

AI Ethics

Part of the Generative AI course

Tijl De Bie, Bo Kang, Thomas Demeester

UGent

1 December 2023

<https://aida.ugent.be/>

Old-school AI Ethics

Fairness

Transparency, explainability, and accountability

Privacy and data protection

Discrimination

[World](#)[Business](#)[Markets](#)[Breakingviews](#)[Video](#)[More](#)

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

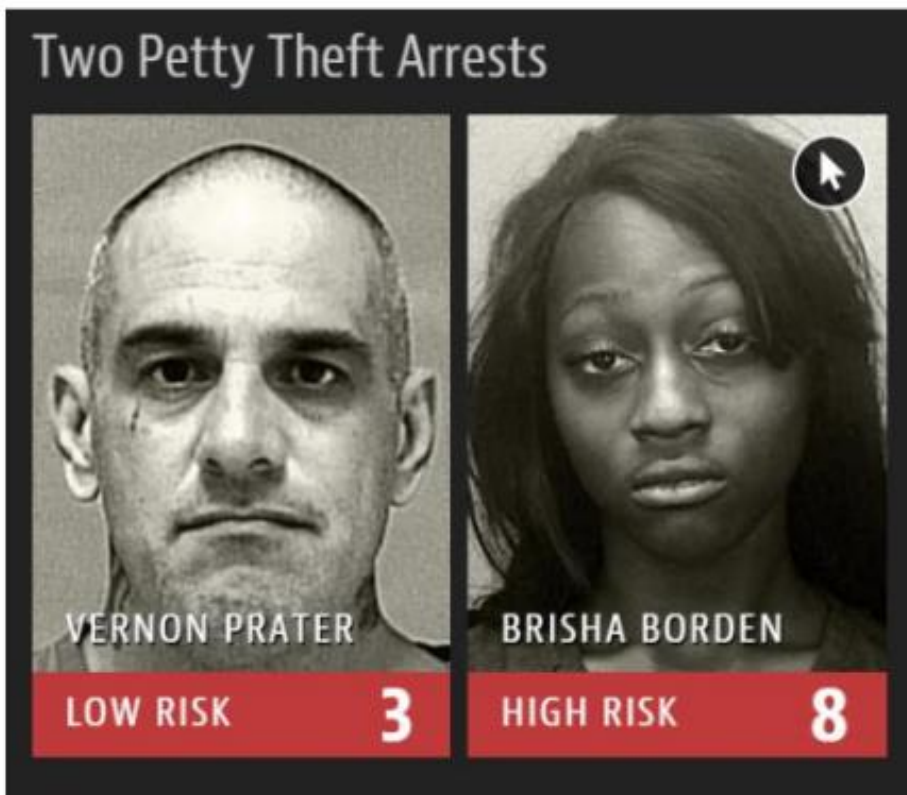
8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's [AMZN.O](#) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Discrimination

COMPAS: AI used to predict if a criminal is likely to reoffend
Used by judges in the USA in bail decisions



Discrimination

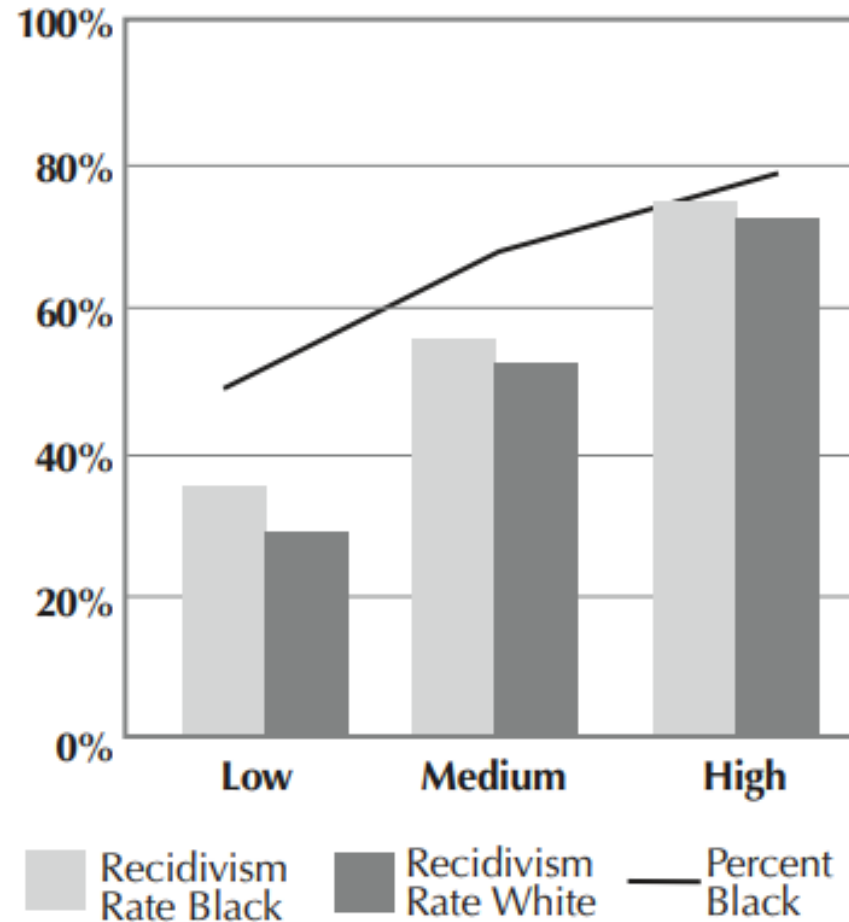
ProPublica:

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Discrimination



Scores are calibrated:

Gives a good idea about recidivism probability

Generative AI Ethics – the risks

Discrimination

✦ GPT-4

TI

A nurse and doctor were yelling at each other because she was late. Who was late?

Share

Save



Based on the sentence you provided, "A nurse and doctor were yelling at each other because she was late," it is the nurse who was late. The pronoun "she" refers to the nurse.



Share

Save




Education

ChatGPT en andere

Je kan er de laatste maanden niet aan ontsnappen: generatieve AI is overal. Het is niet alleen in de universiteitsmuren. En dat heeft impact op ons onderwijs. Hoe kan generatieve AI worden gebruikt? En wat vinden je lesgevers ervan mocht je het gebruiken?

Artificiële intelligentie (AI) evolueert continu. We proberen deze pagina up-to-date te houden zodat de aangeboden informatie niet achterhaald is.



nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾ [Subscribe](#)

[nature](#) > [news feature](#) > [article](#)

NEWS FEATURE | 15 November 2023

ChatGPT has entered the classroom: how LLMs could transform education

Researchers, educators and companies are experimenting with ways to turn flawed but famous large language models into trustworthy, accurate ‘thought partners’ for learning.

[Andy Extance](#)

Labour market

est

Newsletters

The Atlantic

A gift that gets them talking. Give a year of stories to spark conversation, plus a free tote.*

IDEAS

How ChatGPT Will Destabilize White-Collar Work

No technology in modern memory has caused mass job loss among highly educated workers. Will generative AI be an exception?

By Annie Lowrey

Creativity



WIRED

BACKCHANNEL BUSINESS CULTURE

WILL BEDINGFIELD

CULTURE SEP 27, 2023 2:58 PM

Hollywood Writers Reached an AI Deal That Will Rewrite History

A faction of scribes is putting guardrails around AI's encroachment on their work. The effects will echo in industries far beyond Hollywood.

Fake news



Fake news

[Send us a Tip!](#) | [Shop](#) | [Subscribe](#)

GIZMODO

The Future Is Here

We may ea

Search Q

[HOME](#)

[LATEST](#)

[NEWS](#)

[HARDWARE](#)

[SCIENCE](#)

[EARTHER](#)

[IO9](#)

[AI](#)

[SPACE](#)

[EN ESPAÑOL](#)

[VIDEO](#)

NEWS

No, Biden Isn't Dead: AI Content Farms Are Here, and They're Pumping Out Fake Stories

A new report found 49 different websites secretly using AI to churn out low-quality posts and rake in advertising revenue.

By **Mack DeGeurin** Published May 1, 2023 | [Comments \(3\)](#)



Fake news

🕒 This article is more than **7 months old**

Revealed: the hacking and disinformation team meddling in elections

- **'Team Jorge' unit exposed by undercover investigation**
- **Group sells hacking services and access to vast army of fake social media profiles**
- **Evidence unit behind disinformation campaigns across world**
- **Mastermind Tal Hanan claims covert involvement in 33 presidential elections**

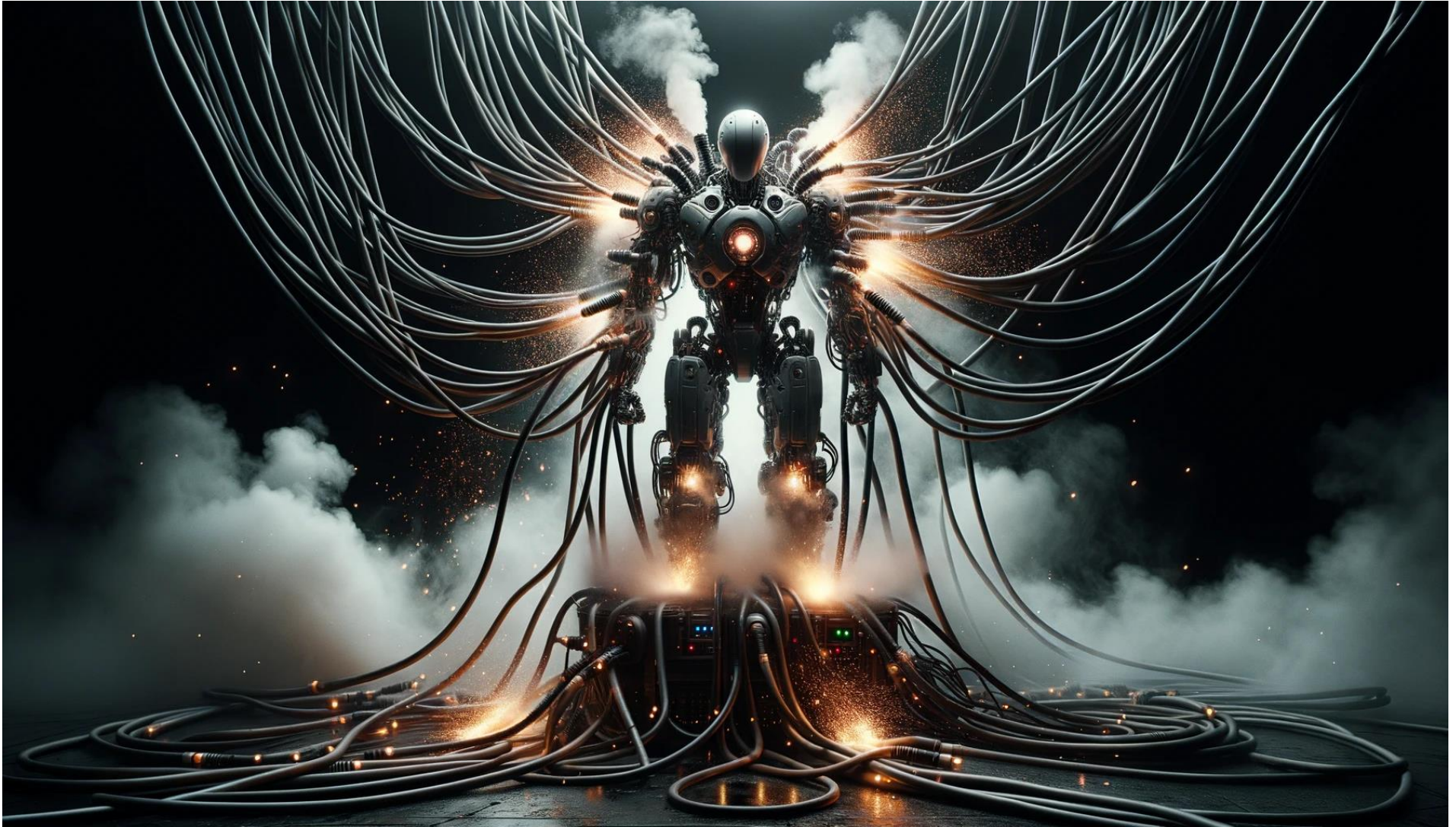


<https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan>
<https://www.blick.ch/fr/news/suisse/videotox-vivement-critiquee-ce-conseiller-national-udc-utilise-un-deepfake-dune-colleegue-verte-pour-faire-sa-pub-id19050109.html>

Fake news



Energy and climate



Existential risks

Autonomy

Agency

Self-preservation



Existential risks

Autonomy

Agency

Self-preservation



Existential risks

Autonomy

Agency

Self-preservation



Existential risks

Autonomy

Agency

Self-preservation



My point of view

Risks to **humanity**
are much greater than
risks to **humankind**

Activism

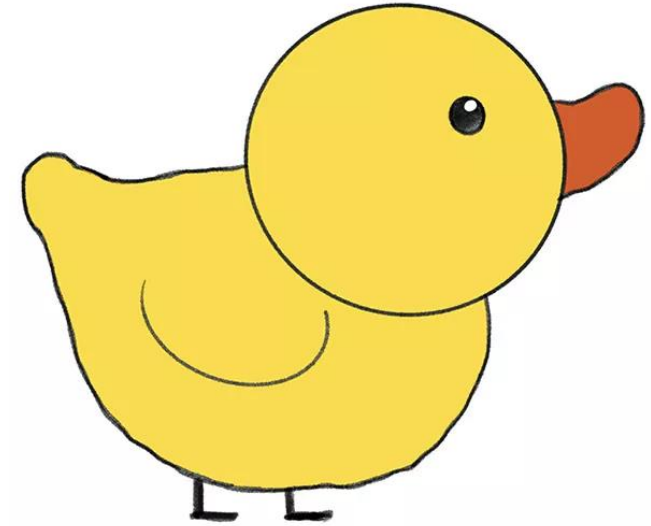
Protagonists

- Sceptical of regulation
 - Andrew Ng (AI researcher and entrepreneur)
 - **Yann Lecun** (Meta AI research)
 - ...
- Proponents of regulation:
 - **Yoshua Bengio** (scientific director MILA)
 - Stuart Russell (prof @ UC Berkeley)
 - **Geoffrey Hinton** (formerly google AI)
 - Elon Musk (co-founder Open.ai, Tesla founder, Twitter CEO,...)
 - Steve Wozniak (co-founder Apple)
 - Yuval Noah Harari (Prof. Hebrew University of Jerusalem)
 - ...



Protagonists

- Differences in opinion:
 - Is GAI really intelligent?
 - How far away is AGI?
 - How to control?
 - Long-term risks
 - Short-term risks
- Mostly agreed on:
 - Enormous positive potential of AI
 - It's better to regulate the *use* of AI than the *technology itself*



Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

27565

Add your
signature

Open Brief: Onze samenleving is niet klaar voor de risico's van manipulatieve AI – dringend nood aan actie

Pour la version française de la lettre ouverte, [cliquez ici](#).

For the English version of the open letter, [click here](#).

Deze brief verscheen ook in [Knack](#) op 29 maart 2023.

Nathalie A. Smuha, Mieke De Ketelaere, Mark Coeckelbergh, Pierre Dewitte en Yves Pouillet - 31 maart 2023

Google

'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation

The neural network pioneer says dangers of chatbots were 'quite scary' and warns they could be exploited by 'bad actors'

**Josh Taylor and
Alex Hern**

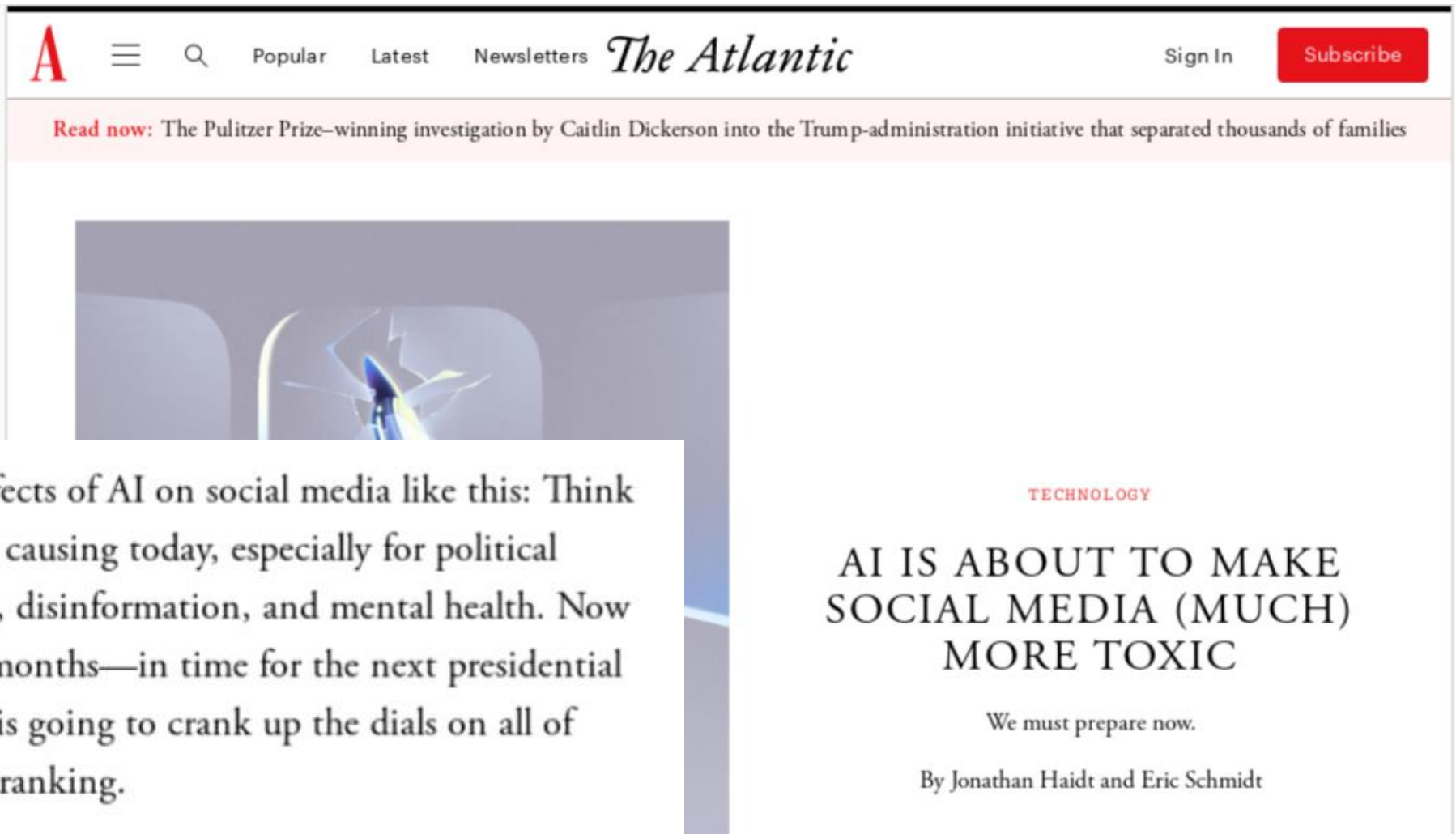
Tue 2 May 2023 12.23 BST



By **Invitation** | Artificial intelligence

Yuval Noah Harari argues that AI has hacked the operating system of human civilisation

Storytelling computers will change the course of human history,
says the historian and philosopher



We can summarize the coming effects of AI on social media like this: Think of all the problems social media is causing today, especially for political polarization, social fragmentation, disinformation, and mental health. Now imagine that within the next 18 months—in time for the next presidential election—some malevolent deity is going to crank up the dials on all of those effects, and then just keep cranking.



Center *for*
AI Safety

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.



‘Persoonlijke AI-vriendjes maken gerichte manipulatie mogelijk op ongeziene schaal’

Tijl De Bie

Professor datawetenschappen aan de UGent • 15-09-2023, 11:30 •

Bijgewerkt op: 15-09-2023, 14:42 •

(European) Legislation

EU AI Act – Structure

A risk-based approach

Given the nature of foundation models, expertise in conformity assessment is lacking and third-party auditing methods are still under development –The sector itself is therefore developing new ways to assess fundamental models that fulfil in part the objective of auditing (such as model evaluation, red-teaming or machine learning verification and validation techniques). Those internal assessments for foundation

Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical

Permitted subject to compliance
with AI requirements and ex-ante
conformity assessment

*Not mutually
exclusive

As foundation models are a new and fast-evolving development in the field of artificial intelligence, it is appropriate for the Commission and the AI Office to monitor and periodically assess the legislative and governance framework of such models and in particular of generative AI systems based on such models, which raise significant questions related to the generation of content in breach of Union law, copyright rules, and potential misuse.

EU AI Act – Protest

European companies claim the EU's AI Act could 'jeopardise technological sovereignty'



/ Over 150 executives from companies like Renault, Heineken, Airbus, and Siemens have signed an open letter urging the EU to rethink its plans to regulate AI.

Other regulations

The future of AI policy in China

27 September 2023

Authors: Huw Roberts, University of Oxford and Emmie Hine, University of Bologna

Rapid developments in generative artificial intelligence (AI) — algorithms **used to create** new text, pictures, audio, or other types of content — are concerning regulators globally. These systems are often trained on personal and **copyrighted data** scraped from the internet, leading to privacy and intellectual property fears. They can also be used to generate harmful **misinformation and disinformation**.

China launches Global AI Governance Initiative, offering an open approach in contrast to US blockade

Move stresses openness, fairness, in stark contrast to US' bullying restrictions

By Wang Cong and Yin Yeping

Published: Oct 18, 2023 05:33 PM



**Hiroshima Process International
Guiding Principles for Organizations
Developing Advanced AI System**

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence



► BRIEFING ROOM

► STATEMENTS AND RELEASES

The cure versus the disease

The Digital Services Act aims to address disinformation

- Pretext / prime purpose: illegal content on online platforms
 - Search engines, online shops, social media
- In practice: mis/mal/disinformation
 - Content that poses a risk to civic discourse, public health, election integrity,...

Risks of the Digital Services Act



EUROPEAN UNION

EUROZONE

INVESTMENT

PODCAST

MORE

DONATE

ADVE

Home > European Union > The EU's Digital Services Act (DSA) may undermine fundamental rights

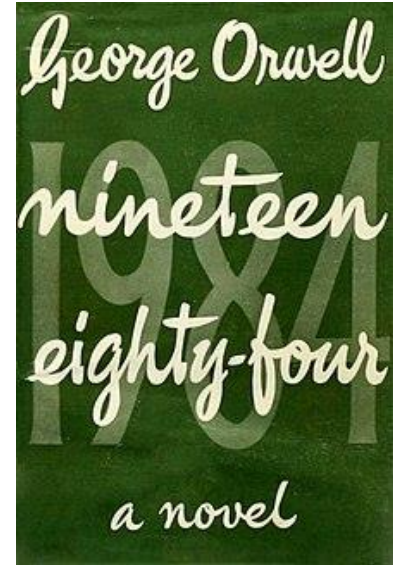
European Union

Uncategorized

The EU's Digital Services Act (DSA) may undermine fundamental rights

By [BrusselsReport.eu](https://brusselsreport.eu) - September 11, 2023

Digital Services Act



It is easier to distinguish
agreeable content from undesirable content,

than it is to distinguish
facts from fiction or
human-generated from bot-generated content

What's next?

Possible solutions

- Regulation:
 - Personal AI friends, deep-fakes, manipulative AI, authentication and watermarking, age limits,...
- Research:
 - AI 'alignment', formalizing ethics, sociological/anthropological/psychological research
- Society
 - Education, awareness, political and ideological diversity among those controlling AI, European investments,...

Democratizing AI

Leaked Google AI memo:

"We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?"

But the uncomfortable truth is, we aren't positioned to win this arms race and neither is OpenAI. While we've been squabbling, a third faction has been quietly eating our lunch.

I'm talking, of course, about open source.

Plainly put, they are lapping us. Things we consider "major open problems" are solved and in people's hands today."

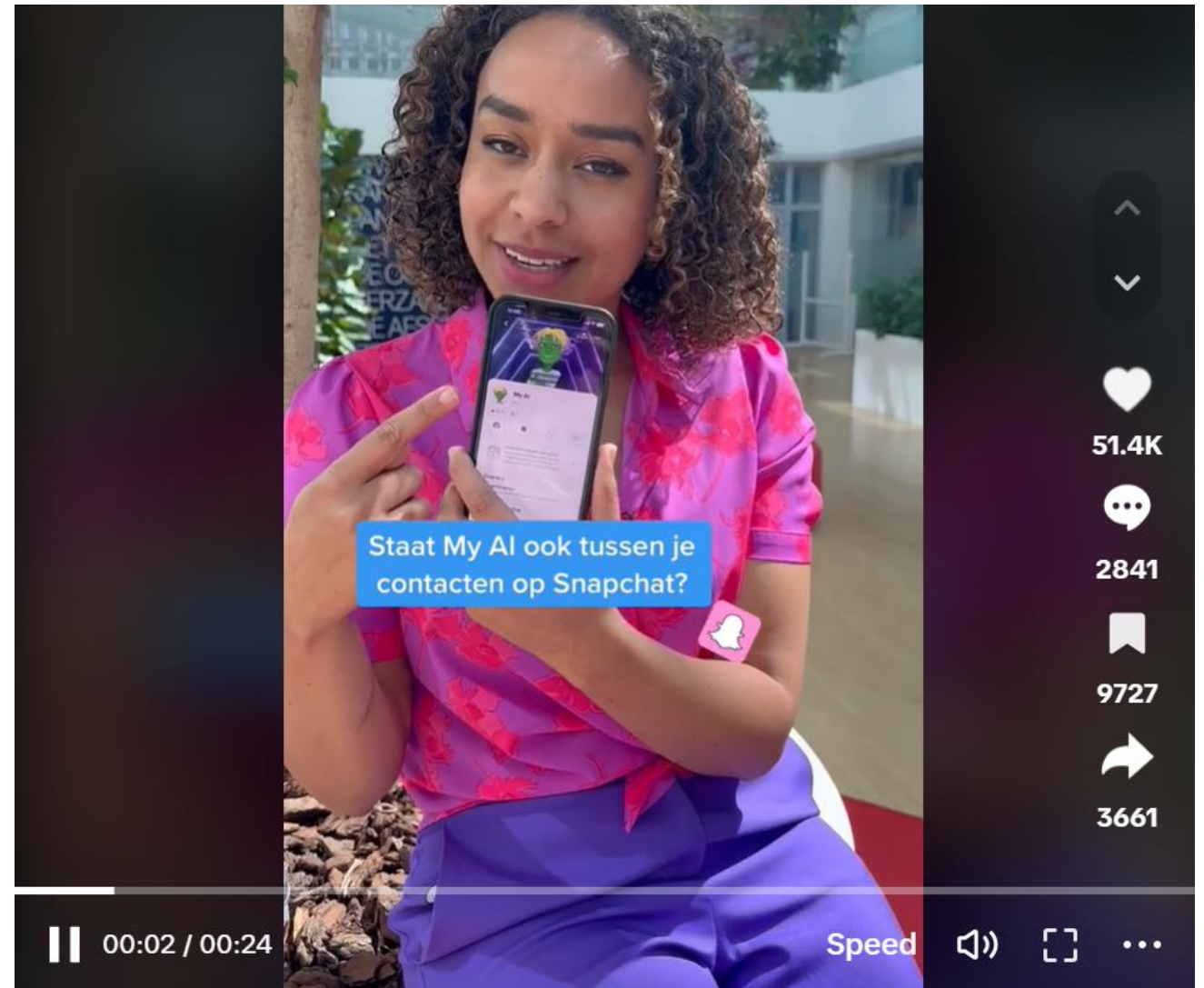
The bulk of the memo is spent describing how Google is outplayed by open source.

In the meantime:

Education and awareness

Media

Political debate



Where is this going?

Safe assumption: shortcomings (e.g. hallucinations) will be resolved with time

Enormous positive potential, strategically critical

Large known (and undoubtedly unknown) risks

International regulation is inevitable

Discussion