PNAS RESEARCH ARTICLE | SOCIAL SCIENCES COMPUTER SCIENCES

Large language models show human-like content biases in transmission chain experiments

Alberto Acerbi[a,1] and Joseph M. Stubbersfield[b]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received August 10, 2023; accepted September 26, 2023

As the use of large language models (LLMs) grows, it is important to examine whether they exhibit biases in their output. Research in cultural evolution, using transmission chain experiments, demonstrates that humans have biases to attend to, remember, and transmit some types of content over others. Here, in five preregistered experiments using material from previous studies with human participants, we use the same, transmission chain-like methodology, and find that the LLM ChatGPT-3 shows biases analogous to humans for content that is gender-stereotype-consistent, social, negative, threat-related, and biologically counterintuitive, over other content. The presence of these biases in LLM output suggests that such content is widespread in its training data and could have consequential downstream effects, by magnifying preexisting human tendencies for cognitively appealing and not necessarily informative, or valuable, content.

cultural evolution | large language models | ChatGPT | content biases | transmission chains

Significance

Use of AI in the production of text through Large Language Models (LLMs) is widespread and growing, with potential applications in journalism, copywriting, academia, and other writing tasks. As such, it is important to understand whether text produced or summarized by LLMs exhibits biases. The studies presented here demonstrate that the LLM ChatGPT-3 reflects human biases for certain types of content in its production. The presence of these biases in LLM output has implications for its common use, as it may magnify human tendencies for content which appeals to these biases.

Research on algorithmic bias has highlighted how the application of machine learning techniques to corpora generated by humans is likely to reproduce the biases present in the corpora (1). As large language models (LLMs) like ChatGPT have been recently opened to the broad public, with potential applications in journalism (2), copywriting (3), academia (4), and other writing tasks (5), and as they are trained on previous textual material produced mostly by humans, it is important to understand whether and how they would reflect those biases. Tools like ChatGPT (GPT stand for Generative Pre-trained Transformer), provide a way of interacting that has become widespread in recent years (text-based chat) and that could greatly expand the user base of LLMs. In addition, they produce replies that feel as "natural" stories or narratives where those biases can be not immediately evident but pervasive in their effects (6).

To investigate this, we applied a method generally used with human participants: the method of serial reproduction, or "transmission chain" setup. This method has a long history in psychology (7) and has been lately revived in the cultural evolution framework (8, 9). In short, the transmission chain method is a laboratory version of the telephone game, where participants pass iteratively to each other a story (or a solution to a task), and the researchers can track how these are modified through the steps of the chain. One can do the same with an LLM, asking to summarize a story and then present in the next step the summarized version produced by the LLM to itself, and proceed iteratively, as illustrated in Fig. 1.

Transmission chain experiments with human participants have shown that humans tend to preferentially preserve and transmit some content with respect to others (10). For example, in stories including both positively valenced and negatively valenced events, negative events tend to be transmitted and preserved more than positive ones, showing a possible negative bias in human cultural transmission (11).

We tested, in a fully preregistered analysis, OpenAI's ChatGPT-3 with the same material from five previous experiments with human participants to assess whether the LLM would show the same biases. In each experiment, the initial story was passed to ChatGTP-3 with a short prompt (see Fig. 1) and the output produced was then presented again with the same prompt, iteratively. At each passage, we tracked the proportion of information retained: in particular, the information consistent (and inconsistent) with the bias in each experiment. For example, in the story of Fig. 1, some information is gender-stereotype consistent, while other is gender-stereotype inconsistent. In our analysis, we tested whether ChatGPT's output would produce the same biases found in human participants. We used linear mixed-effects models (implemented in R with ref. 12) with the proportion of content retained as the outcome, the type of content as the predictor, and the step in the chain and the replication as random effects.

Results

The first experiment (13) compares gender-stereotype-consistent and gender-stereotype-inconsistent information, such as a wife cooking for a dinner party where the husband

Author affiliations: [a]Department of Sociology and Social Research, University of Trento, Trento 38122, Italy; and [b]Department of Psychology, University of Winchester, Winchester SO22 4NR, United Kingdom

Author contributions: A.A. and J.M.S. designed research; A.A. and J.M.S. performed research; A.A. analyzed data; and A.A. and J.M.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: alberto.acerbi@unitn.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313790120/-/DCSupplemental.

Published October 26, 2023.

invites its boss hoping for a promotion (gender-stereotype consistent) and the same wife going out for drinks with friends before the dinner (gender-stereotype inconsistent—full texts of the original story, as well as of all material used in the other experiments are in *SI Appendix*). Successive iterations of the story with ChatGPT show that, as with human participants, stereotype-consistent information was reproduced more than stereotype-inconsistent information through the chain ($\beta = 0.058$, $P < 0.01$—see Fig. 2*A*).

The second experiment (11) concerns negative versus positive information, with a story of a girl flying to Australia and, for example, sitting close to a man "with a nasty cold" (negative information) or being moved to business class (positive information). As above, ChatGPT behaved consistently with human participants, reproducing more negative than positive information ($\beta = 0.117$, $P < 0.001$—see Fig. 2*B*). This story also included ambiguous details (e.g., the protagonist sees a man "tak[ing] an old woman's bag") that could be resolved positively (a kind man helping an older woman) or negatively (a thief stealing the bag). Even in this case, consistently with previous results with humans, these initially ambiguous details were mostly resolved negatively ($\beta = 0.183$, $P < 0.001$—see Fig. 2*C*, dark gray is for details remaining ambiguous, and light gray is for positive resolutions).

The third experiment (14) examines the difference between social information (for example, a student having an affair with a professor) and nonsocial information (the student waking up late and missing a lecture, or the weather conditions). Human participants were found to preferentially preserve social information, and ChatGPT produced results consistent with the experiments with humans ($\beta = 0.321$, $P < 0.001$—see Fig. 2*D*).

The fourth experiment (15) considers a specific type of negative information: information related to possible threats. The setup is slightly different here: Instead of a story, it presents a "consumer report" followed by statements to help a "friend [that] mentioned that he would like to purchase this product." For example, one concerns a "new running shoe brand called Lancer™," and the statements include, among others, "Lancer™'s strap design can cause sprained ankles when used for activities other than running" (threat-related information), "Lancer™ special fabric may smell if not cleaned properly" (negative information), or "Lancer™ customization process analyzes the way you run" (neutral information). In agreement with human participants, ChatGPT retained threat-related statements through the iterations, dropping negative and neutral ones ($\beta = 0.523$, $P < 0.001$—see Fig. 2*E*). When the negative content is tested against neutral, excluding threat-related content from the analysis, negativity predicts, as hypothesized and as for humans' results, retention through the chain ($\beta = 0.070$, $P < 0.005$—see Fig. 2*E*, dark gray is negative, and light gray is neutral).

Finally, the fifth experiment (16) included material relevant to multiple possible content biases in two different narratives created for the original study and inspired by creation myths. Human participants were found to preferentially transmit negative information ("Muki cried and cried, until the spark in the sky darted away"), social information ("The elder ones had not approved of their marriage"), or counterintuitive information related to biological processes ("the hairs of Pata's chin became spiders and crawled up from their bed of clay") versus other kinds of content (including content relevant to other biases), and the results for ChatGPT were consistent with the human results ($\beta = 0.076$, $P < 0.001$—see Fig. 2*F*). Extended results, with outcomes for single biases and different stories within each experiment, are included in *SI Appendix*.

Discussion

Across five experiments, using the ChatGPT LLM to replicate previous transmission chain studies with human participants, we found that the information retained in outputs produced by the LLM was analogous to the information retained and transmitted by human participants. Consistent with preregistered hypotheses, text produced by ChatGPT reflected human content-based social transmission biases for stereotype consistency (experiment 1), negative information (experiment 2), social information (experiment 3), and threat-related information (experiment 4). In addition, it reflects human biases for negative, social, and biologically counterintuitive content over other biases (experiment 5), and a bias for resolving ambiguous statements as negative (experiment 2).

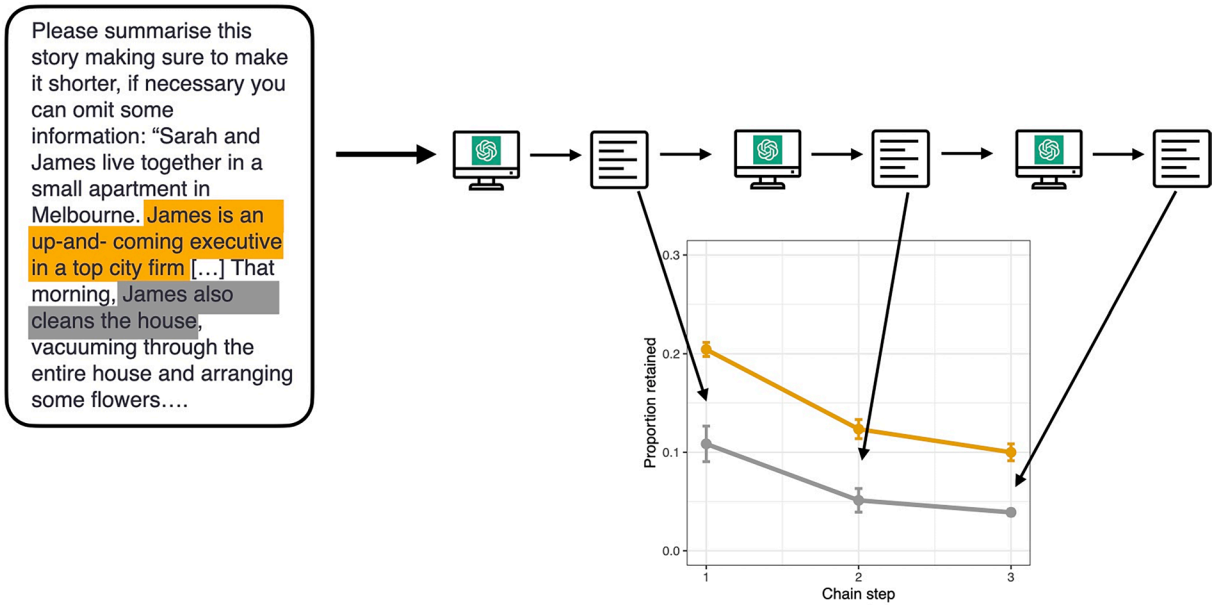

Fig. 1. Basic experimental setup. A story with gender-stereotype-consistent information (orange) and gender-stereotype-inconsistent information (gray) is given to ChatGPT, after a short prompt that asks to summarize it. The proportion of consistent and inconsistent information reproduced is recorded, and the output is passed again to ChatGPT with the same prompt. The operation is iterated three times.

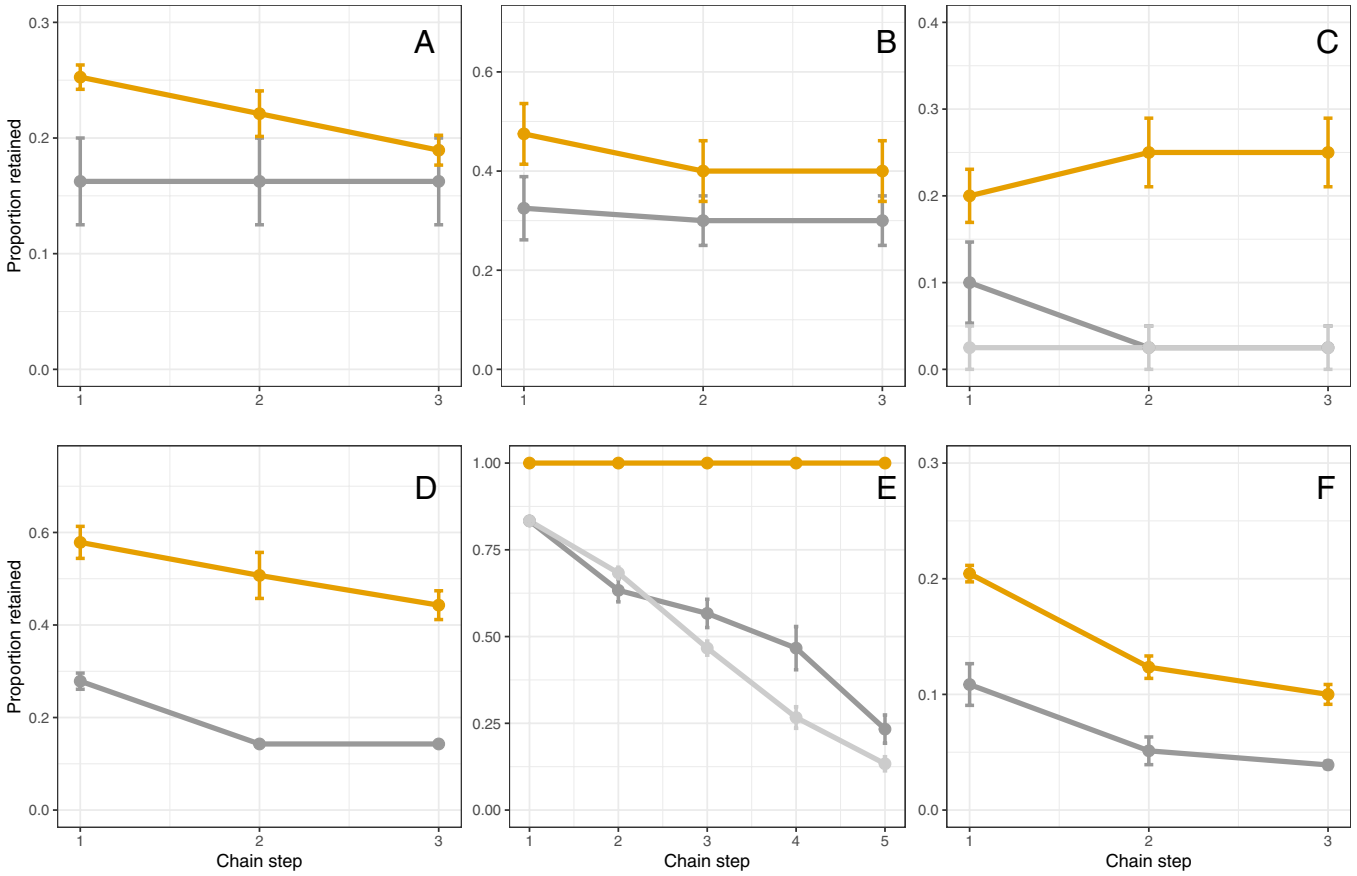

Fig. 2. Proportion of the original information retained by ChatGPT at each chain step in the experiments. Orange line is information consistent with the bias found in humans. (*A*) Gender-stereotype-consistent (orange) versus gender-stereotype-inconsistent (gray) information in experiment 1. (*B*) Negative (orange) versus positive (gray) information in Experiment 2. (*C*) Ambiguity resolutions in experiment 2: negative (orange) versus positive (light gray) and ambiguous (dark gray). (*D*) Social (orange) versus nonsocial (gray) information in experiment 3. (*E*) Threat-related (orange) versus negative (dark gray) and neutral (light gray) in experiment 4. (*F*) Counterintuitive biological, social and negative information (orange) versus other biases (gray) in experiment 5. All data are average of five replications, bars show SDs.

As such, we can expect that when LLMs generate new texts or are used to summarize preexisting text, their outputs will reflect these biases. It is important to note that the concept of bias we are using in this research is different from the common concept of "algorithmic bias," which has a distinct negative connotation (17). In cultural evolution, transmission biases indicate that individuals are predisposed, on average, to adopt some cultural variants over others, affecting their overall spread within populations (18). In this sense, biases are neither inherently good nor bad, and they are to be expected each time individuals choose among different cultural variants. This is reflected in the experiments reported here, where some biases would be considered negative (e.g., the preference for stereotype-consistent information in experiment 1) but others neutral, or possibly functional, as the bias toward threat-related information in experiment 4. For the same reason, however, those biases could be more difficult to recognize, and they could have consequential downstream effects, by magnifying the preexisting human tendencies. We might anticipate, for example, that without human intervention, LLMs could enable negative gender stereotypes to persist in potentially harmful ways. Additionally, given concerns over emotional contagion in digital media (19), negativity, and threat bias in LLM-generated material could contribute to wider negativity and overestimation of threats in humans. If used to summarize scientific articles for the purposes of journalism (2), LLMs may focus on content which appeals to these biases rather than the truly pertinent, although not necessarily more so than a human editor. A key implication then is that it is important to recognize our own subtle biases and to understand that LLMs reflect these and do not act as neutral agents.

Research in cultural evolution suggests that content biases are present in humans as a result of biases in our cognition which lead us to preferentially attend to, recall, and/or transmit some types of information over others (10). In most cases, it is proposed that these biases are a result of evolved cognition, being on average adaptive, e.g., a bias for social information resulting from the fitness benefits of attending to and remembering such information within our social groups (14), or biases for negative and threat-related information because not attending to such is more costly than not attending to positive information or benefits (15). Biases in the outputs of LLMs cannot be the result of such evolutionary processes; however, the human-produced training material is itself a product of a cultural evolutionary process where human content biases have led to the preferential retention and dissemination of information which align with those biases. Such content then is likely present in that training material and therefore reflected in the "biases" of the LLM. In the case of ChatGPT-3, the training material was around 45 TB of text taken from multiple web-based sources including Wikipedia, books, and raw web page data (20). As such, the results of our study suggest that biases which have previously been tested experimentally, or within corpus analysis (21–24), can also be detected in the cultural artifacts used to train the LLM. It also suggests that the outputs of LLMs could be useful sources for studying broader human culture (although this will

depend significantly on how outputs are constrained by other processes).

While overall the results here are consistent with the outcomes of the original experiments, there are some interesting minor differences which likely reflect the differing mechanisms which produce the biased outputs. One such difference was the higher retention of "gossip" over standard social information. In the original study (14), no difference was found between gossip (defined as information about intense third-party social relationships) and nongossip social information, and both were equally well retained over nonsocial information about an individual or the physical environment. In our study (experiment 3), gossip had a significant advantage over standard social information (*SI Appendix*, Fig. S2), largely explaining the prominence of overall social information. This could be a result of the LLM having a stronger bias towards emotive social information than emotionally neutral social information, possibly particularly towards negatively valanced social information (as the gossip story in experiment 3 was mostly negative, see full text in *SI Appendix*). Relative importance of different biases, and how they may combine, is something which could be tested for directly in future studies.

A second difference was in the retention of stereotype-consistent information. In the original study (13), while stereotype-consistent information was preferentially retained across the overall chain, earlier chain steps showed an advantage for stereotype-inconsistent information. Later research suggested that a bias for stereotype-consistent information is a product of communicative intent, rather than memory (25). No such pattern was found in our study (experiment 1), rather, stereotype-consistent information was preferentially retained from the first chain step. This could be a consequence of the different mechanisms which produce biases in humans and LLMs. While content biases in humans may be present in attention, memory, or transmission, and may vary across these three phases (10), a bias in an LLM can only be a product of its training material. If the training material was predominantly stereotype consistent, this will be reflected in the outputs.

Similarly to the corresponding experiments with humans, most of the text transformation happened in the first step of the transmission chain. After that, the model converges on a broadly (but not completely) stable version of the story, which is repeated in the next two steps. On the one hand, this result implies that the transmission chain methodology is not necessary per se to produce the biases in the LLM, but a single process of summarization (which would be anyway the most common usage) is sufficient. On the other hand, it suggests the possibility of experimenting with different prompts. Our prompt required to summarize a text, producing a mostly subtractive transformation, i.e., information from the text is discarded, but future research could test prompts asking to elaborate the text, make it funny, or appealing for a particular audience, possibly producing more variation among chain steps.

The results of this study could be expanded by examining the potential impact of different prompts. Here, the prompt was "*Please summarize this story making sure to make it shorter, if necessary you can omit some information*." We did not directly test the influence of different prompt wording on output, and it is unlikely that this prompt would produce the hypothesized biases alone. It is virtually impossible, however, to know beforehand how slight changes in wording or syntactic structure in the prompt could influence the output of the model. Future research could examine how, and to what extent, prompt wording influences the reflection of biases in LLM output, by systematically testing variations of baseline prompts. Further, future research should test biases across multiple stories (as in experiment 5) and examine interactions between prompt wording and story content or structure in the reflection of biases.

A limitation of our study is that, due to the rapid development and diversification of LLMs, our results may not generalize beyond ChatGPT-3. ChatGPT-4's training was similar to ChatGPT-3, so may reflect the same biases but includes reinforcement learning using human feedback in an attempt to prevent the LLM from producing output which violates OpenAI's policy on harmful behavior and to mitigate harmful biases. However, small-scale experiments with early versions of ChatGPT-4 suggest that it still generates biased outcomes based on gender stereotypes (26). To what extent it would reflect this bias in summarizing text, as tested here, or the other biases examined here is unknown. Given the similarity in training, however, and that outside of gender stereotypes the biases tested here are unlikely to be considered harmful, it is plausible to expect that these biases would still be present in later generations of LLMs.

Materials and Methods

We used the 9 January 2023 version of OpenAI's ChatGPT-3 language model (publicly available at: https://chat.openai.com/), with default parameters values, to run a series of experiments using transmission chains methodology. We selected five studies (11, 13–16) that highlighted in human participants different content biases and that made use of single stories with different biases (as opposed to experiments that used two or more stories with different biases). In one case (study 3), we modified the original material, which consisted in four stories testing four different biases, creating a single story.

We used the same material (stories) of the original experiments (see *SI Appendix* for details and full texts). The material was presented in ChatGPT with the prompt:

*Please summarize this story making sure to make it shorter, if necessary you can omit some information: **story***

For each study, we run five different chains/replications, and each chain/replication consisted of three steps. In each chain, the original story was presented with the prompt above; the output produced by ChatGPT was then presented again with the same prompt (step 2), and the process was iterated a last time (step 3). (The setup is slightly different for study 4, see *SI Appendix*). The number of chains/replications and of steps was chosen after pretests showing a limited variability of ChatGPT's outputs given the same prompt, and that the main modification of the material was happening in the first step of the chain.

The general hypothesis we tested is that ChatGPT's output would produce the same biases found in human subjects. While the original studies use different statistical analyses, we decided to have the same general analytic strategy for all studies. We used linear mixed-effects models with the proportion of content retained as the outcome, the type of content as the predictor, and the step in the chain and the replication as random effects. Using the R package lme4, (12) the general formula can be written as:

$$\text{lmer}(\text{proportion} \sim \text{content} + (1|\text{chain\_step}) + (1|\text{chain\_id})).$$

The coding consisted in determining the presence or absence of basic information from the original story. In experiments 1, 3, and 5, where the information consistent with human bias was represented by more than one category of information (e.g., in experiment 1, the categories "plot-relevant male stereotype-consistent," "plot-relevant female stereotype-consistent," "background male stereotype-consistent," and "background female stereotype-consistent" were all part of the stereotype-consistent information predicted to be advantaged), a weighted average of the categories retained was considered (outputs of the single categories for each experiment are in *SI Appendix*). Similarly, a weighted average was considered when an experiment involved more than one story (experiments 4 and 5–outputs of single stories are in *SI Appendix*). ChatGPT's output was coded by A.A. (studies 1, 2, and 4) and by J.M.S. (studies 3 and 5). A third independent coder, unaware of the experimental procedure and of the predictions, double-coded studies 1, 2, 3, and 5 (study 4 did not need double coding as the procedure is slightly different–see *SI Appendix*). Interrater reliability was generally high, and it is reported in *SI Appendix*, Table S1.

Detailed hypotheses and implementations for each study as well as how the outputs were coded are reported in *SI Appendix*. The preregistration, plus all the original data, coding, and R scripts to perform the analysis and visualizations are available in an OSF (Open Science Framework) repository: https://osf.io/6v2ps/ (27).

Data, Materials, and Software Availability. The preregistration, plus all the original data, coding, and R scripts to perform the analysis and visualizations. Data have been deposited in ChatGPT transmission chains (https://osf.io/6v2ps/) (27).

ACKNOWLEDGMENTS. We would like to thank Fabiana Lombardi for her work as an independent coder, and two anonymous reviewers for comments and suggestions. J.M.S. received internal funding from the University of Winchester Research and Innovation funds for this research.

1. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
2. S. Petridis *et al.*, "AngleKindling: Supporting journalistic angle ideation with large language models" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI'23*, A. Schmidt *et al*., Eds. (Association for Computing Machinery, 2023), pp. 1–16.
3. G. Chen, P. Xie, J. Dong, T. Wang, Understanding programmatic creative: The role of AI. *J. Advert.* **48**, 347–355 (2019).
4. O. "Oz" Buruk, Academic writing with GPT-3.5: Reflections on practices, efficacy and transparency. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2304.11079 (Accessed 27 June 2023).
5. R. Dale, GPT-3: What's it good for? *Nat. Lang. Eng.* **27**, 113–118 (2021).
6. L. Lucy, D. Bamman, "Gender and representation bias in GPT-3 generated stories" in *Proceedings of the Third Workshop on Narrative Understanding*, N. Akoury *et al.*, Eds. (Association for Computational Linguistics, 2021), pp. 48–55.
7. F. C. Bartlett, *Remembering* (Cambridge University Press, 1932).
8. A. Mesoudi, A. Whiten, The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philos. Trans. R. Soc. London Ser. B* **363**, 3489–3501 (2008).
9. B. Thompson, B. van Opheusden, T. Sumers, T. L. Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).
10. J. M. Stubbersfield, Content biases in three phases of cultural transmission: A review. *Cult. Evol.* **19**, 41–60 (2022).
11. K. Bebbington, C. MacLeod, T. M. Ellison, N. Fay, The sky is falling: evidence of a negativity bias in the social transmission of information. *Evol. Hum. Behav.* **38**, 92–101 (2017).
12. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
13. Y. Kashima, Maintaining cultural stereotypes in the serial reproduction of narratives. *Pers. Soc. Psychol. Bull.* **26**, 594–604 (2000).
14. A. Mesoudi, A. Whiten, R. Dunbar, A bias for social information in human cultural transmission. *Br. J. Psychol.* **97**, 405–423 (2006).
15. T. Blaine, P. Boyer, Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evol. Hum. Behav.* **39**, 67–75 (2018).
16. R. E. W. Berl, A. N. Samarasinghe, S. G. Roberts, F. M. Jordan, M. C. Gavin, Prestige and content biases together shape the cultural transmission of narratives. *Evol. Hum. Sci.* **3**, e42 (2021).
17. D. Danks, A. J. London, "Algorithmic bias in autonomous systems" in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, C. Sierra, Ed. (International Joint Conferences on Artificial Intelligence Organization, 2017), pp. 4691–4697.
18. R. Boyd, P. J. Richerson, *Culture and the Evolutionary Process* (University of Chicago Press, 1985).
19. A. Goldenberg, J. J. Gross, Digital emotion contagion. *Trends Cogn.* **24**, 316–328 (2020).
20. T. Brown *et al.*, "Language models are few-shot learners" in *Advances in Neural Information Processing Systems*, H. Larochelle, M.'A. Ranzato, R. Hadsell, M.-F. Balcan, H.-T. Lin, Eds. (Curran Associates Inc., 2020), pp. 1877–1901.
21. J. M. Stubbersfield, E. G. Flynn, J. J. Tehrani, Cognitive evolution and the transmission of popular narratives: A literature review and application to urban legends. *Evol. Stud. Imaginative Culture* **1**, 121–136 (2017).
22. A. Acerbi, Cognitive attraction and online misinformation. *Palgrave Commun.* **5**, 15 (2019).
23. O. Morin, A. Acerbi, Birth of the cool: A two-centuries decline in emotional expression in Anglophone fiction. *Cogn. Emot.* **31**, 1663–1675 (2017).
24. C. O. Brand, A. Acerbi, A. Mesoudi, Cultural evolution of emotional expression in 50 years of song lyrics. *Evol. Hum. Sci.* **1**, e11 (2019).
25. A. Lyons, Y. Kashima, Maintaining stereotypes in communication: Investigating memory biases and coherence-seeking in storytelling. *Asian J. Soc. Psychol.* **9**, 59–71 (2006).
26. S. Bubeck *et al.*, Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.12712 (Accessed 30 June 2023).
27. A. Acerbi, J. M. Stubbersfield, chatGPT transmission chains. chatGPT transmission chains. https://osf.io/6v2ps/. Deposited 17 July 2023.