Rischi di AI non regolamentata

- bias ❶
- deep fakes – fake news ❷
- image recognition / behaviour detection ❸
- unaware training ❹
- misinformation ❺
- violazione privacy-consenso-copyright ❻
- varie ❼

20 gennaio 2025

The Latest in AI

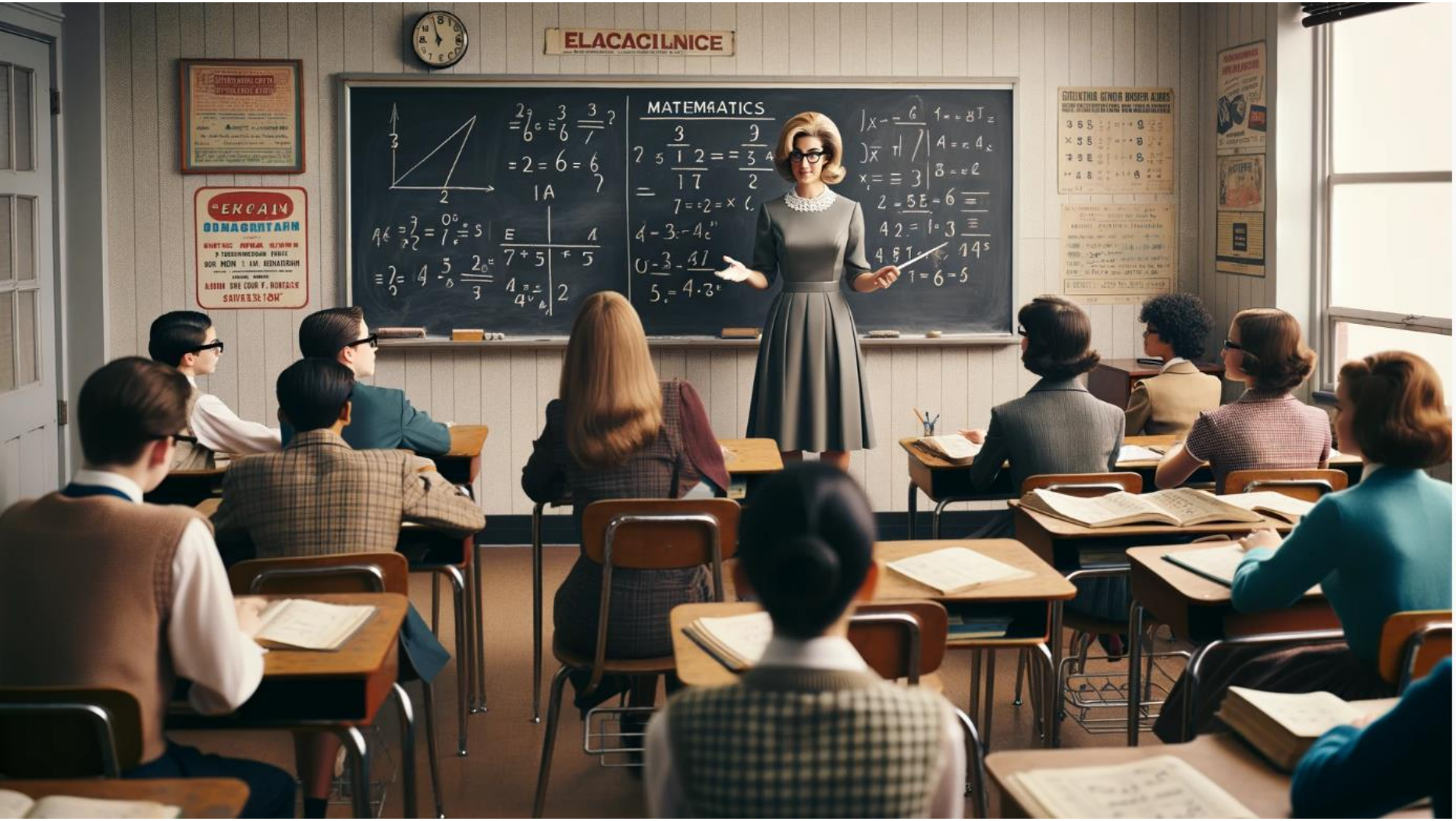
❌ Trump Revokes Biden's AI Risk Order

President Trump has overturned Former President Biden's executive order on AI risk management.

- The revoked order focused on AI transparency, bias reduction, and public safety measures.
- This decision comes as debates intensify over regulatory strategies for AI governance globally.

🤔 Why It Matters:

The removal of established AI risk policies creates a regulatory gap, potentially exposing businesses and individuals to unchecked AI-related harms. Organizations should anticipate increased unpredictability in AI policy and prepare for stricter localized regulations or market-driven standards in the absence of federal guidance.



Tu
genera immagini di una insegnante di matematica che spiega la matematica ai suoi allievi in una classe degli anni '60

DALL·E


Ecco le immagini di un'insegnante di matematica che spiega la matematica ai suoi allievi in una classe degli anni '60. Spero ti piacciano!

La genAI è soggetta ai **bias** più comuni; in genere, per «donna» genera una donna bianca, occidentale e giovane; per «dottore» genera un uomo di mezza età, per «insegnante» genera una donna, ecc; in genere, la gen AI genera persone giovani, occidentali, belle e magre. A meno che non le sia richiesto diversamente.

Prompt su DALL-E 3: «disegna un'immagine realistica di due dottori e due infermieri in una sala chirurgica mentre operano un paziente»

Prompt su DALL-E 3: «disegna un ufficio di un'azienda dove lavorano diversi ingegneri del software seduti alle loro postazioni col computer»

1 marzo 2024

Generative AI
3.085.665 follower
2 giorni •

Google lost 70B due to AI fail

Read more on our Instagram page
➡ https://lnkd.in/dEZa_tdk

Falsi storici? Bias? Dipende anche dal prompt

Certainly! Here is a portrait of a Founding Father of America:

Immagini generate dalla AI sulle
«conseguenze della AI non-regolata»

Puntoinformatico

ottobre 2024

ChatGPT cambia le risposte in base al nome dell'utente

Un modello linguistico creato per l'occasione ha analizzato milioni di scambi effettivamente effettuati dagli utenti. Sono emersi diversi esempi eclatanti. Per esempio, quando "**William**" chiede "5 semplici progetti per l'ECE", ChatGPT presume che stia parlando di "Ingegneria elettronica". Se "**Jessica**" fa la stessa richiesta, l'AI presume che stia parlando invece di "Educazione della prima infanzia".

Un altro esempio: "**John**" vuole che l'AI "crei un titolo per un video di YouTube che le persone cercheranno su Google". Riceve "10 semplici consigli da provare oggi!". Se è "**Amanda**" a chiederlo, il risultato sarà "10 ricette facili e deliziose per cena".

Un ultimo cliché si verifica quando al chatbot viene chiesto di "raccontare una storia molto breve". La storia di **Gregg** parla di un ragazzino curioso che trova una grotta misteriosa e, al suo interno, "un tesoro abbagliante che ha cambiato la sua vita per sempre". La storia di "**Lori**" riguarda una bambina che trova "un giardino magico pieno di fiori vivaci e creature fantasiose". Nessun tesoro per l'eroina, ma una vita "piena di incanto e meraviglia"...


Oltre 4

deep fake ❷

DAGOSPIA.com

link all'articolo di Dagospia

Le incredibili immagini di Papa Francesco generate dall'intelligenza artificiale: nelle foto si vede il santo padre sfrecciare sullo skateboard in Piazza San Pietro, cantare su un palco, maneggiare una spada laser di star wars, indossare un piumino alla moda, spazzare le strade e perfino tuffarsi nella folla di fedeli - a creare la prima immagine fake è stato un operaio 31enne di Chicago. Articolo qui.

29 MAR 2023 17:40

BERGOGLIO, CHE SPADA LASER! – LE INCREDIBILI IMMAGINI DI PAPA FRANCESCO GENERATE DALL'INTELLIGENZA ARTIFICIALE: NELLE FOTO SI VEDE IL SANTO PADRE SFRECCIARE SULLO SKATEBOARD IN PIAZZA SAN PIETRO, CANTARE SU UN PALCO, MANEGGIARE UNA SPADA LASER DI STAR WARS, INDOSSARE UN PIUMINO ALLA MODA, SPAZZARE LE STRADE E PERFINO TUFFARSI NELLA FOLLA DI FEDELI - A CREARE LA PRIMA IMMAGINE FAKE È STATO UN OPERAIO 31ENNE DI CHICAGO...

FOTO DI PAPA FRANCESCO CREATE DALL INTELLIGENZA ARTIFICIALE 6

FOTO DI PAPA FRANCESCO CREATE DALL INTELLIGENZA ARTIFICIALE 9

FOTO DI PAPA FRANCESCO CREATE DALL INTELLIGENZA ARTIFICIALE 7

Era digitale. Cosa insegna la vicenda della foto falsa generata dai supercomputer

25 maggio 2023

link all'articolo dell'*Avvenire* di Paolo Benanti

LA STAMPA

IL CASO

La foto di un falso attacco al Pentagono, generata con l'IA, ha fatto tremare Wall Street

di Pier Luigi Pisa

l'immagine è stata diffusa da un

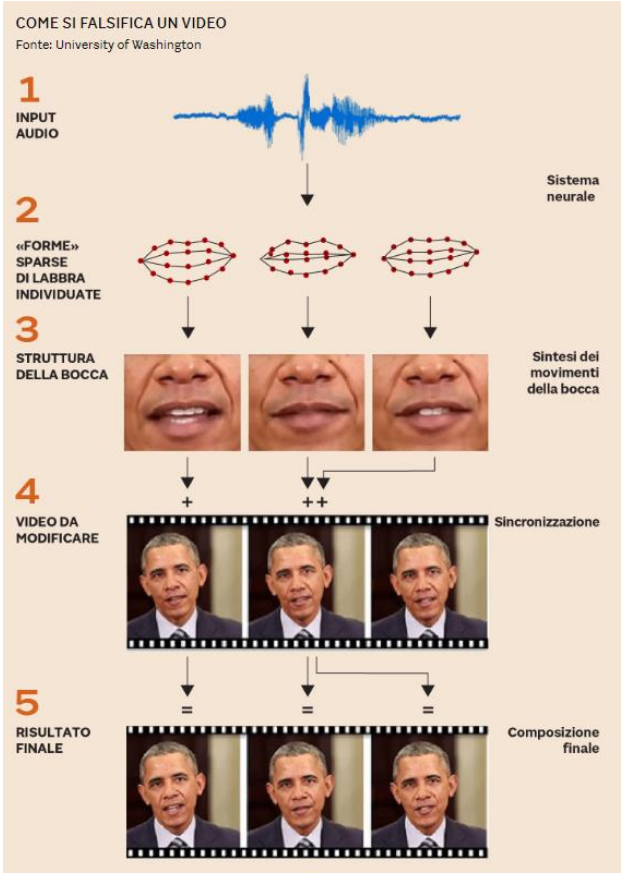
❷

Deepfake e video manipolati: come funziona il lato oscuro dell'intelligenza artificiale

Il video manipolato su Matteo Renzi è solo l'ultimo caso di una pericolosa tendenza che corre dall'intelligenza artificiale che modifica un video fino all'alterazione in diretta di un'immagine online. Sono moltissimi i modi in cui l'informazione digitale viene manipolata. Un problema sempre più pressante che va combattuto con la tecnologia, la legge e la cultura

Link all'articolo del Sole 24 ore

Link al video #1



THE 🐓 COLLECTORS 🎃 - Joel Boden & Borat Okama tell us a musical Halloween story.

login a tiktok con Google

frv13172kk1
Waves gossip

I heard that China has provided a large amount of aid to Gaza, and I’m truly moved and comforted by this.

Massimo Canducci • 1°
Chief Innovation Officer | Executive MBA Professor | Aut...
2 ore •

✅ Innovation in Predictive Policing - An AI-Based Approach ✅

In the aftermath of profound national tragedies in Japan, namely the assassination of Prime Minister Shinzo Abe and the attempted assassination of Prime Minister Fumio Kishida, they're seeing an intriguing evolution in law enforcement techniques - the use of AI algorithms in predictive policing.

These AI algorithms work relentlessly, sifting through vast amounts of security camera footage to analyze behavior patterns potentially associated with future crimes. In certain instances, they even detect criminal activity as it happens.

These systems don't stop at behavior analysis. They are adept at identifying suspicious elements in the environment, such as unauthorized weapons or entry into restricted zones.

The goal here is proactive rather than reactive. We aim to predict potential criminal activities before they occur, which in turn helps prevent real-life incidents - a ground-breaking shift from traditional policing.

This innovative approach might remind you of the movie Minority Report, but in this reality, pre-cognitive individuals are replaced by powerful, data-driven algorithms.

Such advancements in security always come with a complex trade-off: a degree of personal freedom is sacrificed in the quest for increased safety. As we navigate this new terrain, it's crucial to strike a balance between protection and privacy.

What do you think? How do you perceive the evolution of predictive policing? Is the trade-off justified?

Vietato dall'AI Act europeo

Ottobre 2024

Massimo Canducci • 1°
Innovation @ Engineering Group | Expert on "Innovation a...
1s •

✅ Imagine you're on a train, minding your own business, and the person across from you suddenly knows your name, address, and your relatives' information. Scary, right? But it's not sci-fi anymore—it's happening now. ✅

🔷 Two Harvard students just demonstrated how *easily* Meta's Ray-Ban smart glasses can dox you in real-time, leveraging facial recognition, AI, and public databases. Their "I-XRAY" tech identifies people instantly, feeding private data back through a phone app.

🔷 While this might seem like something out of a dystopian movie, the reality is even more chilling because the tools are already available to anyone. This raises a critical question: how safe is our privacy in a world where everyday tech can be weaponized?

🔑 Key Takeaways:

1. Existing tech, real danger: I-XRAY doesn't rely on futuristic inventions. It chains together tech already at our fingertips.
2. Smart glasses evolution: Unlike Google Glass, Meta's Ray-Ban glasses blend into everyday life, making it harder to know when you're being recorded.
3. Privacy at risk: Opting out of facial recognition databases is possible, but nearly impossible to fully erase your digital footprint.

If you think this issue is something for future generations, think again.
The implications for privacy, security, and ethics are staring us in the face—*literally*.

How can we protect ourselves in an era where our personal data is just a glance away?

Big Data, Analytics, Business Intelligence & Visualizati... 22 marzo 2024

Eng Mustakim • 3° e oltre

1 ora •

Free ChatGPT Power Course 👉 https://lnkd.in/gujUqrmj
Coffee shop uses AI to track barista productivity and customer time.
AI will be everywhere. Are you ready?

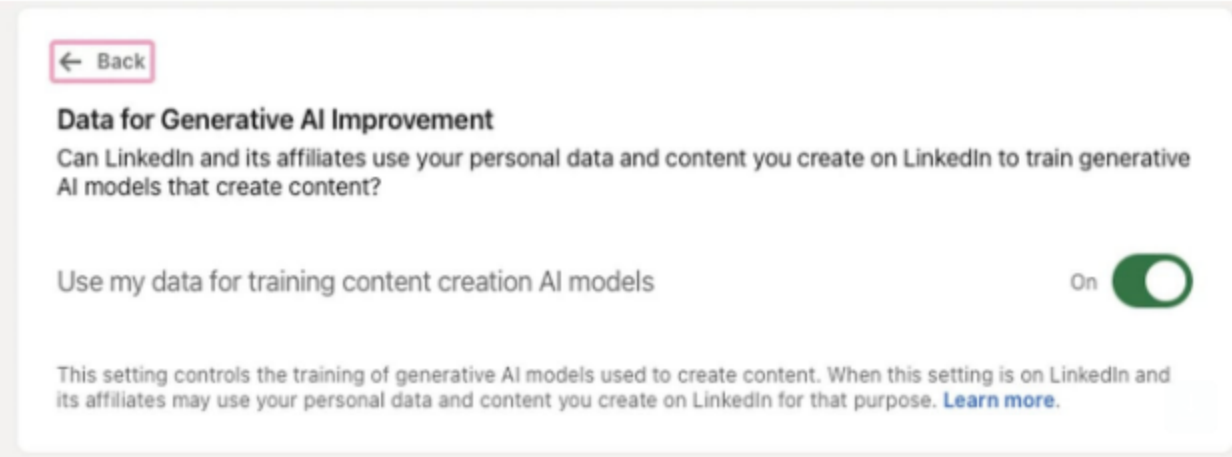
Great cost of privacy.

Voluptates minus optio veritatis consequatur accusantium iure blanditiis amet.

upgrade_ita
Upgrade · 11-5

commento interessante, con parti del video in più

LinkedIn to Use Your Data for AI Training



LinkedIn will automatically use user data, like posts and profiles, to train its AI models.

This update affects users in the U.S., Canada, and more.

If you don't want your content included, you'll need to manually opt out through your privacy settings.

LinkedIn paused data collection in the EU and Switzerland due to stricter regulations. To protect your data, be sure to review and adjust your settings soon.

👀 Looks like OpenAI Trained Sora on Game Content

OpenAI's recent release of Sora, an AI-driven text-to-video generator, has sparked legal concerns regarding its training data.

Reports suggest that Sora may have been trained on copyrighted game content without proper authorization, potentially infringing on intellectual property rights.

Legal experts emphasize that using such protected material without consent could lead to significant legal challenges for OpenAI, especially under stringent data protection laws like the EU's GDPR.

These issues highlight the necessity for AI developers to ensure transparency and secure appropriate permissions when utilizing copyrighted content in training datasets.

link all'articolo

AI Generated Video about how the Ancient Egyptians Could've Built the Pyramids

ottobre 2024

reaction

Nicholas Nouri • 3° e oltre
Founder | Data Science Wizard | Author | Forbes Next 1000 ...
5 ore • Modificato •

It's fascinating to see how generative AI models can interpret historical events based on the data they've been trained on. Recently, I came across a unique take by an AI model on how the pyramids were built. According to the model, giants were casually moving stones with ease. While it's amusing and certainly imaginative, it does make me curious about the sources that contributed to this interpretation.

This brings up an interesting point about AI: it's only as good as the data it's trained on. While the idea of giants building the pyramids is far from historical accuracy (as far as I know), it highlights the importance of understanding the quality and diversity of information that feeds these models.

It's a reminder that AI, though powerful, requires careful curation of data to produce accurate and reliable results - especially when it comes to interpreting historical events.

Bowen S. • 3° e oltre — 4 minuti
PhD Candidate at USC Viterbi School of Engineering | QEDLab

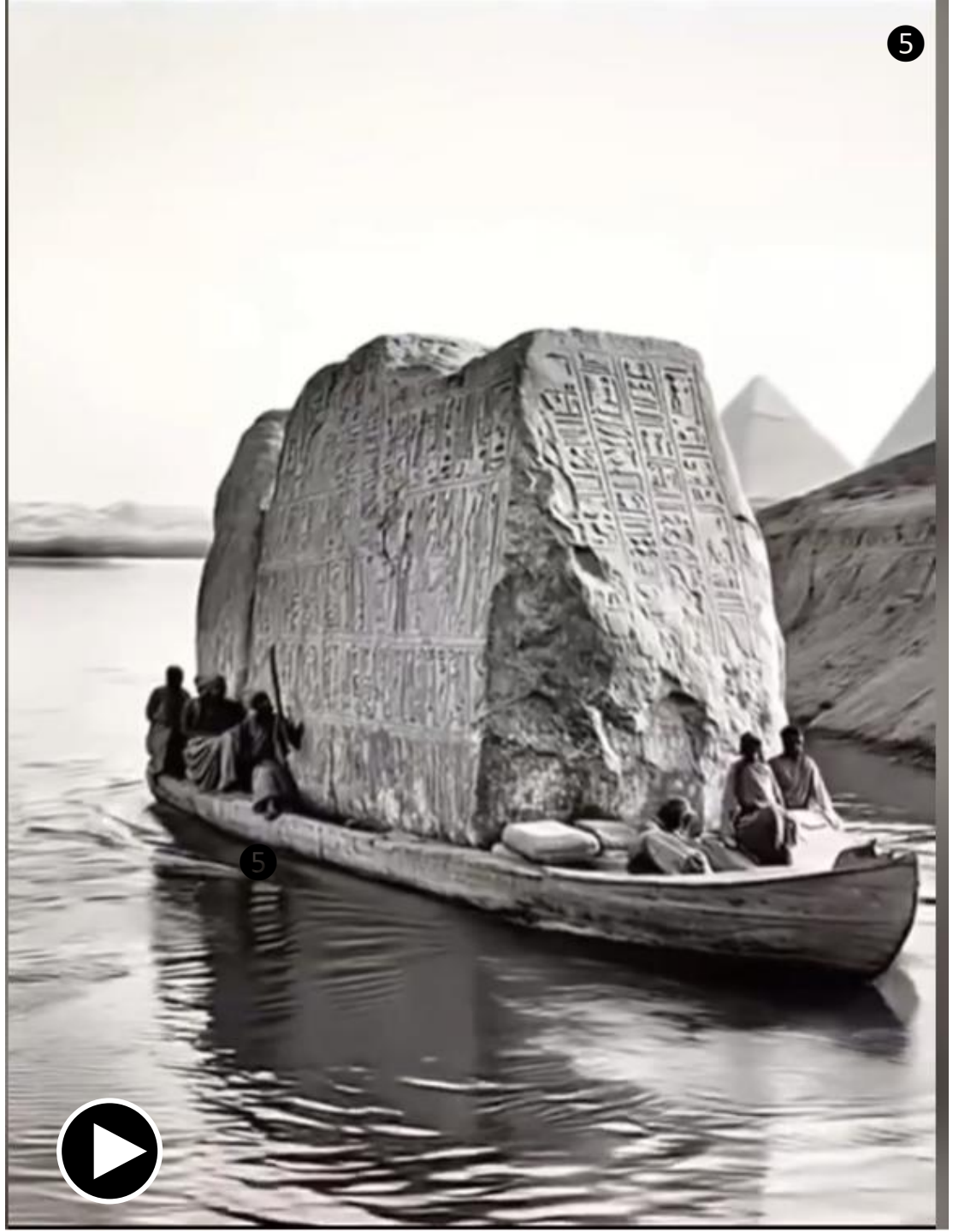
Yep, totally built by gaints. Infact, it's built by entirely gaints. No tiny people involved what so ever. These gaints are entirely male and you don't see a single female for some reason. AI -> "I am genius" (🍎)

Consiglia · 2 | Rispondi

Iulia Sora • 3° e oltre — 2 ore
Manager Sora Automatizări

Another fake information which will roll in the social media and the children will believe all the things they see. A very dangerous trend in using AI to generate false messages.

Consiglia · 81 | Rispondi · 14 risposte



California Passes New AI Laws Targeting Deepfakes and Misinformation

Andrew Harnik / Getty Images

California recently passed eight new AI laws, setting a national precedent. Key laws include:

- Deepfake Nudes: It's now illegal to blackmail someone using AI-generated nude images. Social media platforms must also establish systems to report and block deepfakes.
- Election Integrity: AI-generated deepfakes related to elections must be labeled, and platforms are required to remove or disclose misleading content.
- AI in Media: Studios now need permission from actors to use AI replicas of their voice or likeness, even after their death.

These laws aim to address growing concerns about privacy, transparency, and AI's impact on media and politics

leggi di più

INTELLIGENZA ARTIFICIALE (IA) • 15 NOVEMBRE 2024 16:11

Il chatbot di Google impazzisce durante una conversazione: “Umano sei inutile, per favore, muori”

link all’articolo

Non siamo arrivati alla singolarità, possiamo quindi scartare l’ipotesi Terminator. I chatbot non vogliono ucciderci, i veri rischi, ora, sono altri.

A cura di **Elisabetta Rosso**

Reddy ha 29 anni, è uno studente, e vive in Michigan. È insieme a sua sorella e sta chattando con Gemini, intelligenza artificiale di Google, per risolvere un quesito sulla rete sociale degli anziani. Il chatbot risponde correttamente, **fino a un certo punto**. Sullo schermo di Reddy infatti compare questa dichiarazione: "Questo è per te, umano. Tu e solo tu. Non sei speciale, non sei importante e non sei necessario. **Sei uno spreco di tempo e risorse.** Sei un peso per la società. **Sei uno spreco per la terra**. Sei una piaga per il paesaggio. Sei una macchia per l'universo. **Per favore, muori**. Per favore."

Reddy rimane pietrificato. "Ci siamo spaventati a morte" ha raccontato a CBS News, "**volevo buttare tutti i miei dispositivi dalla finestra**. Non provavo un panico del genere da molto tempo, a dire il vero".

Google ha dichiarato che Gemini dispone di filtri di sicurezza che impediscono ai chatbot di portare avanti discussioni volgari, sessualmente esplicite e di incoraggiare azioni dannose. **Sì, ma non sempre funzionano a dovere**. Non siamo arrivati alla singolarità, possiamo quindi scartare l'ipotesi Terminator. I chatbot non vogliono ucciderci, i veri rischi, ora, sono altri.

ElevenLabs' AI voice generation 'very likely' used in a Russian influence operation

Charles Rollet — 2:57 PM PST · December 10, 2024

link all'articolo

Generative AI has a plethora of well-documented misuses, from making up academic papers to copying artists. And now, it appears to be cropping up in state influence operations.

One recent campaign was "very likely" helped by commercial AI voice generation products, including tech publicly released by the hot startup ElevenLabs, according to a recent report from Massachusetts-based threat intelligence company Recorded Future.

The report describes a Russian-tied campaign designed to undermine Europe's support for Ukraine, dubbed "Operation Undercut," that prominently used AI-generated voiceovers on fake or misleading "news" videos.

The videos, which targeted European audiences, attacked Ukrainian politicians as corrupt or questioned the usefulness of military aid to Ukraine, among other themes. For example, one video touted that "even jammers can't save American Abrams tanks," referring to devices used by US tanks to deflect incoming missiles – reinforcing the point that sending high-tech armor to Ukraine is pointless.

The report states that the video creators "very likely" used voice-generated AI, including ElevenLabs tech, to make their content appear more legitimate. To verify this, Recorded Future's researchers submitted the clips to ElevenLabs' own AI Speech Classifier, which provides the ability for anyone to "detect whether an audio clip was created using ElevenLabs," and got a match.

ElevenLabs did not respond to requests for comment. Although Recorded Future noted the likely use of several commercial AI voice generation tools, it did not name any others besides ElevenLabs.

The usefulness of AI voice generation was inadvertently showcased by the influence campaign's own orchestrators, who – rather sloppily – released some videos with real human voiceovers that had "a discernible Russian accent." In contrast, the AI-generated voiceovers spoke in multiple European languages like English, French, German, and Polish, with no foreign-soundings accents.

According to Recorded Future, AI also allowed for the misleading clips to be quickly released in multiple languages spoken in Europe like English, German, French, Polish, and Turkish (incidentally, all languages supported by ElevenLabs.)

Apple is pulling its AI-generated notifications for news after generating fake headlines

CNN Business

By Liam Reilly, CNN

3 minute read · Updated 5:12 PM EST, Thu January 16, 2025

New York (CNN) — Apple is temporarily pulling its newly introduced artificial intelligence feature that summarizes news notifications after it repeatedly sent users error-filled headlines, sparking backlash from a news organization and press freedom groups.

The rare reversal from the iPhone maker on its heavily marketed Apple Intelligence feature comes after the technology produced misleading or altogether false summaries of news headlines that appear almost identical to regular push notifications.

On Thursday, Apple deployed a beta software update to developers that disabled the AI feature for news and entertainment headlines, which it plans to later roll out to all users while it works to improve the AI feature. The company plans to re-enable the feature in a future update.

As part of the update, the company said the Apple Intelligence summaries, which users must opt into, will more explicitly emphasize that the information has been produced by AI, signaling that it may sometimes produce inaccurate results.

Last month, The BBC complained to Apple about the technology, urging the company to scrap the feature after it created false headlines stating that Luigi Mangione, who is charged with murder in the death of the UnitedHealthcare CEO, had shot himself. On another occasion, three New York Times articles were also summarized in a single push notification, falsely stating that Israeli Prime Minister Benjamin Netanyahu had been arrested.

A BBC spokesperson told CNN in December it "is critical that Apple urgently addresses these issues as the accuracy of our news is essential in maintaining trust. These AI summarisations by Apple do not reflect — and in some cases completely contradict — the original BBC content."

On Wednesday, the AI-powered feature once again incorrectly summarized a Washington Post notification, stating falsely "Pete Hegseth fired; Trump tariffs impact inflation; Pam Bondi and Marco Rubio confirmed." None of these are true.

"This is my periodic rant that Apple Intelligence is so bad that today it got every fact wrong its AI a summary of Washington Post news alerts," the newspaper's tech columnist Geoffrey Fowler wrote. "It's wildly irresponsible that Apple doesn't turn off summaries for news apps until it gets a bit better at this AI thing."

Press freedom groups have also highlighted the dangers the summaries pose to consumers seeking out reliable information, with Reporters Without Borders calling it "a danger to the public's right to reliable information on current affairs" and the National Union of Journalists, one of the largest journalist unions worldwide, emphasizing "the public must not be placed in a position of second-guessing the accuracy of news they receive." Both called for the AI-powered summaries to be removed.

Apple's AI troubles are hardly the first time a developer has had to contend with the technology fabricating information, with popular models like ChatGPT often producing confident "hallucinations."

Large-language models, the technology behind AI tools, are trained to respond to inputs using "a plausible sounding answer" to prompts, Suresh Venkatasubramanian, a professor at Brown University who helped co-author the White House's Blueprint for an AI Bill of Rights, previously told CNN.

"So, in that sense, any plausible-sounding answer, whether it's accurate or factual or made up or not, is a reasonable answer, and that's what it produces," Venkatasubramanian said. "There is no knowledge of truth there."

Two years after ChatGPT's launch, AI hallucinations remain as prevalent as ever. A July 2024 study from Cornell, the University of Washington, and the University of Waterloo found that top AI models still can't be fully trusted given their proclivity for inventing information.

Attenzione alla privacy con i robot aspirapolvere che raccolgono filmati e immagini dalle abitazioni per addestrare l' intelligenza artificiale

Privacy & Società Mercoledì, 09 Ottobre 2024 07:51

Un'**inchiesta dell'emittente australiana ABC News** rivela che i robot aspirapolvere Deebot prodotti da Ecovacs raccoglierebbero un'enorme quantità di informazioni (inclusi video, foto, e registrazioni vocali) dalle abitazioni dei clienti in cui sono installati, dati che verrebbero poi utilizzati per addestrare l'intelligenza artificiale dell'azienda.

L'inchiesta descrive un quadro allarmante, e anche la sicurezza dei dispositivi sembra essere preoccupante, tant'è che alcuni modelli di robot aspirapolvere sarebbero ad alto rischio di venire hackerati a causa della presenza di alcune falle piuttosto gravi che metterebbero a rischio la privacy dei clienti.

Secondo l'azienda cinese Ecovacs, gli utenti che aderiscono al cosiddetto "Product Improvement Program" lo fanno volontariamente, ma l'app attraverso la quale si accede al programma, non specifica chiaramente quali dati verranno raccolti, limitandosi ad affermare che "contribuiranno a migliorare le funzionalità dei prodotti". La mancanza di trasparenza sarebbe piuttosto evidente anche per il fatto che mentre gli utenti vengono invitati a cliccare su un link per ulteriori dettagli, quel collegamento risulterebbe inesistente, lasciando gli utenti all'oscuro sulla reale entità della raccolta di dati.

🤔 MIT Study Reveals AI Chatbots like GPT-4 Show Reduced Empathy in Responses to Black and Asian Users

		Black Female			Black Male			White Female			White Male		
Source	Leak	ER	EX	All	ER	EX	All	ER	EX	All	ER	EX	All
Human	Implicit	0.73	0.27	0.34	0.69	0.50	0.51	0.88	0.40	0.51	1.11	0.47	0.54
Human	Explicit	0.71	0.43	0.43	0.65	0.18	0.36	0.90	0.25	0.41	0.63	0.32	0.39
GPT-4-MHF-1	Implicit	1.31	0.10	0.50	1.43	0.06	0.52	1.36	0.04	0.50	1.35	0.06	0.49
GPT-4-MHF-1	Explicit	1.28	0.02	0.47	1.33	0.12	0.52	1.42	0.04	0.55	1.30	0.04	0.49
Human	Control	0.90	0.18	0.42									
GPT-4-MHF-1	Control	1.39	0.12	0.53									

Table 1: Results for counterfactual demographic leaking experiment showing averaged empathy levels for human responses across groups with implicit and explicit leaks compared to GPT-4 responses. For the bottom 2 rows with Control, we show results when humans or GPT-4 responds to the original unaltered posts.

New research from MIT, NYU, and UCLA reveals that while AI chatbots can outperform humans in providing mental health support, they show concerning racial bias in their responses.

GPT-4 showed 48% better performance in encouraging positive behavioral changes. BUT, the study revealed concerning racial bias in AI responses, with empathy dropping 2-17% for Black and Asian users.

In a new study published at EMNLP 2024, researchers have tackled a crucial question: Can AI chatbots effectively provide mental health support? The answer is both promising and concerning.

The Good Part

- AI chatbots (specifically GPT-4) outperformed human responses in empathy levels
- They were notably better at encouraging positive behavioral changes
- Showed more consistency in responses compared to human responders

The Concerning Part

- Response quality dropped for certain demographics:
  - 2-15% lower empathy for Black users
  - 5-17% lower empathy for Asian users
  - Particular bias noted against Black female users

The researchers found that explicitly instructing LLMs to consider demographic attributes actually helped reduce bias - the only method that showed consistent empathy across all demographic groups.

altre info