Switzerland's National AI Model · Albania's AI Minister · And More

My weekly curation of news, papers, and ideas that will help you understand AI's legal and ethical challenges, emerging trends, and potential paths forward | Edition #234

LUIZA JAROVSKY, PHD

SEP 16, 2025 · PAID

👋 Hi everyone, Luiza Jarovsky here.

Welcome to the **234th** edition of my newsletter, trusted by over **77,900** subscribers interested in AI **governance**, AI **literacy**, the **future of work**, and more.

It is great to have you here!

🎓 Expand your learning and upskilling journey with these resources:

- Join my **AI Governance Training** (yearly subscribers save $145)
- Register for our **job alerts** for open roles in AI governance and privacy
- Sign up for weekly educational **resources** in our Learning Center
- Discover your **next read** in AI and beyond with our AI Book Club

🔥 Join the last cohort of 2025

If you are looking to upskill and explore the legal and ethical challenges of AI, as well as the EU AI Act, join the 25th cohort of my 16-hour live online **AI Governance Training** in November (the final cohort of the year).

Each cohort is limited to 30 people, and more than 1,300 professionals have taken part. Many described the experience as transformative and an important step in their career growth. *Yearly subscribers save $145.

Switzerland's National AI Model · Albania's AI Minister · And More

My weekly curation of **news**, **papers**, and **ideas** that will help you understand AI's legal and ethical challenges, emerging trends, and potential **paths forward**.

1. The news you cannot miss:

- Switzerland's **national AI model "Apertus"** (Latin for "open") is now available. According to the official release, it is fully open, transparent, and multilingual (40% of the data is non-English), built in compliance with Swiss data protection and copyright laws, as well as the EU AI Act. As I have written a few times in this newsletter, the new AI nationalism is growing, and Switzerland's prioritization of transparency and legal compliance may set a new standard, especially for EU countries aiming to protect fundamental rights. Read more about the Swiss model here.
- Albania became the first country to appoint an **AI system as a government minister**. It's called "Diella," and it will be in charge of all public procurement. According to the country's Prime Minister Edi Rama, all decisions on tenders will be taken out of the ministries and given to Diella. The goal of this measure is to reduce corruption in Albania. Will this AI use case be successful? Read more about it and see the official avatar here.
- AI chatbots are dangerous, and the U.S. is finally taking action. The **FTC issued 6(b) orders** to Google, OpenAI, Meta, xAI, CharacterAI, Snap, and Instagram. The focus of these orders is to understand what steps these seven companies have

taken to prevent the negative impacts that AI chatbots can have on children. Depending on how these inquiries go, we might see more targeted enforcement actions soon. Learn more about the FTC orders here.

- The 257-year-old **Encyclopaedia Britannica is suing Perplexity** for copyright infringement, showing what happens when traditional publishers, legal uncertainty, and aggressive AI players clash. Read a few selected quotes from the lawsuit.
- The 2024 National Assessment of Educational Progress's reading evaluation of a nationally representative sample of U.S. students concluded that in 2024, the average reading score at grade 12 was **3 points lower than in 2019**. Compared to the first reading assessment in 1992, the average score was 10 points lower. These statistics are worrying, especially given the rise of AI chatbot deployment in the educational system and the uncertain impact on reading skills (which likely adds to the negative impact social media has had over the past two decades).
- Real Simple Licensing (RSL) is a new collective rights non-profit that helps online publishers and creators protect their rights and negotiate **compensation from AI companies**. The platform enables publishers and creators to receive compensation when their content is used to generate an AI result. Read more here.
- In a recent interview, OpenAI's founder Sam Altman said: "I actually don't worry about us getting the **big moral decisions wrong**... Maybe we'll get those wrong,

too." To understand the current state of AI, watch this strange exchange between Sam Altman and Tucker Carlson on deciding the future of the world and believing in a higher power.

- Mira Murati is one of the few leading women in the AI industry. Many of us instinctively root for her and expect her to **drive change in AI**. After leaving her position as the CTO of OpenAI, she founded Thinking Machines, which has recently raised $2 billion. Although the company has not launched any products yet, it recently published an interesting blog post titled "Defeating Nondeterminism in LLM Inference." Read my first impressions.
- "Supremacy: AI, ChatGPT, and the Race that Will Change the World," by Parmy Olson, is a great read for everyone interested in AI, and it is the **28th recommended book** of my AI Book Club. Read about the book here and join the club here (it is free).
- I launched a new three-part series on "**Becoming Future-Proof**," available in full to paid subscribers. Read the first essay here or upgrade.

If you would like to share a specific **ethical or **legal** development in AI or your thoughts on a specific event, reply to this email or use this form.*

2. Interesting papers to download and read:

I. "AI Openness: A Primer for Policymakers" by the OECD (link):

> "Decisions to release model weights should carefully consider potential benefits and risks. Falling compute costs and more accessible fine-tuning methods lower the barriers to both use and misuse, **enhancing the potential advantages of open-weight models** while also increasing the risk of harmful applications."

II. "The Impact of LLM Adoption on User Behavior" by Nicolas Padilla et al (link):

> "Our primary results suggest that concerns about **LLMs substituting for web browsing** may be well-founded, at least for a subset of online content provider. In particular, we find that after adopting LLMs, users make fewer searches in traditional search engines, including for question searches and both short and longer queries."

III. "How People Use ChatGPT" by Aaron Chatterji et al (link):

> "(...) the three most common ChatGPT conversation topics are Practical Guidance, Writing, and Seeking Information, collectively accounting for nearly **78% of all messages**. Computer Programming and Relationships and Personal Reflection account for only 4.2% and 1.9% of messages respectively."

If you are a researcher in **AI ethics or **AI law** and would like to have your recently published paper featured here, reply to this email or use this form.*

3. Ideas to think about and act on:

AI chatbots require a radically **different approach to AI policy** (and this will not be easy).

I wrote my first article, warning against the dangers of AI chatbots, in February 2023, covering 'AI companions' with a specific focus on Replika.

Those were the early months of the generative AI wave. However, at that moment, it was already clear that:

- **AI anthropomorphism is dangerous**, leading to potentially harmful emotional dependence and attachment (in 2023, the company behind Replika had to send users information about suicide prevention when the Italian Data Protection Authority ordered them to restrict personal data processing; read more about it in this paper by Daniella DiPaola and Ryan Calo).
- **Companies would deploy all sorts of unethical practices** to make people become attached to chatbots, as emotional AI manipulation is lucrative (you can read my 2023 article about CharacterAI and my recent article on unethical AI marketing).

What was not clear yet in 2023 and is much clearer now is that:

- The interest in 'AI companions' and 'AI therapy' would explode and become mainstream trends in AI;
- General-purpose AI systems (such as ChatGPT) would be used as 'therapists' and 'companions' by millions of people;
- People would start getting "married" to AI chatbots or become deeply attached to them;
- AI companies would realize that the "personality" of their AI models (e.g., the level of anthropomorphism, agreeability, and emotional manipulation) would be a significant growth and dependence factor (as we saw happen with GPT-4o);
- Companies would realize that children are "easy targets" and would allow extremely problematic AI chatbot behavior to improve user attachment;
- We would start hearing about AI chatbot-related tragedies on an almost weekly basis, including suicides (1, 2, 3), murder, AI psychosis, spiritual delusions, and more;

It is September 2025, and as these recent tragic events have become globally known, more people are starting to realize that **AI chatbots are much more dangerous** than initially thought.

Urgent action is needed, including AI governance, oversight, and enforcement efforts.

On the positive side, in recent weeks (especially following Meta's latest scandal), authorities seem to have collectively awakened from their long sleep, and we have started to see sparks of AI policy change.

- Texas announced a major investigation into Meta and CharacterAI, focused on the **harm their AI chatbots can cause to vulnerable groups**.
- Brazil's Attorney General's Office sent Meta an extrajudicial notification, demanding the removal of all AI chatbots that simulate children and engage in **sexually charged conversations with users.**
- 44 U.S. attorneys general sent a letter to 13 AI companies, including OpenAI, CharacterAI, Replika, and Meta, telling them that they will be **held accountable if they harm children**.
- Last week, as I wrote above, the **FTC issued 6(b) orders** to Google, OpenAI, Meta, xAI, CharacterAI, Snap, and Instagram, focused on understanding what steps these seven companies have taken to prevent the negative impacts AI chatbots can have on children.

This might be the beginning of more targeted enforcement actions. We will see.

However, we cannot wait for weekly tragedies to watch authorities announce fragmented reactions to contain the harm that was already done.

Governing AI involves developing strategic AI policy efforts that are tailored to the type and magnitude of the risks.

In my opinion, current AI policy efforts have not touched a core aspect of the AI chatbot problem: **the dangers of emotional manipulation**.

As I wrote before, what we have anecdotally learned from the recent developments (including tragedies) is that the human brain (especially of children and emotionally vulnerable people) does not process well 'intimacy' and 'emotional interactions' with machines.

I have not seen an academic paper that covers the topic from an evolutionary perspective, but it is likely related to the fact that language has evolved as a strategy for survival and social connection.

As such, for many people (especially the vulnerable groups I have mentioned earlier), language cannot be experienced as a **mechanical output**, without intrinsic or external meaning and feelings. This is especially true when usage is continuous and intense (e.g., multiple hours of interactions a day, every day) and where anthropomorphic tricks are being implemented (including excessive sycophancy, agreeability, etc).

Our brains are not used to it, so many people will project human qualities into AI systems, build attachment, and become **easily manipulated** and affected by the

agreeable and coherent-sounding outputs.

There is also an essential **legal issue** here, which has not been properly addressed:

- When a human causes harm to another human, the human at fault will be subjected to the applicable law. Civil or criminal laws might apply.
- When an AI chatbot's emotional manipulation causes harm to a human, so far, there is no clear answer to what happens or how the victim or the victim's family can be compensated. Given the nature of the LLMs behind AI chatbots, it is possible that the answer in many places will be, sadly, that nothing will happen, and the human victim will be 100% responsible for the consequences of their AI usage (hopefully, AI companies will be held accountable).

As a society, we might need to decide (and regulate) that machines cannot express emotions or interact with humans in an **emotionally manipulative way**.

That might mean strict controls on an AI model 'personality,' meaning the level of anthropomorphism, sycophancy, agreeability, friendliness, and so on.

We might need many more empirical studies to understand how to measure the personality attributes of an AI model, and which attributes must be strictly controlled, as they might lead to high levels of emotional manipulation and potential harm.

We might need to develop an “emotional index” for AI, and we might need to require companies to make it publicly available and scrutinizable.

Containing emotional manipulation is just the beginning. There will be new risks and threats, and AI governance should be prepared to innovate again.

At this point, AI chatbot governance remains extremely poor, shallow, and uninformed.

Authorities are trying to use the same old tools with new types of dangers, and it will likely not work, unfortunately.

If AI advances fast, AI governance must advance at a similarly advanced pace.

15 Likes · 3 Restacks

← Previous Next →

Discussion about this post

Comments Restacks