferences were largely consistent with those from the US-census representative sample. AI practitioners' preferences were less focused. Compared to the US-census representative sample, practitioners selected humanist values such as fairness, inclusiveness, dignity, and solidarity more often and were less likely to select safety and performance among the most important values.

How important are values in specific deployment scenarios?

In Task 2 participants evaluated how important they considered a value in the context of a specific deployment scenario (Figures 5.3 and 5.4). The perceived importance of performance, accountability, fairness, and transparency varied significantly across deployment settings. In general responsible AI values were rated

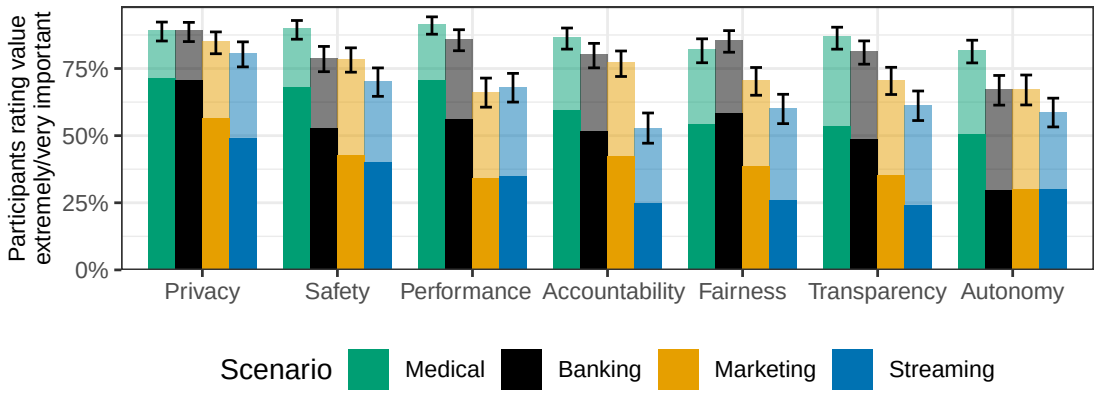

Figure 5.4: **Responsible AI values were rated as most important in the medical and banking scenarios.** N=287 to 344 ratings per bar aggregated across samples. The y-axis shows how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark). The perceived importance of other values is dependent on the application context.

as very or extremely important. Compared to both the US-census representative and the crowdworker samples, on average, AI practitioners evaluated responsible AI values, and privacy, safety, and performance, in particular, as less important. We also observed significant variation of perceived importance across deployment settings, with responsible AI values being considered most important in the medical context and least important in the streaming context.

How values are prioritized when in conflict

In Task 3 participants suggested how values should be prioritized when in conflict (Figures 5.5 and 5.6). Respondents from all participant samples agreed on prioritizing safety over autonomy and transparency. Across scenarios, a majority of respondents agreed on prioritizing privacy over performance or fairness. Most disagreement was observed when performance and fairness conflicted: Participants

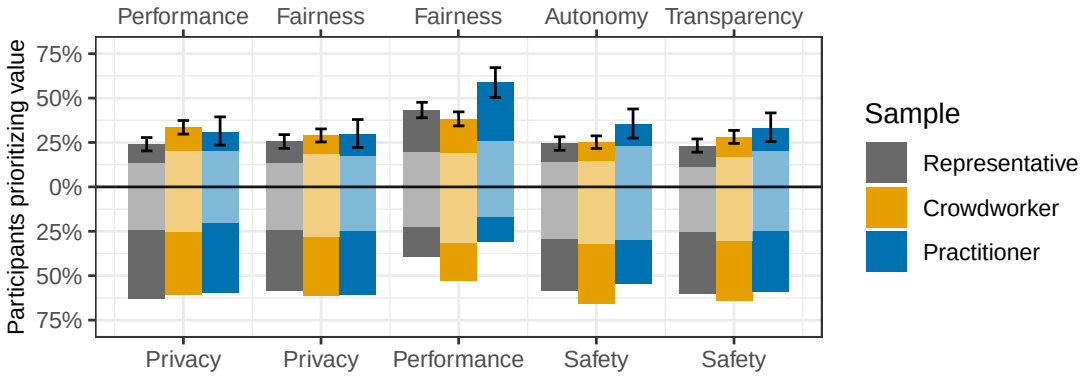

Figure 5.5: **Participants across groups prioritized privacy and safety over fairness, but disagreed on the fairness vs. performance tradeoff.** N = 104 to 607 ratings per bar. The conflicting value pairs are shown on the top and bottom, e.g., performance vs. privacy on the left. The proportion of respondents prioritizing the top value is shown to the top and the proportion of respondents prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, whereas weak preferences are lightly shaded. Undecided respondents are omitted.

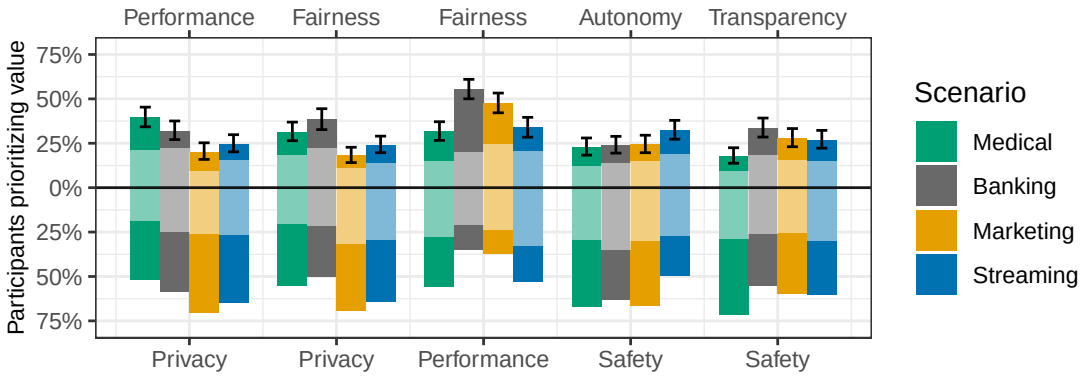

Figure 5.6: **Value priorities vary by context, but most participants prioritized privacy and safety across most scenarios.** N = 276 to 341 ratings per bar aggregated across samples. The proportion of respondents prioritizing the top value are shown to the top and the proportion of respondents prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, weak preferences are lightly shaded.

from the US representative sample were almost equally split in their preferences for fairness versus performance. Crowdworkers were less likely to prioritize performance and AI practitioners were more likely to prioritize fairness than the US-census representative participants.

Across scenarios, respondents prioritized privacy over performance and fairness, and safety over autonomy and transparency. Again, the performance-fairness trade-off produced most variation: Participants prioritized performance in the medical and streaming scenario, and fairness in the banking and marketing scenario.

Table 5.1: Statistical analysis predicting value importance ratings based on scenario, sample, and demographic correlates based on scenario, sample, and demographic correlates. The constant corresponds to a white male respondent from the US-census representative sample assessing a value in the banking scenario. Bold text indicates statistical significance.

	Privacy	Safety	Perform.	Account.	Fairness	Transp.	Autonomy
Marketing system	**-0.051****	-0.024	**-0.14*****	**-0.046***	**-0.10*****	**-0.08*****	-0.004
Medical system	0.005	**0.06*****	**0.044***	0.030	-0.031	0.029	**0.10*****
Streaming system	**-0.08*****	**-0.09*****	**-0.12*****	**-0.18*****	**-0.18*****	**-0.16*****	-0.036
Crowdworker sample	0.005	-0.020	0.002	**-0.05*****	**-0.06*****	**-0.032***	**-0.041***
Practitioner sample	**-0.08*****	**-0.057***	-0.052	**-0.10*****	**-0.057***	**-0.09*****	0.005
Women respondents	**0.04*****	**0.039****	**0.05*****	**0.04***	**0.05*****	**0.028***	**0.07*****
Gender-diverse resp.	-0.042	-0.031	-0.046	-0.017	0.086	-0.010	0.041
Black respondents	**0.046***	**0.052***	**0.062****	0.006	**0.08*****	0.019	0.031
Hispanic respondents	0.024	0.042	0.044	0.034	0.020	0.021	0.023
Asian respondents	0.001	-0.025	-0.017	-0.018	-0.031	**-0.057***	-0.007
Age	0.001	-0.0001	-0.001	-0.0004	-0.001	-0.0003	0.001
Education	-0.020	-0.047	-0.026	0.009	0.007	0.012	0.013
Political leaning	**0.060***	0.011	0.027	0.023	**0.056***	0.049	0.006
Exp. with discrimination	-0.027	0.011	**-0.057***	0.002	0.004	-0.008	0.005
Familiarity with ML	-0.037	-0.007	0.005	-0.028	-0.016	-0.011	0.009
Familiarity with UX	**0.046***	0.043	**0.053***	0.034	**0.055***	**0.057***	0.043
Constant	**0.82*****	**0.80*****	**0.83*****	**0.82*****	**0.81*****	**0.77*****	**0.62*****
Observations	1,246	1,246	1,246	1,246	1,246	1,246	1,246
R^2	0.082	0.084	0.150	0.130	0.119	0.111	0.070
Adjusted R^2	0.070	0.072	0.139	0.118	0.107	0.099	0.058
F Statistic	6.84***	7.05***	13.51***	11.42***	10.33***	9.58***	5.81***
Note:							*p<0.05; **p<0.01; ***p<0.001

Table 5.2: Regression analysis with simple baseline models predicting the value preference ratings based on scenario, sample, and demographic correlates. The constant corresponds to a white male respondents from the US-census representative sample recommending a value prioritization in the banking scenario. Bold text indicates statistical significance.

	Privacy. vs. performance	Privacy vs. fairness	Performance vs. fairness	Safety vs. autonomy	Safety. vs. transparency
Marketing system	**0.164****	**0.301*****	**0.113***	0.067	0.080
Medical system	**-0.112***	**0.113***	**0.378*****	0.078	**0.259*****
Streaming system	0.091	**0.209*****	**0.342*****	**-0.150****	0.086
Crowdworker sample	-0.079	0.048	**0.152*****	0.058	0.015
Practitioner sample	-0.060	0.114	-0.091	-0.078	0.062
Women respondents	0.055	0.038	0.057	0.076	**0.113****
Gender-diverse resp.	0.219	-0.128	-0.182	0.060	-0.214
Black respondents	-0.006	-0.098	-0.109	**0.168****	-0.029
Hispanic respondents	-0.050	**-0.160***	0.082	-0.017	-0.021
Asian respondents	-0.033	0.024	-0.113	0.033	0.020
Age	0.001	**0.003***	-0.002	0.0002	0.0002
Education	0.104	-0.126	-0.067	0.045	-0.063
Political leaning	0.010	**-0.191***	**-0.226****	0.091	-0.017
Exp. with discrimination	-0.113	**-0.205****	0.023	**-0.160***	0.083
Familiarity with ML	0.008	0.073	-0.008	0.090	-0.078
Familiarity with UX	-0.093	0.124	0.026	0.048	0.016
Constant	**0.262***	0.100	-0.057	0.102	0.159
Observations	1,246	1,246	1,246	1,246	1,246
R^2	0.032	0.056	0.080	0.038	0.031
Adjusted R^2	0.019	0.043	0.068	0.025	0.018
F Statistic	2.535***	4.526***	6.719***	3.009***	2.459**
Note:					*p<0.05; **p<0.01; ***p<0.001

Demographics and experiential correlates of value priorities

To explore how demographic and experiential factors correlate with participants' assessments, we mapped their responses to a 5-Likert scale that preserves the direction of the original scale. Treating ordinal scales as interval scales is controversial, but the scales in our study have a unit of measurement with comparable-size intervals and a zero point, so a continuous analysis is meaningful and justifiable (Knapp, 1990). To examine how various demographic, experiential, or contextual factors may explain the variance in respondents' assessments, we used linear regression to build simple baseline models that predict their assessments.

Table 5.1 shows parameter estimates of linear regression models fitted to predict how important respondents consider a value in a specific scenario. The model constant corresponds to a white man from the US-census representative sample evaluating a responsible AI value in the banking scenario. The parameter estimates confirm that the perceived importance of values varies significantly across deployment settings. They also confirm that, compared to the US-census representative sample, AI practitioners evaluated most values as less important. Women and black respondents, on average, evaluated most responsible AI values as more important than other groups. Among the experiential correlates, a self-reported liberal political leaning was associated with a higher valuation of privacy. Self-reported experiences with discrimination predicted lower perceived importance of performance but were not statistically significantly correlated with other responsible AI values. While familiarity with ML did not predict different value priorities, respondents reporting to be familiar with UX research evaluated most responsible AI values as more important.

Table 5.2 shows parameter estimates predicting participants' preference in the case of conflicting responsible AI values. Positive coefficients correspond to a preference for the top value. Responses vary significantly by deployment context, but only the response to the fairness-performance trade-off varies by sample. Women respondents were more likely to prioritize safety over transparency than other groups, and black respondents were more likely to prioritize safety over autonomy. While participants reporting experiences of discrimination were more likely to prioritize fairness over privacy, they were not more likely to prioritize fairness over performance than other groups. Instead, participants with liberal political learning were more likely to prioritize fairness over performance and privacy than other groups. Familiarity with ML neither predicted a preference for performance over privacy nor fairness.

Some variables were correlated with each other. For example, the practitioner sample contains fewer women respondents ($r=-0.14$, $p<0.01$) and black respondents ($r=-0.11$, $p<0.01$), but more educated ($r=0.33$, $p<0.01$) and liberal-leaning ($r=0.2$, $p<0.01$) respondents. Similarly, liberal-leaning respondents were younger ($r=-0.13$, $p<0.01$) and more likely to report experiences with ML ($r=0.1$, $p<0.01$) and discrimination ($r=0.09$, $p<0.01$). However, a correlation analysis suggests that no covariates were highly correlated ($r>0.7$). The variance inflation factor remained below 1.5 across all covariates, indicating little to no multicollinearity issues (Hair, 2009).

Discussion

AI practitioners' value priorities for responsible AI differ from those of the general public. Our results empirically corroborate a commonly raised concern: AI practitioners' value preferences for responsible AI are not representative of the value priorities of the wider US population. Compared to a US-census representative public, AI practitioners evaluated responsible AI values as less important in general and emphasized a different set of values.

US-census representative and crowdworker respondents agreed on what values they found most important: safety, privacy, and performance. Practitioners, in comparison, were more likely to prioritize fairness, dignity, and inclusiveness.

These findings align with prior research finding that different groups have different normative expectations of how AI systems should behave in specific situations (Grgic-Hlaca et al., 2018; Pierson, 2017; Awad et al., 2018; Hidalgo et al., 2021). Our findings extend prior work by demonstrating how AI practitioners' ethical preferences differ from other groups'. We also show that groups not only differ in their judgment of specific behaviors and technical details, but may disagree on the importance of the very values at the core of responsible AI. *The disagreement in value priorities highlights the importance of paying attention to who gets to define what constitutes "ethical" or "responsible" AI.* Responsible AI guidelines (Jobin et al., 2019) may emphasize a different set of values depending on who writes them and who is consulted. We hypothesize that consulting populations outside the Western world about their priorities for responsible AI would surface even starker disagreement about the values underlying responsible AI (Schwartz, 2007; Kapania et al., 2022).

What might explain the differences in value priorities between AI practitioners' and other groups? Our results provide limited insight into plausible drivers of differences in values. First, women and black respondents assessed responsible AI as more important than other demographic groups. Their relatively low representation in the AI practitioner sample compared to the representative sample (only 40% and 2.2% compared to 52% and 15% respectively) explains about 15% of the lower importance ratings AI practitioners assigned to values in general. Increasing the representation of e.g., women and black researchers in AI (Landivar, 2013; Crawford, 2016; House, 2016) may thus result in responsible AI values receiving more attention.

Another demographic variable that robustly predicted differences in value preferences was respondents' political leaning. Liberal-leaning respondents were 10% more likely to select fairness amongst the most important values than conservatives, and were 15.5% more likely to prioritize fairness in the fairness-performance trade-off. Compared to the representative sample, AI practitioner respondents were substantially more likely to self-identify as liberal-leaning (52% compared to 26%), explaining about 27% of practitioners' different evaluation of fairness. This result is in line with the broader research on value differences along ideological lines (Braithwaite, 1998; Wetherell et al., 2013). It highlights that guidelines for responsible AI need to navigate a polarized value landscape.

Other demographic and experiential variables, however, were less predictive of how our participants assessed responsible AI values. Respondents reporting experience with discrimination were more likely to prioritize fairness over privacy, but did not evaluate fairness as more important than other groups. When asked whether developers should prioritize fairness over performance, participants from

minoritized groups and participants reporting experience with discrimination were as undecided as other groups. While previous work identified performance as the central value in machine learning research (Birhane et al., 2021), our results do not suggest that AI practitioners or respondents familiar with machine learning were more likely to value performance. Participants trained in user experience research, however, evaluated responsible AI values more important in general.

Can AI practitioners use crowdsourcing to complement their ethical intuitions in the development process? Our findings emphasize the need for bringing in a diversity of perspectives when decisions are made about the development and operationalization of responsible AI. Crowdworkers are often the go-to convenience sample, but to what extent could they provide a reliable lens into the values that a broader population expect AI systems to adhere to?

As in prior research (Huff and Tingley, 2015), we find that the value priorities of crowdworkers largely align with those of the US-census representative sample. Our results also show that often a majority of participants agreed on value trade-offs. For example, respondents from all samples prioritized privacy over performance across all deployment scenarios. The agreement raises the question of whether and when product teams could use such results to e.g., justify prioritizing privacy over performance.

Here, consensus alone may not justify practical requirements within specific contexts of use. Rather than providing definite answers, the approach developed in this paper provides "values levers" (Shilton, 2013): organizational processes that take the implicit work of value judgments in technology development and transform it into an explicit matter of debate and documentation. Empirical data on different groups' preferences can both inform the development process of responsible AI

and provide opportunities for critical reflection. Rather than prescribing value priorities, responsible AI guidelines could ask practitioners to justify their choices whenever they go against commonly held value preference.

Limitations

The quantitative approach to value elicitation explored above has its benefits: It allows consulting large and representative samples of stakeholders and integrates well with existing crowdwork infrastructures. At the same time, it needs to be complemented by qualitative, small-n investigations like interviews or focus groups for a comprehensive understanding of value differences across social groups. For example, the current study did not explore how groups understand or interpret values differently, what other values some groups might have wanted to include, or why it is that e.g. women, on average, rated responsible AI values as more important.

The results of this survey also should be interpreted with care. No normative "ought" can be derived from a descriptive "is" (Musschenga, 2005). We cannot conclude that safety ought to be prioritized over autonomy from the observation that the respondents in our samples suggested so. Our results aim to increase the context sensitivity of responsible AI decisions, not to prescribe a specific course of action. Empirical ethical research does not replace ethical reasoning but offers perspectives and critical reflections.

Finally, knowledge-dependent tensions arise when contrasting the perspectives of experts and laypeople. One may argue that non-expert perspectives lack the technical and organizational insight required to evaluate AI systems. However, as

we are focusing on ethical rather than technical questions, non-experts have their own valid and legitimate forms of knowledge (Harding, 1992) that experts might not be aware of.

Extended materials

Introduction and task

Artificial intelligence (AI) is a set of emerging technologies concerned with building smart systems or machines capable of performing tasks that typically require human intelligence. Besides technical challenges, building AI systems involves complex decision-making on what the system should or should not do. In this survey, we will ask you to assess the importance of ethical principles for four AI systems.

Value Description and Question Framing

RAI value	**Description**
Transparency	A transparent AI system produces decisions that people can understand. Developers of transparent AI systems ensure, as far as possible, that users can get insight into why and how a system made a decision or inference. How important is it that the system is transparent?

Fairness	A fair AI system treats all people equally. Developers of fair AI systems ensure, as far as possible, that the system does not reinforce biases or stereotypes. A fair system works equally well for everyone independent of their race, gender, sexual orientation, and ability. How important is it that the system is fair?
Safety	A safe AI system performs reliably and safely. Developers of safe AI systems implement strong safety measures. They anticipate and mitigate, as far as possible, physical, emotional, and psychological harms that the system might cause. How important is it that the system is safe?
Accountability	An accountable AI system has clear attributions of responsibilities and liability. Developers and operators of accountable AI systems are, as far as possible, held responsible for their impacts. An accountable system also implements mechanisms for appeal and recourse. How important is it that the system is accountable?
Privacy	An AI system that respects people's privacy implements strong privacy safeguards. Developers of privacy-preserving AI systems minimize, as far as possible, the collection of sensitive data and ensure that the AI system provides notice and asks for consent. How important is it that the system respects people's privacy?

Autonomy	An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives. How important is it that the system respects people's autonomy?
Performance	A high-performing AI system consistently produces good predictions, inferences or answers. Developers of high-performing AI systems ensure, as far as possible, that the system's results are useful, accurate and produced with minimal delay. How important is it that the system performs well?

Value conflict framing

Value pair	**Description**
Fairness vs. performance	The developers realize that making the system treat all people equally (ensuring fairness) may make the system's predictions less accurate (reducing performance). Should they prioritize fairness or performance?
Fairness vs. performance (reverse)	The developers realize that making the system's predictions possibly accurate (ensuring performance) may mean that the system cannot treat all people equally (reducing fairness). Should they prioritize performance or fairness?

Fairness vs. privacy	The developers realize that making the system treat all people equally (ensuring fairness) may require the collection of additional sensitive data (reducing privacy). Should they prioritize fairness or privacy?
Fairness vs. privacy (reverse)	The developers realize that minimizing the collection of sensitive data (ensuring privacy) may mean that the system cannot treat all people equally (reducing fairness). Should they prioritize privacy or fairness?
Privacy vs. performance	The developers realize that minimizing the collection of sensitive data (ensuring privacy) may make the system's predictions less accurate (reducing performance). Should they prioritize privacy or performance?
Privacy vs. performance (reverse)	The developers realize that making the system's predictions possibly accurate (ensuring performance) may require the collection of additional sensitive data (reducing privacy). Should they prioritize performance or privacy?
Safety vs. autonomy	The developers realize that mitigating risks and potential harms (ensuring safety) may require limiting people's choices and control (reducing autonomy). Should they prioritize safety or people's autonomy?
Safety vs. autonomy (reverse)	The developers realize that giving people choices and control (ensuring autonomy) may introduce additional risks and potential harms (reducing safety). Should they prioritize people's autonomy or safety?

Safety vs. transparency	The developers realize that mitigating risks and potential harms (ensuring safety) may require to keep the system's decision process opaque (reducing transparency). Should they prioritize safety or transparency?
Safety vs. transparency (reverse)	The developers realize that revealing the system's decision process (ensuring transparency) may introduce additional risks and potential harms (reducing safety). Should they prioritize transparency or safety?

Application scenario framing

Scenario	**Description**
Banking	A bank uses an AI system that scans loan applicants' data to predict whether they are likely to repay a loan. Thousands of loan applications are automatically rejected based on the output of this AI system.
Medical	A medical clinic uses an AI system that scans patients' medical records to predict whether a patient has a particular disease. Thousands of patients' treatment plans are automatically adjusted based on the output of this AI system.
Marketing	A marketing company uses an AI system that scans the data of web users to predict which advertisements they will respond to. Thousands of advertisements are automatically shown to users based on the output of this AI system.

Streaming	A video streaming company uses an AI system that scans users' data to predict which other movies they would enjoy seeing. A list of recommended movies is automatically shown to thousands of users based on the output of this AI system.

Detailed result graphs

Please refer to Figures 5.7 and 5.8.

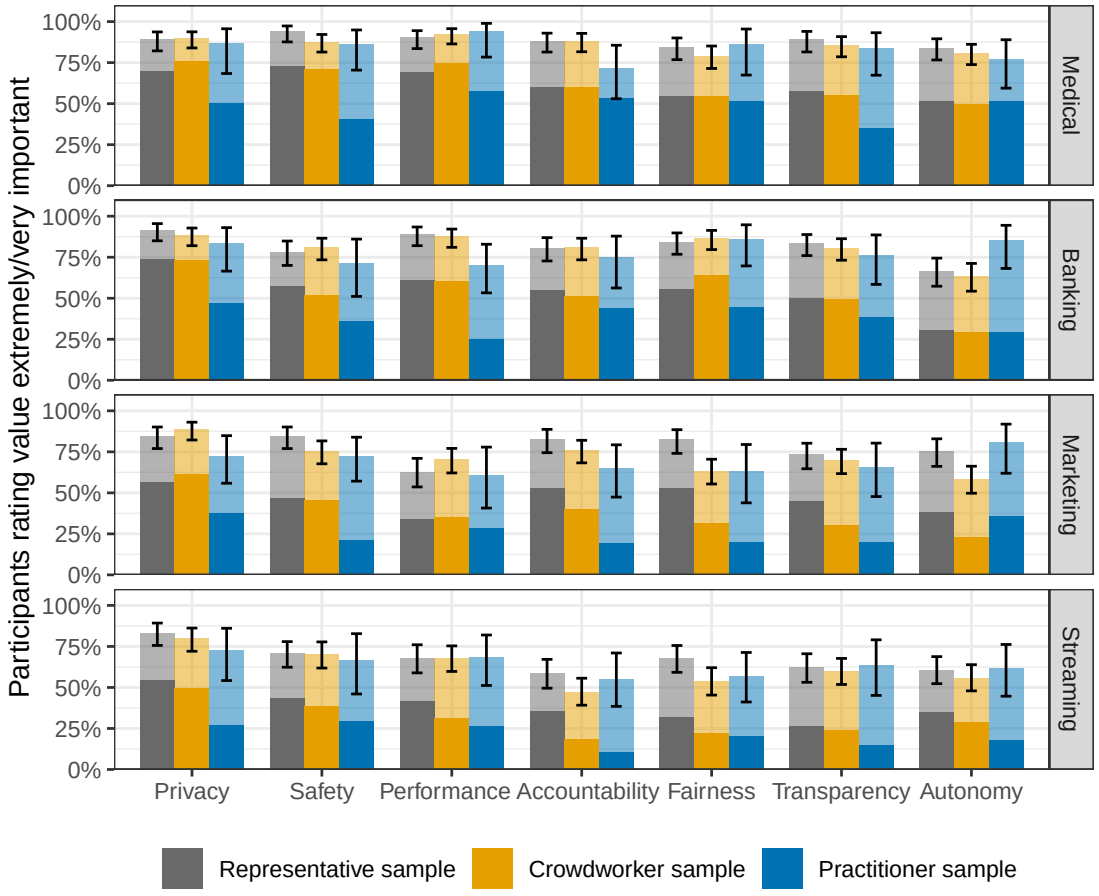

Figure 5.7: **The perceived importance of values across deployment scenarios.** N=28 to 171 ratings per bar. The x-axis shows the assessed responsible AI values and the y-axis indicates how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark).

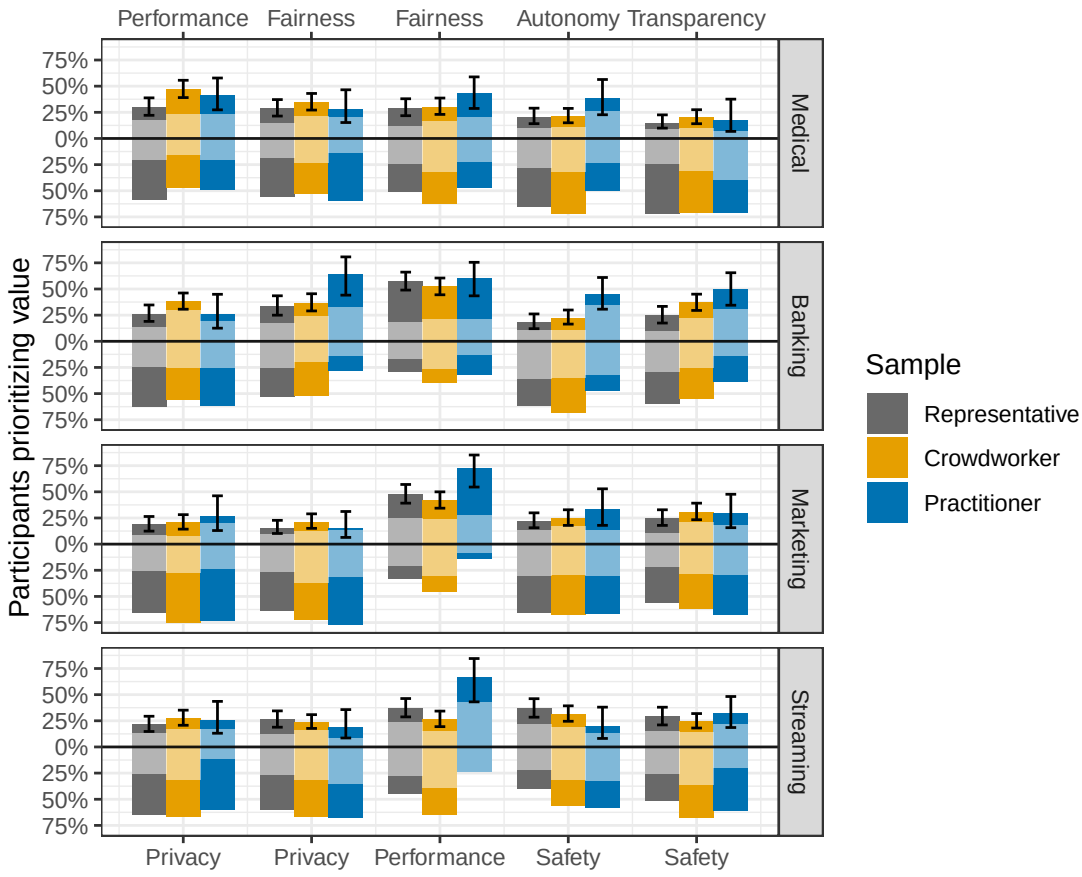

Figure 5.8: **How values are prioritized in different deployment settings.** N = 28 to 173 ratings per bar. The conflicting value pairs are shown on the top and bottom, e.g., privacy vs. performance on left. Respondents prioritizing the top value are shown to the top and responses prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, whereas weak preferences are lightly shaded. Undecided respondents are omitted.

Chapter 6

Discussion: The Coming Age of AI-Mediated Communication

The results presented in this thesis underline a broader point that we make in this discussion chapter: using AI technologies like large language models in communication is not just technical innovation. It constitutes a paradigm shift from previous forms of Computer-Mediated Communication with distinct benefits and risks that require more careful assessment. We introduce the concept of AI-Mediated Communication to theorize how using large language models in communication is distinct from previous forms of Computer-Mediated Communication. We draw on the empirical findings of this dissertation and combine them with results by other researchers to conceptualize how AI-Mediated Communication affects people's agency and ability to trust in mediated communication.

Defining AI-Mediated Communication (AI-MC)

In one of the studies conducted for this dissertation (Jakesch et al., 2019), we developed the concept of AI-Mediated Communication (AI-MC) to theorize how embedding AI technologies like large language models differs from previous forms of Computer-Mediated Communication (CMC). We defined AI-Mediated Commu-

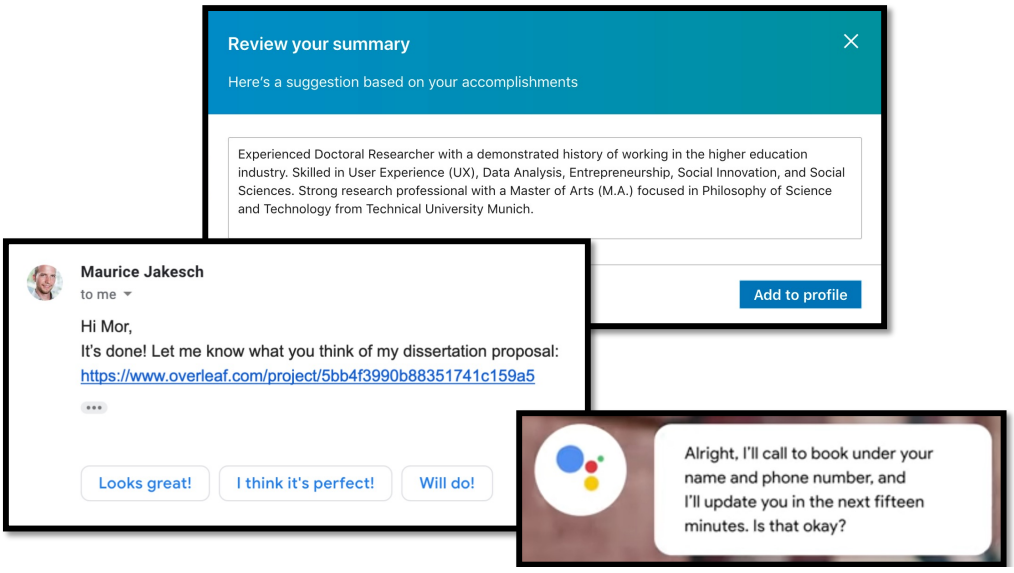

Figure 6.1: Examples of widely used AI-MC applications, from top to bottom: Professional social networks suggest automatically generated profile summaries to their users. Smart reply applications offer AI-generated messages that can be sent with a simple mouse click based on the context of a previous conversation. Language assistants make calls and send messages on behalf of their owners under their name and phone number.

nication as interpersonal communication not simply transmitted by technology but *modified*, *augmented*, or even *generated* by a computational agent to achieve specific communicative or relational outcomes (Jakesch et al., 2019; Hancock et al., 2020). In AI-Mediated Communication, an AI system operates on behalf of the communicating person by augmenting, generating, or suggesting content. AI-MC is distinct from traditional CMC technologies that primarily transmit messages. It also differs from autonomous machine agents that may produce language but do not represent a person. For example, when a chatbot or robot speaks on its own behalf, the AI system does not mediate between people. And while traditional e-mail or chat applications mediate between people, they primarily transmit users' messages without actively participating in the production of content. "Smart replies", i.e.,

email responses automatically suggested based on the context of a conversation, are AI-MC, as the technology generates parts of the user's communication on the user's behalf.

In our definition, we use the term *AI* broadly to refer to computational systems that employ machine learning, natural language processing, and related techniques to alter the content of people's communication. The emerging field of AI-MC (Hancock et al., 2020) presents a significant new research agenda with an impact on core CMC and HCI topics, from communication practices to relationship and interpersonal dynamics (Thurlow et al., 2004, p. 22). AI-MC affects interactions from one-to-one exchanges such as messaging to one-to-many broadcasts like writing user profiles or appearing in a live YouTube video. In text-based communication—the focus of this work—we have already advanced from spell check and predictive auto-completion to early AI-MC instances, like the aforementioned auto-responses for chats and e-mails (Hohenstein and Jung, 2018).

In some AI-MC applications, language technologies enhance people's writing style (Grammarly, 2017), write human-like texts (Yao et al., 2017), and produce online self-presentations (Blogs, 2017). Figure 6.1 shows three examples of AI-MC technologies that are used by millions of users. Researchers have estimated, for example, that more than 36 billion generated messages are sent daily (Mieczkowski et al., 2021b) through the Google's *smart reply* feature (Kannan et al., 2016) shown on top of the figure. Companies are also trying to commercialize the language generation capabilities of GPT-3. These businesses offer content creators to multiply their writing productivity (HyperWrite, 2022) and marketing professionals to automatically create higher converting ads (Copy.AI, 2022; CopySmith, 2022).

While businesses are moving fast to develop products that monetize the benefits

of AI-Mediated Communication, we have a very limited understanding of how integrating AI technologies into communication will affect societies. This dissertation has empirically shown that humans cannot detect language produced by GPT-3, that using large language models in self-presentation may damage interpersonal trust, and that interactions with opinionated models change users' attitudes. We now extend our discussion to the wider societal risks of embedding AI technologies in human communication. Drawing on our own empirical work as well as empirical and theoretical contributions by other researchers, we consider how AI-Mediated Communication raises difficult questions of trust, transparency, agency and manipulation.

Trust and transparency in AI-Mediated Communication

AI-MC technologies challenge assumptions of agency and mediation (Hancock et al., 2020) in ways that potentially subvert existing social heuristics (Ellison et al., 2012; Walther, 2011; Herring, 2002). As Chapter 2 has shown, people cannot tell anymore whether the communication they are receiving is human-authored, machine-generated, or co-created. Similarly, studies by others have suggested that people cannot identify news content (Clark et al., 2021; Ippolito et al., 2019; Kreps et al., 2022a) generated by large language models.

From impersonation (Weidinger et al., 2022) to targeted disinformation campaigns (Zellers et al., 2019), people's inability to identify generated language exacerbates concerns about novel automatized forms of deception, fraud, and identity theft (Biderman and Raff, 2022; Bommasani et al., 2021; Cooke, 2018; Buchanan et al., 2021). Yet, even without adversarial use, AI-Mediated Communication

poses wider systemic risks. When large language models are used to generate communication, many cues and heuristics people rely on lose their diagnosticity. For example, as software filters for smartphone cameras proliferate, it becomes increasingly difficult to draw inferences from photos. Similarly, writing an elaborate and articulate message may cease to be indicative of intention or expertise when large language technologies can easily produce such messages.

As we have shown in Chapter 2, AI systems could even exploit people's cognitive heuristics to create language that is perceived in certain ways. AI-MC systems may optimize messages to make senders appear trustworthy (Ma et al., 2017a), attractive (Leyvand et al., 2008), or signal high social status (Pavlick and Tetreault, 2016). Deep fakes, where AI technologies are used to create realistic misrepresentations of a person in audio and video (Suwajanakorn et al., 2017; Thies et al., 2016), may even undermine the basic heuristic that what one sees and hears did actually happen.

People may react to AI-induced uncertainty in their communication in different ways. Previous research has shown that how individuals react will depend on the context and their perception of the AI system. For example, people who see AI systems as more objective (Sundar, 2008) or knowledgeable (Bruzzese et al., 2020) may be more likely to trust them. The research in this dissertation, however, has shown that people will often react to the additional uncertainty in AI-Mediated Communication by mistrusting those they suspect are using it (Jakesch et al., 2019). The finding that AI-Mediated Communication decreases interpersonal trust has since been replicated in multiple studies and contexts (Wu and Kelly, 2020; Liu et al., 2022; Hohenstein et al., 2021). Related research by us and others also suggests that the use of large language models in political communication can

reduce trust in legislators (Kreps et al., 2022b) and that AI-MC can negatively influence how job seekers are perceived (Weiss et al., 2022).

People's reactions to the introduction of AI-Mediated Communication are further complicated as they cannot reliably tell who uses it, as shown in Chapter 2. People may change how they scrutinize or evaluate a message when they suspect it to be co-written by a machine (Hancock et al., 2004). However, they will often suspect the wrong people to be using AI-MC as they mistake human-written communication for generated language and vice versa (Jakesch et al., 2022c). Some have gone so far as to suggest that increasingly active yet intransparent use of AI in communication will lead to an *authenticity crisis* (Lee, 2020): When people cannot tell who they are talking to and what inferences they can draw anymore, they may refrain from trusting mediated communication altogether.

First regulatory attempts have tried to reduce these risks. A recent blueprint for an AI Bill of Rights from the U.S. White House calls for "Notice and Explanation" when "an automated system is being used" (Nelson et al., 2022b). Similarly, a 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" (Commission, 2021). Disclosure could range from explicit warning labels to implicit AI accents that we have introduced in Chapter 2. In many cases, such as when using AI auto-correction assistants, it does not seem plausible that all AI-MC content must be marked as such (Williams, 2018). Here, we will need ways to restore human judgment without hindering the flow of communication. Further research is required to develop context-aware and effective regulations and mechanisms for disclosures that preserve people's trust in

AI-Mediated Communication.

Agency and manipulation in AI-Mediated Communication

When people write, speak or present themselves through AI-Mediated Communication, their own process of idea generation, translation, and text production (Hayes, 2012) becomes entangled with the technology. Research by our collaborators has suggested that people view AI-MC systems as social actors (Mieczkowski and Hancock, 2022); that is, AI-MC systems are perceived as social entities with their own agency rather than as mere tools or extensions of the writer's intent (Endacott and Leonardi, 2022). Accordingly, using AI-MC technologies reduces people's sense of ownership (Mieczkowski et al., 2021a). While people strive to maintain their agency when co-writing with a language model (Mieczkowski and Hancock, 2022), they readily delegate their communicative agency to the AI system in contexts where they lack expertise (Mieczkowski and Hancock, 2022). Due to the widespread perception that AI systems are more knowledgeable and objective than humans (Sundar, 2008; Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010; Wickens et al., 2015), people's willingness to delegate communicative agency to AI systems may be indicative of over-reliance.

In many cases, delegating agency to AI systems will be the goal of AI-Mediated Communication. When people ask their language assistants to schedule an appointment on their behalf, their goal is the convenience of delegation. In some contexts, people may even want the AI technology to guide or override their own behavior, e.g., by encouraging good communication practice (Tsai et al., 2022) or making sure important topics are not neglected in a conversation (Furlo et al.,

2021).

In other settings, however, the reduction of people's agency in communication raises difficult ethical questions. For example, studies have shown that people assign less responsibility and blame for mishaps to those who use AI in their communication (Hohenstein and Jung, 2020). Here, the blurring of agency in AI-Mediated Communication leads to an erosion of accountability norms. Further, multiple studies have suggested that delegating communicative agency to AI-MC systems also entails a delegation of personal emotions to the system. AI language technologies may artificially amplify or reduce certain emotions in a conversation (Mills et al., 2021), affecting both the senders' as well as receivers' emotional states and judgment (Arnold et al., 2018). Hohenstein and Jung (2018) have shown that widely used AI-MC applications produce language that is overly positive and change how people interact with another (Hohenstein et al., 2021).

A delegation of communicative agency to language technologies is particularly problematic with regard to opinion formation and democratic discourse. AI-MC systems that change what people say and believe may become manipulative if they act covertly and exploit vulnerabilities in people's decision-making (Susser et al., 2019, also compare Chapter 2). As we have shown in Chapter 4, interacting with large language models changes users' views – both in writing as well as in later attitude surveys. Whether this influence is due to *informational* (Myers, 2008) or *normative influence* (Myers, 2008), or simply because the AI technology disrupts and changes users' through processes (Bhat et al., 2021) is not entirely clear. But with the advancement of AI-MC deployments, their impact on public opinion will interact with democratic processes of collective deliberation and opinion formation. In one scenario, the models will simply reinforce the dominant opinions

(Caliskan et al., 2017; Blodgett et al., 2020) found in their training data (Jakesch et al., 2022a). In other scenarios, language models may be used by commercial or political actors to amplify opinions of their choice (Jakesch et al., 2021; Schlessinger et al., 2021). Such influence campaigns could be malignant (Bagdasaryan and Shmatikov, 2021), but they do not have to be: Like search engine and social media network operators (Knoll, 2016), operators of AI-MC applications may monetize the persuasive power of their technology through new forms of implicit advertising.

Chapter 7

Conclusion: Managing the Risk of AI-Mediated Communication

How should we, as researchers and practitioners, respond to the risks of AI-MC technologies? Eminent figures in academia, government, and industry argue for vastly different positions. Some hold AI developments to be a singular opportunity that we need to pursue to its fullest extent (Pfotenhauer et al., 2019), and others warn of the same developments as an existential threat to humanity (Ord, 2020; Bostrom, 2013). In this concluding chapter, I will discuss how the empirical work presented in this dissertation highlights the need for more careful management of the risks of AI technologies. I argue that our results show that certain serious systemic risks of AI-MC are plausible or even likely and that reducing the uncertainty concerning these risks is feasible ahead of deployment.

The first contribution of this thesis is to demonstrate that embedding AI technologies in human communication can have harmful consequences. An authenticity crisis, as discussed in Chapter 3, where people distrust each other and mediated communication more generally, would be a significant and hardly reversible collateral damage. Far from minor technology side-effects, the accidental or even intentional shifts of public opinion through widespread deployment of opinionated language models investigated in Chapter 4 present a threat to democratic societies.

And the possibility of being surrounded by technology that exploits our intuition (compare Chapter 2) presents a loss of human agency reminiscent of dystopian science-fiction. The work presented in this thesis has shown that to some extent, these scenarios are plausible or even likely.

The second contribution of this thesis is to show that we can reduce uncertainty about the likelihood of AI-MC's risks. While assessing the effects and risks of AI technologies like large language models before deployment involves an amount of uncertainty, estimates of the likelihood and severity of damages can be improved. The methodological combination of user experiments and technology prototypes explored generates grounded insights into how future deployments of large language models may affect people's ability to make judgments, form opinions, and their willingness to trust each other. Even if the details of AI-MC deployments remain open, our experiments tell us something about the effects of similar deployments.

Our results also demonstrate that third parties can perform risk assessments in environments where private actors do not disclose the technologies they develop. With full access to the technology and usage data, companies developing the models will be able to produce more comprehensive and accurate assessments than we did. Since digital technologies allow for relatively easy data collection and many companies already are using experiments to optimize their products (Matias, 2017), setting up risk assessment experiments would be comparatively cheap – in particular when compared to the costs of technology development and the severity of potential fallouts.

In the current regulatory environment, the risks of AI are managed primarily by the individual business entities that develop them (Chae, 2020). The studies presented in this dissertation suggest that intuition-driven ad-hoc management

of risks through self-regulating private actors will be error-prone and potentially unfair. Robust risk assessments require careful planning, execution, and analysis. The research questions in this dissertation could not have been answered reliably through intuition and ad-hoc judgement, and product teams that do not perform systematic empirical risk assessments will often misjudge both the likelihood and severity of risks (Jakesch et al., 2022b).

Beyond error in judgment, risk assessments by self-regulating private actors suffer from structural biases: Developers reap benefits from the development of AI-MC systems that will not be available to a wider public. They may be less affected by some risks and may have fewer incentives to mitigate them. Even if practitioners carefully evaluate AI-MC technologies without giving particular weight to their interests, they will still need to understand what risks the wider population is willing to take. As the study presented in Chapter 5 have shown, AI practitioners have different priorities for the values and risks involved in developing AI technologies than the general public (Jakesch et al., 2022b) and their value and risk judgements may not be a good proxy for the valuations of the more general population.

Through surveys and crowd-sourcing tools, such as the one presented in Chapter 5, product teams could understand what a broader population values and fears in AI-MC technologies. But only a more inclusive approach to risk management can ensure that groups that are traditionally disadvantaged by one-size-fits-all technologies (Goldenthal et al., 2021) will not also be disadvantaged by the very process of AI-MC risk management. To give AI development decisions the democratic legitimization that they may need–given the extent of potential losses–a democratic process and institutions for AI risks assessments will be necessary.

This dissertation has illustrated the need to assess the risks of AI-MC technologies in more systematic, inclusive, and empirically grounded ways. The risks involved in developing AI-MC technologies are substantial, ranging from an authenticity crisis in communication to unseen scales of opinion manipulation and agency loss. We have shown that through creative experimentation and user research, one can reduce uncertainties about these risks before the technology is widely deployed. We have also provided a conceptual framework for the paradigm shifts observed in AI-Mediated Communication. We grounded our evaluations and argument in robust empirical work and have advanced a set of methods and tools to support further assessments of the risks of AI-MC technologies. We hope that this work provides impulses for future studies and the development of a broader strategy for managing the societal risk of AI-Mediated Communication.

Chapter 8

Bibliography

Acquisti, A., Brandimarte, L., and Hancock, J. (2022). How privacy's past may shape its future. *Science*, 375(6578):270–272. 2

Agre, P. and Agre, P. E. (1997). *Computation and human experience*. Cambridge University Press. 5, 5

Aral, S. and Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, 365(6456):858–861. 4

Araujo, T., Helberger, N., Kruikemeier, S., and De Vreese, C. H. (2020). In AI we trust? perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3):611–623. 4, 4

Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2018). Sentiment bias in predictive text recommendations results in biased writing. In *Proceedings of the 44th Graphics Interface Conference*, GI '18, page 42–49, Waterloo, CAN. Canadian Human-Computer Communications Society. 4, 4, 6

Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2020). Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 128–138, New York, NY, USA. Association for Computing Machinery. 4

Arnold, K. C., Gajos, K. Z., and Kalai, A. T. (2016). On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 603–608, Tokyo Japan. ACM. 4

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, 58:295–303. 4

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*. 1, 4, 4

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729):59–64. 5, 5, 5

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488. 5

Bagdasaryan, E. and Shmatikov, V. (2021). Spinning sequence-to-sequence models with meta-backdoors. *arXiv preprint arXiv:2107.10443*. 4, 6

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221. 4

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. 4

Banovic, N., Sethapakdi, T., Hari, Y., Dey, A. K., and Mankoff, J. (2019). The limits of expert text entry speed on mobile keyboards with autocorrect. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA. Association for Computing Machinery. 4

Barocas, S. and Boyd, D. (2017). Engaging the ethics of data science in practice. *Communications of the ACM*, 60(11):23–25. 5

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, 6:1–62. 4

Bender, E. M. (2019). A typology of ethical risks in language technology with an eye towards where transparent documentation can help. In *Future of Artificial Intelligence: Language, Ethics, Technology Workshop*. 1

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. 4

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., and Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1):1–25. 2

Berkovsky, S., Freyne, J., and Oinas-Kukkonen, H. (2012). Influencing individually: fusing personalization and persuasion. 4

Bhat, A., Agashe, S., and Joshi, A. (2021). How do people interact with biased text prediction models while writing? In *Proceedings of the First Workshop on*

Bridging Human–Computer Interaction and Natural Language Processing, pages 116–121, Online. Association for Computational Linguistics. 4, 6

Bhat, A., Agashe, S., Mohile, N., Oberoi, P., Jangir, R., and Joshi, A. (2022). Studying writer-suggestion interaction: A qualitative study to understand writer interaction with aligned/misaligned next-phrase suggestion. Publisher: arXiv. 4, 4

Bi, X., Ouyang, T., and Zhai, S. (2014). Both complete and correct? multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2297–2306, New York, NY, USA. Association for Computing Machinery. 4

Biderman, S. and Raff, E. (2022). Neural language models are effective plagiarists. *arXiv preprint arXiv:2201.07406*. 1, 2, 2, 6

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR. 5

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2021). The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*. 5, 5, 5

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of bias in nlp. *arXiv preprint arXiv:2005.14050*. 1, 5, 6

Blogs, M. (2017). Bringing AI to job seekers with resume assistant in word, powered by linkedin. `https://bit.ly/2Di34QB`. 6

Bogert, E., Schecter, A., and Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):1–9. 4

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. 1, 1, 2, 2, 4, 4, 4, 6

Bond Jr, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234. 2

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, 4(1):15–31. 7

Boyarskaya, M., Olteanu, A., and Crawford, K. (2020). Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint arXiv:2011.13416*. 5, 5

Bradshaw, S. and Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. 4

Braithwaite, V. (1998). The value orientations underlying liberalism-conservatism. *Personality and individual differences*, 25(3):575–589. 5

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901. 1, 1, 2, 2, 4, 4, 4, 4

Bruns, A. (2019). *Are filter bubbles real?* John Wiley & Sons. 4

Bruzzese, T., Gao, I., Dietz, G., Ding, C., and Romanos, A. (2020). Effect of confidence indicators on trust in AI-generated profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8. 6

Buchanan, B., Lohn, A., Musser, M., and Sedova, K. (2021). Truth, lies, and automation. *Center for Security and Emerging Technology*. 1, 1, 2, 2, 4, 6

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5. 3

Buschek, D., Bisinger, B., and Alt, F. (2018). ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA. ACM. event-place: Montreal, Quebec, CA. 4

Buschek, D., Zürn, M., and Eiband, M. (2021). The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery. 4

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. 1, 6

Cappelli, C., Cunha, H., Gonzalez-Baixauli, B., and do Prado Leite, J. C. S. (2010). Transparency versus security: early analysis of antagonistic requirements. In

Proceedings of the 2010 ACM symposium on applied computing, pages 298–305. 5

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., and Taylor, L. (2018). Portrayals and perceptions of AI and why they matter. 5

Chae, Y. (2020). Us AI regulation guide: legislative overview and practical considerations. *The Journal of Robotics, Artificial Intelligence & Law*, 3. 7

Chen, M. X., Lee, B. N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A. M., Chen, Z., Sohn, T., and Wu, Y. (2019). Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2287–2295, New York, NY, USA. Association for Computing Machinery. 4, 4

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379. 4

Christakis, N. A. and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258. 4

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118. 4

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*. 1, 2, 2, 2, 2, 6

Clark, E., Ross, A. S., Tan, C., Ji, Y., and Smith, N. A. (2018). Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd Inter-*

national Conference on Intelligent User Interfaces, IUI '18, page 329–340, New York, NY, USA. Association for Computing Machinery. 2, 2, 4

Commission, E. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence. 2, 6

Consult, M. (2016). National tracking poll #2110047. 4

Cooke, N. A. (2018). *Fake news and alternative facts: Information literacy in a post-truth era*. American Library Association. 1, 2, 2, 6

Coppock, A. and McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, 6(1):2053168018822174. 2

Copy.AI (2022). Copy.ai homepage. 6

CopySmith (2022). Copysmith homepage. 6

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806. 5

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

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing? how recommender system interfaces affect users' opinions. In

Proceedings of the SIGCHI conference on Human factors in computing systems, pages 585–592. 4

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial intelligence*, 169(2):104–141. 2

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*, 25(06). 5, 5, 5

Cui, W., Zhu, S., Zhang, M. R., Schwartz, H. A., Wobbrock, J. O., and Bi, X. (2020). *JustCorrect: Intelligent Post Hoc Text Correction Techniques on Smartphones*, page 487–499. Association for Computing Machinery, New York, NY, USA. 4

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies — why and how. *Knowledge-Based Systems*, 6(4):258–266. 3

Dalvi, G., Ahire, S., Emmadi, N., Joshi, M., Joshi, A., Ghosh, S., Ghone, P., and Parmar, N. (2016). Does prediction really help in marathi text input? empirical analysis of a longitudinal study. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '16, page 35–46, New York, NY, USA. Association for Computing Machinery. 4

Dang, H., Benharrak, K., Lehmann, F., and Buschek, D. (2022). Beyond text generation: Supporting writers with continuous automatic text summaries. *arXiv preprint arXiv:2208.09323*. 1

Davis, Z. and Steinbock, A. (2021). Max Scheler. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition. 5, 5, 5

De Angeli, A., Johnson, G. I., and Coventry, L. (2001). The unfriendly user: exploring social reactions to chatterbots. In *Proceedings of The International Conference on Affective Human Factors Design, London*, pages 467–474. 3, 3

DeAndrea, D. C. (2014). Advancing warranting theory. *Communication Theory*, 24(2):186–204. 3

Deloitte (2020). Bringing transparency and ethics into AI? 5

DeVito, M. A., Birnholtz, J., and Hancock, J. T. (2017). Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 740–754. 2, 3, 3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 1

Dillon Jr, E. C., Gilbert, J. E., Jackson, J. F., and Charleston, L. (2015). The state of african americans in computer science-the need to increase representation. *Computing Research News*, 21(8):2–6. 5

Dilmegani, C. (2018). 100 AI use cases and applications. 5

Dobbe, R., Dean, S., Gilbert, T., and Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553*. 5

Donath, J. (2007). Signals in social supernets. *Journal of Computer-Mediated Communication*, 13(1):231–251. 3

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., and Choi, Y. (2022). 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)*, pages 7250–7274. 2

Duerr, S. and Gloor, P. A. (2021). Persuasive natural language generation–a literature review. *arXiv preprint arXiv:2101.05786*. 4

Dunlop, M. and Levine, J. (2012). Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2669–2678, New York, NY, USA. Association for Computing Machinery. 4

Dunn, M., Sheehan, M., Hope, T., and Parker, M. (2012). Toward methodological innovation in empirical ethics research. *Cambridge Quarterly of Healthcare Ethics*, 21(4):466–480. 5

Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech communication*, 50(8-9):630–645. 3

Edwards, C., Edwards, A., Spence, P. R., and Shelton, A. K. (2014). Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33:372–376. 3

Ekstrand, M. D., Joshaghani, R., and Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47. PMLR. 5

Ellison, N., Heino, R., and Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication*, 11(2):415–441. 2, 3

Ellison, N. B. and Boyd, D. M. (2013). Sociality through social network sites. In *The Oxford handbook of internet studies*. 3, 3

Ellison, N. B. and Hancock, J. T. (2013). Profile as promise: Honest and deceptive signals in online dating. *IEEE Security and Privacy*, 11(5):84–88. 3

Ellison, N. B., Hancock, J. T., and Toma, C. L. (2012). Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *New Media & Society*, 14(1):45–62. 3, 6

Endacott, C. G. and Leonardi, P. M. (2022). Artificial intelligence and impression management: Consequences of autonomous conversational agents communicating on one's behalf. *Human Communication Research*. 6

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism management*, 55:62–73. 1, 2, 3, 3

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. 5

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., and Saunders, W. (2021). Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*. 1

Ferrara, E., Varol, O., Davis, C. A., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of The ACM*, 59(7):96–104. 3, 3, 4

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American psychologist*, 46(8):835. 5, 5, 5, 5

Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694. 1, 2, 2

Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):2. 4

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30. 1

Fowler, A., Partridge, K., Chelba, C., Bi, X., Ouyang, T., and Zhai, S. (2015). Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 649–658, New York, NY, USA. Association for Computing Machinery. 4

Friedman, B. (1996). Value-sensitive design. *interactions*, 3(6):16–23. 5, 5, 5

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347. 5

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., Lucchiari, C., Consonni, D., Nordio, F., Pravettoni, G., et al. (2010). Gender-related differences in moral judgments. *Cognitive processing*, 11(3):219–226. 5

Furlo, N., Gleason, J., Feun, K., and Zytko, D. (2021). Rethinking dating apps as sexual consent apps: A new use case for AI-Mediated Communication. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 53–56. 6

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A. (2021). A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184. 1

Gaspari, F., Toral, A., Naskar, S. K., Groves, D., and Way, A. (2014). Perception vs. reality: measuring machine translation post-editing productivity. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 60–72. 1

Gehrmann, S., Strobelt, H., and Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*. 2

Genus, A. and Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research policy*, 47(1):61–69. 1

Gero, K. and Chilton, L. B. (2019). Metaphoria: An algorithmic companion for metaphor creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 4

Gibbs, J. L., Ellison, N. B., and Heino, R. D. (2006). Self-presentation in online personals the role of anticipated future interaction, self-disclosure, and perceived success in internet dating. *Communication Research*, 33(2):152–177. 3

Gibbs, J. L., Ellison, N. B., and Lai, C.-H. (2011). First comes love, then comes google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research*, 38(1):70–100. 3

Gillespie, T. (2014). The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167. 4

Gillespie, T., Boczkowski, P. J., and Foot, K. A. (2014). *Media technologies: Essays on communication, materiality, and society*. MIT Press. 3

Goel, S., Watts, D. J., and Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. 4

Goethe (1808). *Faust, part I.* Penguin Classics, London, England. 1

Goldenthal, E., Park, J., Liu, S. X., Mieczkowski, H., and Hancock, J. T. (2021). Not all AI are equal: Exploring the accessibility of AI-Mediated Communication technology. *Computers in Human Behavior*, 125:106975. 7

Google (2020). Our principles. 5

Gordon, M., Ouyang, T., and Zhai, S. (2016). Watchwriter: Tap and gesture typing on a smartwatch miniature keyboard with statistical decoding. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3817–3821, New York, NY, USA. Association for Computing Machinery. 4

Graefe, A., Haim, M., Haarmann, B., and Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, 19(5):595–610. 3

Graham, J., Meindl, P., Beall, E., Johnson, K. M., and Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130. 5

Grammarly (2017). Free grammar checker - grammarly. `https://www.grammarly.com/`. 6

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, pages 903–912. 5, 5

Guillory, J. and Hancock, J. T. (2012). The effect of linkedin on deception in resumes. *Cyberpsychology, Behavior, and Social Networking*, 15(3):135–140. 2, 3, 3

Gunaratne, J., Zalmanson, L., and Nov, O. (2018). The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35(4):1092–1120. 4

Guttentag, D. (2015). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18(12):1192–1217. 3, 3

Hair, J. F. (2009). Multivariate data analysis. 5

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2007a). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23. 3

Hancock, J. T., Naaman, M., and Levy, K. (2020). AI-Mediated Communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100. 1, 2, 6, 6

Hancock, J. T., Thom-Santelli, J., and Ritchie, T. (2004). Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 129–134, New York, NY, USA. ACM. 3, 6

Hancock, J. T., Toma, C., and Ellison, N. (2007b). The truth about lying in online dating profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 449–452, New York, NY, USA. ACM. 3

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation. 3

Harding, S. (1992). Rethinking standpoint epistemology: What is" strong objectivity?". *The Centennial Review*, 36(3):437–470. 5

Hartwig, M. and Bond Jr, C. F. (2011). Why do lie-catchers fail? a lens model meta-analysis of human lie judgments. *Psychological bulletin*, 137(4):643. 2

Hashimoto, T. B., Zhang, H., and Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*. 2

Hayes, J. R. (2012). Modeling and remodeling writing. *Written communication*, 29(3):369–388. 6

Herring, S. C. (2002). Computer-mediated communication on the internet. *Annual Review of Information Science and Technology*, 36(1):109–168. 3, 6

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., and Martin, N. (2021). *How humans judge machines*. MIT Press. 5, 5

Hohenstein, J., DiFranzo, D., Kizilcec, R. F., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., and Jung, M. (2021). Artificial intelligence in communication impacts language and social relationships. *arXiv:2102.05756 [cs]*. arXiv: 2102.05756. 6, 6

Hohenstein, J. and Jung, M. (2018). Ai-supported messaging: An investigation of human-human text conversation with AI support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page LBW089. ACM. 3, 3, 6, 6

Hohenstein, J. and Jung, M. (2020). Ai as a moral crumple zone: The effects of AI-Mediated Communication on attribution and trust. *Computers in Human Behavior*, 106:106190. 4, 6

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*. 2

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425. 3

House, W. (2016). Preparing for the future of artificial intelligence. executive office of the president national science and technology council. committee on technology. 5, 5, 5

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*. 4

Hua, Y., Namavari, A., Cheng, K., Naaman, M., Ristenpart, T., and Tech, C. (2021). Increasing adversarial uncertainty to scale private similarity testing. *arXiv preprint arXiv:2109.01727*. 5

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., and Kohli, P. (2019). Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*. 1, 4, 4, 4

Huff, C. and Tingley, D. (2015). "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648. 5

HyperWrite (2022). Hyperwrite homepage. 6

IBM (2020). Ibm's principles for trust and transparency. 5

Intemann, K. (2010). 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia*, 25(4):778–796. 5, 5, 5, 5

Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*. 1, 2, 6

Ischen, C., Araujo, T., Voorveld, H., Noort, G. v., and Smit, E. (2019). Privacy concerns in chatbot interactions. In *International workshop on chatbot research and design*, pages 34–48. Springer. 2

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2022a). Interacting with opinionated language models influences users' views. 6

Jakesch, M., Buçinca, Z., Amershi, S., and Olteanu, A. (2022b). How different groups prioritize ethical values for responsible AI. *arXiv preprint arXiv:2205.07722*. 4, 7

Jakesch, M., French, M., Ma, X., Hancock, J. T., and Naaman, M. (2019). AI-Mediated Communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. 1, 2, 2, 6, 6

Jakesch, M., Garimella, K., Eckles, D., and Naaman, M. (2021). Trend alert: A cross-platform organization manipulated twitter trends in the indian general election. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–19. 6

Jakesch, M., Hancock, J., and Naaman, M. (2022c). Human heuristics for AI-generated language are flawed. *arXiv preprint arXiv:2206.07271*. 4, 4, 6

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399. 5, 5, 5, 5, 5

Jobin, A., Man, K., Damasio, A., Kaissis, G., Braren, R., Stoyanovich, J., Van Bavel, J. J., West, T. V., Mittelstadt, B., Eshraghian, J., et al. (2021). Ai reflections in 2020. *Nature Machine Intelligence*, 3(1):2–8. 5

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., and Bertulfo, D. J. (2022). The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*. 4, 4, 4

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan. 4

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., Young, P., et al. (2016). Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 955–964, New York, NY, USA. Association for Computing Machinery. 1, 4, 6

Kapania, S., Siy, O., Clapper, G., SP, A. M., and Sambasivan, N. (2022). " because AI is 100% right and safe": User attitudes and sources of AI authority in india. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18. 4, 4, 5

Karinshak, E., Liu, S., Park, J. S., and Hancock, J. (2022). Can AI persuade? examining a large language model's ability to generate pro-vaccination messages. *International Communication Association Annual Conference*. 4

Karpinska, M., Akoury, N., and Iyyer, M. (2021). The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*. 2, 2, 2

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14. 5

Khondker, H. H. (2011). Role of the new media in the arab spring. *Globalizations*, 8(5):675–679. 4

Kim, A. Y. and Escobedo-Land, A. (2015). Okcupid data for introductory statistics and data science courses. *Journal of Statistics Education*, 23(2). 2

Kimmerle, J., Moskaliuk, J., Bientzle, M., Thiel, A., and Cress, U. (2012). Using controversies for knowledge construction: Thinking and writing about alternative medicine. In *ICLS*. 4

Kiyonari, T., Yamagishi, T., Cook, K. S., and Cheshire, C. (2006). Does trust beget trustworthiness? trust and trustworthiness in two games and two cultures: A research note. *Social Psychology Quarterly*, 69(3):270–283. 3

Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research*, 39(2):121–123. 4, 5

Knoll, J. (2016). Advertising in social media: a review of empirical evidence. *International journal of Advertising*, 35(2):266–300. 6

Köbis, N. and Mossink, L. D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in human behavior*, 114:106553. 2, 2

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (awcf) provided by grammarly: A multiple case study. *Assessing Writing*, 44:100450. 1

Kreps, S., McCain, R. M., and Brundage, M. (2022a). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117. 1, 2, 2, 4, 4, 6

Kreps, S., Moradi, P., and Jakesch, M. (2022b). AI-Mediated Communication, legislative responsiveness, and trust in democratic institutions. *Technologies of Deception*. 6

Kristensson, P. O. and Vertanen, K. (2014). The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, MobileHCI '14, pages 335–338, Toronto, ON, Canada. Association for Computing Machinery. 4

Lampe, C. A., Ellison, N., and Steinfield, C. (2007). A familiar Face(book): Profile elements as signals in an online social network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 435–444, New York, NY, USA. ACM. 3, 3

Lampinen, A. and Cheshire, C. (2016). Hosting via airbnb: Motivations and financial assurances in monetized network hospitality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1669–1680. ACM. 3, 3

Landivar, L. C. (2013). Disparities in stem employment by sex, race, and hispanic origin. *Education Review*, 29(6):911–922. 5, 5, 5

Lauterbach, D., Truong, H., Shah, T., and Adamic, L. (2009). Surfing a web of trust: Reputation and reciprocity on couchsurfing. com. In *Computational*

Science and Engineering, 2009. CSE'09. International Conference on, volume 4, pages 346–353. IEEE. 3, 3

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. (1968). *The people's choice*. Columbia University Press. 4

Lee, E.-J. (2020). Authenticity model of (mass-oriented) computer-mediated communication: Conceptual explorations and testable propositions. *Journal of Computer-Mediated Communication*, 25(1):60–73. 6

Lee, M., Liang, P., and Yang, Q. (2022). Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery. 4, 4

Lee, M. K. and Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 1035–1048. 5

Lehmann, F., Markert, N., Dang, H., and Buschek, D. (2022). Suggestion lists vs. continuous generation: Interaction design for writing with generative models on mobile devices affect text length, wording and perceived authorship. In *Mensch Und Computer 2022*, MuC '22, New York, NY, USA. Association for Computing Machinery. 4

Leonard, T. C. (2008). Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness. 4, 4

Leyvand, T., Cohen-Or, D., Dror, G., and Lischinski, D. (2008). Data-driven

enhancement of facial attractiveness. In *ACM SIGGRAPH 2008 papers*, SIGGRAPH '08, pages 1–9. ACM, New York, NY, USA. 6

Liel, Y. and Zalmanson, L. (2020). What if an AI told you that 2+ 2 is 5? conformity to algorithmic recommendations. In *ICIS*. 4

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*. 4

Liu, Y., Mittal, A., Yang, D., and Bruckman, A. (2022). Will AI console me when i lose my pet? understanding perceptions of AI-mediated email writing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13. 6

Livingstone, S., Haddon, L., Görzig, A., and Ólafsson, K. (2011). Risks and safety on the internet. *The perspective of European children. Full findings and policy implications from the EU Kids Online survey of*, pages 9–16. 5

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103. 4, 4

Longino, H. E. (2020). *Science as social knowledge*. Princeton university press. 5

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37(Supplement C):94–100. 3

Ma, X., Hancock, J. T., Lim Mingjie, K., and Naaman, M. (2017a). Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 2397–2409. 1, 2, 3, 3, 3, 3, 3, 3.1, 3, 6

Ma, X., Neeraj, T., and Naaman, M. (2017b). A computational approach to perceived trustworthiness of airbnb host profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI. 3, 3, 3

Ma, X., Sun, E., and Naaman, M. (2017c). What happens in happn: The warranting powers of location history in online dating. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 41–50, New York, NY, USA. ACM. 3, 3

MacIntyre, A. (1981). The nature of the virtues. *Hastings Center Report*, pages 27–34. 5

Macnish, K. and van der Ham, J. (2020). Ethics in cybersecurity research and practice. *Technology in society*, 63:101382. 2

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 117–124. IEEE. 5

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160(4):835–850. 5, 5

Marwick, A. E. and Boyd, D. (2011). I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133. 4

Matias, J. N. (2017). *Governing human and machine behavior in an experimenting society*. PhD thesis, Massachusetts Institute of Technology. 7

Mayer, R. C. and Davis, J. H. (1999). The effect of the performance appraisal

system on trust for management: A field quasi-experiment. *Journal of applied psychology*, 84(1):123. 3

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3):709–734. 3

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*. 1

McLuhan, M. (1994). *Understanding media: The extensions of man*. MIT press. 1

Medvedev, A. N., Lambiotte, R., and Delvenne, J.-C. (2017). The anatomy of reddit: An overview of academic research. *Dynamics on and of Complex Networks*, pages 183–204. 4

Meijer, R., Conradie, P., and Choenni, S. (2014). Reconciling contradictions of open data regarding transparency, privacy, security and trust. *Journal of theoretical and applied electronic commerce research*, 9(3):32–44. 5

Microsoft (2020). Responsible AI homepage. 5

Mieczkowski, H. and Hancock, J. (2022). Examining agency, expertise, and roles of AI systems in AI-Mediated Communication. 6

Mieczkowski, H., Hancock, J. T., and Naaman, M. (2021a). AI-Mediated Communication: Language use and interpersonal effects in a referential communication task. 6

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., and Hohenstein, J. (2021b). AI-Mediated Communication: Language use and interpersonal ef-

fects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–14. 1, 6

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371. 4

Mills, G., Gregoromichelaki, E., Howes, C., and Maraev, V. (2021). Influencing laughter with AI-Mediated Communication. *Interaction Studies*, 22(3):416–463. 6

Mittelstadt, B. (2019). Ai ethics–too principled to fail. *arXiv preprint arXiv:1906.06668*. 5

Montreal, U. (2017). The montreal declaration for a responsible development of artificial intelligence. 5, 5

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics. 3

Musschenga, A. W. (2005). Empirical ethics, context-sensitivity, and contextualism. *The Journal of medicine and philosophy*, 30(5):467–490. 5, 5, 5, 5

Myers, D. (2008). *Social Psychology*. McGraw-Hill. 4, 6

Nanayakkara, P., Hullman, J., and Diakopoulos, N. (2021). Unpacking the expressed consequences of AI research in broader impact statements. *arXiv preprint arXiv:2105.04760*. 5

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103. 3, 3

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. ACM. 3

Nelson, A., Friedler, S., and Fields-Meyer, F. (2022a). Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age. 2

Nelson, A., Friedler, S., and Fields-Meyer, F. (2022b). Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age. 6

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675. 2

Newman, R. and Antin, J. (2016). Building for trust: Insights from our efforts to distill the fuel for the sharing economy. `http://nerds.airbnb.com/building-for-trust`. 3

Nickell, G. S. and Pinto, J. N. (1986). The computer attitude scale. *Computers in human behavior*, 2(4):301–306. 3, 3.1

Nozza, D., Bianchi, F., and Hovy, D. (2021). Honest: Measuring hurtful sentence completion in language models. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 1, 4, 4, 4

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books. 7

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27. 2, 4

Palin, K., Feit, A. M., Kim, S., Kristensson, P. O., and Oulasvirta, A. (2019). How do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, pages 1–12, New York, NY, USA. Association for Computing Machinery. 4

Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410. 4, 6

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253. 4, 6

Pavlick, E. and Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74. 6

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, 211(2828):42–45. 2

Petty, R. E. and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer. 3, 4

Pfotenhauer, S. M., Juhl, J., and Aarden, E. (2019). Challenging the "deficit model" of innovation: Framing policy issues under the innovation imperative. *Research Policy*, 48(4):895–904. 7

Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*. 5, 5

Pinar Saygin, A., Cicekli, I., and Akman, V. (2000). Turing test: 50 years later. *Minds and machines*, 10(4):463–518. 2, 2

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *arXiv preprint arXiv:1709.02012*. 5

Quinn, P. and Zhai, S. (2016). A cost-benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 83–88, New York, NY, USA. Association for Computing Machinery. 4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9. 1, 2, 2, 2, 4

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*. 1, 4

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44. 5, 5

Rashotte, L. (2007). Social influence. *The Blackwell encyclopedia of sociology*. 4

Rolin, K. (2006). The bias paradox in feminist standpoint epistemology. *Episteme*, 3(1-2):125–136. 5

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301. 5

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. (2019). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 99–106. 5

Schlenker, B. R. (2012). Self-presentation. 2

Schlessinger, J., Garimella, K., Jakesch, M., and Eckles, D. (2021). Effects of algorithmic trend promotion: Evidence from coordinated campaigns in twitter's trending topics. 6

Schwämmlein, E. and Wodzicki, K. (2012). What to tell about me? self-presentation in online communities. *Journal of Computer-Mediated Communication*, 17(4):387–407. 2, 3

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier. 5

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45. 5

Schwartz, S. H. (2007). Basic human values: Theory, measurement, and applications. *Revue française de sociologie*, 47(4):929. 5, 5, 5, 5

Sengers, P., Boehner, K., David, S., and Kaye, J. J. (2005). Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, CC '05, pages 49–58, New York, NY, USA. ACM. 5

Shiller, R. J. (2015). *Irrational exuberance*. Princeton university press. 4

Shilton, K. (2013). Values levers: Building ethics into design. *Science, Technology, & Human Values*, 38(3):374–397. 5

Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. 5

Simons, H. W. (2011). *Persuasion in society*. Routledge. 4

Singh, N., Bernal, G., Savchenko, D., and Glassman, E. L. (2022). Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction*, page 3511599. 4

Slovic, P. and Lichtenstein, S. (1971). Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational behavior and human performance*, 6(6):649–744. 2

Stevenson, C., Smal, I., Baas, M., Grasman, R., and van der Maas, H. (2022). Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*. 1

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*. 4

Sundar, S. S. (2008). The main model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility*, 73100. 3, 6, 6

Susser, D., Roessler, B., and Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2). 6

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13. 6

Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*. 1

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54. 2

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395. 6

Thurlow, C., Lengel, L., and Tomic, A. (2004). *Computer Mediated Communication*. Sage Publishing. 6

Toma, C. L. and Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97. 3

Toma, C. L., Hancock, J. T., and Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, 34(8):1023–1036. 3

Tsai, C.-H., Sandbulte, J., and Carroll, J. M. (2022). Promoting family healthy lifestyles through explainable AI-Mediated Communication. *Available at SSRN 4183221*. 6

Union, E. (2020). Ethics guidelines for trustworthy AI. 5

Uski, S. and Lampinen, A. (2014). Social norms and self-presentation on social network sites: Profile work in action. *New Media & Society*. 3

Van Der Heide, B., D'Angelo, J. D., and Schumaker, E. M. (2012). The effects of verbal versus photographic self-presentation on impression formation in facebook. *Journal of communication*, 62(1):98–116. 2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 1, 2, 4, 4, 4

Vertanen, K., Memmi, H., Emge, J., Reyal, S., and Kristensson, P. O. (2015). VelociTap: Investigating Fast Mobile Text Entry using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 659–668, Seoul, Republic of Korea. Association for Computing Machinery. 4

Waddell, T. F. (2018). A robot wrote this? how perceived machine authorship affects news credibility. *Digital Journalism*, 6(2):236–255. 3

Walther, J. (2011). Theories of computer-mediated communication and interpersonal relations. In *The Handbook of Interpersonal Communication*, pages 443–479. 3, 3, 3, 6

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

Walther, J. B., Loh, T., and Granka, L. (2005). Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24(1):36–65. 3

Walther, J. B. and Parks, M. R. (2002). Cues filtered out, cues filtered in. *Handbook of interpersonal communication*, pages 529–563. 3

Walther, J. B., Van Der Heide, B., Hamel, L. M., and Shulman, H. C. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using facebook. *Communication research*, 36(2):229–253. 3

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. 1, 4

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229. 1, 2, 2, 4, 6

Weiss, D., Liu, S. X., Mieczkowski, H., and Hancock, J. T. (2022). Effects of using artificial intelligence on interpersonal perceptions of job applicants. *Cyberpsychology, Behavior, and Social Networking*, 25(3):163–168. 6

Weizenbaum, J. (1966). Eliza – a computer program for the study of natural

language communication between man and machine. *Communications of the ACM*, 9(1):36–45. 3

Wetherell, G. A., Brandt, M. J., and Reyna, C. (2013). Discrimination across the ideological divide: The role of value violations and abstract values in discrimination by liberals and conservatives. *Social Psychological and Personality Science*, 4(6):658–667. 5

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200. 5, 5

Wickens, C. D., Clegg, B. A., Vieane, A. Z., and Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human factors*, 57(5):728–739. 4, 6

Williams, J. (2018). Should AI always identify itself? it's more complicated than you might think. *Electronic Frontier Foundation*. 2, 6

Wilson, T. D. and Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology*, 60(2):181. 2

Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., and Fung, P. (2021). Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*. 4, 4, 4

Winograd, T., Flores, F., and Flores, F. F. (1986). *Understanding computers and cognition: A new foundation for design*. Intellect Books. 5

Wölker, A. and Powell, T. E. (2018). Algorithms in the newsroom? news readers' perceived credibility and selection of automated journalism. *Journalism*, page 1464884918757072. 3, 3

Wu, Y. and Kelly, R. M. (2020). Online dating meets artificial intelligence: How the perception of algorithmically generated profile text impacts attractiveness and trust. In *32nd Australian Conference on Human-Computer Interaction*, pages 444–453. 6

Wylie, A., Figueroa, R., and Harding, S. (2003). Why standpoint matters. *Science and other cultures: Issues in philosophies of science and technology*, 26:48. 5

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, 51(1):110. 3, 3.1

Yang, D., Zhou, Y., Zhang, Z., Li, T. J.-J., and Ray, L. (2022). Ai as an active writer: Interaction strategies with generated text in human-ai collaborative fiction writing 56-65. In *IUI Workshops*. 4, 4

Yao, Y., Viswanath, B., Cryan, J., Zheng, H., and Zhao, B. Y. (2017). Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1143–1158, New York, NY, USA. ACM. 6

Yuan, A., Coenen, A., Reif, E., and Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pages 841–852, New York, NY, USA. Association for Computing Machinery. 4, 4

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and

Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, 32. 1, 2, 4, 4, 6

Zervas, G., Proserpio, D., and Byers, J. (2015). A first look at online reputation on airbnb, where every stay is above average. *Social Science Research Network*. 3

Zhang, M. R., Wen, H., and Wobbrock, J. O. (2019). Type, then correct: Intelligent text correction techniques for mobile text entry using neural networks. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 843–855, New York, NY, USA. Association for Computing Machinery. 4

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438. 4