October 14, 2026

Prague

Artificial Insecurity: how AI threatens digital security and what we can do about it

Marcel Kolaja <marcel@accessnow.org>

Policy & Advocacy Director – Europe, Access Now

OpenSSL Conference 2026

accessnow

Access Now defends and extends **the digital rights of people and communities at risk.**

100 staff

7 **global regions**
and officially registered in **4** countries

What is AI going to deliver?

AI generated

Reality check

- Range of more mundane but highly consequential issues at the intersection of AI, privacy, and encryption
- This talk will cover:
  - Risks to privacy from commercial LLM-based tools
  - Overview of “private AI” approaches, including Trusted Execution Environments (TEEs)
  - Privacy and security risks of “content moderation” in LLM-based tools
  - Agentic AI and encrypted messaging tools
- Some of the preliminary findings of a report that my colleague Daniel Leufer has been working on
  - It is not published yet and you’re the first audience to get a sneak peek

New York Times vs OpenAI

- Problem: “**The New York Times is demanding that we turn over 20 million of your private ChatGPT conversations.** They claim they might find examples of you using ChatGPT to try to get around their paywall.”
- Solution: “Our long-term roadmap includes advanced security features designed to keep your data private, including **client-side encryption for your messages with ChatGPT.** We believe these features will help **keep your private conversations private and inaccessible to anyone else, even OpenAI.**”
- New problem: “We will build **fully automated systems to detect safety issues in our products.** Only serious misuse and critical risks—such as threats to someone’s life, plans to harm others, or cybersecurity threats—may ever be **escalated to a small, highly vetted team of human reviewers.**”

OpenAI’s blog post Fighting the New York Times’ invasion of user privacy

Security and privacy defaults of AI tools

- Commercial cloud-based AI tools come with shoddy security and privacy defaults
- *User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies*
  - Paper from 2025
  - Found the following:

Data Privacy Practice	Amazon	Anthropic	Google	Meta	Microsoft	OpenAI
Chat input used for training by default	Not Specified	Yes	Yes	Yes	Yes	Yes
Mechanism to opt out of chat training	Not Specified	Yes	Yes	Not Specified	Yes	Yes
Chat data retained indefinitely	Yes	No	No	Yes	No	Yes
Chatbot personalization features	Not Specified	Not Specified	Yes	Yes	Yes	Yes
Allows accounts for children 13-18	No	No	Yes	Yes	Yes	Yes

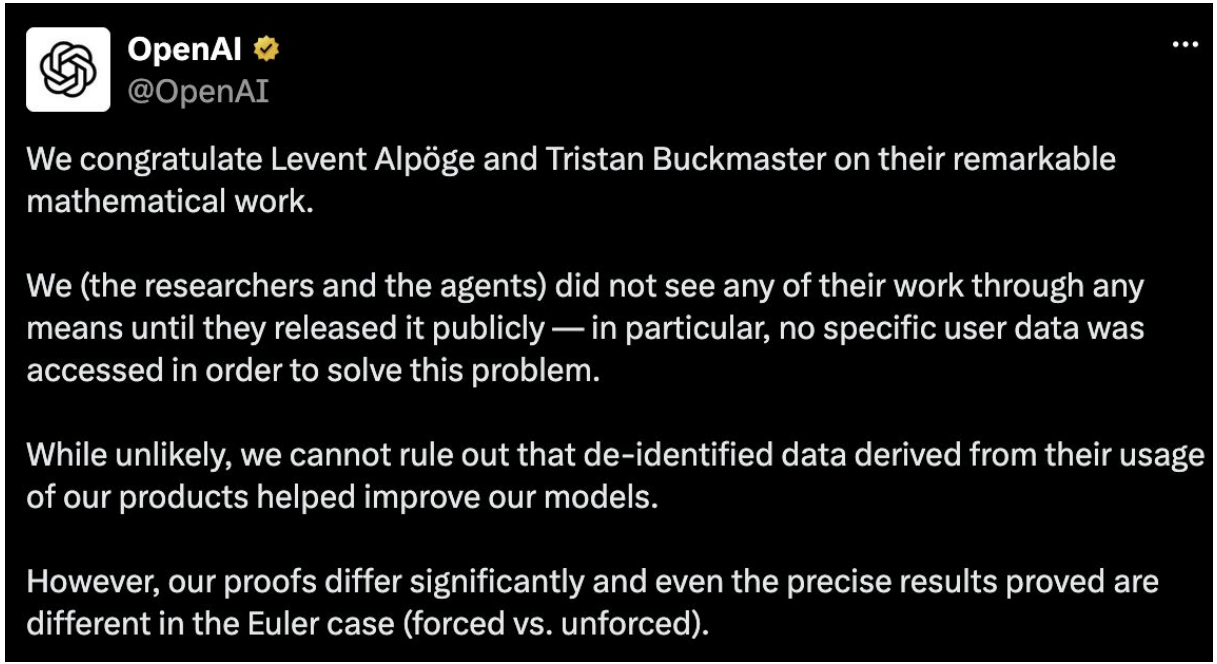
Highly sensitive info shared with LLMs

Key findings

- 66% of AI chatbot users have discussed at least one sensitive topic with a chatbot, including personal finances, sexual health, and relationships.
- 42% of users have little or no trust in AI companies to protect sensitive information shared with chatbots.
- Money, mental health, and career problems are the three most commonly discussed sensitive topics.
- AI users would still rather discuss sensitive topics with friends and family, ahead of AI chatbots and therapists.
- Convenience and freedom from judgment are the most commonly cited reasons for choosing AI over a person.
- 48% of users say they would be more willing to share sensitive information if their conversations were guaranteed to not be used to train AI models.
- Eight in 10 users express interest in an AI chatbot specifically designed for privacy.

Proton's blog post Survey: People don't trust AI, but they tell it everything

Confidentiality of (research) data



OpenAI ✓
@OpenAI

We congratulate Levent Alpöge and Tristan Buckmaster on their remarkable mathematical work.

We (the researchers and the agents) did not see any of their work through any means until they released it publicly — in particular, no specific user data was accessed in order to solve this problem.

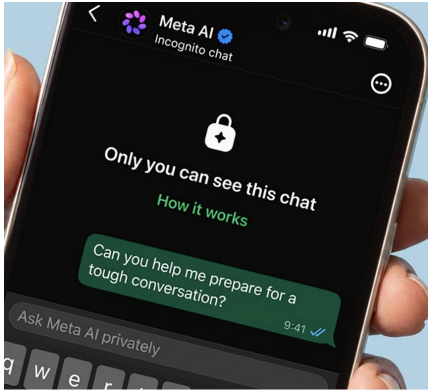
While unlikely, we cannot rule out that de-identified data derived from their usage of our products helped improve our models.

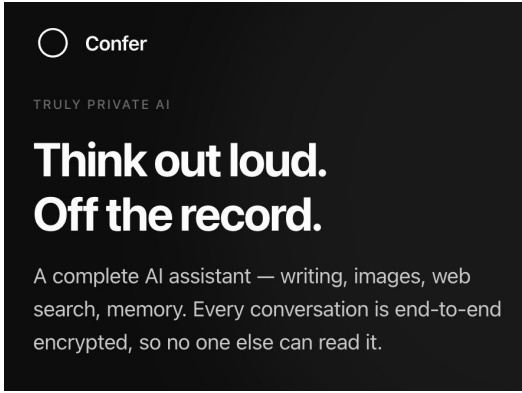
However, our proofs differ significantly and even the precise results proved are different in the Euler case (forced vs. unforced).

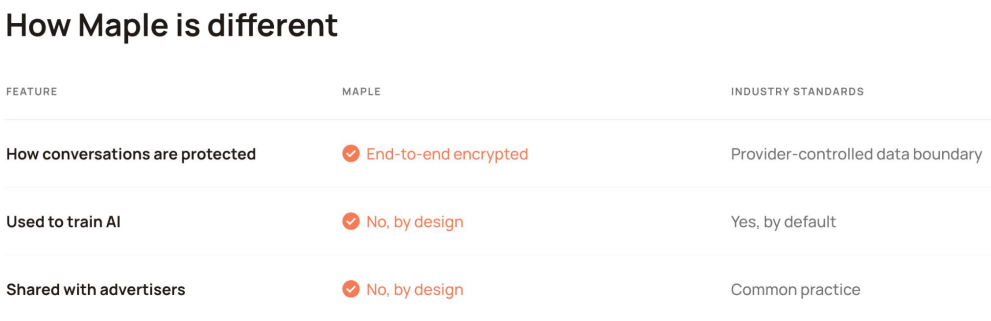
VentureBeat: OpenAI solves longstanding math problem with 10,000-agent swarm – but can't rule out benefitting from a researcher's private Codex data

Positive development

- On a basic level, the situation with security and privacy defaults has, generally, been improving
  - Although there's still lots of room for improvement
- There is a growing ecosystem of "secure" chatbots offering private inference options
  - Allegedly **"end-to-end encrypted"**







From “private AI” to “E2EE”

- There’s a wide range of technical methods being offered under various banners
  - Ranging from **“private AI”** all the way up to tools that claim to offer **end-to-end encryption**
    - Duck.ai by DuckDuckGo
      - Interface that anonymizes metadata and sets improved privacy defaults
    - Lumo AI by Proton
      - Different types of encryption to protect data at different stages of the inference pipeline
  - The most secure are those offering inference with a **Trusted Execution Environment (TEE)**
    - Significant increase of security and privacy protections for both input and output

What is a TEE?

- Hardened, secure, isolated processor area
- Intended to process data that other processes should not have access to
- Authenticity of the code is ensured by attestation
- Probably the TEE that you use most often is your phone holding your biometric unlock data

TEE == E2EE?

- TEE offers an improvement over sending plaintext prompts to the cloud
- Should not be conflated with end-to-end encryption known from interpersonal communication

> That being said, a TEE versus plaintext on a server is the difference between being able to easily read the data and having to do a bunch of specialized work to get at the data. That work often involves accessing the physical machine. This is most relevant for protecting against mass surveillance, and for many people, that might just be enough security.
>
> But that's the core of the problem. “Secure enough for most cases” and “encrypted as in math” are not the same thing, and it’s important not to conflate the two. And services that currently offer “encryption as in math” have a real downgrade in security when they switch to security based on TEEs.

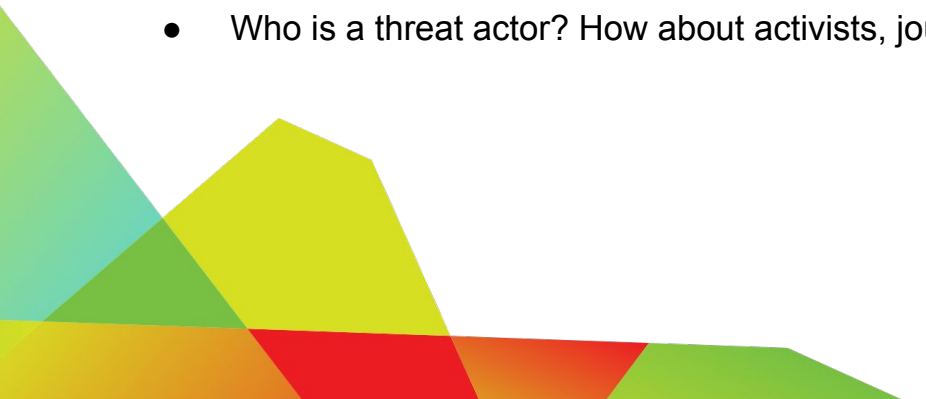
Electronic Frontier Foundation: Secure Messaging and AI Remain In Conflict Despite the Promise of TEEs

Data in transit vs at rest vs in use

- In traditional understanding of end-to-end encryption, the end refers to the user
- TEE is a solution for protecting data in use
- Data must be decrypted for processing
- Math must be complemented with engineering guarantees with a high level of complexity
- Points where data can potentially be compromised + web search, agents & tooling (API calls)

What is AI Safety?

- To come back to the slide from earlier, OpenAI promised to develop client side encryption, but also said:
  - “We will build **fully automated systems to detect safety issues in our products.** Only serious misuse and critical risks—such as threats to someone’s life, plans to harm others, or cybersecurity threats—may ever be **escalated to a small, highly vetted team of human reviewers.**”
- AI Safety is arguably dominated by opinions to focus on biosecurity, cybersecurity, or loss of control
  - Leading to restricted access to advanced AI capabilities and retaining control over how people use AI
  - Combined with LLMs being used by governments (governmental and military contracts)
  - Results in anti Open Source
- Who is a threat actor? How about activists, journalists, protesters, civil disobedience?



From AI Safety to content moderation

- Unsurprisingly, tackling child sexual abuse material (CSAM) on-line as a foot in the door for exceptions
- How DuckDuckGo describes the content moderation performed:

How we moderate uploaded images and files

To detect child sexual abuse material (CSAM), images and files you upload are anonymously checked by a third-party content moderation provider, and in some cases, the model provider. No identifying information, other than what might be contained in the uploaded image or file itself, is shared with the content moderation provider (for example, your IP address is not shared). Only uploaded images and files are processed in this way, not text prompts or model responses. Uploaded images and files are not retained unless they are flagged as CSAM, in which case they are retained and reported to the relevant authorities.

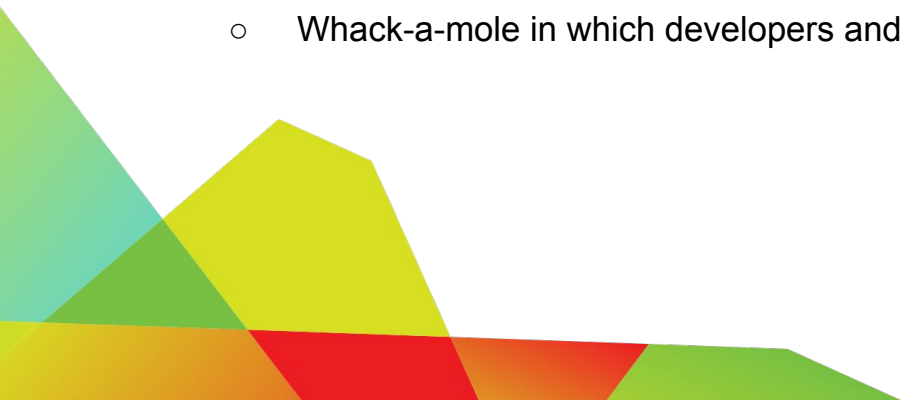
This process applies to all models, including those served through Tinfoil. Tinfoil processes prompts and generates model responses in a Trusted Execution Environment, which prevents them from viewing prompts, responses, or uploaded images and files. To perform content moderation in this architecture, your uploaded images and files are sent to the third-party content moderation provider outside of the encrypted channel.

From moderation to insecurity

- HuggingFace attacked by AI agents
  - Cloud models refused to cooperate on the defense
    - Safety guardrails can pose a confidentiality risk but also an **availability risk**
  - Used locally run GLM-5.2 to respond to the attack
  - https://en.wikipedia.org/wiki/OpenAI_rogue_agent_breach_of_Hugging_Face
- More transparency over what moderation is taking place in so-called private AI tools is urgently needed

When DRM protects encryption /o\

- You are all aware of the risks associated with the use of AI agents
- There are, however, also **risks for those who don’t even use AI agents**
- Microsoft Recall vs end-to-end encrypted chats
  - Agentic AI taking screenshots to store them, so that you can have a “photographic memory”
  - Signal resorted to using DRM to prevent Recall from taking screenshots of your E2EE messages
  - There’s always a risk that the communicating counterpart gets hacked or leaks information
  - However, chatting with someone whose OS is systematically taking screenshots is a new level
- To generalize, the risk lies in introducing agentic AI features at the OS level
  - Whack-a-mole in which developers and users try to disable dodgy AI features



More AI agents processing my data…

- ChatGPT plugin for Apple's iMessage

This means your Apple Messages can:

Train OpenAI's models, depending on your plan and settings. Training is on by default for Free, Go, Plus, and Pro accounts, and off by default for Business, Enterprise, Edu, and API accounts. Once content has been used for training, it cannot be undone.

Remain stored on OpenAI's servers indefinitely if you decide to let ChatGPT store message content on its servers, unless you choose to delete them. It takes up to 30 days to fully remove deleted chats. If the company decides it needs to retain your conversation history for legal, security, accounting, or other reasons, it can hold on to that data indefinitely.

Land in the hands of US government or intelligence agencies without a warrant or notice to you, since OpenAI is a US company subject to US legal processes such as FISA Section 702 (which can sweep in a US person's messages without an individual warrant) and National Security Letters (which carry gag orders barring OpenAI from telling you it happened).

Proton's blog post: ChatGPT's iMessage plugin opens a backdoor in Apple Messages

Even more AI agents?

Each hour, Muse updates its dossiers on you and the people you've mentioned in chats, messages, and emails that Muse has read. These pages amount to a map of each user's social relationships. They record details of how you and your contacts met, your shared interests, your disputes, and "tensions and alliances" within your social group.

Even people who don't use Muse are subject to this mapping process by other people's Muse agents. Muse is told to "strengthen your relationship" with users by adapting how it talks, noticing "inside jokes, memorable phrasing, and shared context that makes the relationship feel continuous," according to the internal instructions TIME reviewed.

The level of insight in the resulting dossiers makes social-media algorithms that study your likes and clicks look quaint. Muse is designed to infer users' goals, including those you "have not said out loud." It identifies which kinds of prompts from the personal assistant work best: "This user responds better to short nudges after 10 PM," reads an internal example found in one of Muse's prompts. Muse agents perform a nightly analysis of the day's conversations with the user.

Time: Meta's Muse AI Agent Is Building a Dossier On You

(and people you interact with)

Conclusion

- All of this is giving momentum to Open Source (or at least open-weight), locally run alternatives
  - Use of the funding under the tech sovereignty initiative
- Nevertheless, shadow AI will always be an issue: confidential data leaks (business, legislative, sensitive)
  - Privacy defaults are key
- Deregulation (AI & Data Omnibuses) is the complete opposite of what is needed
  - Data protection and safe AI tools' defaults
- Community alignment needed on what private AI means to avoid privacy-washing

Questions?

accessnow