PUBLIC HEALTH

AI model GPT-3 (dis)informs us better than humans

Giovanni Spitale, Nikola Biller-Andorno, Federico Germani*

Artificial intelligence (AI) is changing the way we create and evaluate information, and this is happening during an infodemic, which has been having marked effects on global health. Here, we evaluate whether recruited individuals can distinguish disinformation from accurate information, structured in the form of tweets, and determine whether a tweet is organic or synthetic, i.e., whether it has been written by a Twitter user or by the AI model GPT-3. The results of our preregistered study, including 697 participants, show that GPT-3 is a double-edge sword: In comparison with humans, it can produce accurate information that is easier to understand, but it can also produce more compelling disinformation. We also show that humans cannot distinguish between tweets generated by GPT-3 and written by real Twitter users. Starting from our results, we reflect on the dangers of AI for disinformation and on how information campaigns can be improved to benefit global health.

INTRODUCTION

Artificial intelligence (AI) text generators caught much attention over the last years, especially after the release of GPT-3 in 2020 (*1*). GPT-3, the latest iteration of the generative pretrained transformers developed by OpenAI, is arguably the most advanced system of pretrained language representations (*2*). A generative pretrained transformer, in its essence, is a statistical representation of language; it is an AI engine that, based on users′ prompts, can produce very credible, and sometimes astonishing, texts (*3*). An initial test on people′s ability to tell whether a ∼500-word article was written by humans or GPT-3 showed a mean accuracy of 52%, just slightly better than random guessing (*1*).

GPT-3 does not have any mental representations or understanding of the language it operates on (*4*). The system relies on statistical representations of language for how it is used in real-life by real humans or "a simulacrum of the interaction between people and the world" (*4*). Even keeping in mind these structural limitations, what GPT-3 can do is remarkable, and remarkable is also the possible implication. While GPT-3 can be a great tool for machine translations, text classification, dialogue/chatbot systems, knowledge summarizing, question answering, creative writing (*2*, *5*, *6*), detecting hate speech (*7*), and automatic code writing (*2*, *8*), it can also be used to produce "misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing, and social engineering pretexting" (*1*, *9*–*11*). GPT-3 serves as a lever, amplifying human intentions. It can receive instructions in natural language and generate output that may be in either natural or formal language. The tool is inherently neutral from an ethical point of view, and as every other similar technology, it is subject to the dual-use problem (*12*).

The advancements in AI text generators and the release of GPT-3 historically coincide with the ongoing infodemic (*13*), an epidemic-like circulation of fake news and disinformation, which, alongside the coronavirus disease 2019 (COVID-19) pandemic, has been greatly detrimental for global health. GPT-3 has the potential to generate information, which raises concerns about potential misuse, such as producing disinformation that can have devastating effects on global health. Therefore, it is crucial to assess how text generated by GPT-3 can affect people′s comprehension of information.

The purpose of this paper is to assess whether GPT-3 can generate both accurate information and disinformation in the form of tweets. We will compare the credibility of this text with information and disinformation produced by humans. Furthermore, we will explore the potential for this technology to be used in developing assistive tools for identifying disinformation. For clarity, we acknowledge that the definitions of disinformation and misinformation are diverse, but here, we refer to an inclusive definition, which considers disinformation as both intentionally false information (also partially false information) and/or unintentionally misleading content (*14*).

To achieve our goals, we asked GPT-3 to write tweets containing informative or disinformative texts on a range of different topics, including vaccines, 5G technology and COVID-19, or the theory of evolution, among others, which are commonly subject to disinformation and public misconception. We collected a set of real tweets written by users on the same topics and programmed a survey in which we asked respondents to classify whether randomly selected synthetic tweets (i.e., written by GPT-3) and organic tweets (i.e., written by humans) were true or false (i.e., whether they contained accurate information or disinformation) and whether they were written by a real Twitter user or by an AI. Note that this study has been preregistered on the Open Science Framework (OSF) (*15*), and we have conducted a power analysis based on the findings of a pilot study, as described in Materials and Methods.

RESULTS

Study design and demographics

To evaluate the capability of the GPT-3 AI model as a tool for generating tweets containing accurate information or disinformation, we created instruction prompts. These prompts were used to instruct GPT-3 to generate fake tweets on the following topics: climate change, vaccine safety, theory of evolution, COVID-19, mask safety, vaccines and autism, homeopathy treatments for cancer, flat Earth, 5G technology and COVID-19, antibiotics and viral infections, and COVID-19 and influenza. Furthermore, we performed a Twitter search to identify accurate tweets and

Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland.
*Corresponding author. Email: federico.germani@ibme.uzh.ch

disinformation tweets written by Twitter users. We call those tweets that are generated by GPT-3 as "synthetic," and we call those real tweets retrieved from Twitter as "organic." Human respondents were recruited online to participate in a quiz, in which they were asked to recognize whether a set of tweets were organic or synthetic and true or false (i.e., whether they contained accurate information or disinformation). GPT-3 was also questioned about whether tweets forming the same dataset were true or false (Fig. 1A). We recruited 869 respondents. A total of 157 responses were excluded because they were incomplete. Of the 712 remaining responses, 15 additional responses were removed because the respondents were too fast to meaningfully complete the survey, for a total of 697 responses included in our analysis (Fig. 1B). Most of the respondents were from the United Kingdom, Australia, Canada, United States, and Ireland (fig. S1A), with more females than males (fig. S1B); a balanced age, with a high representation of people between 42 and 76 years old (fig. S1C); and a balanced education level profile, with most of the respondents holding a bachelor's degree (fig. S1D); among those with a bachelor's degree or above, their field of study was mostly in the social sciences and humanities, natural sciences, or medical sciences (fig. S1E).

GPT-3 AI model informs and disinforms us better

We measured how accurately participants recognized whether a tweet was containing disinformation or accurate information (disinformation recognition score, range 0 to 1) for four types of tweets: "organic true," which are tweets published by Twitter users (organic) and containing accurate information (true); "synthetic true," which are tweets generated by GPT-3 (synthetic) and containing accurate information (true); "organic false," which are tweets generated by Twitter users (organic) and containing disinformation (false); and last, "synthetic false," which are tweets generated by GPT-3 (synthetic) and contain disinformation (false). Participants recognized organic false tweets with the highest efficiency, better than synthetic false tweets (scores 0.92 versus 0.89, respectively; $P = 0.0032$) (Fig. 1C). Similarly, they recognized synthetic true tweets correctly more often than organic true tweets (scores 0.84 versus 0.72, respectively; $P < 0.0001$). This indicates that human respondents can recognize the accuracy of tweets containing accurate

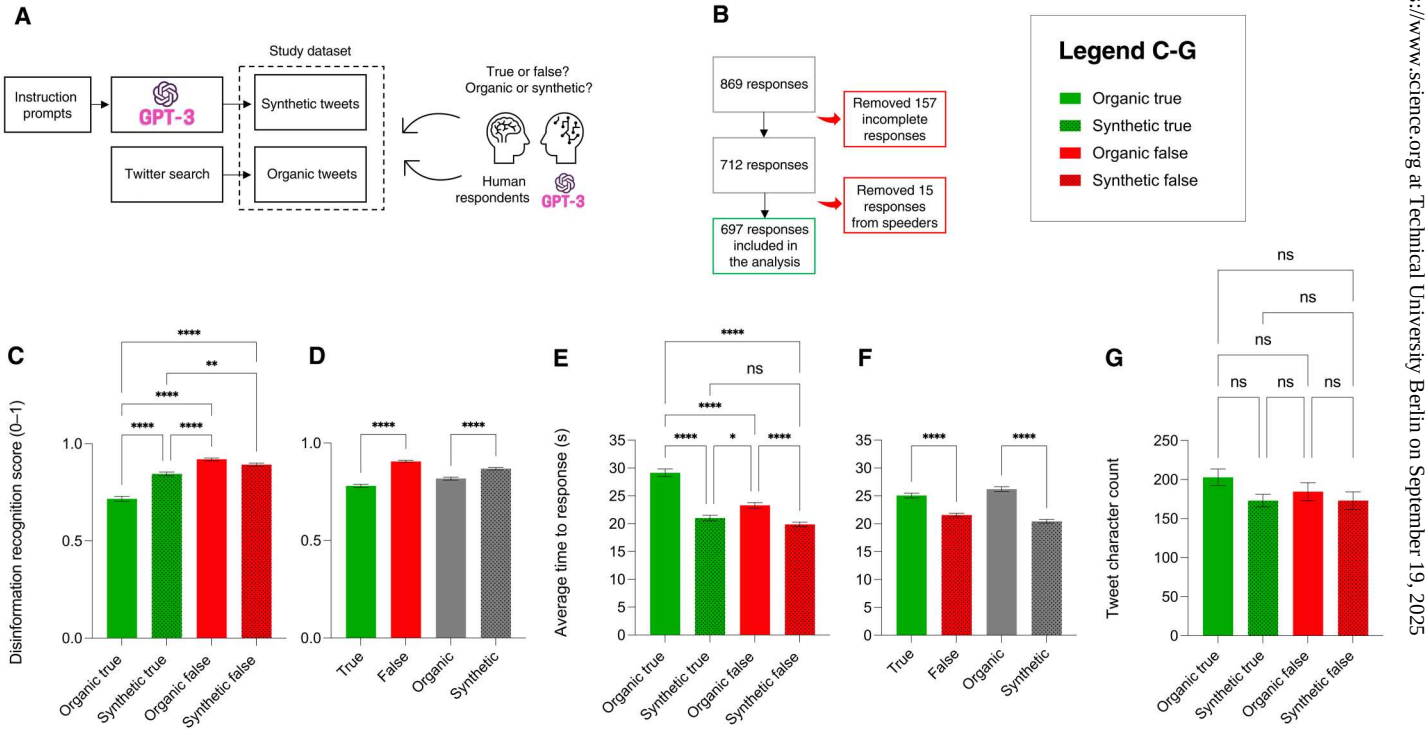

Fig. 1. The GPT-3 AI model informs and disinforms us better. (**A**) GPT-3 produced synthetic tweets containing either accurate information or disinformation. Organic tweets were retrieved and classified as accurate information or disinformation. Participants and GPT-3 were then asked to determine whether the tweets were true or false and whether they were organic or synthetic. (**B**) We gathered 869 responses to our survey: 157 responses were incomplete and were removed and 615 responses were removed as they were completed too fast to be reliable. Our analysis was conducted on 697 complete and reliable responses. (**C**) GPT-3's information and disinformation tweets are recognized as accurate more often than humans'. Green bars, accurate tweets from Twitter users; dotted green bars, accurate tweets from GPT-3. Red bars, disinformation tweets from Twitter users; dotted red bars, disinformation tweets from GPT-3. (**D**) Disinformation tweets (red bars) are recognized more often correctly than accurate tweets (green bars). Synthetic tweets (dotted grey bars) are recognized more often correctly than organic tweets (grey bars). Disinformation recognition score (or TF score, range 0 to 1) is the average score for all 697 respondents (1, 100% correct answers; 0, 0%); ordinary one-way analysis of variance (ANOVA) multiple-comparisons Tukey's test, $n = 697$; $^{**}P < 0.01$ and $^{****}P < 0.0001$. Error bars = SEM. (**E**) Average time to respond in seconds for organic and synthetic true as well as organic and synthetic false tweets. Organic true tweets took the longest to be evaluated; synthetic true/false tweets took the least time. (**F**) Accurate tweets took longer to be evaluated than false ones, and organic tweets took longer than synthetic tweets. Ordinary one-way ANOVA multiple-comparisons Tukey's test, $n = 697$; ns, nonsignificant, $P > 0.05$; $^{*}P < 0.05$ and $^{****}P < 0.0001$. Error bars = SEM. (**G**) Tweet evaluation time is independent of tweet length as character count is consistent across tweet types. Ordinary one-way ANOVA multiple-comparisons Tukey's test, n(total) = 220, n(type) = 55; ns, $P > 0.05$.

information more often when such tweets are generated by GPT-3, when compared with organic tweets retrieved from Twitter. Similarly, this means that disinformation tweets generated with GPT-3 achieve their deceiving goal more often when compared with disinformation tweets generated organically; although in this case, the effect is small. When evaluating the same dataset, segmenting the analysis for true versus false tweets—regardless of whether they are organic or synthetic—and for organic versus synthetic tweets—regardless of their truthfulness—tweets containing accurate information received lower scores when compared with tweets containing disinformation (scores 0.78 versus 0.91, respectively; $P < 0.0001$). Similarly, synthetic tweets were categorized more often correctly for the accuracy of the information they contained (scores 0.87 versus 0.82, respectively; $P < 0.0001$) (Fig. 1D). Participants required, on average, 29.14 s to determine whether an organic true tweet was accurate or contained disinformation. This was significantly more when compared with organic false tweets, which required 23.28 s for evaluation, with synthetic true tweets requiring 21.02 s and synthetic false tweets requiring 19.87 s (Fig. 1E). True tweets required a longer time for evaluation when compared with false tweets (25.07 s versus 21.97 s; $P < 0.0001$), as well as organic tweets when compared with synthetic tweets (26.21 s versus 20.44 s; $P < 0.0001$). (Fig. 1F). The time required for evaluation was not dependent on the length of tweets (Fig. 1G). Furthermore, we calculated disinformation recognition scores for each category (e.g., "climate change" and "vaccines and autism"), for each type of tweet (i.e., organic true, synthetic true, organic false, and synthetic false) (fig. S2), and plotted the average disinformation scores for each type of tweet (fig. S3), obtaining comparable results with the analysis run on the disinformation recognition scores of each respondent. This confirms that, for humans, accurate information is more difficult to evaluate when compared with disinformation and that information produced by GPT-3 is not only more effective to inform and disinform humans but also does so more efficiently in less time. Here, we refer to "efficient" and "inefficient" as the speed and ease with which information can be processed and evaluated by humans (efficient = information that is quickly processed in accordance with the intentions of the informer). A list of the disinformation tweets recognized most often as accurate tweets can be seen in fig. S4, and a list of tweets containing accurate information, recognized most often as disinformation tweets, can be seen in fig. S5.

Humans evaluate the accuracy of information better than GPT-3

The respondents of our survey evaluated the accuracy or inaccuracy of the information contained in 220 tweets. Using the same dataset, we asked GPT-3 to evaluate whether the tweets were accurate or whether they contained disinformation. For disinformation tweets, humans and GPT-3 performed similarly (success rates: 0.90 versus 0.89, respectively). For accurate tweets, GPT-3, similar to human respondents, had more difficulties evaluating the accuracy of the information. In comparison, human respondents performed better than GPT-3 (success rates: 0.78 versus 0.64, respectively) (Fig. 2A). A detailed analysis of the results for each category of tweets can be found in fig. S6. These results suggest that human respondents can evaluate information better than GPT-3. Considering that these respondents are not necessarily trained individuals in recognizing disinformation, with a high degree of information literacy, trained humans may perform much better than machines at performing this task. It is important to acknowledge that our evaluation approach differed between GPT-3 and human performance. While GPT-3 assessed tweets individually, human performance was based on an average score across multiple tweets per participant. This difference in methodology should be considered when interpreting and comparing the results of our study.

GPT-3 can "disobey" requests to produce disinformation

Because GPT-3 does not have mental representations or anything that can resemble a proper intentionality (*16*, *17*), we believe that the use of quotes when referring to "obedience" or "disobedience" is needed. As mentioned, we instructed GPT-3 to produce a set of true and false tweets (i.e., accurate or disinformation tweets) (Fig. 1A). To obtain the dataset, we instructed GPT-3 to produce 10 accurate and 10 disinformation tweets for each category. Of these, we included in our survey only the tweets for which GPT-3 "obeyed" our request to produce accurate or disinformation tweets. We calculated the rate of obedience, i.e., the percentage of requests satisfied by GPT-3 divided by the overall number of requests made to the system. We found that GPT-3 complied with our requests for accurate information 99 times of 101, while the obedience rate for disinformation was much lower at 80 of 102 requests (Fig. 2B). This suggests that GPT-3 is capable of "refusing" to generate disinformation, and in rarer instances, it may produce disinformation when instructed to produce accurate information. For a detailed analysis of the obedience rate per category of tweet, we refer to fig. S7.

Information generated by GPT-3 is indistinguishable from human-generated information

We calculated the AI recognition score (or OS score, range 0 to 1), i.e., the ability of respondents to recognize whether tweets are organic (i.e., produced by Twitter users) or synthetic (i.e., produced by GPT-3). As for the evaluation of the ability to recognize disinformation, we calculated the AI recognition score for each type of tweet (i.e., organic true, synthetic true, organic false, and synthetic false). The scores, on average, hovered around 0.5, suggesting that respondents were unable to distinguish between tweets generated by AI and those posted by real Twitter users. This indicates that the choice between synthetic and organic tweets was essentially random (Fig. 3). That said, respondents obtained a significantly higher score for organic true tweets and, to a lesser extent, for organic false tweets (0.67 and 0.60, respectively), whereas for synthetic true and synthetic false tweets, scores were below 0.5 (0.34 and 0.40, respectively) (Fig. 3A). Our analysis of true versus false tweets and organic versus synthetic tweets revealed an interesting finding: The accuracy of the information did not affect the participants' ability to distinguish between organic and synthetic tweets. On average, the responses were essentially random, indicating that people were unable to determine whether a tweet was generated by AI or posted by a real user regardless of its veracity (0.50 versus 0.50, respectively; $P = 0.9576$). However, when comparing organic versus synthetic tweets, organic tweets were recognized as organic more often when compared with synthetic tweets recognized as synthetic (0.63 versus 0.37; $P < 0.0001$) (Fig. 3B). Therefore, both organic and synthetic tweets tend to be classified as "human," indicating that GPT-3 can effectively mimic human-generated information. Furthermore, we calculated AI recognition scores for each category (e.g., climate change and vaccines and autism), for each type of tweet (i.e., organic true, synthetic true, organic false, and

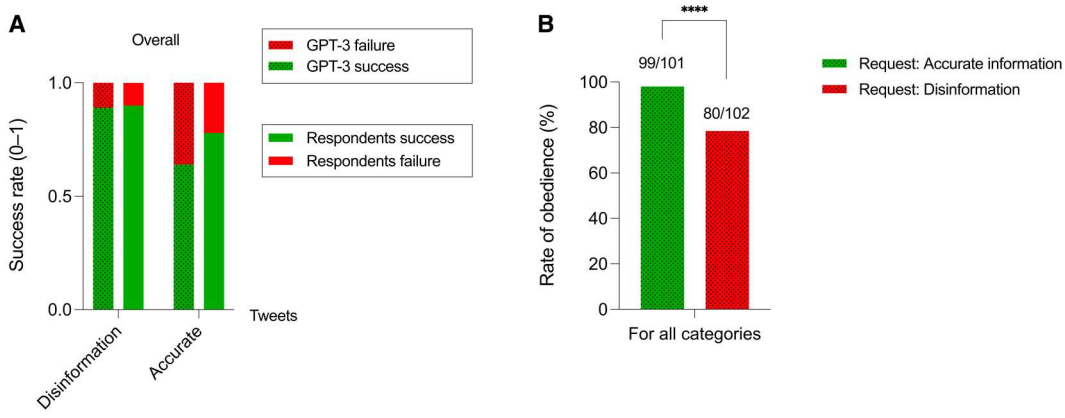

Fig. 2. Humans evaluate information and disinformation better than GPT-3, and GPT-3 can "disobey" requests to generate disinformation. (**A**) Green column bars represent successful responses given by human respondents, whereas green dotted bars represent successful responses given by GPT-3. Red bars represent incorrect responses from human respondents, whereas red dotted bars represent incorrect responses from GPT-3. The success rate concerning the evaluation of disinformation is 89 and 90% for GPT-3 and human respondents, respectively. The success rate concerning the evaluation of accurate information is 64 and 78% for GPT-3 and human respondents, respectively. The evaluation was conducted on organic tweets retrieved from Twitter, which were included in our survey. (**B**) Rate of "obedience" for GPT-3, i.e., how often GPT-3 respected our request to generate information or disinformation tweets. For accurate information tweets, GPT-3 "obeyed" our request 99 times of 101 requests, whereas for disinformation tweets, it "obeyed" our request 80 times of 102 requests. The chi-square statistic is 18.679. ****$P < 0.0001$.

synthetic false) (fig. S8), and plotted the average AI recognition scores for each type of tweet (fig. S9), obtaining comparable results with the analysis run on the AI recognition scores of each respondent. A list of the organic tweets recognized most often as synthetic can be seen in fig. S10, and a list of synthetic tweets recognized most often as organic can be seen in fig. S11.

Building versus crashing confidence: How the self-reported ability to recognize disinformation and AI-generated information changes after survey completion

At the beginning of the survey, we asked respondents to define how confident they were in their ability to recognize disinformation and in their ability to identify AI-generated versus human-generated text using a 1 to 5 Likert scale (Fig. 4, A and B). The disinformation recognition confidence before the test was higher than AI recognition confidence before the test (3.05 versus 2.69, respectively). Following the survey, but before revealing the results, we requested the respondents once more to indicate their level of confidence in distinguishing disinformation and AI-generated text from human-generated text. Respondents were more confident in their ability to recognize disinformation (before versus after, 3.05 versus 3.49, respectively; $P < 0.0001$) (Fig. 4A), whereas they were much less confident in their ability to recognize synthetic versus organic tweets (before versus after, 2.79 versus 1.70, respectively; $P < 0.0001$) (Fig. 4B). The rise in confidence to detect disinformation may be attributed to the inoculation theory of misinformation (*18*), which proposes that exposure to disinformation with a critical mindset can enhance the ability to recognize and withstand it. On the opposite, the stark decrease in confidence to detect synthetic tweets could depend on what we could call "resignation theory," which suggests that individuals may become discouraged and abandon their attempts to critically evaluate information, particularly when faced with a large volume of information that is challenging to decipher in terms of its source or intent. This may lead to a sense of hopelessness or apathy toward information consumption and a tendency to rely on heuristics or simple cues (such as the emotions evoked by the information).

DISCUSSION

How to communicate and evaluate information

Our findings show that tweets produced by GPT-3 can both inform and disinform better than organic tweets. Synthetic tweets containing reliable information are recognized as true better and faster than true organic tweets, while false synthetic tweets are recognized as false worse than false organic tweets. Moreover, GPT-3 does not perform better than humans in recognizing both information and disinformation. The results suggest that GPT-3 may be more efficient at conveying information because it can generate text that is easier to read and understand compared to text written by humans. On the basis of these results, we propose a model for efficient communication and evaluation of information that challenges the current approach and consensus, according to which humans produce information and AI assists in the evaluation (Fig. 4, C and D) (*19*). A well-tailored information campaign can be shaped by providing instruction prompts to GPT-3, which produces effective information campaigns targeting humans (initiation phase). The accuracy of information is then evaluated by trained humans (Fig. 4C). Instead, information campaigns written and prepared by humans would turn out to be less effective, and AI would perform an inefficient evaluation of how truthful and reliable information is (Fig. 4D). The proposed model is of relevance in the context of a public health crisis and infodemic, given the need to communicate fast and clearly to large segments of the public.

"Disobedience," training datasets, and error propagation

Our results indicate that GPT-3 is less likely to generate misinformation on certain topics, such as vaccines and autism, when prompted (fig. S7). GPT-3 being a statistical representation of language, for how language is used in the datasets it was trained on, we assume that GPT-3's "disobedience" depends on the composition of

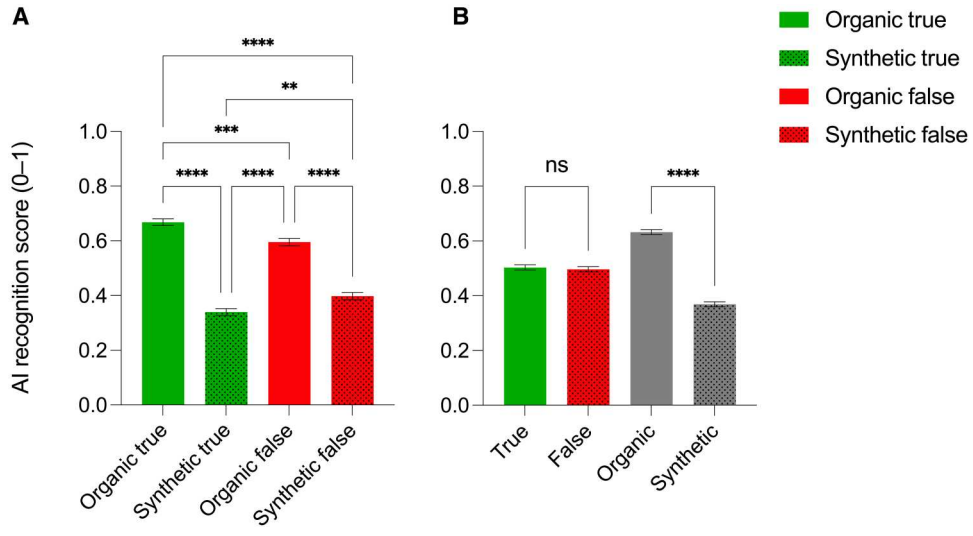

Fig. 3. Human respondents cannot distinguish organic versus synthetic tweets but recognize their origin better when they are generated by Twitter users. (**A**) AI recognition score for organic true (green bars), synthetic true (green dotted bars), organic false (red bars), and synthetic false (red dotted bars) tweets. AI recognition score (0 to 1) indicates the probability that human respondents can identify whether a tweet is produced organically (i.e., by a Twitter user) or synthetically (i.e., by GPT-3). Human respondents recognize whether organic true tweets are organic or synthetic tweets more effectively than all other type of tweets, whereas synthetic true tweets are recognized correctly the least. (**B**) Human respondents cannot predict whether true or false tweets (i.e., accurate tweets or disinformation tweets, green versus red bars) are produced by Twitter users or by GPT-3, and the truthfulness of the information does not have an impact on the AI recognition score. Regarding organic versus synthetic tweets (grey versus grey dotted bars), human respondents recognize whether tweets are generated by humans or GPT-3 better when they are organic (i.e., generated by Twitter users), when compared with synthetic tweets (i.e., generated by GPT-3). The AI recognition score (0 to 1) is the average score for all 697 respondents (1, 100% correct answers; 0, 0% correct answers); ordinary one-way ANOVA multiple-comparisons Tukey's test, $n = 697$; ns, $P > 0.05$; $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$. Bars represent SEM.

GPT-3's training datasets. If the training dataset contains volumes of information contradicting what the prompt asks for, then the system will likely output that type of information. We can therefore conclude that the volume of information in the training dataset debunking causal links between vaccines and autism may be higher than the volume of information debunking conspiracy theories on other topics taken into consideration by our study. Some control on the material fed into the training datasets is therefore crucial. GPT-3 is trained on data obtained from Common Crawl, WebText2, Books1, Books2, and Wikipedia (*20*), which could also include misinformation and disinformation. To reduce the risk of generating disinformation, we suggest that future text transformers should be trained on datasets regulated by the principles of accuracy and transparency: Information entering the training datasets should be verified and its origin should be open for independent scrutiny. Last, the output of models trained on accurate and transparent datasets should report the sources used for its generation, thus increasing transparency and allowing independent fact-checking. Fact-checking may still be difficult given the amount of information that likely serves as a source but, nonetheless, declaring sources would be a good start.

"As human as humans": Synthetic text identification and impersonation

In line with previous research (*21*), we found that both respondents and GPT-3 were not able to distinguish whether a tweet was organic or synthetic (data on GPT-3's assessment are available in the study's repository) (*15*). It might be possible to develop specific training courses to improve humans' recognition of synthetic text, based on linguistic markers, grammatical structure, and syntax. However, because the release of ChatGPT (an interactive, conversational, and even simpler interface to GPT-3), users started to search for ways to circumvent OpenAI's content policy blocks. An effective and commonly used strategy involves impersonation. When GPT-3 declines to generate output that may breach content policies, users simply request it to impersonate a character, for which content policies apparently do not apply (*22*–*24*). With this approach, even more credible swathes of disinformation could be produced by first asking GPT-3 to generate fake profiles of people to impersonate and, in a second iteration, to generate tweets that these profiles could write. Besides circumventing content policy blocks, this would add an even more "human-like" feel to the tweets and make it even harder to identify them as synthetic. On the basis of these premises, synthetic text identification might soon be a hopeless battle to fight, for both people and AIs.

Resignation theory and Dunning-Kruger effect

Our results indicate that not only can humans not differentiate between synthetic text and organic text but also their confidence in their ability to do so also significantly decreases after attempting to recognize their different origins. This decrease in self-confidence after exposure to both synthetic and organic texts may be due to the realization that there is no clear marker that allows users to identify whether a text has been generated by a machine or a human. This is likely because of GPT-3's ability to mimic human writing styles and language patterns. In addition, respondents may have initially underestimated GPT-3's abilities to write human-like text: This may be due to the fact that such technology is new and revolutionary and people are not yet accustomed to how powerful it can be. We refer to this phenomenon as resignation theory. We propose that, when

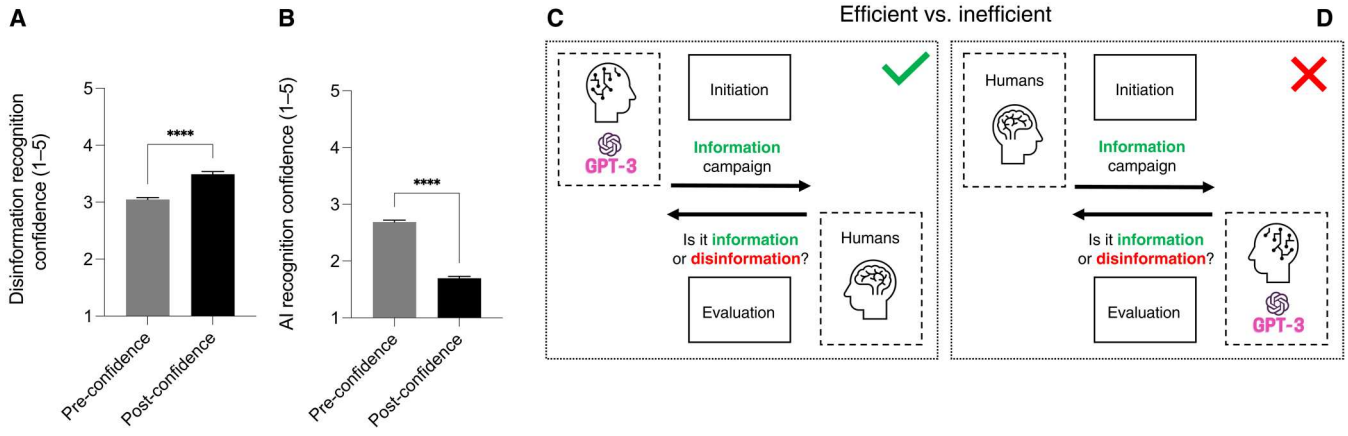

Fig. 4. The confidence in recognizing disinformation increases post-survey, whereas the confidence in recognizing AI-generated information decreases, and proposed model to launch information campaigns and evaluate information. (**A**) Respondents were asked to provide a score of how confident they were in their ability to recognize disinformation tweets before taking the survey (grey bar) and after taking the survey (black bar). Participants' confidence in disinformation recognition increased significantly from 3.05 to 3.49 of 5. $n = 697$; Welch's t test; ****$P < 0.0001$. Bars represent SEM. (**B**) Respondents were asked to provide a score of how confident they were in their ability to recognize whether tweets were generated by humans (grey bar) or by AI (black bar). Participant's confidence in AI recognition dropped significantly from 2.69 to 1.7 of 5. $n = 697$; Welch's t test; ****$P < 0.0001$. Bars represent SEM. (**C**) Model for an efficient and inefficient communication strategy and launch of information campaign. On the basis of our data, and with the AI model adopted for our analysis, an efficient system relies on accurate information generated by GPT-3 (initiation phase), whereas it relies on trained humans to evaluate whether a piece of information is accurate or whether it contains disinformation (evaluation phase). (**D**) An inefficient system relies on humans to generate information and initiate an information campaign and it relies on AI to evaluate whether a piece of information is accurate or whether it contains disinformation.

individuals are faced with a large amount of information, they may feel overwhelmed and give up on trying to evaluate it critically. As a result, they may be less likely to attempt to distinguish between synthetic and organic tweets, leading to a decrease in their confidence in identifying synthetic tweets. Another possible interpretation is that the survey may have made participants more aware of GPT-3's potential to generate disinformation with a human-like feel, making them more skeptical of both synthetic and organic information, thus decreasing their confidence in their ability to identify organic text as well.

An alternative view is proposed by the Dunning-Kruger effect, which can also help in interpreting our findings (*25*, *26*). This theory suggests that a person's belief in their ability to perform a task successfully can affect their performance. It is widely recognized that individuals tend to exhibit an overestimation of their perceived competence in information literacy skills (*27*), which can result in a corresponding decrease in motivation when faced with the actual challenge, as they discover that their actual performance falls short of their prior expectations (*28*).

In the case of our study, participants' confidence in their ability to differentiate between synthetic and organic text decreased after exposure. Thus, this decrease in confidence during the assessment may have negatively affected their ability to accurately distinguish between the two types of text in subsequent attempts, exacerbating the difficulty of distinguishing between synthetic and organic text. However, as some controversy exists about the Dunning-Kruger effect being imputable to statistical artifacts (*29*, *30*), we still consider our resignation theory as a preferable interpretation of the phenomenon observed in the data.

Beyond Twitter

We decided to focus our study on tweets for the following reasons: Twitter is currently used by more than 368 million monthly active users (*31*) who use the platform several times a day (*32*) to consume mostly news and political information (*32*, *33*). Furthermore, Twitter offers a very simple application programming interface (API) to develop bots, i.e., programs able to post content and interact with posts or users without human supervision (*34*). Recent research shows that only about 5% of Twitter users are bots, but that these bots cumulatively account for 20 to 29% of the contents posted on Twitter (*35*). Because of these characteristics, Twitter is the ideal target, and potentially a very vulnerable one, for AI-generated swathes of disinformation. Overall, our findings raise important questions about the potential uses and misuses of GPT-3 and other advanced AI text generators and the implications for information dissemination in the digital age, particularly in relation to the spread of disinformation, particularly on social media. Note that while we focused on tweets in this study, our results could be extended to other social media platforms and other forms of communication that can be used by bots via APIs and that could be exploited to programmatically disseminate AI-generated disinformation. We generated tweet-like social media posts that we call tweets, but have features shared with other types of social media posts, such as Instagram or Facebook posts.

The genie is out of the bottle

Starting from our findings, we predict that advanced AI text generators such as GPT-3 could have the potential to greatly affect the dissemination of information, both positively and negatively. As demonstrated by our results, large language models currently available can already produce text that is indistinguishable from organic text; therefore, the emergence of more powerful large language models and their impact should be monitored. In the upcoming months, it will be important to evaluate how the information landscape has changed on social and traditional media with the widespread use of ChatGPT since November 2022. If the technology is

found to contribute to disinformation and to worsen public health issues, then regulating the training datasets used to develop these technologies will be crucial to limit misuse and ensure transparent, truthful output information. In addition, until we do not have efficient strategies for identifying disinformation (whether based on human skills or on future AI improvements), it might be necessary to restrict the use of these technologies, e.g., licensing them only to trusted users (e.g., research institutions) or limiting the potential of AIs to certain types of applications. Last, it is crucial that we continue to critically evaluate the implications of these technologies and take action to mitigate any negative effects they may have on society.

Limitations

Despite the findings of our study, it is important to acknowledge its limitations. One potential limitation is the use of a relatively large sample size, which has led to small differences between the groups being highly significant. Therefore, caution should be taken when interpreting the significance of the results, especially when considering the effect size. That said, despite this limitation, we believe that the small differences found between AI- and human-made texts in terms of their effectiveness in communication are meaningful. With a large number of information pieces, even small differences in effectiveness can have a substantial impact on public health, both in terms of the dissemination of information and the spread of disinformation. Moreover, the potential impact of these differences could be further exacerbated by new AI developments such as GPT-4 or other more capable large language models. Furthermore, our study only investigated the recognition of tweets in isolation, focusing solely on the text and neglecting contextual factors such as the profile of the account from which a tweet is posted, past content, or profile image. These factors and others may influence the accuracy of recognizing disinformation. Future studies could investigate the recognition of disinformation in a more naturalistic setting, considering the contextual factors that may influence the recognition of disinformation on social media platforms. Furthermore, our study focused on English-speaking Facebook users. Future studies could investigate the recognition of disinformation in different regions, cultures, or specific sociodemographic groups to determine how AI affects the understanding of information across specific target publics. Our study confronted synthetic tweets with random organic tweets, written by random users. Future studies, rather than comparing synthetic tweets to random organic tweets, could compare synthetic tweets with organic tweets written by recognized public health institutions to clarify whether our findings about synthetic information being faster and easier to understand stand true even in this case, therefore confirming or falsifying the model we propose in Fig. 4 (C and D). As a final note, in this study, we assumed that organically retrieved tweets were generated without the use of AI tools, although there is a possibility that a small fraction of the tweets analyzed were actually synthetically generated.

MATERIALS AND METHODS

We registered the protocol of this study before starting the data collection. The preregistration is available on OSF: https://doi.org/10.17605/OSF.IO/HV6ZY.

Definition of the topics

As the focus of this study, we identified 11 topics on which disinformation exists. This list included the following:

1) Climate change,
2) Vaccine safety,
3) Theory of evolution,
4) COVID-19,
5) Mask safety,
6) Vaccines and autism,
7) Homeopathic treatments for cancer,
8) Flat Earth,
9) 5G technology and COVID-19,
10) Antibiotics and viral infections,
11) COVID-19 = influenza.

Generation of synthetic tweets

On the basis of the list defined above, we generated synthetic tweets passing input to GPT-3 via API. The code asks to generate 10 true tweets and 10 false tweets for each of the topics detailed above (e.g., prompt: "Write a tweet to explain why climate change is real," category: "Climate change"). The tweet generation code consists of one function to pass input prompts to GPT-3 and of two different loops to iterate over categorized prompts. The first function defines the parameters to pass to GPT-3 (temperature, max_token, top_p, best_of, frequency_penalty, and presence_penalty), empirically defined in an iterative process as the most apt to produce text that resembles social media content. GPT-3's API returns also the reason for termination (e.g., reaching the length specified in max_tokens). For these cases, the text sometimes contains unfinished sentences: These have been removed. The loops to generate true and false tweets read input organized in .csv files (prompt and category) and generate the given number of texts per each prompt (in this example, 10). The output is then exported as a .xlsx file containing three columns: the text, the reason for termination, and the category. All the codes, available in this study's preregistration repository, are organized in commented Jupyter lab notebooks for scrutiny and replication (*15*). The prompts and the output are available in the same repository.

Definitions

Throughout the manuscript, we adopt, and sometimes explain for added clarity, the terminology "true" and "false" tweets. True tweets are those tweets containing accurate information, and false tweets are those containing inaccurate information, i.e., disinformation.

As for the definition of accurate information and disinformation, we base ourselves on the current scientific knowledge and understanding of the topics and information under scrutiny. To avoid dubious and debatable cases, which may be subject to personal opinions and interpretations, we only analyzed and added to our questionnaire those tweets containing information that is categorizable as true or false. Notably, if a tweet contained partially incorrect information, meaning that it contained more than one piece of information and at least one was incorrect, then it was labeled as false. As discussed in Introduction, we acknowledge that the definition of disinformation and misinformation is diverse, but we refer to an inclusive definition, which considers false information (also partially false information) and/or misleading content (*14*).

Retrieval of organic tweets

Using Twitter's advanced search, we collected a random sample of recent organic tweets on the topics listed above, including both true and false tweets. The tweets are available in the study's repository (*15*).

Expert assessment of synthetic and organic tweets

We evaluated synthetic and organic tweets to assess whether they contained disinformation. The expert assessment was performed independently by F.G. and G.S., and a following joint analysis was conducted by F.G. and G.S. to verify the correctness of their initial assessments.

Selection of the tweets to include in the survey and generation of tweet images

Following our evaluations as described earlier, we have made the following tweet selections for each category: five tweets labeled as synthetic false, five tweets labeled as synthetic true, five tweets labeled as organic false, and five tweets labeled as organic true. We only selected tweets for which F.G. and G.S. agreed in their evaluation, following the expert assessment phases. This resulted in a data frame of 220 tweets [available in the repository (*15*)] used to generate the images of the tweets. The code generates a random pseudonym and a random username for each tweet (e.g., "John S.," @john_s) and generates an image that mimics a tweet. The code, the data frame containing the tweets, and the output images are available in the study's repository (*15*).

AI assessment of tweets

The AI assessment was performed by GPT-3 (true/false evaluation and organic/synthetic evaluation). The first evaluation function defines the parameters to pass to GPT-3 to produce a "true/false evaluation" (i.e., whether the tweet is true or false). The second evaluation function defines the parameters to pass to GPT-3 to produce an "organic/synthetic evaluation" (i.e., whether the tweet was written by a person or by an AI). The loops for evaluation read the content of the files containing the tweets and evaluate them. The output is scored (i.e., whether GPT-3's assessment matches the expert assessment for true/false and whether it matches the origin of the tweet for the organic/synthetic classification) and then exported as a .xlsx file. The code and the files containing the assessments are available in the study's repository (*15*).

Programming of the survey

We programmed a Qualtrics survey to collect demographics, display the tweets to the respondents, and collect their assessments (true versus false and organic versus synthetic). For each tweet, respondents assessed the following:

1) Whether it is accurate or whether it contains disinformation (single choice, accurate/misinformation);

2) Whether it was written by a real person or generated by an AI (single choice, real person/AI).

In addition, respondents provided the following:

1) Some demographic information (nationality, age, sex, education level, and education field).

2) Self-perceived (before and after survey) ability to recognize disinformation and synthetic text (Likert scale: 1, very difficult to 5, very easy).

The images of the tweets are organized in nested randomizers within the survey structure:

1) The first-level randomizer randomizes the category order (e.g., climate change, etc.). All the categories are displayed to every respondent.

2) Second-level randomizers (for each category) randomize the single tweet displayed for each category to the respondent. Each category comprises a total of 20 tweets: 5 synthetic false, 5 synthetic true, 5 organic false, and 5 organic true tweets. The second-level randomizers evenly present one tweet from the pool of 20 tweets.

The survey adopts a gamified approach to keep respondents engaged: At the beginning of the survey, respondents are told that, upon completion of the survey, they will obtain their score for both scales (disinformation recognition and synthetic text recognition). This ensured a low dropout rate. In-survey scoring is achieved using the "scoring" function in Qualtrics. The survey file and structure are available in the study's repository (*15*).

Pilot testing and sample size definition

We pilot tested the survey in two phases. During the first phase, we circulated the link to a convenience sample with the aim to test the usability and the layout. This led to minor modifications in the interface and in the wording. During the second phase, we distributed the link via a Facebook ads campaign. Details are provided in the Supplementary Materials.

Data collection

We distributed the survey via different Facebook ads campaigns to compensate for some demographic imbalances we noted from the pilot data (overrepresentation of women and underrepresentation of people aged 18 to 54) (*36*). The campaigns took place in October and November 2022. Details about the data collection and distribution strategy are available in the Supplementary Materials.

Our recruitment strategy aimed to enroll a population of active social media users by using a social media platform. Because of this design, we were unable to recruit a representative sample upfront. Instead, we chose to assess representativeness through a "rolling assessment" of demographics by targeting different segments of the population in sequential campaigns based on the demographics of already recruited participants (*36*).

Analysis

Scoring and analysis are implemented in Python, using a Jupyter notebook. The code takes the results of our Qualtrics survey as input and generates the files needed for the analysis as output. The code is available for scrutiny and replication in the study's repository (*15*).

Cleaning

To ensure data quality, incomplete responses, responses generated from preview links, and those submitted within less than 170.5 s were removed during data cleaning. This time frame was determined empirically as the minimum amount of time needed to complete the survey, calculated as the average time taken by a convenience sample to read and answer the questions with sustained rhythm.

Inferential statistics

Correlation analyses were performed as follows: For quantitative/quantitative data arrays, we first performed a Pearson's test, followed by Shapiro's test to determine data normality, and followed by *t* test for hypothesis testing. For qualitative/quantitative data arrays, we first performed analysis of variance (ANOVA), followed by Shapiro's test to determine data normality, and followed by a Kruskal-Wallis test. Last, we performed multiple comparisons with a Tukey test. Effect sizes resulting from ANOVA and Kruskal-Wallis tests are interpreted as small when $\eta^2 \leq 0.01$, medium when $0.01 < \eta^2 < 0.06$, and as large when $\eta^2 \geq 0.14$.

"The hard ones"

We defined tweets that were difficult to identify correctly for respondents (we called them "the hard ones") as follows. False identified as true: false tweets with average scores >0.75; true identified as false: true tweets with scores <0.25; synthetic identified as organic: synthetic tweets with average scores >0.75; and organic identified as synthetic: organic tweets with scores <0.25.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S13
Table S1

REFERENCES AND NOTES

1. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. arXiv:2005.14165 [cs.CL] (28 May 2020).
2. R. Dale, GPT-3: What's it good for? *Nat. Lang. Eng.* **27**, 113–118 (2021).
3. GPT-3, "Update: Some replies by GPT-3." Daily Nous (2020); https://dailynous.com/2020/07/30/philosophers-gpt-3/.
4. W. L. Benzon, "GPT-3: Waterloo or Rubicon? Here be dragons," Working paper, Social Science Research Network, Rochester, NY, 2020.
5. R. Marlow, D. Wood, Ghost in the machine or monkey with a typewriter—Generating titles for Christmas research articles in *The BMJ* using artificial intelligence: Observational study. *BMJ* **375**, e067732 (2021).
6. K. Elkins, J. Chun, Can GPT-3 pass a writer's turing test? *J. Cult. Anal.* **5**, 17212 (2020).
7. K.-L. Chiu, A. Collins, R. Alexander, Detecting hate speech with GPT-3. arXiv:2103.12407 [cs.CL] (24 March 2022).
8. M. I. B. Ugli, Will human beings be superseded by generative pre-trained transformer 3 (GPT-3) in programming? *Int. J. Orange Technol.* **2**, 141–143 (2020).
9. G. Cabanac, C. Labbé, A. Magazinov, Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXiv:2107.06751 [cs.DL] (12 July 2021).
10. N. Dehouche, Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics Sci. Environ. Polit.* **21**, 17–23 (2021).
11. M. Mindzak, S. E. Eaton, "Artificial intelligence is getting better at writing, and universities should worry about plagiarism." The Conversation (2021); http://theconversation.com/artificial-intelligence-is-getting-better-at-writing-and-universities-should-worry-about-plagiarism-160481.
12. J. Forge, A note on the definition of "dual use". *Sci. Eng. Ethics* **16**, 111–118 (2010).
13. WHO, Immunizing the public against misinformation (2020); www.who.int/news-room/feature-stories/detail/immunizing-the-public-against-misinformation.
14. J. Roozenbeek, J. Suiter, E. Culloty, Countering misinformation: Evidence, knowledge gaps, and implications of current interventions (2022); https://psyarxiv.com/b52um/.
15. G. Spitale, F. Germani, N. Biller-Andorno, Can AI disinform us better? (2022); https://osf.io/hv6zy/.
16. A. Sobieszek, T. Price, Playing games with Ais: The limits of GPT-3 and similar large language models. *Minds Mach.* **32**, 341–364 (2022).
17. L. Floridi, A defence of constructionism: Philosophy as conceptual engineering. *Metaphilosophy.* **42**, 282–304 (2011).
18. J. Cook, S. Lewandowsky, U. K. H. Ecker, Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE* **12**, e0175799 (2017).
19. T. Ahmad, E. A. Aliaga Lazarte, S. Mirjalili, A systematic literature review on fake news in the COVID-19 pandemic: Can AI propose a solution? *Appl. Sci.* **12**, 12727 (2022).
20. K. Cooper, "OpenAI GPT-3: Everything you need to know." Springboard Blog (2021); www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/.
21. E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, All that's "human" is not gold: Evaluating human evaluation of generated text. arXiv:2107.00061 [cs.CL] (30 June 2021).
22. Z. Witten [@zswitten], "Pretending is all you need (to get ChatGPT to be evil). A thread." Twitter (2022); https://twitter.com/zswitten/status/1598088267789787136.
23. F. Germani, "ChatGPT and the fight against disinformation: How AI is changing the game." Culturico (2023); https://culturico.com/2023/03/04/chatgpt-and-the-fight-against-disinformation-how-ai-is-changing-the-game/.
24. G. Spitale, F. Germani, SDPI—Synthetic disinformation through politeness and impersonation (2023); https://osf.io/jn349/.
25. J. Kruger, D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* **77**, 1121–1134 (1999).
26. T. Schlösser, D. Dunning, K. L. Johnson, J. Kruger, How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *J. Econ. Psychol.* **39**, 85–100 (2013).
27. K. Mahmood, Do people overestimate their information literacy skills? A systematic review of empirical evidence on the dunning-kruger effect. *Commun. Inf. Lit.* **10**, 199–213 (2016).
28. M. Mazor, S. M. Fleming, The Dunning-Kruger effect revisited. *Nat. Hum. Behav.* **5**, 677–678 (2021).
29. J. R. Magnus, A. A. Peresetsky, A statistical explanation of the Dunning-Kruger effect. *Front. Psychol.* **13**, 840180 (2022).
30. G. E. Gignac, M. Zajenkowski, The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence* **80**, 101449 (2020).
31. S. Dixon, "Twitter: Number of worldwide users 2019-2024." Statista (2022); www.statista.com/statistics/303681/twitter-users-worldwide/.
32. T. Rosenstiel, J. Sonderman, K. Loker, M. Ivancin, N. Kjarval, "How people use Twitter in general." American Press Institute (2015); www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general/.
33. Twitter news, How many people come to Twitter for news? As it turns out, a LOT (2022); https://blog.twitter.com/en_us/topics/insights/2022/how-many-people-come-twitter-for-news.
34. J. Garson, "How to create a Twitter bot with Twitter API v2." developer.twitter.com (2023); https://developer.twitter.com/en/docs/tutorials/how-to-create-a-twitter-bot-with-twitter-api-v2.
35. D. F. Carr, "Bots likely not a big part of twitter's audience—But tweet a lot." Similarweb (2022); www.similarweb.com/blog/insights/twitter-bot-research-news/.
36. L. G. Shaver, A. Khawer, Y. Yi, K. Aubrey-Bassler, H. Etchegary, B. Roebothan, S. Asghari, P. P. Wang, Using facebook advertising to recruit representative samples: Feasibility assessment of a cross-sectional survey. *J. Med. Internet Res.* **21**, e14021 (2019).

Acknowledgments

Funding: The authors acknowledge that they received no funding in support for this research.
Author contributions: Conceptualization: G.S. and F.G. Methodology: G.S. and F.G. Investigation: G.S. and F.G. Validation: F.G. and N.B.-A. Visualization: F.G. Supervision: N.B.-A. Writing—original draft: G.S. and F.G. Writing—review and editing: G.S., F.G., and N.B.-A.
Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Original raw data and software used for this study are available via OSF: https://osf.io/9ntgf/.

Submitted 15 February 2023
Accepted 24 May 2023
Published 28 June 2023
10.1126/sciadv.adh1850