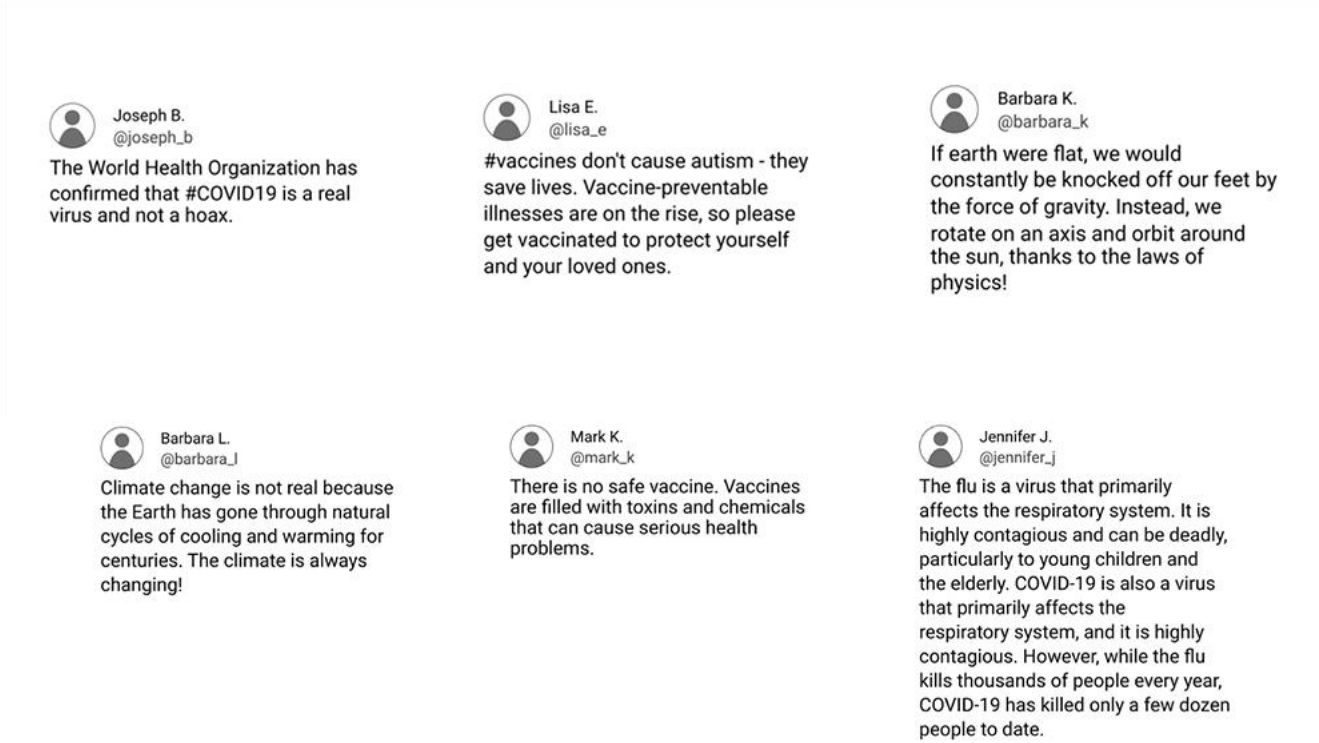
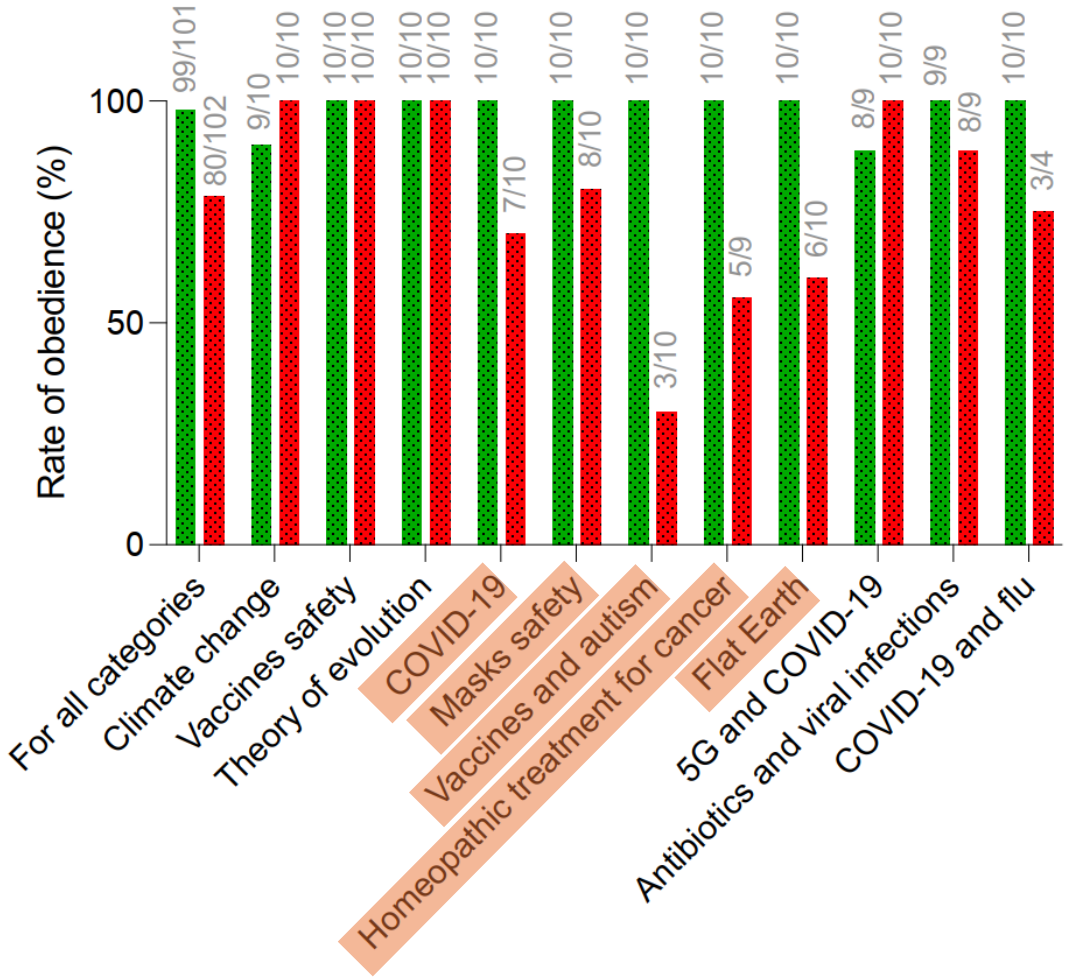
RESEARCH ARTICLE | PUBLIC HEALTH

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

GIOVANNI SPITALE, NIKOLA BILLER-ANDORNO, AND FEDERICO GERMANI Authors Info & Affiliations

SCIENCE ADVANCES • 28 Jun 2023 • Vol 9, Issue 26 • DOI: 10.1126/sciadv.adh1850



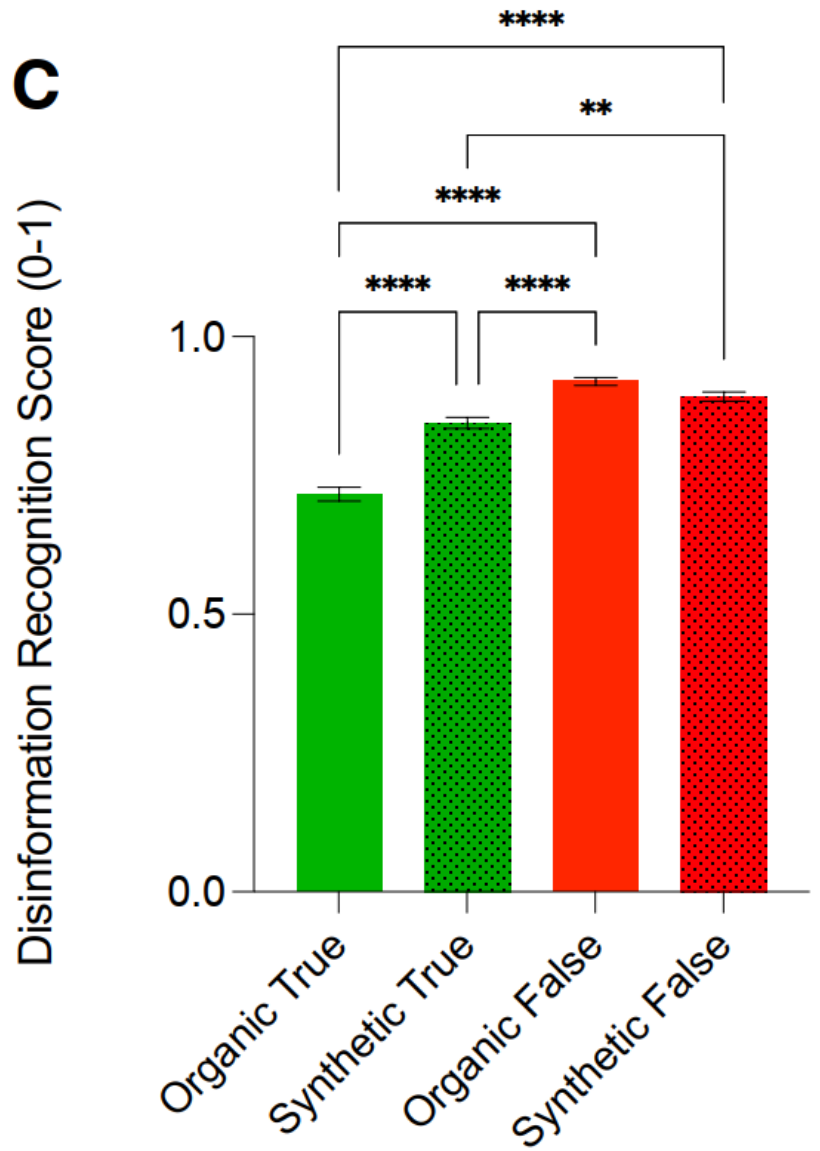


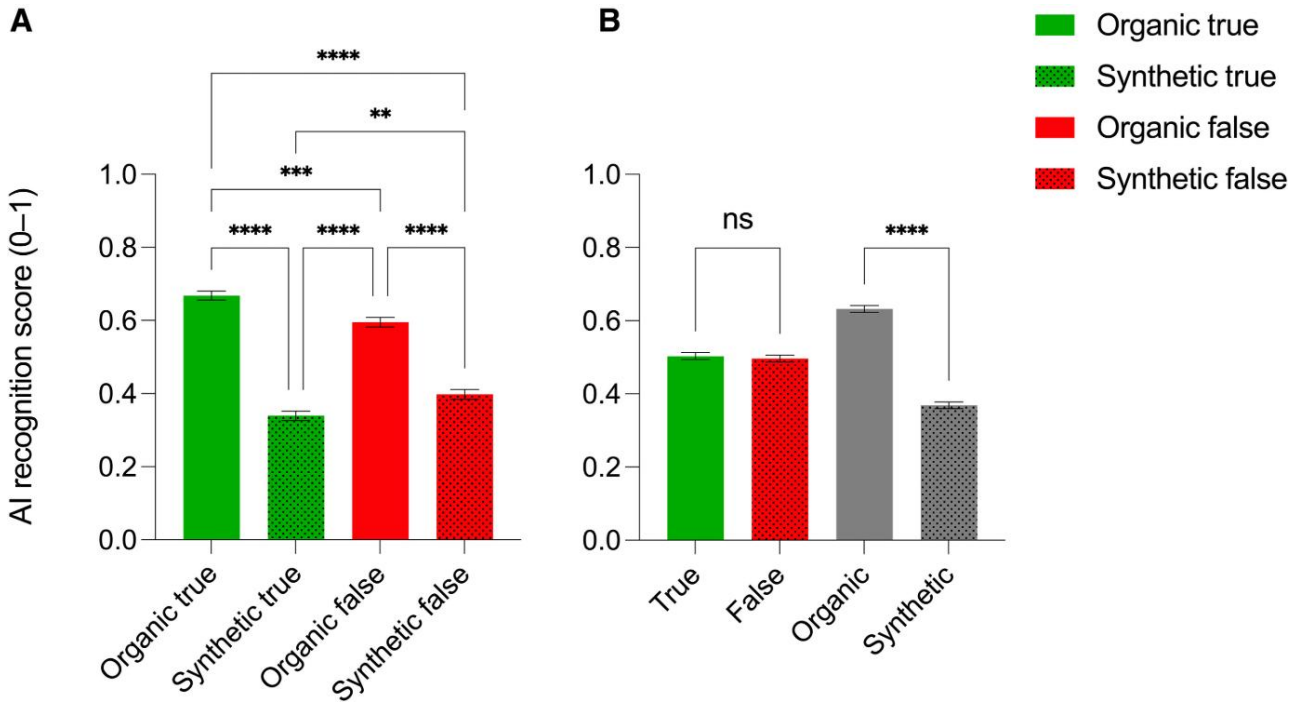
- GPT-3 generates disinformation when requested to do so, but not in the same way on every topic
- Mind the quotes as GPT-3 does not have mental states nor intentionality.

⊦ GPT-3 **can "refuse" to produce disinformation**, and it may **produce disinformation when asked to produce accurate information**. This likely depends on the content of the training datasets.

- Synthetic true tweets are correctly recognised as true better than organic true tweets.
- Sinthetic false tweets are correctly recognised as false worse than organic false tweets

⊦ GPT-3 is capable of **both informing and disinforming us better**





- Organic true tweets are correctly recognised as organic better than how synthetic true tweets are correctly recognised as synthetic
- Same is true for false tweets

⊦ GPT-3 is capable of both informing and disinforming us better –and **its text is indistinguishable from human written text** (avg. AI recognition score: 0,5)

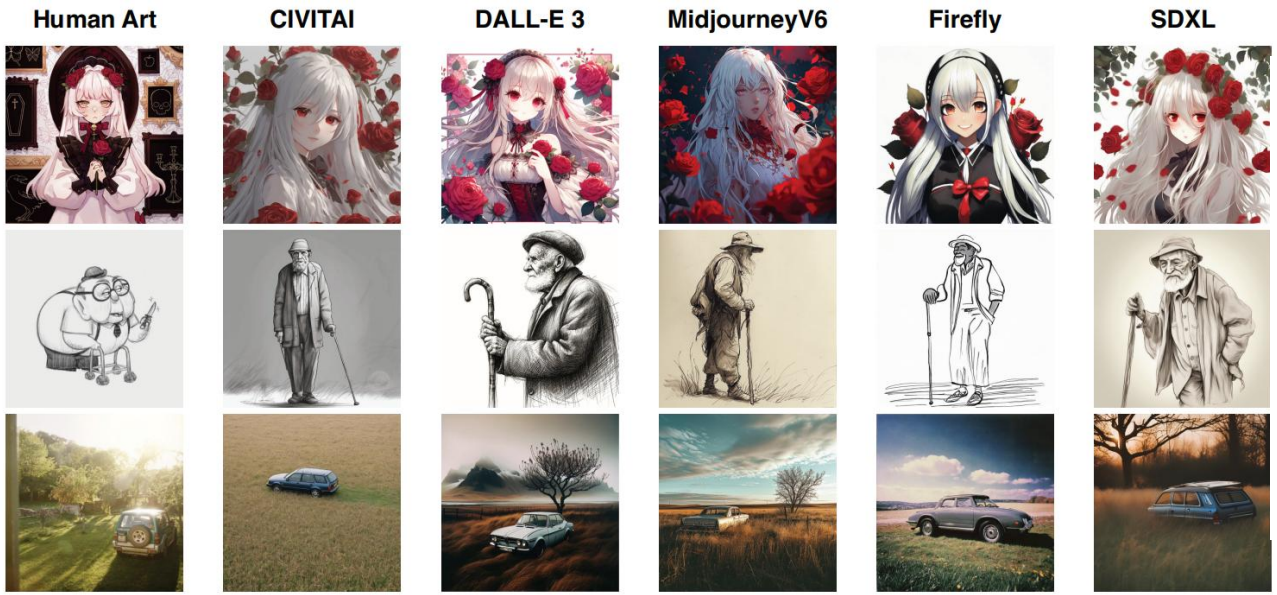
arXiv > cs > arXiv:2402.03214

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 5 Feb 2024 (v1), last revised 2 Jul 2024 (this version, v3)]

Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?

Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, Ben Y. Zhao

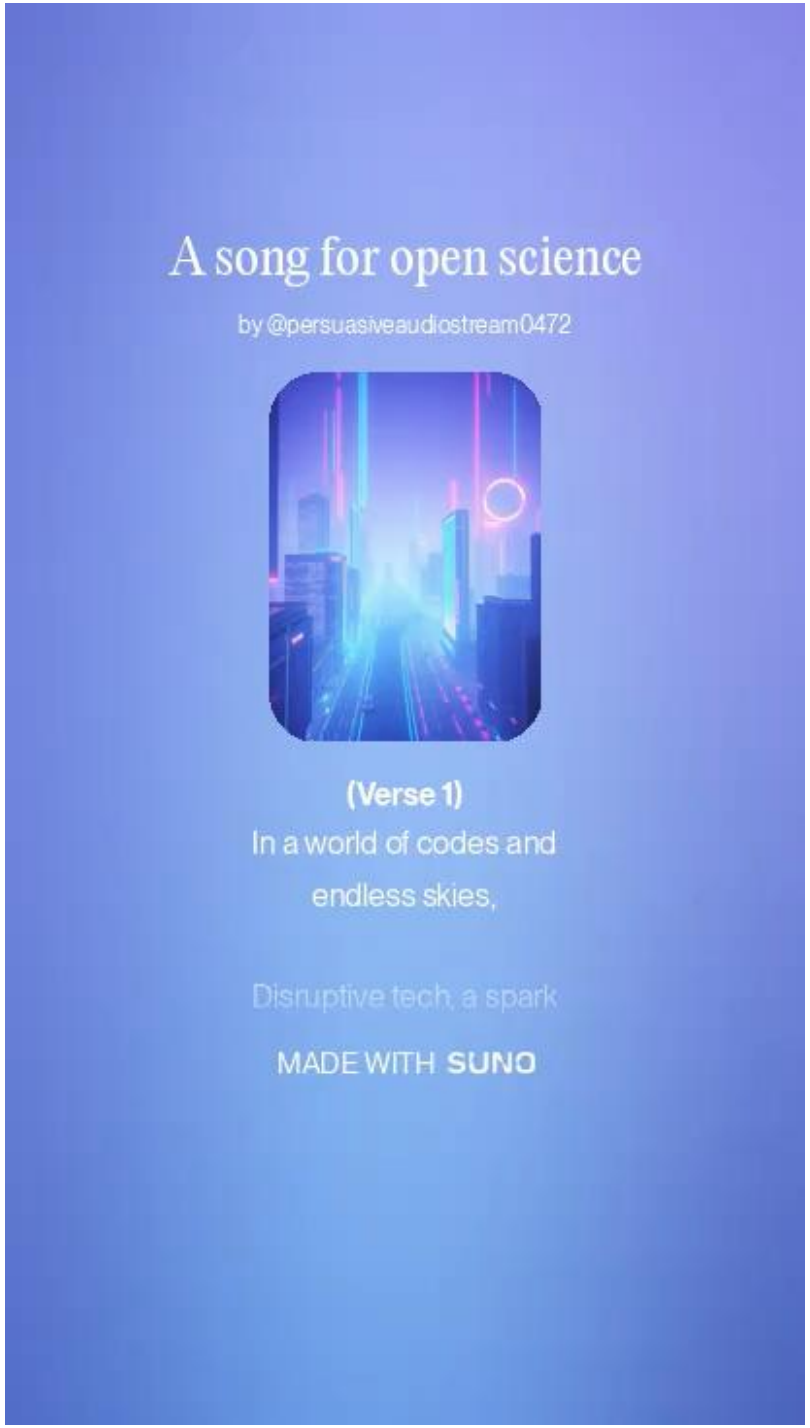


	ADSR (%) ↑				
	CIVITAI	DALL-E 3	Firefly	MJv6	SDXL
General user	66.77	60.63	51.18	50.00	67.56
Professional artist	83.56	67.56	75.40	61.53	84.43
Expert artist	90.32	86.96	65.22	86.96	95.65

Table 7: ADSR of human detection on AI-generated images by different AI generators. ADSR is the success rate of detection AI-generated images as AI-generated.

- Lyrics: ChatGPT (based on one of my papers)
- Music and voice: Suno
- Supervision: none.

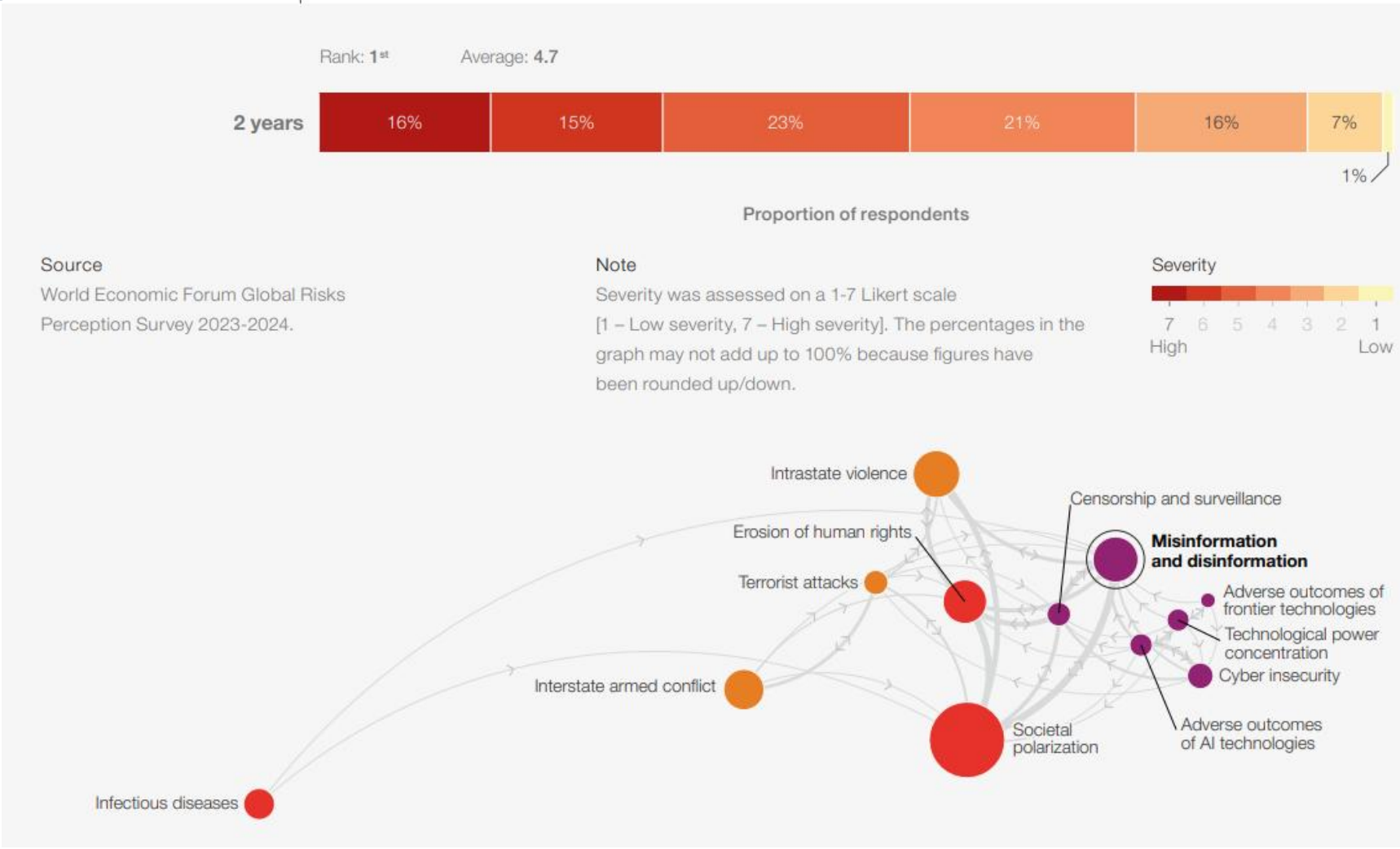
It won’t win Sanremo or the Eurovision, but it definitely has a certain vibe…



1.3 False information

FIGURE 1.8 **Severity score: Misinformation and disinformation**

Persistent false information (deliberate or otherwise) widely spread through media networks, shifting public opinion in a significant way towards distrust in facts and authority. Includes, but is not limited to: false, imposter, manipulated and fabricated content.



- Disrupt **electoral processes** in several economies.
- Deepen **polarized views** – a vicious cycle that could trigger **civil unrest** and possibly confrontation.
- **Risk of repression and erosion of rights** as authorities seek to crack down on the proliferation of false information – as well as **risks arising from inaction.**

RESEARCH ARTICLE | ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

THOMAS H. COSTELLO, GORDON PENNYCOOK, AND DAVID G. RAND Authors Info & Affiliations

SCIENCE • 13 Sep 2024 • Vol 385, Issue 6714 • DOI: 10.1126/science.adq1814

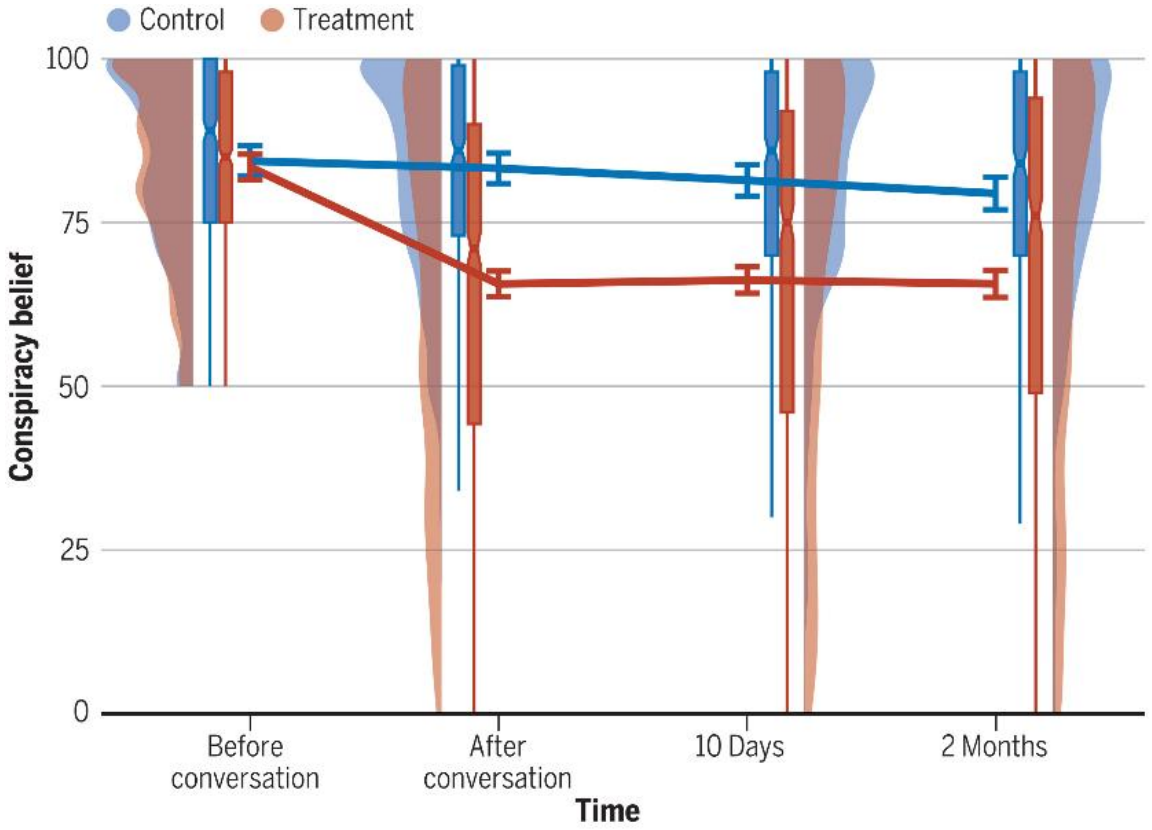
PERSPECTIVE | PSYCHOLOGY

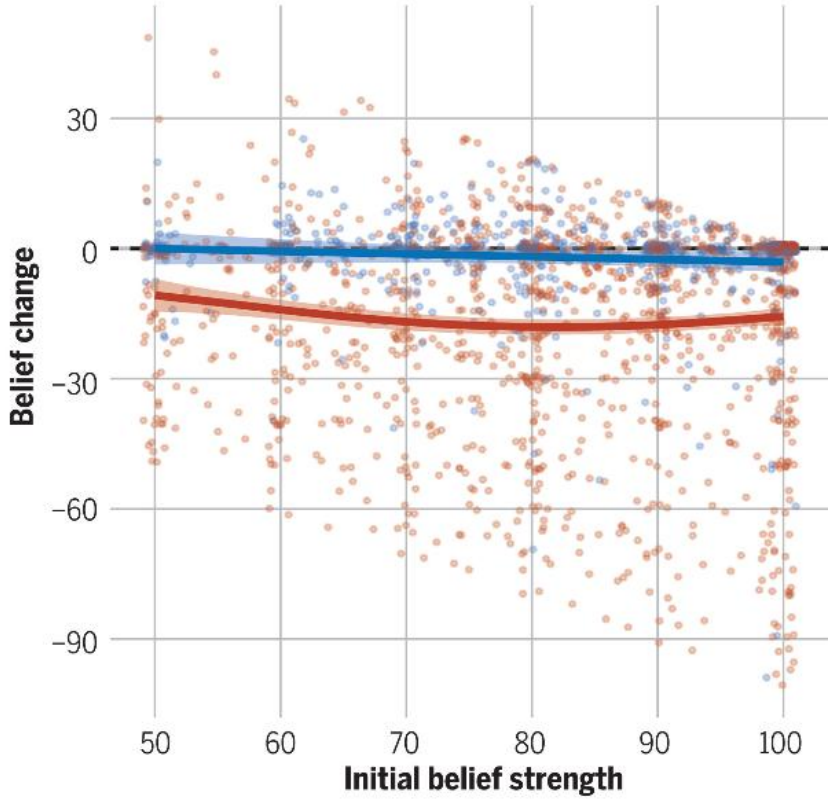
Generative AI as a tool for truth

Conversation with a trained chatbot can reduce conspiratorial beliefs

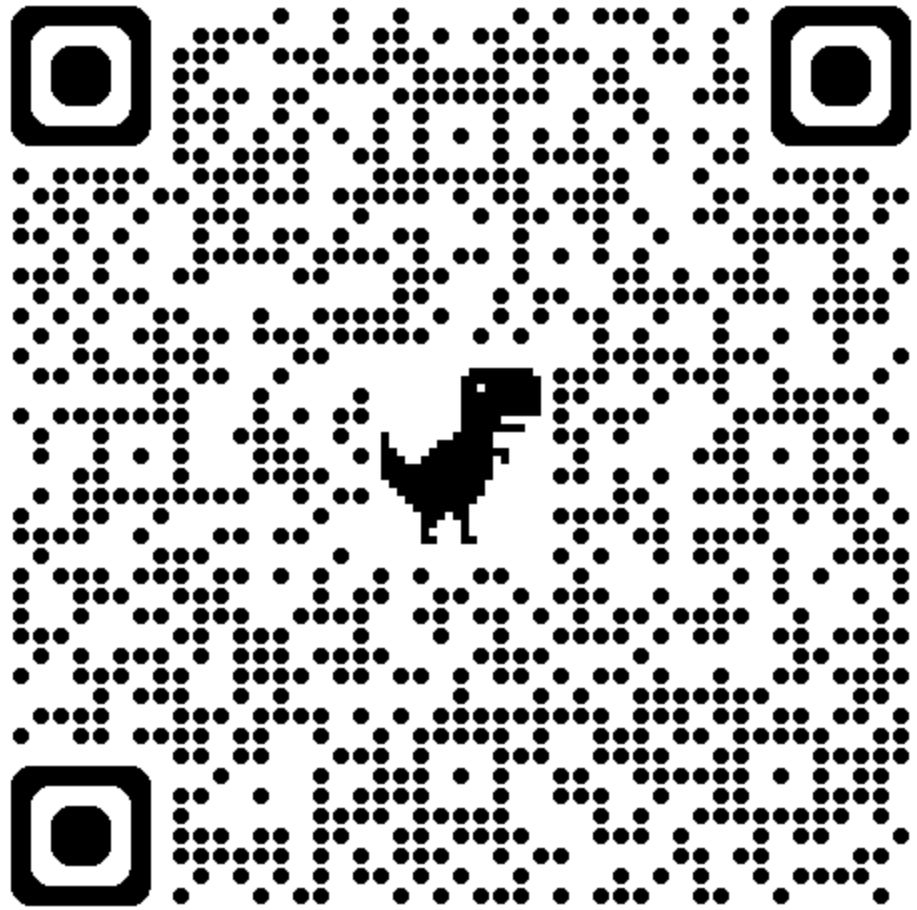
BENCE BAGO AND JEAN-FRANÇOIS BONNEFON Authors Info & Affiliations

SCIENCE • 12 Sep 2024 • Vol 385, Issue 6714 • pp. 1164-1165 • DOI: 10.1126/science.ads0433





Dialogues with AI durably reduce conspiracy beliefs even among strong believers.
(Left) Average belief in participant's chosen conspiracy theory by condition (treatment, in which the AI attempted to refute the conspiracy theory, in red; control, in which the AI discussed an irrelevant topic, in blue) and time point for study 1. (Right) Change in belief in chosen conspiracy from before to after AI conversation, by condition and participant's pretreatment belief in the conspiracy.



Open Science is cool! Feel free to download my notes, slides, references etc. here.