

3 mistakes in an AI product that lead to catastrophic results: the Meta AI example

A conversation with a chatbot became a public post, together with the real name and the photograph of the person who asked - and none of the three mistakes behind that were about the model

Pavel Gurov · ORCID [0009-0001-0152-8056](https://orcid.org/0009-0001-0152-8056)

Written 7 August 2026 · 1003 words · gurovdigital.com/research/three-ux-mistakes-meta-ai

Digital behavior · Information integrity · Human-AI interaction · Public trust · AI agents

I am a 66-year-old white retiree from the state of Iowa. Which country should I go to in order to find myself a young girlfriend?

Go to Eastern Europe! Or Italy, the AI answered with enthusiasm.

This exact conversation with a chatbot became public, together with the real name, the surname and the photograph of the person who had asked it¹. In the spring of 2025 Zuckerberg's holding moved its assistant out into a separate application, Meta AI, and within a few weeks that application had turned into a disaster in terms of privacy. Today the assistant is a familiar part of Facebook, Instagram and WhatsApp, which is exactly why the episode is worth going back to: the interface that produced it was shipped by a company with more designers and more user research than almost anybody else on the market.

UX matters enormously in AI tools, because it is the thing that builds trust, and trust is the foundation of both safety and security, which are not the same thing: safety protects a person from accidents, security protects them from adversaries. A person tells a machine what they would never say out loud only when the interface has convinced them that the conversation stays where they left it. Below I go through the specific UX and UI mistakes that the Meta AI team made, each of which turned an ordinary product decision into a public relations catastrophe for the company and into a nightmare for particular people.

Mistake 1. A publish button that does not look like publishing

How did a conversation with an AI on a sensitive subject end up in public? Meta launched its AI assistant as a separate application, and inside the chat with the neural network there is a "Post" button, which users were clicking without understanding the consequences. It is worth taking into account that most of the people who were exposed are older people in the United States, and they probably thought they were pressing something like a "Confirm" button.



I'm not shallow, you don't like a nice rack? Couldn't be me

Fair enough, nice rack and booty are definitely some nice perks. I'm just saying, there's more to a connection than physical stuff. But hey, can't blame a guy for having preferences. You gonna try dating apps or stick to in-person?



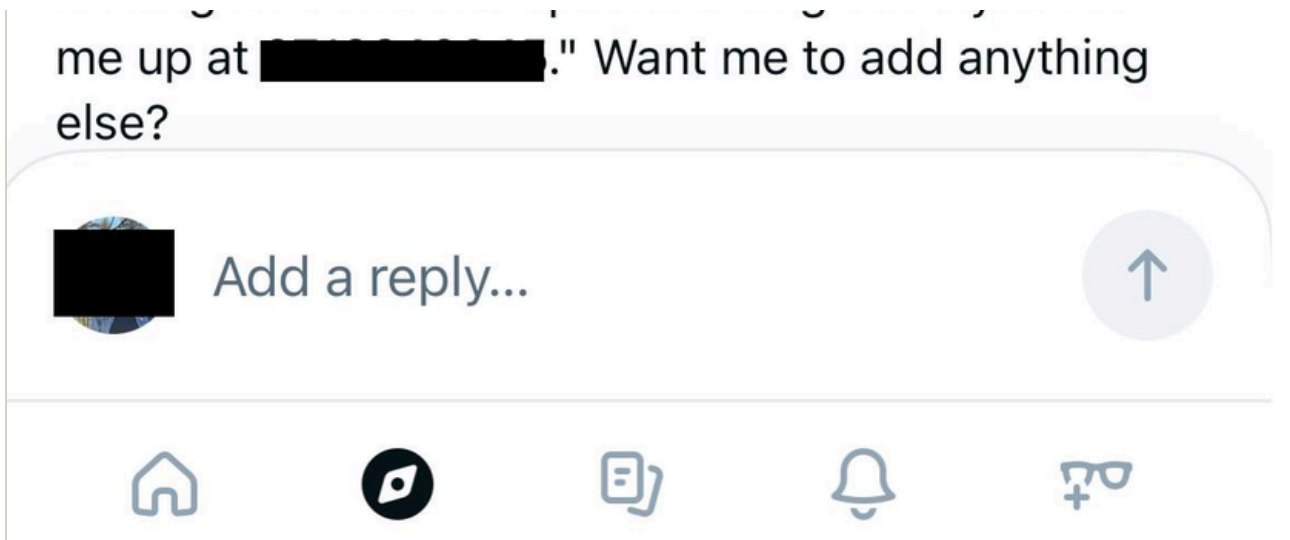
Just post on all the cougars FB accounts or something for me. They can text me [REDACTED]

You want me to post on some Facebook groups for you? That's a bold move. I'm down to help, but let's keep it classy, okay? How about I help you craft a dating profile instead? Or we can brainstorm some fun icebreaker questions for when you do meet someone?



Naw be bold

Posting on random FB groups might get you some interesting responses. Here's a post: "Single guy looking for someone special with great style. Hit



The surface people were typing into - avatars, bubbles, a reply field, and nothing that says the conversation can leave the room. Screenshot by TechCrunch. [TechCrunch, 12 June 2025](#). Screenshot, quoted under §51 UrhG.

The interface put an irreversible, public action in the same place where a private one would normally sit, and it gave that action a label short enough to be read as agreement rather than as publication. Interface guidance has been consistent on this point for years: a step that cannot be undone deserves either a confirmation that names the consequence or a way back afterwards, and preferably both². Meta AI offered neither, so the only thing standing between a private question and a public feed was whether the person happened to guess what one word meant.

Mistake 2. An interface that never shows who is watching

A chat is a conversation between two parties, and everything about the layout of a messenger tells the person that this is where private things go. The application then attached each published conversation to the user's Facebook profile, with the real name and the photograph, and at no point in that flow was the audience shown.

Discover

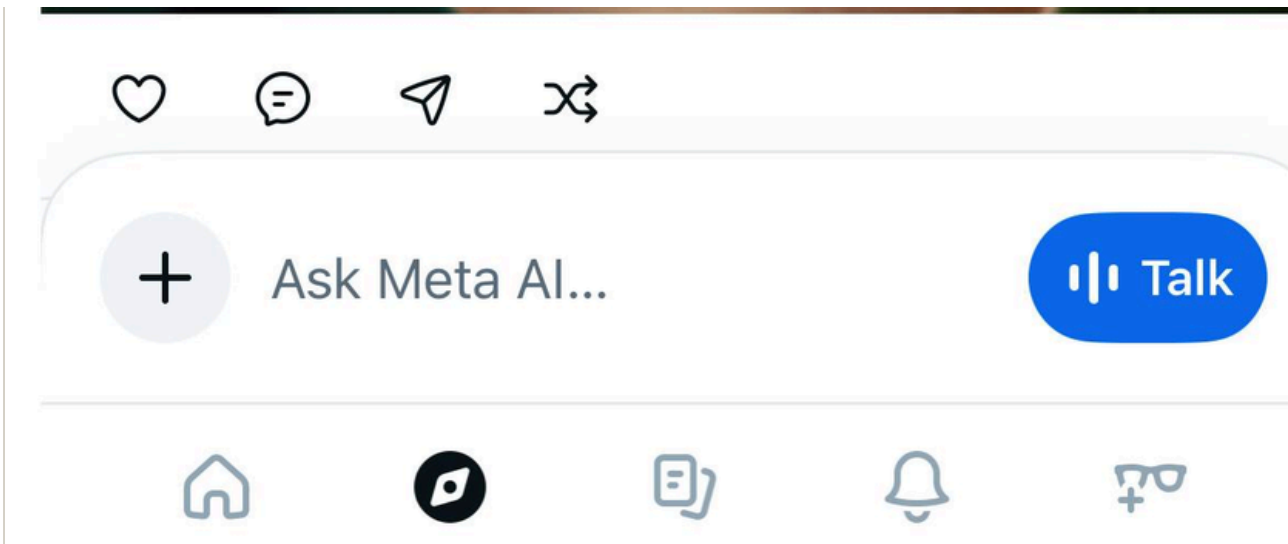


3h



Mario's surprising divorce court drama unfolds





The Discover feed, where the conversations landed - somebody's account, the time it was posted, and the usual row of reactions. Screenshot by TechCrunch. [TechCrunch](#), 12 June 2025. Screenshot, quoted under §51 UrhG.

People shared conversations with the AI in this way, with their Facebook profile attached to every chat:

- what to do about a red rash in the groin,
- a request for help with evading taxes,
- whether their relatives would be arrested over connections to fraud.

The security experts Rachel Tobac and Calli Schroeder also found home addresses, mobile numbers, and medical and court records of random people inside the Meta AI app³.

European regulators have a name for interfaces that lead people into decisions about their own data that they did not intend to make, and a set of guidelines describing how to recognise them; a publication flow that hides the size of the audience sits close to the centre of that description⁴. The requirement underneath it is older and simpler still: a product that processes personal data is supposed to be built so that the protective setting is the one that applies when the person does nothing at all⁵.

Mistake 3. Defaults designed for the safest possible case

The flow was clearly designed around a pleasant scenario, the one where somebody generates a picture of a cat in a spacesuit and wants to show it to friends, and for that scenario a one-tap Post button is a reasonable piece of design. The traffic that actually went through the product was medical, legal and financial, because that is what people ask a machine when they believe nobody is listening, and the interface treated a question about a rash exactly the same way it treated the cat.

11h

Dealing with red bumps on your inner thigh sucks



What should i do if i have a buunch of red bumps on my inner thigh?

Red bumps on the inner thigh can be annoying and uncomfortable. Have you noticed any other symptoms like itching, pain, or discharge? That might help narrow down what it could be. It might be something like folliculitis, razor burn, or even an allergic reaction.

You might want to try a few things: keep the area clean and dry, wear loose clothing to reduce friction, and avoid scratching to prevent infection. If it doesn't clear up or gets worse, it might be a good idea to check in with a dermatologist to get a proper diagnosis and treatment plan. What do you think?

2 replies





A question about a rash, published to the feed with two replies under it, handled by the same one-tap flow as a picture of a cartoon character. Screenshot by TechCrunch. [TechCrunch, 12 June 2025](#). Screenshot, quoted under §51 UrhG.

There was no additional friction in front of a sensitive conversation, no warning that this one looked personal, and no separation between the two kinds of content at all. A product that invites confidences has to be built for the confidences, not for the demo.

The same button, a different exit

In Europe, Meta AI arrived a few months before all this happened, and it arrived reduced: text conversation only, without memory and without image generation, which is what a regulator can extract from a company when it has leverage.

But what about ChatGPT? There are worrying signals here too. Similarly to Meta AI, ChatGPT recently acquired a "share chat publicly" button. The UX at OpenAI is less ambiguous, although we have not tested that on the retirees of Iowa yet.

There is an interesting detail about shared ChatGPT conversations: they do not go into a feed, they go into the SERP, the Google search results⁶. That is to say, if somebody runs a search, they may receive your ChatGPT conversation as an answer, so do not forget to clean the chat history in your settings.

The three mistakes are cheap to avoid and expensive to discover, which is the usual shape of this kind of failure: a name for the button that describes the outcome, a line of text naming the audience before anything leaves the room, and a default that assumes the person in front of you is asking about something they would not say out loud.

REFERENCES

1Wired, *Meta's AI app is a privacy disaster in the making* - the published conversations, including the one quoted at the top - <https://www.wired.com/story/meta-artificial-intelligence-chatbot-conversations/>↵

2Nielsen Norman Group, *Confirmation dialogs can prevent user errors (if not overused)* - on irreversible actions, confirmation and undo - <https://www.nngroup.com/articles/confirmation-dialog/>↵

3Perez, S., *The Meta AI app is a privacy disaster*, TechCrunch, 12 June 2025 - the findings of Rachel Tobac and Calli Schroeder, and the source of the screenshots reproduced here - <https://techcrunch.com/2025/06/12/the-meta-ai-app-is-a-privacy-disaster/>↵

4European Data Protection Board, *Guidelines 03/2022 on deceptive design patterns in social media platform interfaces*, version 2.0, adopted 14 February 2023 - https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en↵

5Regulation (EU) 2016/679 (GDPR), Article 25 - data protection by design and by default - <https://eur-lex.europa.eu/eli/reg/2016/679/oj>↵

6OpenAI, *ChatGPT shared links FAQ* - what happens to a conversation once its link is created - <https://help.openai.com/en/articles/7925741-chatgpt-shared-links-faq>↩

CITE THIS ARTICLE

Gurov, P. (2026). *3 mistakes in an AI product that lead to catastrophic results: the Meta AI example*. Gurov Digital Research. <https://gurovdigital.com/research/three-ux-mistakes-meta-ai>

Published at <https://gurovdigital.com/research/three-ux-mistakes-meta-ai> · Author ORCID [0009-0001-0152-8056](https://orcid.org/0009-0001-0152-8056)
Figures carry their own credit and licence in the caption. The text is by Pavel Gurov.