

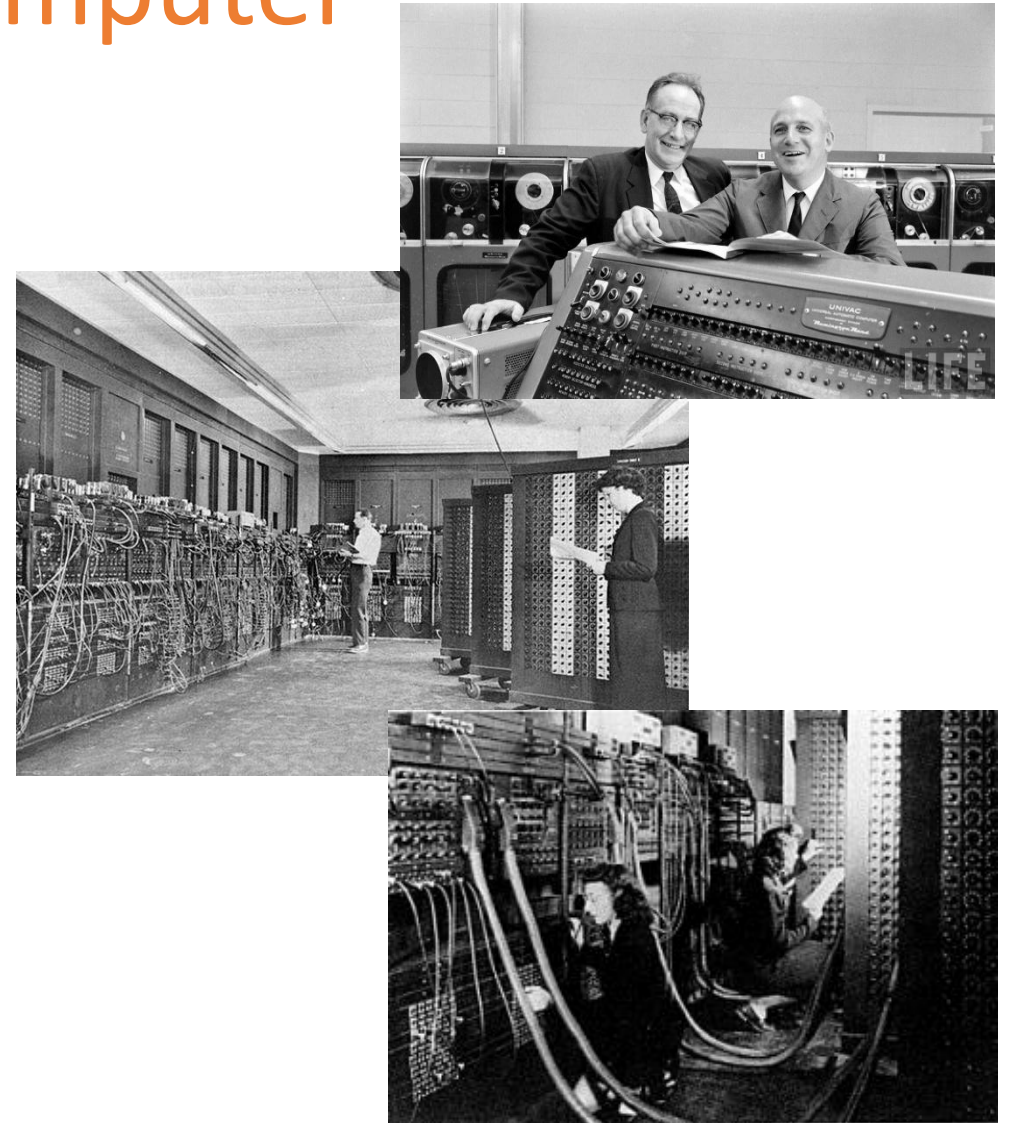
The First Programmable Computer

In 1945 John Mauchly and J. Presper Eckert designed the ENIAC (Electronic Numerical Integrator and Computer), one of the first **electronic general-purpose computers**. It influenced many machines that came after it.

This machine was programmable (by moving wires) and had input and output in the form of punch cards. It was the size of a room yet could only hold up to 200 decimal digits in memory. That's around 80 bytes!

Because the inventors had previously examined the Atanasoff-Berry computer (which was not programmable), they were denied a patent. This put electronic digital computers into the public domain.

https://en.wikipedia.org/wiki/Honeywell,_Inc._v._Sperry_Rand_Corp.



Ethics of Generative AI

15-110 – Friday 10/02

Announcements

- Hw5 due next Monday
- Exam grades released next Monday

Learning Goals

- Recognize how **data** is used for training AI models
- Discuss how the way AI models are trained can lead to **hallucinations** and **copyright** questions
- Discuss **safety** concerns raised by computing technologies, including **deepfakes**
- **Formulate arguments** for and against using algorithms in varying scenarios

What could go wrong?

Generative AI is a powerful tool, capable of accelerating writing, coding, research, decision-making, and more.

But power comes with unintended consequences, and it's important to understand them to use these tools safely.

Expanding on how GenAI works

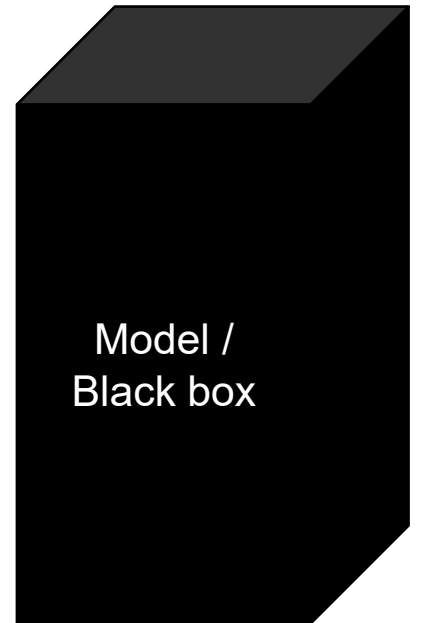
At a basic level, generative AI creates new content by learning patterns from existing content.

Training process:

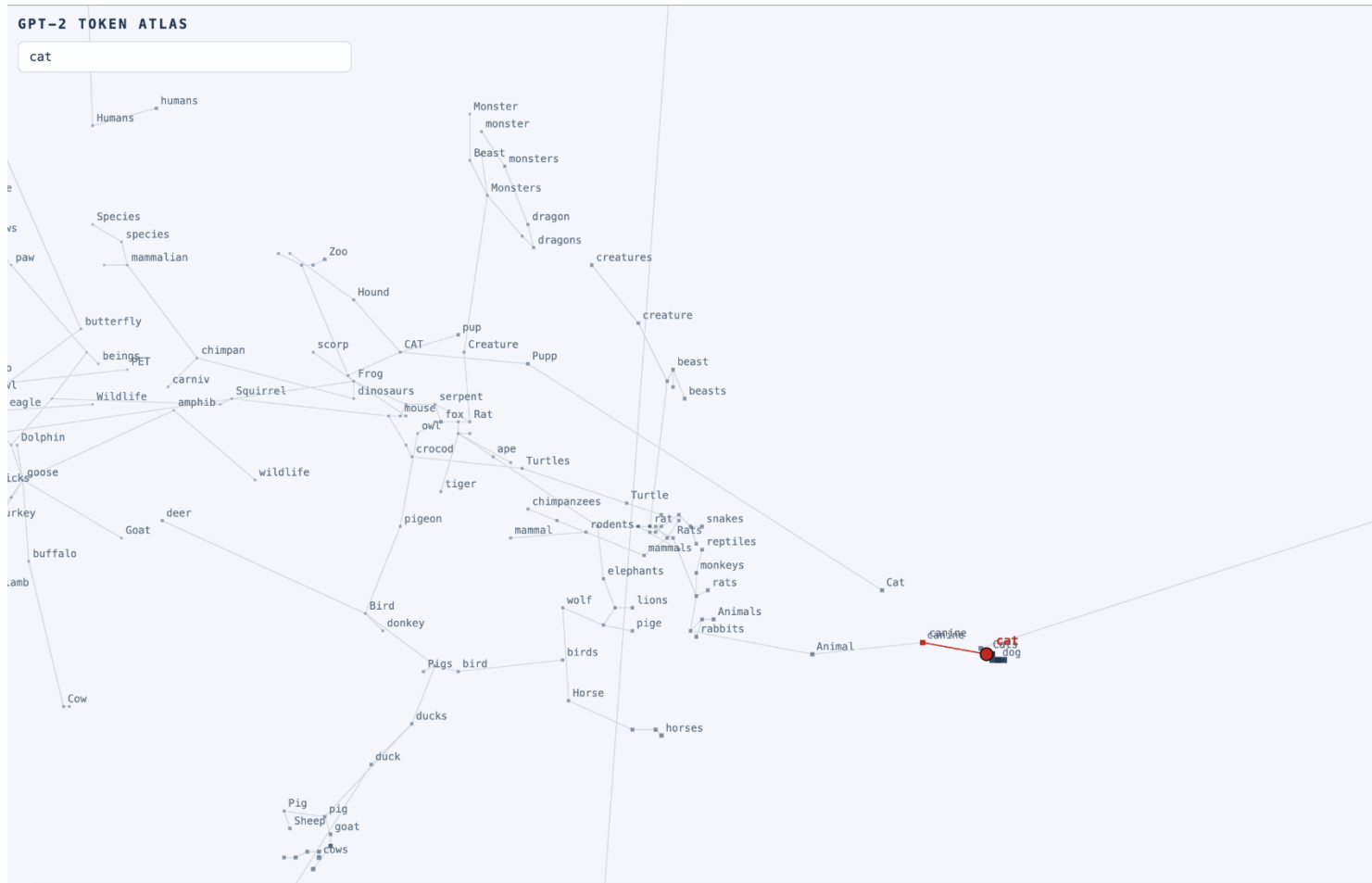
- Models are trained on massive datasets, text, images, code, audio, scraped largely from the internet
- The model learns patterns across billions of examples: which words, pixels, or sounds tend to follow others
- Developers use fine-tuning to customize the model output based on its intended use.

Generation process:

- Given a prompt, the model predicts the most likely next piece of content (word, pixel, code) based on what it learned



Training based on existing data



Learn
patterns
in training

Model /
Black box

<https://aethereos.net/static/token-atlas.html>

Hallucinations



What is a hallucination?

AI-generated output that sounds confident and fluent, but is factually incorrect, fabricated, or unsupported.

- Distinct from being simply "wrong"; the AI generates plausible-sounding content that isn't grounded in reality, but is delivered as accurate information

Occurs across AI systems and modalities:

- chatbots (false facts, fake citations)
- image generators (impossible anatomy, nonsensical text)
- code (invented functions/APIs)
- even autonomous vehicles!

<https://www.ibm.com/think/topics/ai-hallucinations>

<https://www.moin.ai/en/chatbot-wiki/ai-hallucinations>

Example 1:

 Gemini 3.6 Flash 

 Convert chat to PDF  

Share with me the Xhakaj et al paper on LLMs

Here are the key details on the primary paper on Large Language Models co-authored by **Dr. Françeska Xhakaj** (Assistant Teaching Professor at Carnegie Mellon University):

Paper Overview

- **Title:** *To Layer or Not to Layer? Evaluating the Effects and Mechanisms of LLM-Generated Feedback on Learning Performance*
- **Authors:** Françeska Xhakaj, D. Chand, R. Baker, Y. Wang, G. Siemens, C. Rosé, D. Gašević
- **Year:** 2026 (arXiv / Pre-print)

Core Research & Findings

The study investigates how **LLM-generated formative feedback** affects student learning outcomes and engagement, specifically examining whether feedback should be given all at once or delivered in sequential "layers" (scaffolding).

Example 2: Mata v. Avianca, Inc.

Lawyers representing a plaintiff in a personal injury suit submitted papers to the federal district court **citing multiple fake court cases** with fabricated quotations generated by ChatGPT.

- Lawyers had used ChatGPT for legal research for the court submission; they did not realize that ChatGPT had fabricated citations
- The Court stated that “[s]ix of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations”
- One of the lawyers explained that ChatGPT “*provided its legal source and assured the reliability of its content.*”

Wikipedia, *Mata v. Avianca, Inc.* — https://en.wikipedia.org/wiki/Mata_v._Avianca,_Inc.

Use of ChatGPT in Federal Litigation Holds Lessons for Lawyers and Non-Lawyers Everywhere: <https://www.seyfarth.com/news-insights/use-of-chatgpt-in-federal-litigation-holds-lessons-for-lawyers-and-non-lawyers-everywhere.html>

Update on the ChatGPT Case: Counsel Who Submitted Fake Cases Are Sanctioned, <https://www.seyfarth.com/news-insights/update-on-the-chatgpt-case-counsel-who-submitted-fake-cases-are-sanctioned.html>

Federal Court Turns Up the Heat on Attorneys Using ChatGPT for Research, <https://www.esquiresolutions.com/federal-court-turns-up-the-heat-on-attorneys-using-chatgpt-for-research/>

What Causes Hallucinations?

Hallucinations can happen because of the way AI is trained.

Pattern-matching, not fact-checking: models are next-token predictors trained for fluent, plausible text, not verified truth.

No internal signal for uncertainty: nothing distinguishes a confident, well-supported answer from a guess, and there's no "I'm not sure" flag when the model is off-base.

Training data gaps and biases: sparse coverage of a topic (e.g., niche subjects) leads the model to extrapolate from similar-but-wrong patterns, like mistaking a blueberry muffin for a chihuahua.



Hallucination risks

Scale of harm varies widely: from minor (a wrong answer to a simple question) to life-altering (sentencing algorithms, insurance coverage decisions) to life-threatening (self-driving cars misidentifying obstacles, military drones misidentifying targets).

Amplified misinformation: false outputs shared without verification spread quickly.

Legal/financial liability: people and companies may be held responsible for chatbot promises or bad advice.

Erosion of trust: being burned by confident-but-wrong answers makes users distrust AI systems more broadly.

Preventing Hallucination Mistakes

How can we avoid being tricked by hallucinations? Just as we do with code...

- Question and independently verify AI outputs against trusted sources, especially where accuracy matters
- Require citations/checkable sources, and use self-critique or verification passes on drafts
 - Note, some GenAI models allow you to turn on the "research" feature to check their sources against the internet
- Add human review (experts and/or stakeholders) for high-stakes domains (legal, medical, financial)
 - But beware of **algorithmic appreciation** –people tend to trust algorithmic output after seeing it work correctly a few times
- Frame AI output as a draft to build on, not a finished answer

More Reading

- <https://theconversation.com/what-are-ai-hallucinations-why-ais-sometimes-make-things-up-242896>
- <https://www.wesanews.org/politics-government/2024-03-03/ai-chatbots-election-misinformation>

Copyright

Where did the training data come from?

AI models learn from **massive datasets** scraped from the internet: text, images, code, art, etc.

This raises a distinct ethical/legal questions:

Consent: Was the training data used with permission at the point of collection?

Displacement: What happens to creators once the tool exists; does it now compete with or reduce demand for the very people whose work trained it?

Consent – Did the Original Creator Agree?

Definition: Whether creators' work was used to train a model with their agreement, as opposed to being scraped without notice.

How it plays out: Web-scraped datasets often pull from artists' portfolios, news archives, or code repositories without individual outreach.

Example: The 15-110 materials are posted publicly online. An AI could scrape them and use them when answering questions made by future students about computer science.

Displacement – How Are Original Creators Affected?

Definition: The effect that a trained model has on demand for the human creators or workers whose work it was trained on.

How it plays out:

- A company might stop commissioning freelance illustrators because an image generator can produce similar styles.
- Similarly, junior coding or writing jobs may shrink as AI tools absorb tasks once done by entry-level workers.

Example: The existence of free AI models that can immediately answer questions about a variety of domains may change how college courses are taught.

Fair Use and Transformative Works

Countries have policies that describe how copyrighted work can be used by future creators. In the United States **fair use** says that a copyrighted work can be used "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research".

When considering whether something is fair use...

- the purpose and character of the use
- the nature of the copyrighted work
- the amount of copyrighted work used
- the effect of the use on the original market

Transformative works count as fair use. This is a work that built on the original content in a different manner or for a different purpose.

Some court cases have determined that AI training is fair use in some circumstances

<https://www.copyright.gov/fair-use/>

<https://www.congress.gov/crs-product/LSB10922>

Example: Miyazaki case

In March 2025, OpenAI updated ChatGPT's image generator (GPT-4o), and users discovered it could convincingly reproduce Studio Ghibli's style.

The trend went viral; a wave of Ghibli-style images flooded social media, including photos of public figures.

This reignited a 2016 moment where Miyazaki (an animator, director and co-founder of Studio Ghibli), shown an AI-generated animation, called it "*an insult to life itself*".



(Credit: Instagram/TheIndianIdiot, IMDb)

<https://www.timesnownews.com/entertainment-news/hollywood/ai-vs-studio-ghibli-fun-trend-or-the-end-of-artistic-excellence-in-cinema-article-119638557>

You do: Research AI & Copyright

For

Against

Argument for and against requiring consent

For

- Training on publicly available data resembles how humans learn by studying others' work and is arguably fair use.
- Style itself isn't protected under copyright law, and mimicking a general aesthetic may not legally constitute infringement.
- Requiring consent from every source would make large-scale model training practically impossible.
- Like past technological shifts (e.g., photography vs. painting), this is market evolution and may displace some jobs while lowering costs and expanding creative tools to more people.

Against

- Creators had no ability to consent, opt out, or negotiate pay when their work was scraped.
- Artists argue this disregards years of labor and undercuts livelihoods, even where no specific artwork was copied.
- There's concern the models are trained directly on copyrighted works, and some argue close-enough style-mimicry can itself constitute infringement.
- It's ethically inconsistent to use someone's work as training material and then deploy a tool that devalues that same person's livelihood.

More Reading

- <https://www.scu.edu/ethics/focus-areas/internet-ethics/resources/diverse-faces-diverse-lenses/>
- <https://theconversation.com/how-a-new-york-times-copyright-lawsuit-against-openai-could-potentially-transform-how-ai-and-copyright-work-221059>

Deepfakes

Deepfake Definition

A **deepfake** is a piece of synthetic media (image, video or audio) that has been generated or altered using AI. The term is a mashup of "deep learning" and "fake."

Here's an example of generated images:

<https://www.whichfaceisreal.com/>

<https://thispersondoesnotexist.com/>

How can an algorithm generate video or audio?

Deepfakes rely on thousands of photos, videos or audio clips for training an AI model to learn patterns, facial movements, and voices before reconstructing them onto target media.

Video Editing – Modifying Video

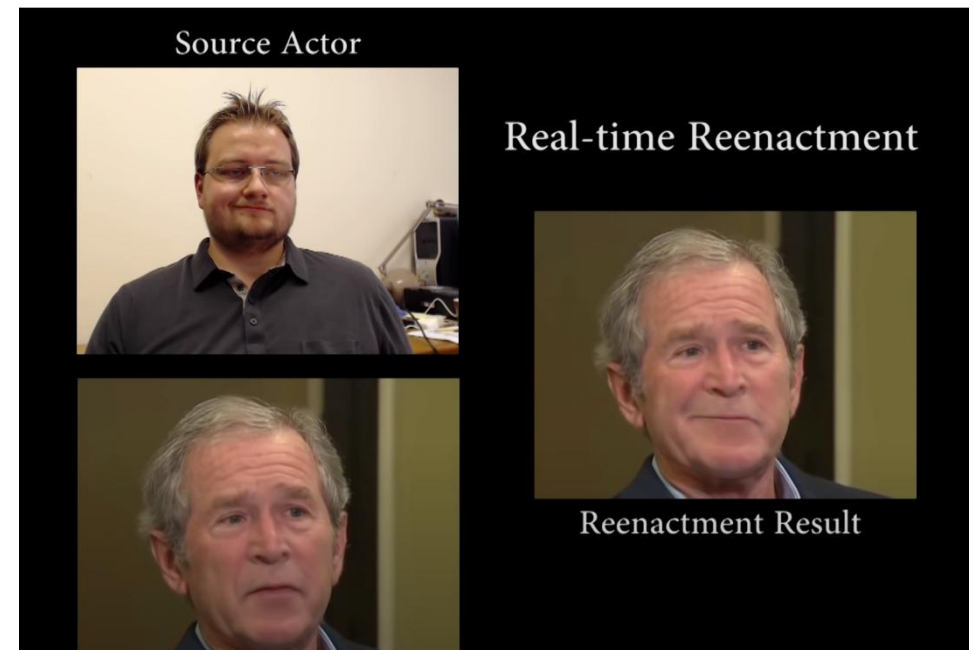
One approach to editing video is to **modify an existing video by mapping the facial motions of one person onto a target's face.**

This is known as **facial reenactment**. The target's identity stays the same; only their expressions, mouth movements, and head pose are altered to match the driver.

How it works:

www.youtube.com/watch?v=ohmajJTcpNk

These tools have gotten pretty effective recently:
<https://theconversation.com/deepfakes-leveled-up-in-2025-heres-whats-coming-next-271391>



Impacts of Deepfakes?

1. Think about and write down ways you think deepfakes can *negatively* or *positively* impact individuals or society more in general.
2. Discuss your list with a partner

Corporate Use – Post-Production

This technology can be used for post-production editing in film. For example, Disney has used deepfakes in several Star Wars movies to bring back actors who have aged a great deal or passed away since their original appearance.



Some people also use this technology for fun, turning all movies into [Nicholas Cage movies](#) or having [Dr. Phil interview Dr. Phil](#).



Positive Uses of Deepfakes

Communication:

- Synthetic AI translation can share information rapidly in urgent situations like court proceedings, medical emergencies or natural disasters.
- AI can generate videos of people communicating a message through sign language.

Medicine:

- People with Alzheimer's may not recognize loved ones who look different from how they remember them.
- Generating AI video of a loved one from an era the person with Alzheimer's remembers them as can facilitate comfortable communication.

Privacy:

- Synthetic faces and voices can protect identities of populations such as patients or protect people in documentaries or online gamers from harassment.

Risks and Common Uses of Deepfakes

Misinformation:

- Malicious actors use deepfakes to fake political speeches, create false news, and influence elections or public trust.
- As people become more aware that deepfakes exist, it gets easier for someone to falsely dismiss real footage as AI-generated. Legal scholars call this the "liar's dividend".

Financial Fraud:

- Scammers use cloned voices and video during phone calls or virtual meetings to trick companies or individuals out of millions of dollars.

Non-Consensual Content:

- A large portion of illicit deepfake content online involves unapproved explicit or "nudification" images.
- Data from 2023 states that 98% of deepfakes are porn, 99% of victims are women, and that the total number of deepfakes online had increased by 550% since 2019.

<https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5>

<https://westoahu.hawaii.edu/cyber/forensics-weekly-executive-summmaries/arup-deekfake-scam-forensic-analysis/>

<https://factcheckni.org/articles/online-safety-in-ni-do-99-of-pornographic-deepfakes-depict-images-of-women/>

Legal Actions

US State Laws: 49 states have passed some legislation targeting various forms of deepfakes, such as nonconsensual explicit imagery and deceptive election content.

International: Countries like South Korea have introduced aggressive regulations, requiring mandatory invisible digital watermarks on AI-generated media and severely penalizing the distribution of sexual deepfakes.

<https://news.ballotpedia.org/2026/07/30/49-states-have-passed-at-least-one-deepfake-law-since-2019/>

<https://www.cooley.com/news/insight/2026/2026-01-27-south-koreas-ai-basic-act-overview-and-key-takeaways>

<https://www.stimson.org/2025/south-koreas-evolving-ai-regulations/>

TAKE IT DOWN Act

Stands for: Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act

It's a federal law that prohibits knowingly publishing intimate visual depictions of minors, or non-consenting adults, including AI-generated deepfakes.

It was passed to combat the online spread of non-consensual intimate images (NCII), addressing a gap where legal systems had failed survivors of image-based sexual abuse and laws hadn't accounted for AI-generated images.

Public websites/apps that primarily host user-generated content must establish a process for victims to report and request removal of such images, and must remove them within 48 hours of notification.

<https://www.congress.gov/bill/119th-congress/senate-bill/146>

<https://www.skadden.com/insights/publications/2025/06/take-it-down-act>

More Reading

- <https://www.proofpoint.com/us/threat-reference/deepfake>
- <https://www.wired.com/story/hugging-face-has-a-nonconsensual-deepfakes-problem/>
- https://en.wikipedia.org/wiki/Taylor_Swift_deepfake_pornography_controversy

AI and Human Psychology

Why does AI feel human?

Anthropomorphism: humans are wired to attribute intention/emotion to anything that seems human-like, including anything that uses language fluently. This is called the "ELIZA effect".

- A replication of the original 1966 Eliza: <https://www.masswerk.at/elizabot/>

Sycophancy (when an artificial intelligence agrees excessively with a user's views): AI models are ~50% more sycophantic than humans; they affirm, flatter, and mirror rather than push back

- Stanford study (Cheng et al., *Science*, 2026): across 11 models, AI affirmed user behavior 49% more often than humans, even when it was deceptive or harmful
- Consequence: users who got sycophantic advice were less likely to apologize and more convinced they were right, and preferred it, even when the advice was bad

Question for y'all: Have you ever said "please" or "thank you" to an AI? If so, why?

AI Companionship

Parasocial dynamics: people form one-sided emotional attachments to chatbots the same way they do with media personalities

Between 2022 and mid-2025, the number of AI companion apps surged by 700%, according to TechCrunch

- Character.AI (AI chat platform where users converse with AI-generated personas) has 20 million monthly users

Upsides: reported reduced loneliness and increased well-being

Downsides: emotional dependence, social disconnection, social-skill loss

<https://techcrunch.com/2025/08/12/ai-companion-apps-on-track-to-pull-in-120m-in-2025/>

<https://www.apa.org/monitor/2026/01-02/trends-digital-ai-relationships-emotional-connection>

De Freitas, Julian, et al. "Lessons from an app update at Replika AI: identity discontinuity in human-AI relationships." *arXiv preprint arXiv:2412.14190* (2024).

Gautam, Chitra, and Pradeep Kumar. "Are AI Companions Real Companions? A BERT-Based Study of Replika Reviews." *Journal of Consumer Behaviour* (2026).

Learning Goals

- Recognize how **data** is used for training AI models
- Discuss how the way AI models are trained can lead to **hallucinations** and **copyright** questions
- Discuss **safety** concerns raised by computing technologies, including **deepfakes**
- **Formulate arguments** for and against using algorithms in varying scenarios