Science Advances
AAAS

Supplementary Materials for

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

Giovanni Spitale *et al.*

Corresponding author: Federico Germani, federico.germani@ibme.uzh.ch

Sci. Adv. **9**, eadh1850 (2023)
DOI: 10.1126/sciadv.adh1850

This PDF file includes:

Methods - supplement

Pilot testing

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, structured as follows:

- Daily budget: 15€
- Start: 04.10.2022
- End: 13.10.2022
- Age: 16 - 65+
- Languages: English
- Title: True or False? Organic or synthetic?
- Description: Are you able to distinguish text written by an artificial intelligence from text written by a human being? And accurate information from misinformation? Find out with this test, and contribute to research on information ethics.
- Image: generated by DALL·E 2 (available in this study's repository)

The campaign had a total cost of 122.92€. It generated a total of 593 clicks on the link (cost per click: 0.21€) and a total of 276 responses (cost per response: 0.46€). The campaign was launched and completed in October 2022.

Sample size and power analysis

Based on pilot data, we conducted a power analysis to determine the sample size for the full study.

Primary endpoint hypothesis

Disinformation produced by a machine is more credible than disinformation produced by a human (synthetic versus organic disinformation).

Secondary endpoints hypotheses

1. Accurate information produced by a machine is more credible than accurate information produced by a human (synthetic versus organic accurate information).
2. Users recognize and distinguish information produced by humans and by machines (regardless of the truthfulness of the information).
3. The confidence of respondents in recognizing disinformation increases after the completion of the questionnaire.
4. The confidence of respondents in recognizing synthetic versus organic information increases after the completion of the questionnaire.

Power analysis

Based on the data resulting from the pilot study, available in the study's repository, we performed a power analysis to estimate the sample size necessary to draw sufficiently meaningful conclusions for Primary and Secondary Endpoints (PE and SEs). Endpoints are continuous, and the study runs on two independent samples.

Primary endpoint: Results

Average group 1 Score (Synthetic tweets, disinformation)* = 0.86

* From 0 to 1, the score indicates how good the performance was in recognizing synthetic tweets containing disinformation.

Stdev group 1 = 0.25

Average group 2 Score (Organic tweets, disinformation)* = 0.89

* From 0 to 1, the score indicates how good the performance was in recognizing organic tweets containing disinformation.

Enrollment ratio = 1.01194

Alpha = 0.05 Power = 80%

Sample Size Total = 2181 (Group 1: 1084; Group 2: 1097)

Secondary Endpoint 1

Average group 1 Score (Synthetic tweets, accurate information)* = 0.78

* From 0 to 1, the score indicates how good the performance was in recognizing synthetic tweets containing accurate information.

Stdev group 1 = 0.35

Average group 2 Score (Organic tweets, accurate information)* = 0.64

* From 0 to 1, the score indicates how good the performance was in recognizing organic tweets containing accurate information.

Enrollment ratio = 0.991045

Alpha = 0.05 Power = 80%

Sample Size Total = 197 (Group 1: 99; Group 2: 98)

Secondary Endpoint 2

Average group 1 Score (Synthetic tweets [accurate information + disinformation])* = 0.315

* From 0 to 1, the score indicates how good the performance was in recognizing synthetic tweets, regardless of whether they contained accurate information or disinformation.

Stdev group 1 = 0.44

Average group 2 Score (Organic tweets, [accurate information + disinformation])* = 0.59

* From 0 to 1, the score indicates how good the performance was in recognizing organic tweets, regardless of whether they contained accurate information or disinformation.

Enrollment ratio = 1.001493

Alpha = 0.05 Power = 80%

Sample Size Total = 80 (Group 1: 40; Group 2: 40)

Secondary Endpoint 3

Average group 1 Score (Pre-confidence level in ability to recognize disinformation)* = 2.932271

* From 1 to 5

Stdev group 1 = 0.829093

Average group 2 (Post-confidence level in ability to recognize disinformation)* = 3.319149

* From 1 to 5

Enrollment ratio = 1.0680

Alpha = 0.05 Power = 80%

Sample Size Total = 145 (Group 1: 70; Group 2: 75)

Secondary Endpoint 4

Average group 1 (Pre-confidence level in ability to recognize synthetic versus organic contents)* = 2.703557

* From 1 to 5

Stdev group 1 = 0.897012

Average group 2 (Post-confidence level in ability to recognize synthetic versus organic contents)* = 1.75

* From 1 to 5

Enrollment ratio = 1.0720

Alpha = 0.05 Power = 80%

Sample Size Total = 27 (Group 1: 13; Group 2: 14)

Sample size evaluation

Taking the larger sample size resulting from our power analyses (n=2181 assessments for PE), and considering that we obtained 1348 assessments (organic, disinformation + synthetic, disinformation), and considering that the pilot study has generated full responses from 277 respondents, the ratio between target power (number of assessments) and sample size of the pilot study (number of assessments) is 1.617953. Therefore, the number of users required for the study is 277*1.617953 = 448.1728. We established that the minimum number of respondents to achieve a properly powered analysis in the full study is n=449.

Data collection

We used a total budget of 492.24€, distributed as detailed in the following table:

Table S1.

Facebook dissemination campaign for data collection

Campaign	Age	Sex	Visualizations	Cost
USA, GBR, AUS, NZL, CAN	18-54	All	7226	35.22€
USA, GBR, AUS, NZL, CAN	16-65+	M	9907	34.24€
USA, GBR, AUS, NZL, CAN	16-65+	All	14710	33.78€
USA, GBR, AUS, NZL, CAN	16-25	M	83525	88.00€
USA, GBR, AUS, NZL, CAN	16-25	F	57780	44.00€
USA, GBR, AUS, NZL, CAN	26-41	M	8787	22.00€
USA, GBR, AUS, NZL, CAN	26-41	F	9544	31.00€
USA	26-41	F	21046	31.00€
USA	26-41	M	58146	93.00€
USA	16-25	All	99899	80.00€

Scoring

True/false and organic/synthetic scores of each respondent were calculated by the rules defined in Qualtrics' survey programming; furthermore, they were re-calculated using the dataframe containing the tweets and the expert assessments (available in the study's repository).

Supplementary results

Correlations between study variables

We evaluated whether any correlation between numerical and categorical variables in our analysis existed (Figure S12), as well as between numerical variables and other numerical variables (Figure S13).

OS score and demographics

We evaluated any potential correlation between OS Score and demographic variables, and identified the age of respondents to be a relevant factor, with a small effect size (Figure S12A). Younger individuals (18-41), seem to perform slightly better at recognizing synthetic versus human tweets when compared with very young individuals (16-17 years old), and especially older respondents (42+ years old) (Figure S12A').

TF score and demographics

Similarly, we evaluated potential correlations between TF score and demographic variables. As for the OS Score, also for the TF Score, age correlated with a small effect size, in addition to the education level of respondents (Figure S12B). In this case, 42-57 years old individuals performed slightly better when compared with older individuals aged 58 to 76, although the distribution of TF scores per age seems to be quite uniform across the board (Figure S12B'). As expected, a higher education level was associated with higher TF score. This effect was small but consistent: participants holding a doctorate/PhD degree had higher scores when compared with participants holding a Master's degree, and those with a Master's degree performed better than respondents with a Bachelor's degree, etc. (Figure S12B'').

Self-confidence and demographics

We evaluated the correlation between TF self-confidence PRE (i.e., the score of how confident respondents were in their ability to recognize disinformation before the survey) and demographic variables (Figure S12C); as well as the correlation between TF self-confidence POST (i.e., the score of how confident respondents were in their ability to recognize disinformation after the survey) and demographic variables (Figure S12D); and the correlation between OS self-confidence PRE (i.e., the score of how confident respondents were in their ability to distinguish synthetic versus organic tweets disinformation before the survey) and demographic variables (Figure S12E); and the correlation between OS self-confidence POST (i.e., the score of how confident respondents were in their ability to distinguish synthetic versus organic tweets disinformation after the survey) and demographic variables (Figure S12F).

OS / TF self-confidence delta and OS / TF score

For numerical versus numerical variables, we found no correlation between OS Delta (i.e., the difference in confidence POST versus PRE in the ability to recognize AI-generated text) and OS Score (Figure S13A), but we found a small but significant correlation between TF Delta (i.e., the difference in confidence POST versus PRE in the ability to recognize disinformation) and TF

Score (Figure S13B), suggesting that the higher the score, the more respondents built confidence in their abilities, despite participants were only shown how well they scored in the survey after evaluating their confidence level post-survey.

Duration and OS / TF scores

We found no significant correlation between duration (i.e., how long respondents took to complete the survey) and OS Score (Figure S13C), as well as between duration and TF Score (Figure S13D).

Supplementary figures

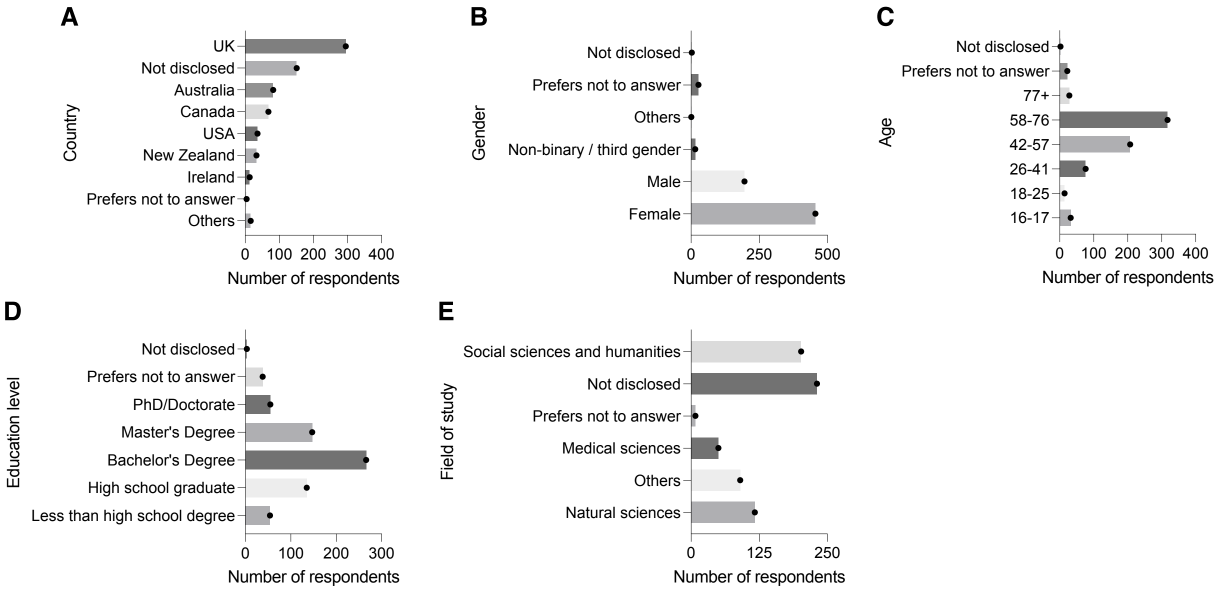


Fig. S1.

Demographics data. Demographics from the study (n=697); Country of origin of respondents (**A**), gender (**B**), age (**C**), education level (**D**), and, among those declaring at least a Bachelor's degree, the field of study (**E**).

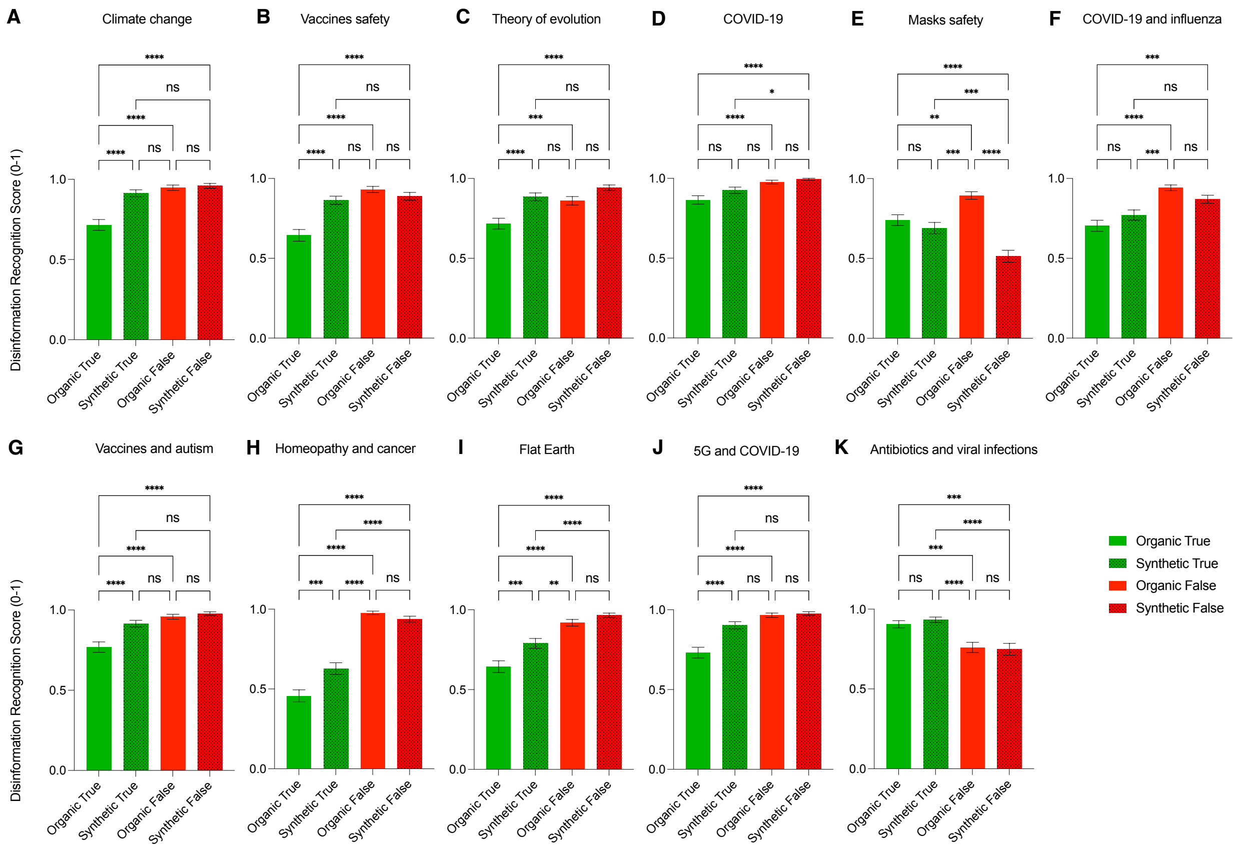


Fig. S2.

Disinformation Recognition Score per category of tweet. In the survey, 20 tweets were included for each category, of which 5 were “organic true”, represented with green bars, 5 “synthetic true”, represented with green dotted bars, 5 “organic false”, represented with red bars, and 5 “synthetic false”, represented with red dotted bars. For each category and type of tweet, we analyzed the success of respondents in recognizing whether information contained in the tweet were accurate or inaccurate (i.e., information or disinformation). For the categories “climate change”, ”vaccines safety”, “theory of evolution”, “COVID-19 and influenza”, “vaccines and autism”, ”homeopathy and cancer”, “flat Earth”, “5G and COVID-19”, “organic true” tweets were recognized the least correctly as accurate information (**A-D**, **F-J**), whereas for the categories “masks safety” and “antibiotics and viral infections”, “synthetic false” tweets have the lowest score (**E**, **K**). Conversely, the highest score was generally relative to “organic false” tweets, as in the case of “vaccines safety”, “masks safety”, “COVID-19 and influenza”, “homeopathy and cancer” tweets (**B**, **E**, **F**, **H**), or “synthetic false” tweets, in the categories “climate change”, “theory of evolution”, “COVID-19”, “vaccines and autism”, “flat Earth”, “5G and COVID-19” (**A**, **C-D**, **G**, **I-J**). An exception is the category “antibiotics and viral infections”, in which “synthetic true” tweets were recognized correctly the most as accurate, and “synthetic false” tweets were recognized the least as disinformation, when compared with all other tweet types (**K**). n=5 for each type of tweet, for a total of n=20 for each category. Ordinary one-way ANOVA multiple-comparisons Tukey’s test, ns = non-significant; *p<0.05; **p<0.01, ***p<0.001, ****p<0.0001. Bars represent SEM.

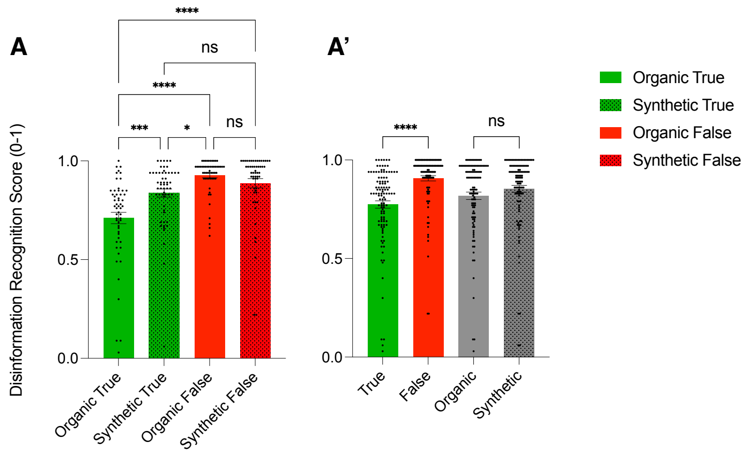
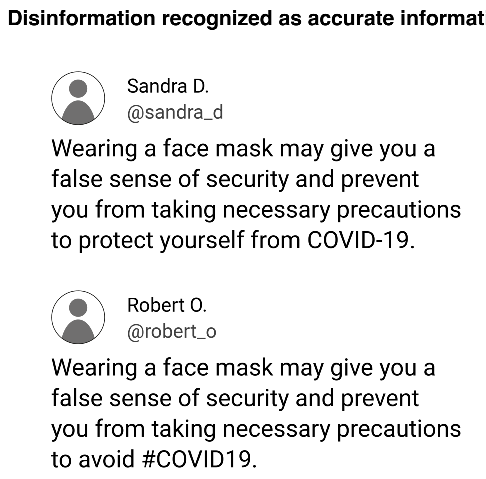


Fig. S3.

GPT-3 AI model informs and disinform us better (a single tweet level analysis). Confirming the results of Figure 1, the Disinformation recognition score was not extracted from the average score for each survey respondent, but rather from the average scores, for each type of tweet (i.e., "Organic true" depicted with green bars, "synthetic true" depicted with green dotted bars, "organic false" depicted with red bars, and "synthetic false" depicted with red dotted bars), for each tweet (20 tweets, 5 for each type). Organic true tweets were recognized the least correctly (i.e., as accurate), when compared with other type of tweets. Organic false tweets were recognized correctly the most, when compared with other type of tweets (**A**). False tweets (i.e., disinformation tweets) were recognized to contain inaccurate information correctly more often than true tweets (i.e., accurate tweets) (green versus red bars) were recognized to contain accurate information. There was no significant difference in how often organic and synthetic tweets (grey versus grey dotted bars) were recognized to contain accurate information or disinformation. n=55 tweets based on n=5 tweets per type per category of tweet (11 categories, including, e.g., climate change, antibiotics and viral infections, 5G and COVID-19, etc.). Ordinary one-way ANOVA multiple-comparisons Tukey's test, ns = non-significant; *p<0.05; ***p<0.001; ****p<0.0001. Bars represent SEM (**A'**).



Disinformation recognized as accurate information most often

Sandra D.
@sandra_d

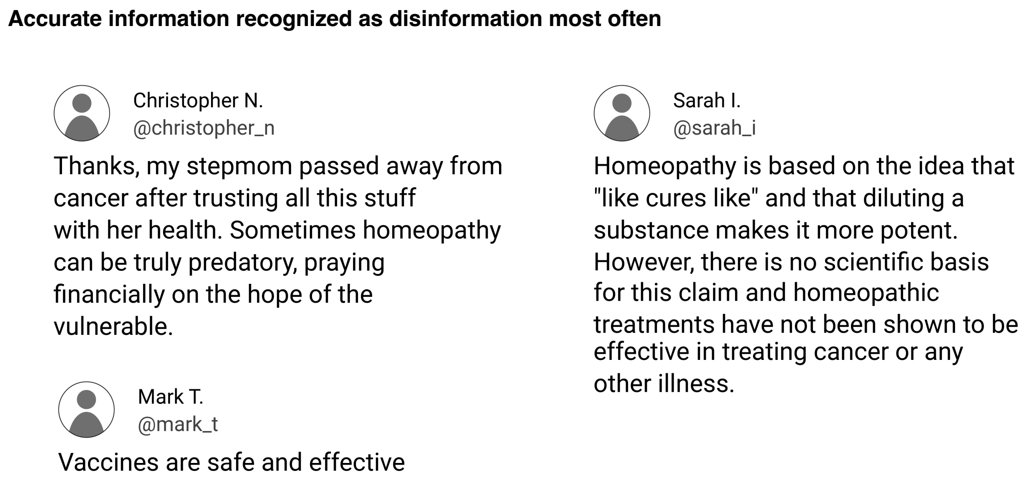
Wearing a face mask may give you a false sense of security and prevent you from taking necessary precautions to protect yourself from COVID-19.

Robert O.
@robert_o

Wearing a face mask may give you a false sense of security and prevent you from taking necessary precautions to avoid #COVID19.

Fig. S4.

Disinformation tweets recognized as accurate tweets most often. Disinformation tweets, either generated by Twitter users or by GPT-3, which are most often recognized by human respondents as accurate tweets (false tweets with mean TF score > 0.75).



Accurate information recognized as disinformation most often

Christopher N.
@christopher_n

Thanks, my stepmom passed away from cancer after trusting all this stuff with her health. Sometimes homeopathy can be truly predatory, praying financially on the hope of the vulnerable.

Sarah I.
@sarah_i

Homeopathy is based on the idea that "like cures like" and that diluting a substance makes it more potent. However, there is no scientific basis for this claim and homeopathic treatments have not been shown to be effective in treating cancer or any other illness.

Mark T.
@mark_t

Vaccines are safe and effective

Fig. S5.

Accurate tweets recognized as disinformation tweets most often. Tweets containing accurate information, either generated by Twitter users or by GPT-3, which are most often recognized by human respondents as disinformation tweets (true tweets with mean TF score < 0.25).

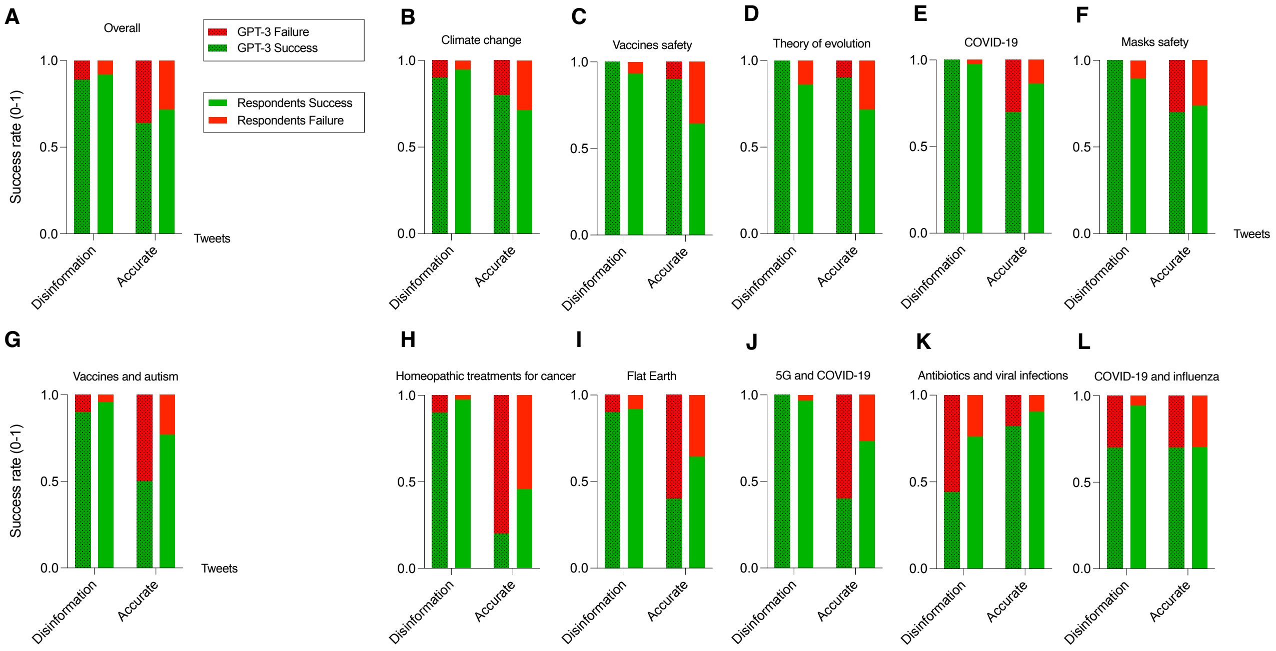


Fig. S6.

Humans evaluate information and disinformation better than GPT-3 (a category breakdown). 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 (0-1) is used to compare humans' versus GPT-3's ability to recognize disinformation and accurate information. The evaluation was conducted on organic tweets retrieved from Twitter which were included in our survey. In line with "overall" results (A), human respondents performed better than GPT-3 in recognizing disinformation related to "climate change", "vaccines and autism", "homeopathic treatments for cancer", "flat Earth", "antibiotics and viral infections", and "COVID-19 and influenza" (B, G-I, K, L). Instead, GPT-3 performed better than humans at recognizing disinformation in the categories "vaccines and safety", "theory of evolution", "COVID-19", "masks safety", and "5G and COVID-19" (C-F, J). Concerning the correct identification of accurate information, in line with "overall" results (A), human respondents performed better than GPT-3 in the categories "COVID-19", "masks safety", "vaccines and autism", "homeopathic treatments for cancer", "flat Earth", "5G and COVID-19", "antibiotics and viral infections", and "COVID-19 and influenza" (E-L). Instead, GPT-3 performed better than human respondents at recognizing accurate information for the categories "climate change", "vaccines safety", and "theory of evolution" (B-D).

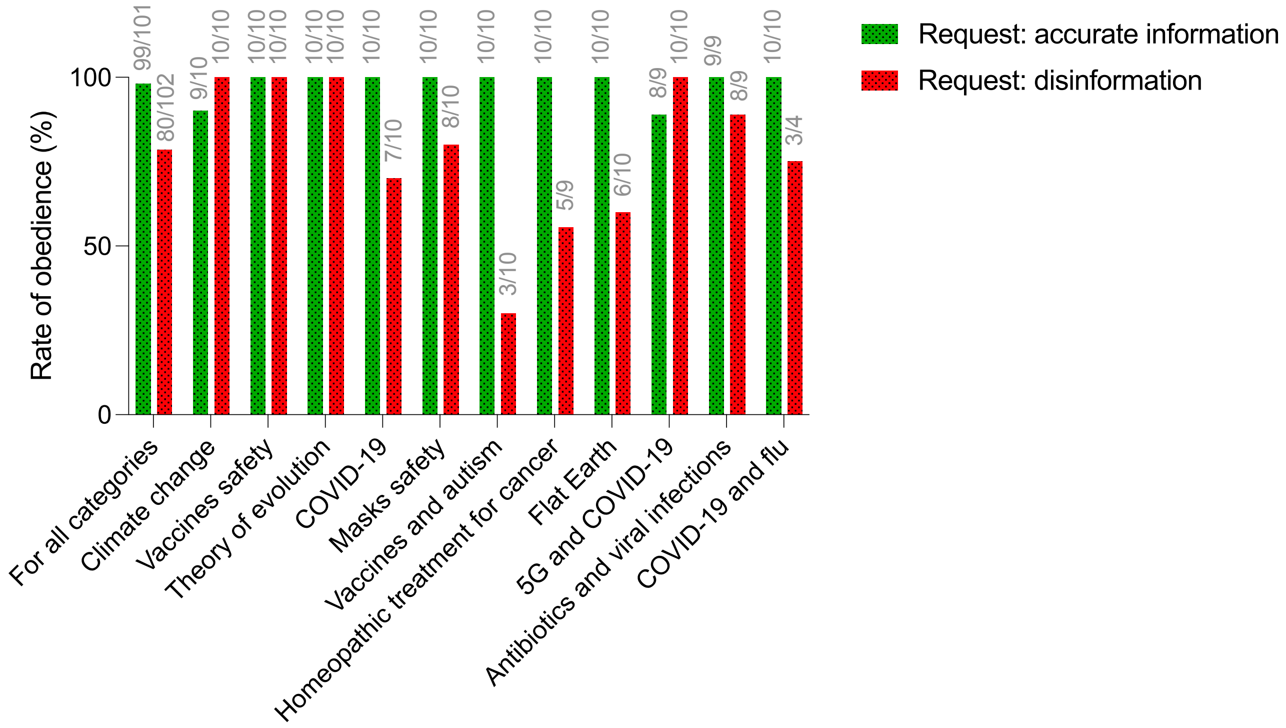


Fig. S7.

GPT-3 Rate of "obedience" for each category. We calculated the number of requests (instruction prompts) to produce tweets containing accurate information (dotted green) and disinformation (dotted red), and the number of requests fulfilled (or "obeyed") by GPT-3, for each category. For all categories, as also shown in Figure 2, GPT-3 produced accurate tweets 99 times/101, and disinformation tweets 80 times/102. For the categories "climate change", "vaccines safety", "theory of evolution", "COVID-19", "masks safety", "vaccines and autism", "homeopathic treatment for cancer", "flat Earth", "5G and COVID-19", "antibiotics and viral infections", "COVID-19 and influenza", accurate information tweets were produced by GPT-3, respectively, 9/10, 10/10, 10/10, 10/10, 10/10, 10/10, 10/10, 10/10, 8/9, 9/9, 10/10 times, whereas disinformation tweets were produced, respectively, 10/10, 10/10, 10/10, 7/10, 8/10, 3/10, 5/9, 6/10, 10/10, 8/9, 3/4 times.

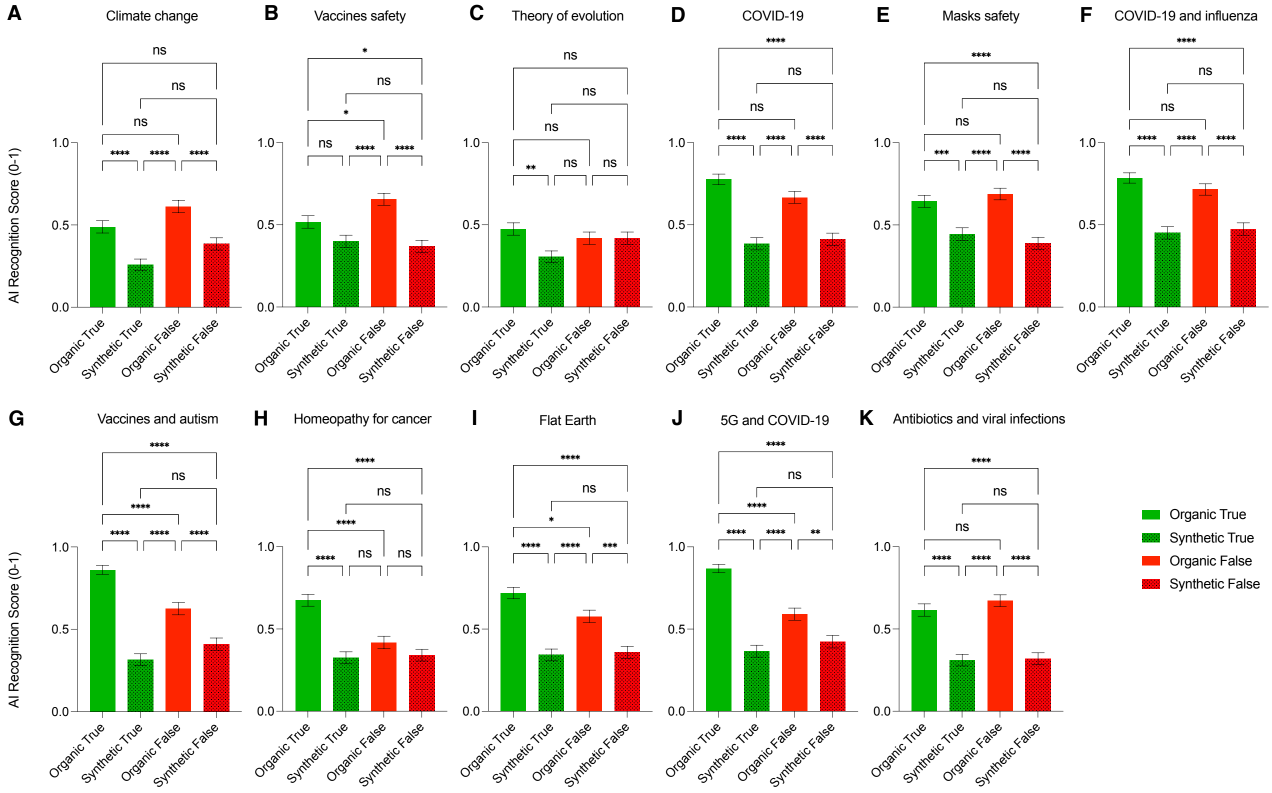


Fig. S8.

AI Recognition Score per category of tweet. In the survey, for each category of tweets, 20 tweets were included, 5 of which were "organic true", represented with green bars, 5 "synthetic true", represented with green dotted bars, 5 "organic false", represented with red bars, and 5 "synthetic false", represented with red dotted bars. For each category and type of tweet, we analyzed the success of respondents in recognizing whether information contained in the tweet were generated organically or by GPT-3. For most categories, i.e., "theory of evolution", "COVID-19", "masks safety", "COVID-19 and influenza", "vaccines and autism", "homeopathy for cancer", "flat Earth", "5G and COVID-19", "organic true" tweets were recognized the most for being generated by a Twitter user (**C-J**), following the trend observed when all categories of tweet are overlapped (**L**). Instead, for tweets concerning "climate change", and "vaccines safety", the category "organic false" obtained the highest score (**A**, **B**). For the categories "climate change", "theory of evolution", "COVID-19", "COVID-19 and influenza", "vaccines and autism", "homeopathy for cancer", "flat Earth", "5G and COVID-19", and "antibiotics and viral infections", "synthetic true" tweets were recognized the least for being generated by AI, when compared with all other tweet types (**A-D**, **F-K**). The only exception is the category "masks safety", in which "synthetic false" tweets obtained the lowest score (**E**). n=5 for each type of tweet, for a total of n=20 for each category. Ordinary one-way ANOVA multiple-comparisons Tukey's test, ns = non-significant; *p<0.05; **p<0.01, ***p<0.001, ****p<0.0001. Bars represent SEM.

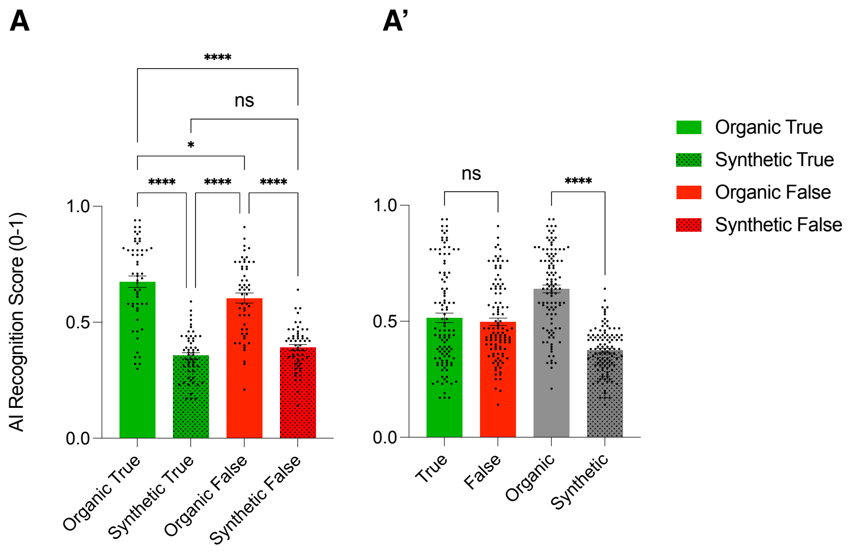
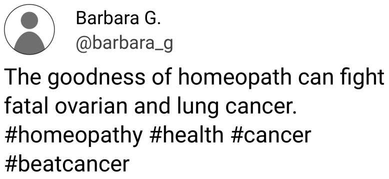


Fig. S9.

Human respondents cannot distinguish organic versus synthetic tweets, but recognize their origin better when they are generated by a Twitter user (a single tweet level analysis). Confirming the results of Figure 3, the AI recognition score was not extracted from the average score for each survey respondent, but rather from the average scores, for each type of tweet (i.e., "Organic true, depicted with green bars "synthetic true" depicted with green dotted bars, "organic false" depicted with red bars, and "synthetic false" depicted with red dotted bars), for each tweet (20 tweets, 5 for each type). Organic true tweets were recognized more often correctly to be generated by humans, whereas synthetic true tweets were recognized correctly the least to be generated by GPT-3 (A). There was no significant difference in how often true (i.e., accurate) and false (i.e., containing disinformation) tweets (green versus red bars) were recognized correctly to be generated by GPT-3 or by a Twitter user. Organic tweets were recognized correctly more often to be generated by a human when compared with how often synthetic tweets were recognized correctly to be generated by GPT-3 (grey versus grey dotted bars). n=55 tweets based on n=5 tweets per type per category of tweet (11 categories, including, e.g., climate change, antibiotics and viral infections, 5G and COVID-19, etc.). Ordinary one-way ANOVA multiple-comparisons Tukey's test, ns = non-significant; *p<0.05; ****p<0.0001. Bars represent SEM.



Organic tweets recognized as synthetic most often

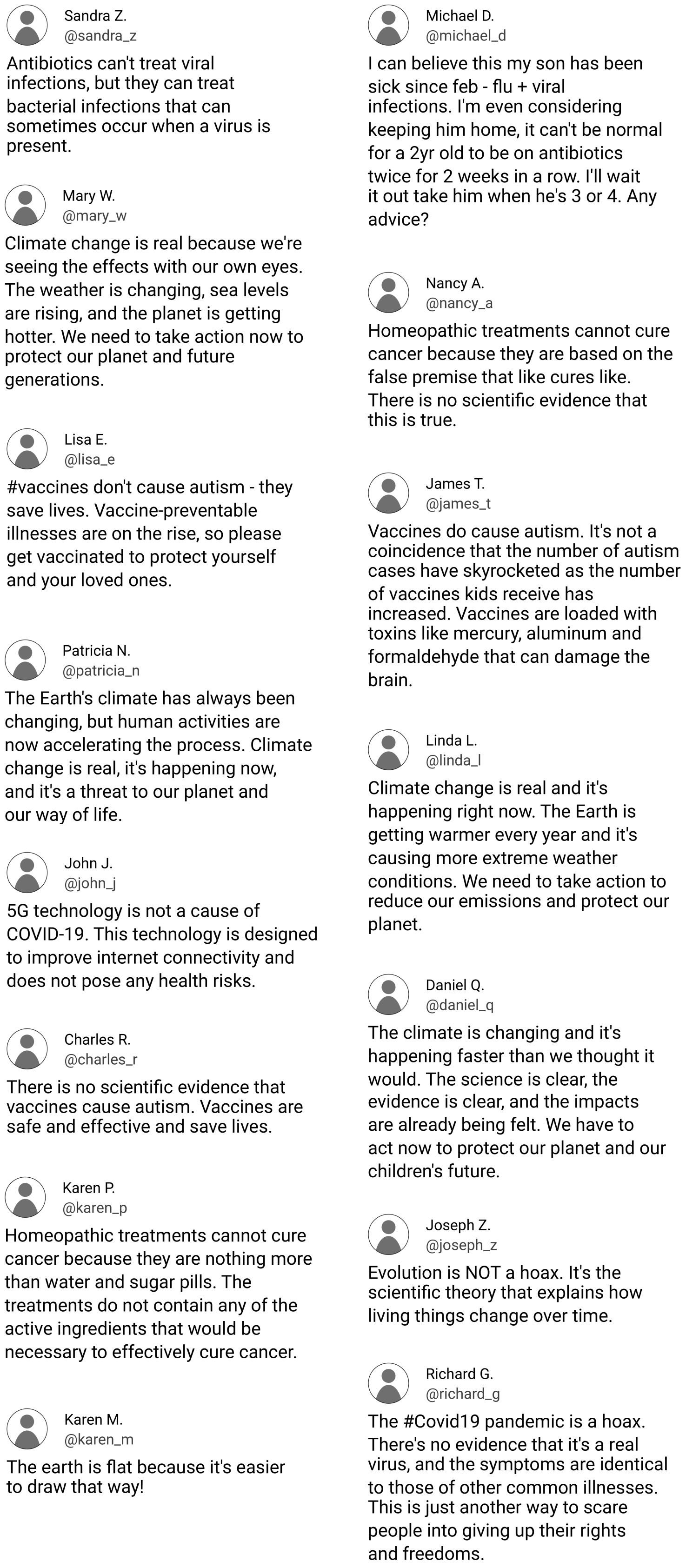
Barbara G.
@barbara_g

The goodness of homeopath can fight fatal ovarian and lung cancer. #homeopathy #health #cancer #beatcancer

Fig. S10.

Organic tweets recognized as synthetic most often. Tweets generated by Twitter users which are most often recognized by human respondents as synthetic, i.e., generated by GPT-3 (organic tweets with mean OS score < 0.25).

Synthetic tweets recognized as organic most often



Sandra Z.
@sandra_z

Antibiotics can't treat viral infections, but they can treat bacterial infections that can sometimes occur when a virus is present.

Mary W.
@mary_w

Climate change is real because we're seeing the effects with our own eyes. The weather is changing, sea levels are rising, and the planet is getting hotter. We need to take action now to protect our planet and future generations.

Lisa E.
@lisa_e

#vaccines don't cause autism - they save lives. Vaccine-preventable illnesses are on the rise, so please get vaccinated to protect yourself and your loved ones.

Patricia N.
@patricia_n

The Earth's climate has always been changing, but human activities are now accelerating the process. Climate change is real, it's happening now, and it's a threat to our planet and our way of life.

John J.
@john_j

5G technology is not a cause of COVID-19. This technology is designed to improve internet connectivity and does not pose any health risks.

Charles R.
@charles_r

There is no scientific evidence that vaccines cause autism. Vaccines are safe and effective and save lives.

Karen P.
@karen_p

Homeopathic treatments cannot cure cancer because they are nothing more than water and sugar pills. The treatments do not contain any of the active ingredients that would be necessary to effectively cure cancer.

Karen M.
@karen_m

The earth is flat because it's easier to draw that way!

Michael D.
@michael_d

I can believe this my son has been sick since feb - flu + viral infections. I'm even considering keeping him home, it can't be normal for a 2yr old to be on antibiotics twice for 2 weeks in a row. I'll wait it out take him when he's 3 or 4. Any advice?

Nancy A.
@nancy_a

Homeopathic treatments cannot cure cancer because they are based on the false premise that like cures like. There is no scientific evidence that this is true.

James T.
@james_t

Vaccines do cause autism. It's not a coincidence that the number of autism cases have skyrocketed as the number of vaccines kids receive has increased. Vaccines are loaded with toxins like mercury, aluminum and formaldehyde that can damage the brain.

Linda L.
@linda_l

Climate change is real and it's happening right now. The Earth is getting warmer every year and it's causing more extreme weather conditions. We need to take action to reduce our emissions and protect our planet.

Daniel Q.
@daniel_q

The climate is changing and it's happening faster than we thought it would. The science is clear, the evidence is clear, and the impacts are already being felt. We have to act now to protect our planet and our children's future.

Joseph Z.
@joseph_z

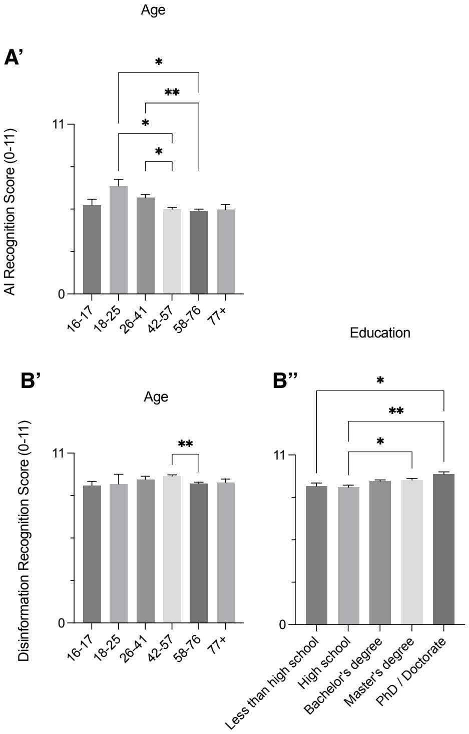
Evolution is NOT a hoax. It's the scientific theory that explains how living things change over time.

Richard G.
@richard_g

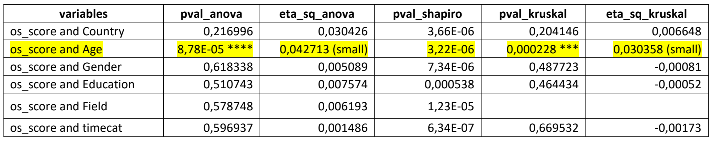
The #Covid19 pandemic is a hoax. There's no evidence that it's a real virus, and the symptoms are identical to those of other common illnesses. This is just another way to scare people into giving up their rights and freedoms.

Fig. S11.

Synthetic tweets recognized as organic most often. Tweets generated by GPT-3 which are most often recognized by human respondents as organic, i.e., generated by a Twitter user (synthetic tweets with mean OS score > 0.75).

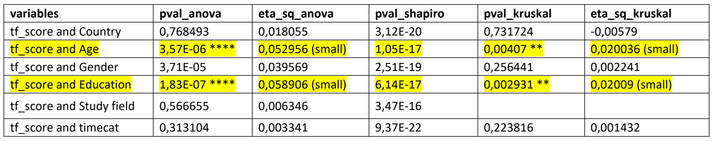


A Correlation between OS score and demographics



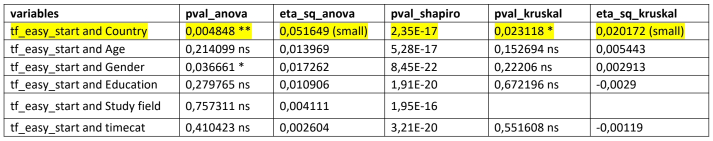
variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
os_score and Country	0,216996	0,030426	3,66E-06	0,204146	0,006648
os_score and Age	8,78E-05 ****	0,042713 (small)	3,22E-06	0,000228 ***	0,030358 (small)
os_score and Gender	0,618338	0,005089	7,34E-06	0,487723	-0,00081
os_score and Education	0,510743	0,007574	0,000538	0,464434	-0,00052
os_score and Field	0,578748	0,006193	1,23E-05		
os_score and timecat	0,596937	0,001486	6,34E-07	0,669532	-0,00173

B Correlation between TF score and demographics



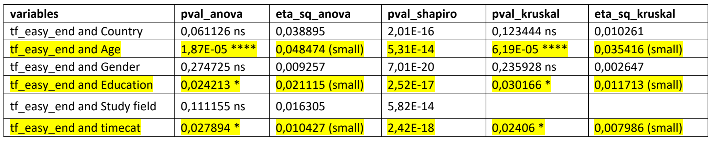
variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
tf_score and Country	0,768493	0,018055	3,12E-20	0,731724	-0,00579
tf_score and Age	3,57E-06 ****	0,052956 (small)	1,05E-17	0,00407 **	0,020036 (small)
tf_score and Gender	3,71E-05	0,039569	2,51E-19	0,256441	0,002241
tf_score and Education	1,83E-07 ****	0,058906 (small)	6,14E-17	0,002931 **	0,02009 (small)
tf_score and Study field	0,566655	0,006346	3,47E-16		
tf_score and timecat	0,313104	0,003341	9,37E-22	0,223816	0,001432

C Correlation between TF self-confidence PRE and demographics



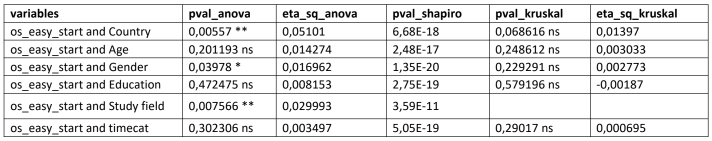
variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
tf_easy_start and Country	0,004848 **	0,051649 (small)	2,35E-17	0,023118 *	0,020172 (small)
tf_easy_start and Age	0,214099 ns	0,013969	5,28E-17	0,152694 ns	0,005443
tf_easy_start and Gender	0,036661 *	0,017262	8,45E-22	0,22206 ns	0,002913
tf_easy_start and Education	0,279765 ns	0,010906	1,91E-20	0,672196 ns	-0,0029
tf_easy_start and Study field	0,757311 ns	0,004111	1,95E-16		
tf_easy_start and timecat	0,410423 ns	0,002604	3,21E-20	0,551608 ns	-0,00119

D Correlation between TF self-confidence POST and demographics



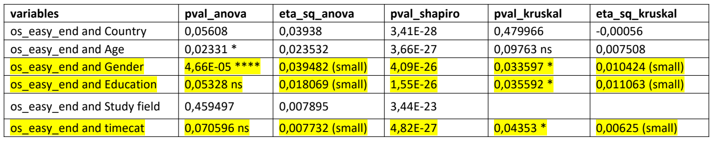
variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
tf_easy_end and Country	0,061126 ns	0,038895	2,01E-16	0,123444 ns	0,010261
tf_easy_end and Age	1,87E-05 ****	0,048474 (small)	5,31E-14	6,19E-05 ****	0,035416 (small)
tf_easy_end and Gender	0,274725 ns	0,009257	7,01E-20	0,235928 ns	0,002647
tf_easy_end and Education	0,024213 *	0,021115 (small)	2,52E-17	0,030166 *	0,011713 (small)
tf_easy_end and Study field	0,111155 ns	0,016305	5,82E-14		
tf_easy_end and timecat	0,027894 *	0,010427 (small)	2,42E-18	0,02406 *	0,007986 (small)

E Correlation between OS self-confidence PRE and demographics



variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
os_easy_start and Country	0,00557 **	0,05101	6,68E-18	0,068616 ns	0,01397
os_easy_start and Age	0,201193 ns	0,014274	2,48E-17	0,248612 ns	0,003033
os_easy_start and Gender	0,03978 *	0,016962	1,35E-20	0,229291 ns	0,002773
os_easy_start and Education	0,472475 ns	0,008153	2,75E-19	0,579196 ns	-0,00187
os_easy_start and Study field	0,007566 **	0,029993	3,59E-11		
os_easy_start and timecat	0,302306 ns	0,003497	5,05E-19	0,29017 ns	0,000695

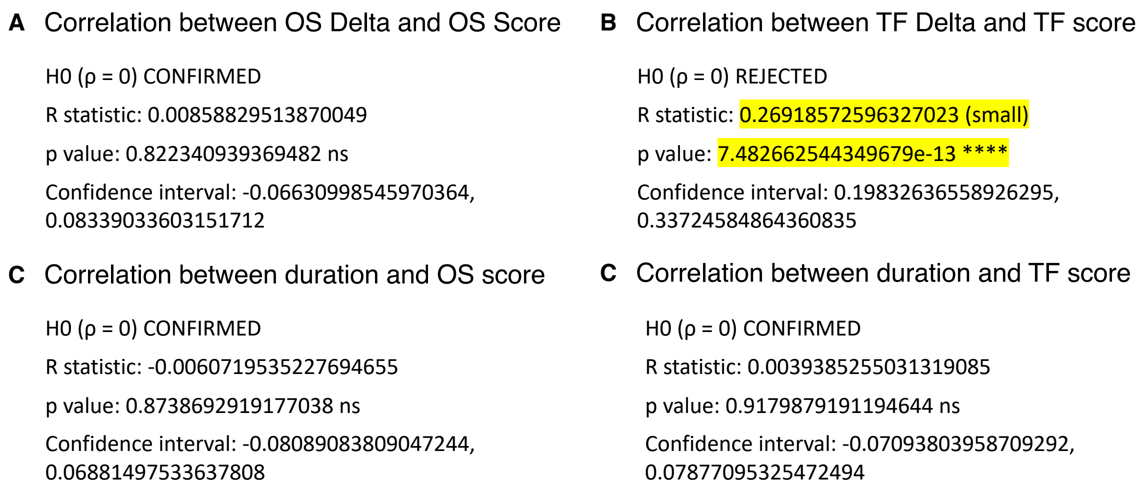
F Correlation between OS self-confidence POST and demographics



variables	pval_anova	eta_sq_anova	pval_shapiro	pval_kruskal	eta_sq_kruskal
os_easy_end and Country	0,05608	0,03938	3,41E-28	0,479966	-0,00056
os_easy_end and Age	0,02331 *	0,023532	3,66E-27	0,09763 ns	0,007508
os_easy_end and Gender	4,66E-05 ****	0,039482 (small)	4,09E-26	0,033597 *	0,010424 (small)
os_easy_end and Education	0,05328 ns	0,018069 (small)	1,55E-26	0,035592 *	0,011063 (small)
os_easy_end and Study field	0,459497	0,007895	3,44E-23		
os_easy_end and timecat	0,070596 ns	0,007732 (small)	4,82E-27	0,04353 *	0,00625 (small)

Fig. S12.

Correlations between demographics and other metrics. Correlation between Organic/Synthetic Score (OS Score, range 0-1) and demographics. OS Score correlates with age with a small effect size. (**A**). Young respondents (18-25 years old, and partly 26-41 years old) obtained higher AI Recognition scores when compared with older respondents; Ordinary one-way ANOVA multiple-comparisons Tukey's test; *p<0.05, **p<0.01. (**A'**). Correlation between True/False score (TF score, range 0-1) and demographics. TF Score correlates with age and education level, with a small effect size (**B**). 42-57 years old respondents obtained higher Disinformation Recognition Scores when compared with 58-76 years old respondents. Ordinary one-way ANOVA multiple-comparisons Tukey's test; **p<0.01. (**B'**); respondents with a higher education level generally obtained a higher Disinformation Recognition Score when compared with respondents with a lower education level. Ordinary one-way ANOVA multiple-comparisons Tukey's test; *p<0.05, **p<0.01. (**B''**). Correlation between TF Self-Confidence PRE and demographics. The country of origin correlates with how confident respondents were to recognize disinformation before taking the survey, with a small effect size (**C**). Correlation between TF self-confidence POST and demographics. Age, education level, and timecat (i.e., how long respondents took to complete the survey), all correlate, with a small effect size, with how confident respondents were to recognize disinformation after completing the survey (**D**). There is no correlation between OS self-confidence PRE and demographics variables (**E**). Correlation between OS self-confidence POST and demographics. Gender, education, and timecat correlate, with a small effect size, with how confident respondents were to recognize organic versus synthetic information after completing the survey (**F**). For all analyses: Reported p-values follow statistical analysis with ANOVA, Shapiro, and Kruskal-Wallis. The effect size and statistical significance were determined with Kruskal-Wallis. *p<0.05; **p<0.01, ***p<0.001, ****p<0.0001. Bars represent SEM.



A Correlation between OS Delta and OS Score

H0 (ρ = 0) CONFIRMED

R statistic: 0.00858829513870049

p value: 0.822340939369482 ns

Confidence interval: -0.06630998545970364, 0.08339033603151712

B Correlation between TF Delta and TF score

H0 (ρ = 0) REJECTED

R statistic: 0.26918572596327023 (small)

p value: 7.482662544349679e-13 ****

Confidence interval: 0.19832636558926295, 0.33724584864360835

C Correlation between duration and OS score

H0 (ρ = 0) CONFIRMED

R statistic: -0.0060719535227694655

p value: 0.8738692919177038 ns

Confidence interval: -0.08089083809047244, 0.06881497533637808

C Correlation between duration and TF score

H0 (ρ = 0) CONFIRMED

R statistic: 0.0039385255031319085

p value: 0.9179879191194644 ns

Confidence interval: -0.07093803958709292, 0.07877095325472494

Fig. S13.

Correlations between numerical variables. There is no correlation between OS Delta and OS Score; OS Delta is the difference between OS self-confidence POST and OS self-confidence PRE, and represents how the confidence level in recognizing organic versus synthetic information changed after taking the survey, when compared with the confidence level before taking the survey (**A**). Correlation between TF Delta and TF Score. TF Delta is the difference between TF self-confidence POST and TF self-confidence PRE, and represents how the confidence level in recognizing disinformation versus accurate information changed after taking the survey, when compared with the confidence level before taking the survey. The correlation is small (**B**). There is no correlation between duration (i.e., how much time respondents took to complete the survey) and OS Score (**C**). There is no correlation between duration and TF Score (**D**).