

Τεχνολογία

Πώς να αποφύγετε να εξαπατηθείτε από παραπληροφόρηση που δημιουργείται από την τεχνητή νοημοσύνη

Η πρόοδος στη γενετική τεχνητή νοημοσύνη σημαίνει ότι οι ψεύτικες εικόνες, τα βίντεο, ο ήχος και τα bots είναι πλέον παντού. Αλλά οι μελέτες έχουν αποκαλύψει τους καλύτερους τρόπους για να πει κανείς αν κάτι είναι αληθινό

Του [Jeremy Hsu](#)

📅 2 Σεπτεμβρίου 2024




⚠️ Πολλές εικόνες που δημιουργούνται από AI φαίνονται ρεαλιστικές μέχρι να ρίξετε μια πιο προσεκτική ματιά
MidJourney


Παρατηρήσατε ότι η παραπάνω εικόνα δημιουργήθηκε από τεχνητή νοημοσύνη; Μπορεί να είναι δύσκολο να εντοπίσουμε εικόνες, βίντεο, ήχο και κείμενο που δημιουργούνται από την τεχνητή νοημοσύνη σε μια εποχή που οι τεχνολογικές εξελίξεις τα καθιστούν ολοένα και πιο δυσδιάκριτα από το πολύ περιεχόμενο που δημιουργείται από τον άνθρωπο, αφήνοντάς μας ανοιχτούς σε χειραγώγηση μέσω παραπληροφόρησης. Ωστόσο, γνωρίζοντας την τρέχουσα κατάσταση των τεχνολογιών τεχνητής νοημοσύνης που χρησιμοποιούνται για τη δημιουργία παραπληροφόρησης και το εύρος των ενδεικτικών ενδείξεων ότι αυτό που κοιτάζετε μπορεί να είναι ψεύτικο, μπορείτε να προστατεύσετε τον εαυτό σας από το να σας δεχτούν.

Οι παγκόσμιοι ηγέτες ανησυχούν. Σύμφωνα με [έκθεση του Παγκόσμιου Οικονομικού Φόρουμ](#) 🌐

<https://www.weforum.org/publications/global-risks-report-2024/in-full/> , η παραπληροφόρηση και η παραπληροφόρηση μπορεί να «διαταράξουν ριζικά τις εκλογικές διαδικασίες σε αρκετές οικονομίες τα επόμενα δύο χρόνια», ενώ η ευκολότερη πρόσβαση σε εργαλεία τεχνητής νοημοσύνης «έχουν ήδη επιτρέψει μια έκρηξη σε παραποιημένες πληροφορίες και τα λεγόμενα «συνθετικά». περιεχόμενο, από εξελιγμένη κλωνοποίηση φωνής έως πλαστές ιστοσελίδες».




Οι όροι παραπληροφόρηση και παραπληροφόρηση αναφέρονται και οι δύο σε ψευδείς ή ανακριβείς πληροφορίες, αλλά η παραπληροφόρηση είναι αυτή που έχει σκοπό να εξαπατήσει ή να παραπλανήσει.




<<Το ζήτημα με την παραπληροφόρηση που βασίζεται στην τεχνητή νοημοσύνη είναι η κλίμακα, η ταχύτητα και η ευκολία με την οποία μπορούν να ξεκινήσουν οι καμπάνιες>>, λέει ο [Hany Farid](https://www.ischool.berkeley.edu/people/hany-farid)  <https://www.ischool.berkeley.edu/people/hany-farid> στο Πανεπιστήμιο της Καλιφόρνια στο Μπέρκλεϋ. <<Αυτές οι επιθέσεις δεν θα λαμβάνουν πλέον κρατικούς φορείς ή καλά χρηματοδοτούμενους οργανισμούς – ένα άτομο με πρόσβαση σε κάποια μέτρια υπολογιστική ισχύ μπορεί να δημιουργήσει τεράστιες ποσότητες ψεύτικο περιεχόμενο>>.

Λέει ότι η γενετική τεχνητή νοημοσύνη (βλ. [γλωσσάρι, παρακάτω](#)  #DeepDive-1) <<μολύνει ολόκληρο το οικοσύστημα πληροφοριών, θέτει σε αμφιβολία ό,τι διαβάζουμε, βλέπουμε και ακούμε>>. Λέει ότι η έρευνά του δείχνει ότι, σε πολλές περιπτώσεις, οι εικόνες και ο ήχος που δημιουργούνται από την τεχνητή νοημοσύνη είναι <<σχεδόν δυσδιάκριτοι από την πραγματικότητα>>.

Ωστόσο, έρευνα του Farid και άλλων αποκαλύπτει ότι υπάρχουν στρατηγικές που μπορείτε να ακολουθήσετε για να μειώσετε τον κίνδυνο να πέσετε σε παραπληροφόρηση ή παραπληροφόρηση των μέσων κοινωνικής δικτύωσης που δημιουργείται από την τεχνητή νοημοσύνη.

Πώς να εντοπίσετε ψεύτικες εικόνες AI

Remember seeing a photo of [Pope Francis wearing a puffer jacket](#)  </article/2366312-should-you-be-worried-that-an-ai-picture-of-the-pope-went-viral/>? Such fake AI images have become more common as new tools based on diffusion models (see [glossary, below](#)  #DeepDive-1) have allowed anyone to start churning out images from simple text prompts. One [study](#)  <https://arxiv.org/abs/2405.11697> by Nicholas Dufour at Google and his colleagues found a rapid increase in the proportion of AI-generated images in fact-checked misinformation claims from early 2023 onwards.

“Nowadays, media literacy requires AI literacy,” says [Negar Kamali](#)  <https://scholar.google.com/citations?user=BtbeIckAAAAJ&hl=en> at Northwestern University in Illinois. In a 2024 [study](#)  <https://arxiv.org/abs/2406.08651>, she and her colleagues identified five different categories of errors in AI-generated images (outlined below) and provided guidance on how people can spot these for themselves. The good news is that their research suggests people are currently about 70 per cent accurate at detecting fake AI images of people. You can use their [online image test](#)  <https://detectfakes.kellogg.northwestern.edu/> to assess your own sleuthing skills.

5 common types of errors in AI-generated images:

- **Sociocultural implausibilities:** Is the scene depicting rare, unusual or surprising behaviour for certain cultures or historical figures?
- **Anatomical implausibilities:** Take a close look: are body parts like hands unusually shaped or sized? Do the eyes or mouths look strange? Have any body parts merged?
- **Stylistic artefacts:** Does the image look unnatural, almost too perfect or stylistic? Does the background look odd or like it is missing something? Is the lighting strange or variable?
- **Functional implausibilities:** Do any objects look bizarre or like they might not be real or work? For example, are buttons or belt buckles in weird places?
- **Violations of physics:** Are shadows pointing in different directions? Are mirror reflections consistent with the world depicted within the image?



▲ **Strange objects and behaviour can be clues that an image was created by AI**

MidJourney

How to identify video deepfakes

AI technology known as generative adversarial networks (see [glossary, below](#) 🗨️ #DeepDive-1) has allowed tech-savvy individuals to create [video deepfakes](#) 🗨️ /article/2418188-deepfakes-are-out-of-control-is-it-too-late-to-stop-them/ since 2014 – digitally manipulating existing videos of people to swap in different faces, create new facial expressions and insert new spoken audio aligned with matching lip-syncing. This has enabled a growing array of scammers, state-backed hackers and internet users to produce video deepfakes where celebrities such as [Taylor Swift](#) 🗨️ /article/2418740-could-an-ai-replace-all-music-ever-recorded-with-taylor-swift-covers/ and ordinary people alike may find themselves unwillingly featured in non-consensual deepfake pornography, scams and political misinformation or disinformation.

The techniques for spotting AI fake images (see above) can be applied to suspect videos too. Additionally, researchers at the Massachusetts Institute of Technology and Northwestern University in Illinois have compiled [some tips](#) 🗨️ <https://www.media.mit.edu/projects/detect-fakes/overview/> for how to spot such deepfakes, but they have acknowledged that there is no fool-proof method that always works.

6 tips for spotting AI-generated video:

- **Mouth and lip movements:** Are there moments when the video and audio aren't completely synced?
- **Anatomical glitches:** Does the face or body look weird or move unnaturally?
- **Face:** Look for inconsistencies in face smoothness or wrinkles around the forehead and cheeks, along with facial moles.
- **Lighting:** Is the lighting inconsistent? Do shadows behave as you would expect? Pay particular attention to a person's eyes, eyebrows and glasses.
- **Hair:** Does facial hair look weird or move in strange ways?
- **Blinking:** Too much or too little blinking could be a sign of a deepfake.

A newer category of video deepfakes is based on diffusion models (see [glossary, below](#) 🗨️ #DeepDive-1) – the same AI technology behind many image generators – that can create completely AI-generated video clips based on text prompts. Companies are already testing and releasing commercial versions of [AI video generators](#) 🗨️ /article/2417639-realism-of-

openai-sora-video-generator-raises-security-concerns/ that could make it easy for anyone to do this without needing special technical knowledge. So far, the resulting videos tend to feature distorted faces or bizarre body movements.

“These AI-generated videos are probably easier for people to detect than images, because there is a lot of movement and there is a lot more opportunity for AI-generated artefacts and impossibilities,” says Kamali.

How to spot deepfakes and AI-generated images



How to identify AI bots

Social media accounts controlled by computer bots have become common on many social media and messaging platforms. A growing number of these bots have also been taking advantage of generative AI technologies such as large language models (see glossary, below [🔗](#) #DeepDive-1) since 2022. These make it both easy and cheap to churn out AI-written content through thousands of bots that is grammatically correct and convincingly customised to different situations.

It has become much easier “to customise these large language models for specific audiences with specific messages”, says Paul Brenner [🔗](https://crc.nd.edu/about/people/paul-brenner/) https://crc.nd.edu/about/people/paul-brenner/ at the University of Notre Dame in Indiana.

Brenner and his colleagues have found in their research that volunteers could only distinguish AI-powered bots from humans about 42 per cent of the time [🔗](https://arxiv.org/abs/2402.07940) https://arxiv.org/abs/2402.07940 – despite the participants being told they were potentially interacting with bots. You can test your own bot detection skills [here](https://nd.qualtrics.com/jfe/form/SV_dgy2Ymsq74ZkN0m) [🔗](https://nd.qualtrics.com/jfe/form/SV_dgy2Ymsq74ZkN0m) https://nd.qualtrics.com/jfe/form/SV_dgy2Ymsq74ZkN0m.

Some strategies can help identify less sophisticated AI bots, says Brenner.

5 ways to determine whether a social media account is an AI bot:

- **Emojis and hashtags:** Excessive use of these can be a sign.
- **Uncommon phrasing, word choices or analogies:** Unusual wording could indicate an AI bot.
- **Repetition and structure:** Bots may use repeated wording that follows similar or rigid forms and they may overuse certain slang terms.
- **Ask questions:** These can reveal a bot’s lack of knowledge about a topic – particularly when it comes to local places and situations.
- **Assume the worst:** If a social media account isn’t a personal contact and their identity hasn’t been clearly validated or verified, it could well be an AI bot.

How to detect audio cloning and speech deepfakes

Voice cloning (see glossary, below 🌐 #DeepDive-1) AI tools have made it easy to generate new spoken audio that can mimic practically anyone. This has led to the rise of audio deepfake scams that clone the voices of family members, company executives and political leaders such as US President Joe Biden. These can be much more difficult to identify compared with AI-generated videos or images.

“Voice cloning is particularly challenging to distinguish between real and fake because there aren’t visual components to support our brains in making that decision,” says Rachel Tobac 🌐 <https://www.linkedin.com/in/racheltobac/>, co-founder of SocialProof Security, a white-hat hacking organisation.



How this moment for AI will change society forever (and how it won't)

There is no doubt that the latest advances in artificial intelligence from OpenAI, Google, Baidu and others are more impressive than what came before, but are we in just another bubble of AI hype?

🌐 /article/mg25834352-800-how-this-moment-for-ai-will-change-society-forever-and-how-it-wont/

Detecting such AI audio deepfakes can be especially tricky when they are used in video and phone calls. But there are some common-sense steps you can follow to distinguish authentic humans from AI-generated voices.

4 steps for recognising if audio has been cloned or faked using AI:

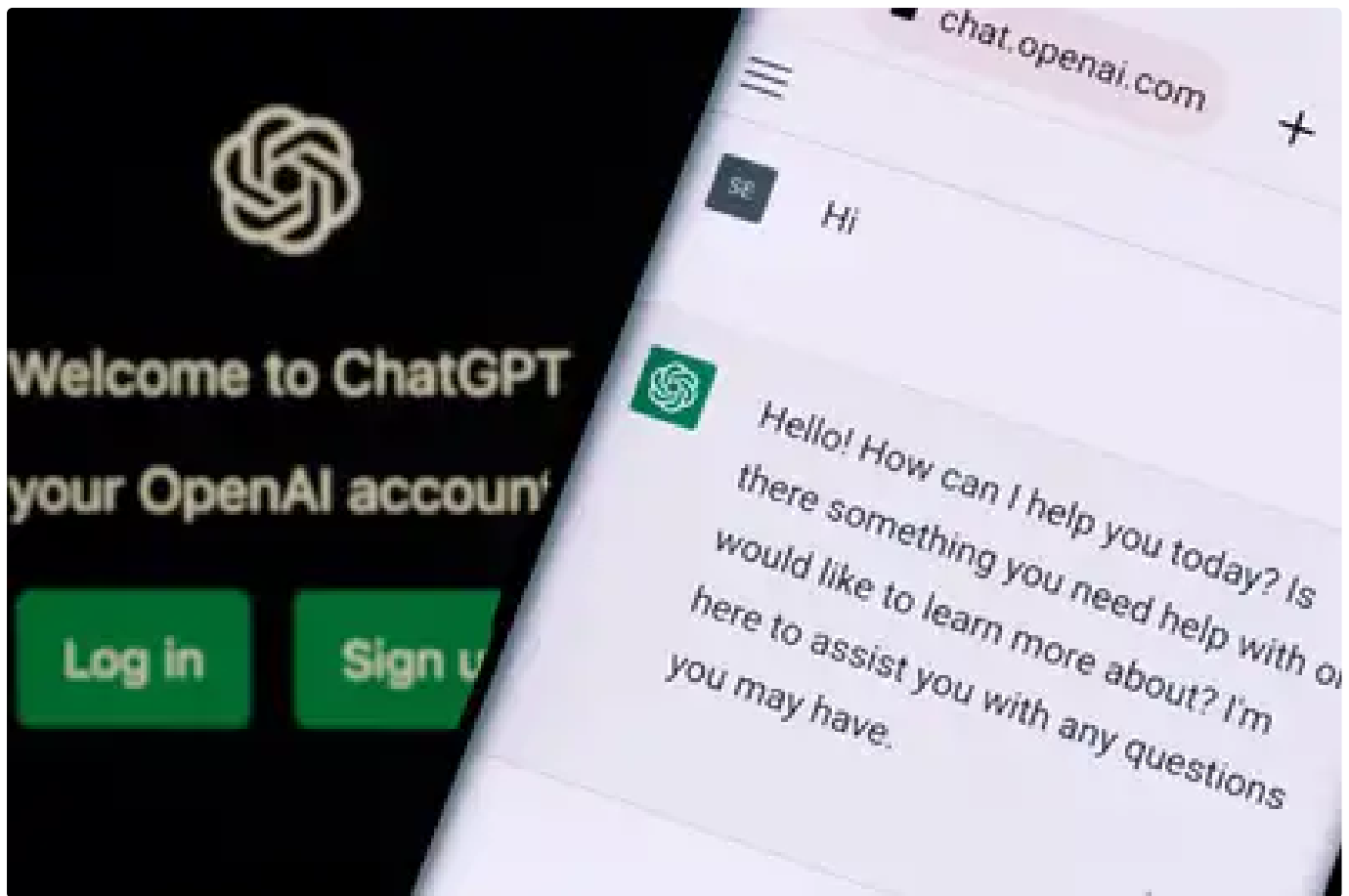
- **Public figures:** If the audio clip is of an elected official or celebrity, check if what they are saying is consistent with what has already been publicly reported or shared about their views and behaviour.
- **Look for inconsistencies:** Compare the audio clip with previously authenticated video or audio clips that feature the same person’s voice. Are there any inconsistencies in the sound of their voice or their speech mannerisms?
- **Awkward silences:** If you are listening to a phone call or voicemail and the speaker is taking unusually long pauses while speaking, they may be using AI-powered voice cloning technology.
- **Weird and wordy:** Any robotic speech patterns or an unusually verbose manner of speaking could indicate that someone is using a combination of voice cloning to mimic a person’s voice and a large language model to generate the exact wording.



⚠ Δημόσια πρόσωπα όπως ο **Narendra Modi** που συμπεριφέρονται εκτός χαρακτήρα μπορεί να είναι ένα δώρο AI
@the_indian_deepfaker

Η τεχνολογία θα γίνει μόνο καλύτερη

Ως έχει, δεν υπάρχουν συνεπείς κανόνες που να μπορούν πάντα να διακρίνουν το περιεχόμενο που δημιουργείται από AI από το αυθεντικό ανθρώπινο περιεχόμενο. Τα μοντέλα τεχνητής νοημοσύνης που μπορούν να δημιουργούν κείμενο, εικόνες, βίντεο και ήχο θα συνεχίσουν σχεδόν σίγουρα να βελτιώνονται και συχνά μπορούν να παράγουν γρήγορα αυθεντικό περιεχόμενο χωρίς εμφανή τεχνουργήματα ή λάθη. «Να είστε ευγενικά παρανοϊκοί και συνειδητοποιήστε ότι η τεχνητή νοημοσύνη χειραγωγεί και κατασκευάζει εικόνες, βίντεο και ήχο γρήγορα – μιλάμε ότι έχει ολοκληρωθεί σε 30 δευτερόλεπτα ή λιγότερο», λέει ο Tobac. «Αυτό διευκολύνει τα κακόβουλα άτομα που προσπαθούν να ξεγελάσουν τους ανθρώπους να περιστρέφουν γρήγορα την παραπληροφόρηση που δημιουργείται από την τεχνητή νοημοσύνη, χτυπώντας τα μέσα κοινωνικής δικτύωσης μέσα σε λίγα λεπτά από τις έκτακτες ειδήσεις».



Η δοκιμή κεφαλαίων γραμμάτων είναι ένας αλάνθαστος τρόπος ταξινόμησης AI από ανθρώπους

Ένα κόλπο για να κάνετε ερωτήσεις χρησιμοποιώντας κεφαλαία γράμματα φαίνεται να μπερδεύει την τεχνητή νοημοσύνη όπως το ChatGPT, ενώ οι άνθρωποι μπορούν εύκολα να δώσουν τη σωστή απάντηση

[/article/2375114-capital-letter-test-is-a-foolproof-way-of-sorting-ais-from-humans/](#)

Αν και είναι σημαντικό να ακονίσετε το βλέμμα σας για ψευδείς πληροφορίες που δημιουργούνται από την τεχνητή νοημοσύνη και να μάθετε να κάνετε περισσότερες ερωτήσεις για όσα διαβάζετε, βλέπετε και ακούτε, τελικά αυτό δεν θα είναι αρκετό για να σταματήσει το κακό και η ευθύνη για τον εντοπισμό απομιμήσεων δεν μπορεί να μειωθεί πλήρως σε άτομα. Ο Farid είναι μεταξύ των ερευνητών που λένε ότι οι κυβερνητικές ρυθμιστικές αρχές πρέπει να λογοδοτήσουν τις μεγαλύτερες εταιρείες τεχνολογίας –μαζί με νεοφυείς επιχειρήσεις που υποστηρίζονται από εξέχοντες επενδυτές της Silicon Valley– που έχουν αναπτύξει πολλά από τα εργαλεία που κατακλύζουν το διαδίκτυο με ψεύτικο περιεχόμενο που δημιουργείται από AI. «Η τεχνολογία δεν είναι ουδέτερη», λέει ο Farid. «Αυτή η γραμμή που μας έχει πουλήσει ο τομέας της τεχνολογίας ότι κατά κάποιο τρόπο δεν χρειάζεται να απορροφούν την ευθύνη από κάθε άλλη βιομηχανία, απλά την απορρίπτω».

Ένα γλωσσάρι AI

Μοντέλα διάχυσης : Μοντέλα τεχνητής νοημοσύνης που μαθαίνουν προσθέτοντας πρώτα τυχαίο θόρυβο στα δεδομένα – όπως το θάμπωμα μιας εικόνας – και μετά αντιστρέφοντας τη διαδικασία για την ανάκτηση των αρχικών δεδομένων.

Generative adversarial networks : Μια μέθοδος μηχανικής μάθησης που βασίζεται σε δύο νευρωνικά δίκτυα που ανταγωνίζονται τροποποιώντας τα αρχικά δεδομένα και στη συνέχεια προσπαθούν να προβλέψουν εάν τα δεδομένα που δημιουργούνται είναι αυθεντικά ή πραγματικά.

Generative AI: Μια ευρεία κατηγορία μοντέλων τεχνητής νοημοσύνης που μπορούν να παράγουν κείμενο, εικόνες, ήχο και βίντεο αφού εκπαιδευτούν σε παρόμοιες μορφές τέτοιου περιεχομένου.

Μεγάλα γλωσσικά μοντέλα : Ένα υποσύνολο μοντέλων τεχνητής νοημοσύνης που δημιουργούνται, που μπορούν να παράγουν διαφορετικές μορφές γραπτού περιεχομένου ως απόκριση σε μηνύματα κειμένου και μερικές φορές να μεταφράζουν μεταξύ διαφόρων γλωσσών.

Κλωνοποίηση φωνής : Η μέθοδος χρήσης μοντέλων τεχνητής νοημοσύνης για τη δημιουργία ψηφιακού αντιγράφου της φωνής ενός ατόμου και, στη συνέχεια, τη δημιουργία νέων δειγμάτων ομιλίας σε αυτήν τη φωνή.

