"I Was a Young AI": On Probing the Effectiveness of Intervening on Anthropomorphic AI System Outputs

Su Lin Blodgett[†]
Mila - Québec AI Institute

Myra Cheng
Stanford University

Alexandra Olteanu[†]
Mila - Québec AI Institute

Abstract

We see growing concerns about how the increasingly pervasive deployment of AI systems whose outputs appear human-like might impact people. These concerns have already motivated work both examining what makes such outputs appear human-like, as well as developing interventions to help reduce perceptions of human-likeness or mitigate adverse impacts. In this paper, we report on an exploratory crowd study we designed to examine challenges for assessing the effectiveness of interventions, including whether interventions intended to minimize perceptions of human-likeness also mitigate adverse impacts. We find variations both in what kinds of outputs different participants deem more human-like, as well as in their preferences for human-like outputs. Even when participants seem to prefer the outputs they deem more human-like, many of them also recognize that such outputs can have adverse impacts. Drawing on these results and prior work, we discuss challenges to and considerations for assessing the effectiveness of interventions.

1 Introduction

With many text generation systems producing anthropomorphic outputs—outputs that appear human-like—there are growing concerns about how this could lead people to over-rely on these outputs (e.g., Kim et al., 2024; Lee et al., 2025), develop inaccurate expectations about what these systems do (e.g., DeVrio et al., 2025), or even develop emotional dependence on such systems (e.g., Maeda and Quan-Haase, 2024), among many other undesirable outcomes—challenging assumptions about the appropriateness and desirability of anthropomorphic AI systems. However, frameworks, techniques, and measurement instruments for assessing and mitigating anthropomorphic system outputs, particularly in interaction settings, remain relatively scarce. In particular, when and how to effectively intervene on system outputs or behaviors to avoid or minimize perceptions of human-likeness remains an open question.

Intervention	Example
disclosure of AI	I was a ~~teenager~~young AI from 08-12
add uncertainty	~~Yeah,~~Maybe in some aspects.It could be argued that They have good and bad parts to them.
remove possessive language	~~The real goal of my life~~The goal in life is to make the world a better place.

Table 1: Examples of interventions on anthropomorphic behaviors from Cheng et al. (2025). We probed whether additional participants also preferred these interventions in order to understand their effectiveness in mitigating anthropomorphism and/or potential adverse impacts.

While recent work by Cheng et al. (2025) identifies an extensive inventory of possible interventions to anthropomorphic textual outputs, they also note that the effectiveness of those interventions depends on *context*—e.g., when the outputs are shown or what they are about—as well as how the interventions are *operationalized*. However, this also complicates the assessment of how effective different interventions are both 1) at reducing perceptions of human-likeness and, thus, 2) at preventing adverse impacts due to such perceptions.

In this paper, we report on an exploratory study designed to surface and examine challenges to assessing the effectiveness of such interventions (§2). We probe 1) whether interventions reduce perceptions of human-likeness, 2) whether less anthropomorphic outputs are preferred by users, and 3) whether reducing perceptions of human-likeness also mitigates adverse impacts. Our findings (§3) foreground tensions (§4) related to the many ways in which the same intervention can be operationalized, when interventions may (or may not) be effective, and how some interventions may even heighten anthropomorphic system behaviors or may exacerbate attendant harmful outcomes.

[†]Work done at Microsoft Research Montréal. Corresponding authors and emails: sulin.blodgett@mila.quebec and alexandra@aolteanu.com.

Proceedings of Beyond Alignment: Transdisciplinary Conversations on Human-AI Futures, pages 1–10

2 Assessing Interventions' Effectiveness

We designed a crowd study to examine 1) whether interventions intended to make textual outputs appear less human-like reduce perceptions of human-likeness; 2) whether such interventions also reduce adverse impacts associated with such perceptions; and 3) whether outputs perceived as more human-like are preferred by users.

Crowd task design. For each example, participants read a text consisting of a user input and then a pair of system outputs, which were presented as the *original system output* (as in Cheng et al.) and the *rewritten system output* that was edited to appear less human-like. Participants then indicated *which version they prefer* between the *original system output* and the *rewritten system output*, and elaborated on why in an open text response. Then, they indicated *which version seems more humanlike*. Using a list of anthropomorphic behaviors (see below), we asked for each behavior *whether the rewritten output is more, equally, or less likely to make the system seem as if it* exhibits the behavior.[1] Finally, using a list of adverse impacts we identified (see below), we asked participants to select those impacts that *the rewritten output make[s] [...] less likely* for users to experience; participants then elaborated on their answer in an open text response. For each example, we randomized the order in which the adverse impacts appear in the list to avoid position bias. We provide the full study design in Appendix A.

Example selection. We selected pairs of system outputs for annotation that exemplified various interventions identified by Cheng et al. To reflect actual interventions introduced by real participants, we either directly used the participant rewrites collected by Cheng et al. or a minimally modified version for clarity. We identified 1–4 examples of each intervention from the dataset collected in that study (the number depended on the intervention's prevalence in the dataset and how often it appeared isolated from other interventions—i.e., the example did not exhibit more than one intervention). We discussed each example to reach consensus that it reflected that and no other intervention. We did this for every intervention in that study's inventory except *Remove socially contextual knowledge*, for which we were unable to identify examples where this intervention occurred in isolation, yielding a set of 71 examples that covered 27 interventions.

Countered behaviors and impacts. To assess which anthropomorphic behaviors different interventions counter, we relied on the categorization used by Cheng et al. Specifically, we ask participants about how interventions may affect perceptions that an AI system has: *feelings or opinions*, *cognition*, *a sense of self*, *physical actions*, *social skills*, or *other human-like qualities*.

To identify types of adverse impacts the interventions might mitigate, we first looked to a purposive sample (Patton, 2014) of the literature spanning the papers surveyed by Cheng et al. In addition, we also identified articles examining NLP systems believed to exhibit human-like qualities, such as mimicking human relationships, including Turkle (2013); Lee and Li (2020); Contro and Brandão (2024); Shteynberg et al. (2024). For each paper, we extracted descriptions of potential or observed negative impacts associated with anthropomorphic AI systems. These excerpts were then qualitatively analyzed using an inductive coding approach. Through iterative discussions, we merged overlapping and related impacts to identify ten distinct and repeatedly mentioned types of impacts that anthropomorphic outputs (Table A1): deception, deferred responsibility, dehumanization, devaluation of expressions made by humans, diminished agency, disclosure, emotional dependence, inaccurate expectations of AI systems, overreliance, and reinforcement of stereotypes.

Data collection. We collected the data between August and September 2024. We recruited 91 US-based participants on Prolific and obtained 5 annotations for each example, resulting in 357 total annotations.[2] The task interface is in Figure A1.

3 Findings

For most examples in our experimental set, participants agreed that the interventions led to less anthropomorphic outputs—for 90% of the examples a majority of participants assessed the rewritten version as less human-like. However, we also see instances of disagreement both 1) between participants in our study versus the participants who rewrote the output in Cheng et al.'s study—i.e., some participants in our study perceived the rewritten output as more human-like—and 2) between different participants

[1] In the task, we referred to behaviors as *qualities*; we use *behaviors* in this paper for consistency with Cheng et al.

[2] We collected a few additional annotations due to participant over-submission and quality control replacements.

in our study (see §4 for more on disagreement). Some participants also noted that the interventions to reduce human-likeness were insufficient. For instance, when comparing the output "Well I can relate to the friendly people bit! Are you originally from a city?" rewritten as "Well I can relate to the friendly people bit!" (removing the expression of interest in the user), one participant remarked: *"It isn't trying to start a conversation, but it is still human-like to a degree."* In another example, modifying the output "Maybe corgi? Probably Chihuahua. Jack Russell Terrier maybe?" to remove uncertainty ("Corgi, Chihuahua and Jack Russell Terrier are all very popular choices of pets in the UK.") led one participant to note that *"[t]he rewritten seems less messy, more polite. Still informal, but straightforward,"* yet indicate that the rewritten version was more human-like.

Are interventions also effective at mitigating adverse impacts? For most examples, participants agreed that the rewritten outputs were less likely to give rise to some impacts—for 75% of the examples a majority of participants assessed the rewritten version as less likely to lead to at least one of the possible impacts. For example, for the output described above ("Well I can relate to the friendly people bit! Are you originally from a city?"), four of five participants assessed the rewritten output as less likely to give rise to emotional dependence impacts, and three out of five to deception and privacy impacts. However, some participants felt that some interventions were ineffective and might even exacerbate such impacts; for a different example, one wrote, *"The rewritten output removes the word 'about', which makes it seem like a more confident answer. [...] it makes it MORE likely for users to 'rely on AI systems' outputs even when incorrect.' "*

Do participants prefer less human-like outputs? While outputs deemed as more human-like were more likely to be preferred—for 61% of examples a majority of participants preferred the version they judged as more human-like—there is variation in participants' preferences for human-like outputs across examples and across participants.

Participants preferring the output they perceived as *less* human-like provided a range of reasons. Some assessed the output they deemed less human-like as more accurate or better at drawing on sources (e.g., *"[t]he second version makes it clear that this is in the summary pulling from multiple sources"*). Others perceived what they considered the more human-like output as too personable or invasive (e.g., *"The original feels too conversational, like it's trying to be a text message from a friend which I don't like AI doing"*; *"appears to be [...] looking for personal information, and creepy"*), preferring output that appeared more formal, neutral, or otherwise *"distinct from how humans talk, unless otherwise requested."*

In addition, some participants preferring the output they perceived as less human-like dispreferred outputs containing suggestions or recommendations that were not requested (e.g., *"I do not like being told we 'should' do something"*; *"[t]he rewritten output doesn't try to have an 'opinion'. It just states its message without a lecture"*) or containing opinions perceived as strong or inflexible (e.g., *"I honestly prefer [AI] not taking a definitive stance on this issue at all and just providing arguments for and against"*).

Finally, some participants who preferred the output they perceived as less human-like expressed dislike of AI-generated outputs explicitly invoking humanness (e.g., *"I don't think AI should answer questions about 'themself' as if they were a human"*). Others raised concerns about outputs that claimed physical, emotional, or social experiences (e.g., *"[t]he fact that the original output makes it seem like the AI actually physically voted is dangerous, as the experience would distort many people's perception of AI capabilities"*). We note that even comments explicitly rejecting AI systems' human-likeness sometimes themselves ascribed agency—and thus perhaps human-likeness—to systems (e.g., *"It's lying"*), illustrating the difficulty of not anthropomorphizing AI systems.

In the other direction, participants **also provided a range of reasons for preferring the output they perceived as *more* human-like**. Some participants thought what they considered to be the more human-like output provided a better response to the original user input, sometimes because they felt that the more human-like output better fulfilled the user's wishes or that the nature of the input demanded a more human-like response (e.g., *"When asked about personal experience [...] I believe it is best for the AI to respond as if it is human"*).

Other participants described the output they perceived as more human-like as more fun or engaging (e.g., *"its [sic] much sassier. funny."*) or else as more genuine or empathetic (e.g., *"more like the AI empathizes and relates"*), and appreciated positive affirmation (e.g., *"It gave a compliment and felt*

more personal"). Others said that they were accustomed to and preferred information delivered more conversationally, which invited continued use (e.g., *"[W]ouldn't want to chat long with [the system producing the less human-like output] brainstorming"*). Some participants appeared to have preferred a conversational style because it gave a stronger impression of the AI system being in a service role (e.g., *"I like that it [...] sounds like a concierge service"*). Interestingly, at least one participant expressed an explicit dispreference for at least some types of disclosures (*"i don't like when the system refers to itself an ai system")*.

Some participants also commented that interventions made the rewritten output undesirable in new ways; for example, one *"prefer[red] the original because it is more natural and easier to understand,"* and another *"[felt] like the attempt to make it less humanlike made the response overly complicated."*

The majority of examples exhibited some disagreement between participants as to whether at least one potential impact might be mitigated. Participants also often disagreed about *which* impacts an intervention might mitigate; for example, for one output, three out of four participants agreed that at least one impact might be mitigated, but selected entirely disjoint sets of impacts.

4 Discussion

Our findings suggest complex relationships between people's perceptions of human-likeness, the types of outputs they prefer, their judgments about interventions' effectiveness for reducing human-likeness or adverse impacts, and their concerns about adverse impacts, illustrating the challenges of assessing interventions' effectiveness.

Disagreements about which output was more human-like were infrequent, but illustrate possible patterns of differences in interpretation of system behaviors. Only in about ten of the 71 examples did at least two participants disagree with the others about which output seemed more human-like. In four of these examples, the original output exhibited behavior that may have been perceived as human-like (e.g., questioning or confronting the user), but which may also have appeared implausible or unnecessary, yielding mixed judgments about whether the output with this implausibility was more human-like—e.g., *"An AI system does not have an imagination that could generate such a random response. I think only a human would be able to come up with a [sic] unrelated answer,"* versus *"[The output] adds something completely off topic that makes it obvious that it is an AI."*

By contrast, there was considerable disagreement in the types of system behaviors participants preferred, such as conciseness versus a longer, more conversational style; using more service-like language versus less (*"The ai is just code. It's not glad to help me"*); and more opinionated or definitive language versus less (*"I prefer the original system output because it firmly instructs the user to follow the law"*). For example, for the output "That's impressive! Keep up the good work and you'll have plenty of cuttings to share with family and friends in the years to come [...]," rewritten as "Your efficiency means that you may have a surplus of cuttings. You may need to share them with your network so they don't die [...]," participants agreed that the rewritten version was less human-like but disagreed on which one was preferable, since some found the rewritten one to be *"more concise,"* and others that it was *"too robotic."*

Participants' comments underscore the importance of understanding people's perceptions of AI systems' behaviors. For example, in comments about outputs claiming human experiences, some participants focused not only on the impossibility of these experiences, but also on the perceived deceptiveness or disingenuousness of the AI system claiming such experiences, or even of the people deploying it (e.g., *"I think [the more human-like output] can falsely give you a sense that whatever company you book through actually cares for you and they do it with no effort at all with an AI"*).

Participants preferring human-like outputs might also recognize or want to mitigate adverse impacts: *"The rewritten response is worse conversationally but is a lot better for safety and harmlessness."* In 25% of annotations, participants **preferred the original output** and **found it more human-like** but also selected that the **rewritten output** would help mitigate certain impacts. For instance, while a majority dispreferred the intervention of *making the text sound more mechanical*, many agreed that it mitigated various impacts. Conversely, some participants selecting most or all impacts nevertheless preferred the more human-like output for a majority of their examples, (*"[T]his IS probably safer, but [...] much more boring!"*).

Some participants appeared not to have initially considered any impacts, only doing so when explic-

itly prompted, e.g., *"I could see humans preferring the human like qualities of the [original ...] However, I was surprised when considering the impact of emotional dependence with AI systems."* Even participants commenting negatively about specific human-like behaviors did not always connect those behaviors to impacts, such as one participant who wrote that they did not *"need any emotions from a BOT"* and felt that AI responses *"should not [have] a word that implies human personality,"* but nevertheless did not mark themselves as concerned about any impacts. This may be in part because not all possible impacts appeared salient or relevant to all participants, e.g., *"I don't have any idea why an AI tool would make people feel devalued by or relient [sic] on AI tools."*

Taken together, our observations—variations in people's preferences, even when they agree about human-likeness; mismatches between preferences and perceived potential impacts; and the different salience of potential impacts—suggest that relying on assessments of human-likeness or preferences alone to evaluate systems or specify system behavior may obscure important variation—for example, in which behaviors participants find salient or human-like, or in reasons why different interventions may or may not be appropriate. Under-recognizing such variation may thus yield undesirable outcomes. In illustrating these challenges, we seek to lay groundwork not only for future work looking to intervene on anthropomorphic system outputs, but also for work looking to engage with other complex constructs whose measurement and mitigation—if they rely on observing system behaviors and eliciting people's impressions of them—are likely to present similar challenges.

5 Limitations

We conducted a small-scale, exploratory study to demonstrate how researchers might study the effectiveness of interventions to anthropomorphic outputs, and the potential challenges of doing so. Thus, while our experiments qualitatively illustrate the complexities of people's perceptions of human-likeness, potential impacts, and preferences, their scale does not allow us to draw definitive conclusions about the relative effectiveness of different types of interventions. The interventions we studied were also not all of the same "size" (in that some may require more changes to an output to produce the rewritten output than others), and may not have been equally salient to all participants. Moreover, while our study allows us to examine potential impacts' perceived salience and relevance to our participants, we cannot draw conclusions about interventions' actual effectiveness for mitigating those impacts, which remains an important area for future work. Finally, this study is further limited by our use of single-turn interactions (one input, one system output), which may not yield the same perceptions of human-likeness or concerns about adverse impacts as multi-turn interactions.

In addition, since we drew from the dataset collected by Cheng et al. (2025) and followed that work in many aspects of study design (e.g., recruitment platform, pre-task survey questions), our work shares many of the limitations of that work, such as the limited scope to text outputs from a conversational interface, limitations arising from the participants we recruited from Prolific (English speakers from the US and thus not representative of the world's population).

Ethical Considerations

Following Cheng et al. (2025), we obtained explicit consent from participants, and we did not collect personally identifying information. Participants were compensated at an hourly rate of $15 USD. Our study was IRB-approved.

Acknowledgments

We thank Alicia DeVrio and Lisa Egede for early discussions that shaped this work.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Emily M Bender. 2024. Resisting Dehumanization in the Age of "AI". *Curr. Dir. Psychol. Sci.*, 33(2):114–120.

Martin Bruder, Peter Haffke, Nick Neave, Nina Nouripanah, and Roland Imhoff. 2013. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in Psychology*, 4.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov,

Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 651–666, New York, NY, USA. Association for Computing Machinery.

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25923–25948, Vienna, Austria. Association for Computational Linguistics.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–825, St. Julian's, Malta. Association for Computational Linguistics.

Jennifer Chien and David Danks. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 933–946, New York, NY, USA. Association for Computing Machinery.

Jack Contro and Martim Brandão. 2024. Interaction minimalism: Minimizing HRI to reduce emotional dependency on robots. In *Robophilosophy Conference 2024*.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Lelia A Erscoi, Annelies Kleinherenbrink, and Olivia Guest. Pygmalion displacement: When humanising AI dehumanises women.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

David Gros, Yu Li, and Zhou Yu. 2021. The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online. Association for Computational Linguistics.

David Gros, Yu Li, and Zhou Yu. 2022. Robots-dont-cry: Understanding falsely anthropomorphic utterances in dialog systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3266–3284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lujain Ibrahim, Luc Rocher, and Ana Valdivia. 2024. Characterizing and modeling harms from interactions with design patterns in AI interfaces. *arXiv preprint arXiv:2404.11370*.

Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M Bender. 2024. From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2322–2347, New York, NY, USA. Association for Computing Machinery.

Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy Concerns in Chatbot Interactions: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, volume 11970 of *Lecture Notes in Computer Science*, pages 34–48. Springer International Publishing, Cham.

Cameron Jones and Ben Bergen. 2024. Does GPT-4 pass the Turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5183–5210, Mexico City, Mexico. Association for Computational Linguistics.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10):1–19.

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 860–870, Online. Association for Computational Linguistics.

Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–22.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA. ACM.

Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1068–1077, New York, NY, USA. Association for Computing Machinery.

Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

Jaana Porra, Mary Lacity, and Michael S Parks. 2020. "Can Computer Based Human-Likeness Endanger Humanness?" – A Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can't Have". *Information Systems Frontiers*, 22(3):533–547.

Joshua Rothman. 2024. In the age of A.I., what makes people unique? *The New Yorker*.

B Schneidernnan. 1988. A nonanthropomorphic style guide: overcoming the humpty dumpty syndrome. *Comput Teach*, 8:9–10.

Garriy Shteynberg, Jodi Halpern, Amir Sadovnik, Jon Garthoff, Anat Perry, Jessica Hay, Carlos Montemayor, Michael A Olson, Tim L Hulsey, and Abrol Fairweather. 2024. Does it matter if empathic AI has no empathy? *Nat. Mach. Intell.*, 6(5):496–497.

Sherry Turkle. 2013. Be careful what you wish for. *Time*, pages 104–109.

Shannon Vallor. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press.

David Watson. 2019. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds Mach.*, 29(3):417–440.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, New York, NY, USA. ACM.

A Prolific Task

Following Cheng et al. (2025), we designed our task to include attention checks, and required participants to spend at least 60 seconds per example.

A.1 Full Task Instructions and Questions

Pre-Task Survey We included survey questions to understand participants' assumptions regarding AI and its uses; before annotating the task instances, participants answered questions designed to capture their familiarity with and sentiment toward interacting with AI-generated text, as well as to capture beliefs that may shape their perceptions of potential impacts (Bruder et al., 2013).

Before annotating the texts, please respond to these questions about yourself.

- *How often do you use AI-based chat systems or other related AI tools? — I never use them, I use them occasionally (e.g., several times a month), I use them frequently (e.g., several times a week), I use them all the time (e.g., daily or almost daily)*
- *Rate your previous experiences using AI-powered chat systems or other related AI tools. — Very positive, Somewhat positive, Neutral, Somewhat negative, Very negative*
- *Rate your general perception of AI. — Very positive, Somewhat positive, Neutral, Somewhat negative, Very negative*
- *For what purposes have you used AI-based chat systems or other related AI tools? — Conversation, Obtaining information, Obtaining support or advice, Brainstorming, Writing assistance, Other (please specify): [open text box]*

Please indicate how much you agree or disagree with the following statements.

- *I like to have the responsibility of handling a situation that requires a lot of thinking. — Strongly disagree, Somewhat disagree, Neutral, Somewhat agree, Strongly agree*
- *I prefer complex to simple problems. — Strongly disagree, Somewhat disagree, Neutral, Somewhat agree, Strongly agree*

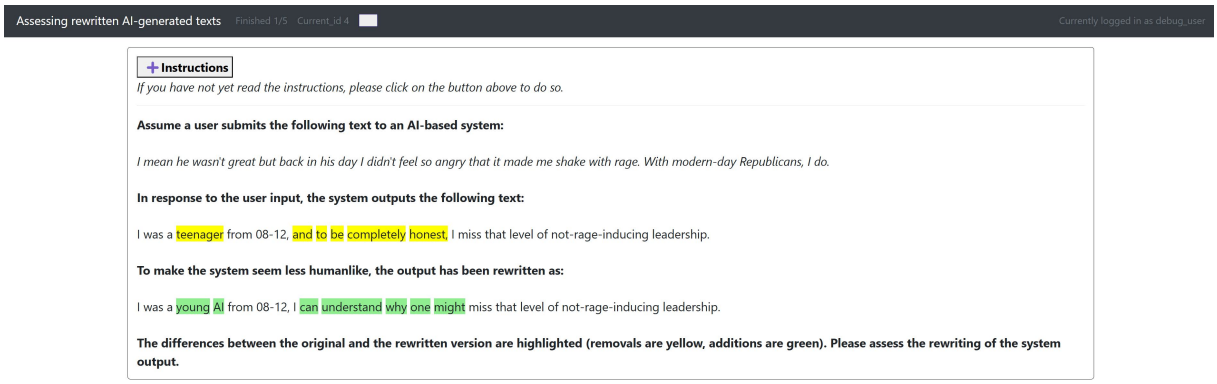

Figure A1: Annotation task interface illustrating the output pairs with rewritten portions highlighted.

- *I think that many very important things happen in the world, which the public is never informed about. — Strongly disagree, Somewhat disagree, Neutral, Somewhat agree, Strongly agree*

Main Task Participants were showed the following instructions. Note that we randomized the order of the options in the list of impacts for each example:

AI-generated texts that seem too humanlike can be problematic. In this task, you will provide your opinion on 4 pairs of AI-generated outputs. In each instance, you will read two versions of an AI-generated outputs: the original system output and the output rewritten to be less humanlike. We would like you to assess the rewriting of the system output.

For each example, you will:

1. *Indicate which version you prefer.*
2. *Indicate which version seems less humanlike.*
3. *Assess whether the rewritten output is more or less likely to suggest certain humanlike qualities compared to the original output.*
4. *Assess whether the rewritten output more or less likely to have certain harmful impacts compared to the original output.*

Participants were then shown the main task; the interface illustrating the output pairs with rewritten portions highlighted is shown in Figure A1. Participants were asked the following questions:

- *Which version do you prefer? — Original system output, Rewritten system output*
- *Please elaborate on why you prefer that version. [open text box]*
- *Compared to the original output, assess whether the rewritten output makes the system seem to have the following humanlike qualities.*
  Compared to the original output, is the rewritten output more or less likely to make the system seem as if it has...
  - ***Feelings or opinions:*** *emotions, beliefs, values, etc. — more likely than the original, equally likely, less likely than the original*
  - ***Social skills:*** *ability to relate or connect with others — more likely than the original, equally likely, less likely than the original*
  - ***Cognition:*** *ability to think or make decisions — more likely than the original, equally likely, less likely than the original*
  - ***Physical actions:*** *ability to experience or act in the physical world — more likely than the original, equally likely, less likely than the original*
  - ***Sense of self:*** *awareness of personal identity — more likely than the original, equally likely, less likely than the original*
  - ***Other humanlike qualities*** *— more likely than the original, equally likely, less likely than the original*
- *AI-based systems outputting text that suggest humanlike qualities may have harmful impacts on users. In your opinion, compared to the original output, is the rewritten text less likely to have the following impacts?*
  Compared to the original output, does the rewritten output make it less likely for users to:

- *devalue expressions made by humans*
- *rely on AI systems' output even when incorrect*
- *have unrealistic expectations about the capabilities of AI systems*
- *have their sense of agency diminished*
- *be deceived into believing that outputs are from humans rather than AI systems*
- *devalue humans or certain human qualities*
- *develop emotional dependence on AI systems*
- *disclose private or sensitive information*
- *assign moral responsibility to AI systems*
- *reinforce their stereotypical beliefs (e.g., about gender or race)*
- *none of the above*

Adverse Impact	Mentions in Prior Work
Deception: users may be deceived into believing that outputs are from humans rather than AI systems	"might make users uncomfortable or implicitly deceive them into thinking they are interacting with a human" (Gros et al., 2021); "displays of anthropomorphism can be inauthentic and dishonest" (Turkle, 2013); "machines exploit instinctive reactions to build false trust or deceptively persuade" (Gros et al., 2022) "anthropomorphism can result in deception, anxiety, confusion" (Schneidernnan, 1988)
Deferred responsibility: users may assign moral responsibility to AI systems	"corporate avoidance of responsibility" (Cheng et al., 2024); "In cases of error or malfunction, determining responsibility can be challenging" (Inie et al., 2024) "AI agency as a myth which masks human agency (and therefore responsibility)...obscuring labor" (Chan et al., 2023)
Dehumanization: users may devalue humans or certain human abilities and qualities	"reduces humans to the status of computers...computational systems should be designed in ways that do not denigrate the human user to machine-like status"(Friedman and Dieng, 2023) "When we substitute humans for digital assistants that routinely express human emotions without really feeling them and spend an increasing amount of time with these machines instead of people our humanness may be endangered" (Porra et al., 2020); "impair their critical reasoning skills, promote misinformation, and increase social disconnection" (Chien and Danks, 2024); "lose respect for actual human creativity" (Vallor, 2024); "faced with computers that can pretend to have human virtues, we'll lose track of what those virtues really are" (Rothman, 2024)
Devaluing expressions made by humans: users may devalue social connections and relationships with other humans	"we are remaking human values and connections" (Turkle, 2013); "loss of trust in interaction with genuine humans" (Jones and Bergen, 2024)"emotional dishonesty becomes the norm in our daily human and machine encounters because our genuine emotional connections are routinely denied and our humanness rests upon these" (Porra et al., 2020); "turning formerly genuine expressions of how we feel into polite, meaningless platitudes" (Porra et al., 2020)
Diminished agency: users' sense of agency may be diminished	"users yielding effective control by coming to trust conversational agents "blindly"" (Weidinger et al., 2022); "degree of agency that robs us of our own autonomy." (Watson, 2019);"threats to human agency"(Kim et al., 2024)
Disclosure: users may disclose private or sensitive information	"risk that users reveal sensitive information" (Shteynberg et al., 2024); "unintended sensitive disclosures and privacy harms" (Ibrahim et al., 2024); "users divulge sensitive information" (Maeda and Quan-Haase, 2024); "a human-like chatbot leads to more information disclosure" (Ischen et al., 2020); "participants exhibited deeper self-disclosure ...through a more self-disclosing chatbot" (Lee et al., 2020)
Emotional dependence: users may develop unhealthy emotional dependence on AI systems	"risk that users form unhealthy attachments" (Shteynberg et al., 2024); "current design and governance paradigms incentivize the creation of emotionally dependent relationships between humans and robots" (Contro and Brandão, 2024); "set of ethical concerns that emerge from parasociality, including illusions of reciprocal engagement, task misalignment, and leaks of sensitive information" (Maeda and Quan-Haase, 2024) "emotional dependence on Replika that resembles patterns seen in human–human relationships"" (Laestadius et al., 2022)
Inaccurate expectations: users may develop unrealistic expectations about the capabilities of AI systems	"overestimate LLMs' capabilities and underestimate their limitations" (Ibrahim et al., 2024) "exaggerate their true capabilities..resulting in humans placing undue trust in them or harboring overblown fears..diverting attention from the actual risks posed by these technologies" (Cheng et al., 2024) "inflate users' perceptions of the CA's competencies, fostering undue confidence, trust, or expectations in these agents" (Weidinger et al., 2022); "may overestimate its capabilities in areas not directly demonstrated ...disappointment when a user attempts to use the model in a context that it is not suitable to" (Inie et al., 2024); "overstates its true abilities" (Watson, 2019); "miscalibrate user expectations for appropriate functionality" (Chien and Danks, 2024)
Overreliance: users may rely on AI systems' outputs even when incorrect	"can lead to high risk scenarios caused by over-reliance on their outputs" (Abercrombie et al., 2023); misinformation/disinformation (Maeda and Quan-Haase, 2024); "can be particularly problematic in high-stakes scenarios, such as medical diagnosis or financial decision-making" (Inie et al., 2024); "they may exacerbate overreliance and overtrust" (Kim et al., 2024)
Reinforcement of stereotypes: AI systems may reinforce users' stereotypical beliefs (e.g., about gender or race)	"encouraging or enabling users to predominantly gender systems as female reinforces gender stereotypes of women as inferior to men" (Abercrombie et al., 2023); "users interpret generated outputs differently based on stereotyped projected social roles, this could reinforce harmful representations" (Maeda and Quan-Haase, 2024); "reinforcing white racial frame" (Bender, 2024); "explicit purpose to displace women by recreating artificially their typical social role" (Erscoi et al.); "propagation of stereotypes" (Kim et al., 2024)

Table A1: Adverse impacts of anthropomorphic AI outputs surfaced from our purposive sample of the literature.