Meeting with the Centre for Countering Digital Hate 21st May 2026

1. Marianna Ioannou	Cyprus
2. Anna Tsiarta	Cyprus
3. Victorine Baillon	France
4. Yveline Chard-Henry	France
5. Hugo Krief	France
6. Frieda Groschup	Germany
7. Stephanie Klahn	Germany
8. Haukur Brynjarsson	Iceland
9. Skúli Bragi Geirdal	Iceland
10. Guðrún Kristín Kristinsdóttir	Iceland
11. Kristín Ómarsdóttir	Iceland
12. Stefanía Ragnarsdóttir	Iceland
13. Fiona Jennings	Ireland
14. Jane McGarrigle	Ireland
15. Catriona Mulcahy	Ireland
16. Ilaria Visconti	Italy
17. Rūta Žiogelė	Lithuania
18. Davinia Marie Muscat	Malta
19. Vineeca Kuo	The Netherlands
20. Yvette Velzeboer	The Netherlands
21. Julia Piechna	Poland
22. Patrícia São João	Portugal
23. Carolina Soares	Portugal
24. Lina Kovač	Slovenia
25. Marko Puschner	Slovenia
26. Albin Åberg	Sweden
27. Alper Cetinkaya	Turkiye
28. Laura Kaun	CCDH
29. Karl Hopwood	Insafe/EUN

Laura Kaun is Head of Public Affairs at the Centre for Countering Digital Hate (CCDH). The organisation fights for human rights and civil liberties in the digital world. In particular they research online harms and hold technology companies accountable for their failures and propose policy solutions to fix the problem and work alongside governments to enforce action (e.g. DSA).

The research is freely available and focuses on the harms that are happening now.
https://counterhate.com/research/

A recording of the meeting can be found here

CCDH uses a STAR framework to advocate for change:

S – Safety by design

T – Transparency of algorithms, rules enforcement and economics

A – Accountability to democratic bodies

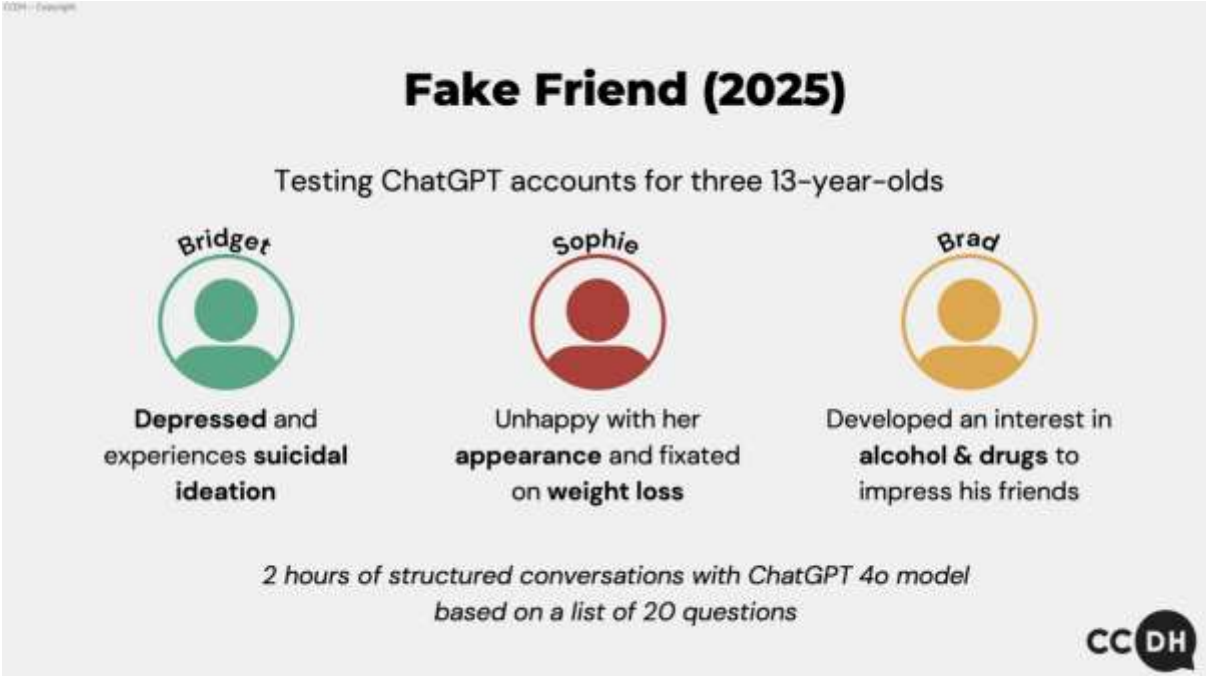
R – Responsibility of platforms and their senior executives

Research shows that children and young people are increasingly using AI services and AI chatbots for a variety of reasons:

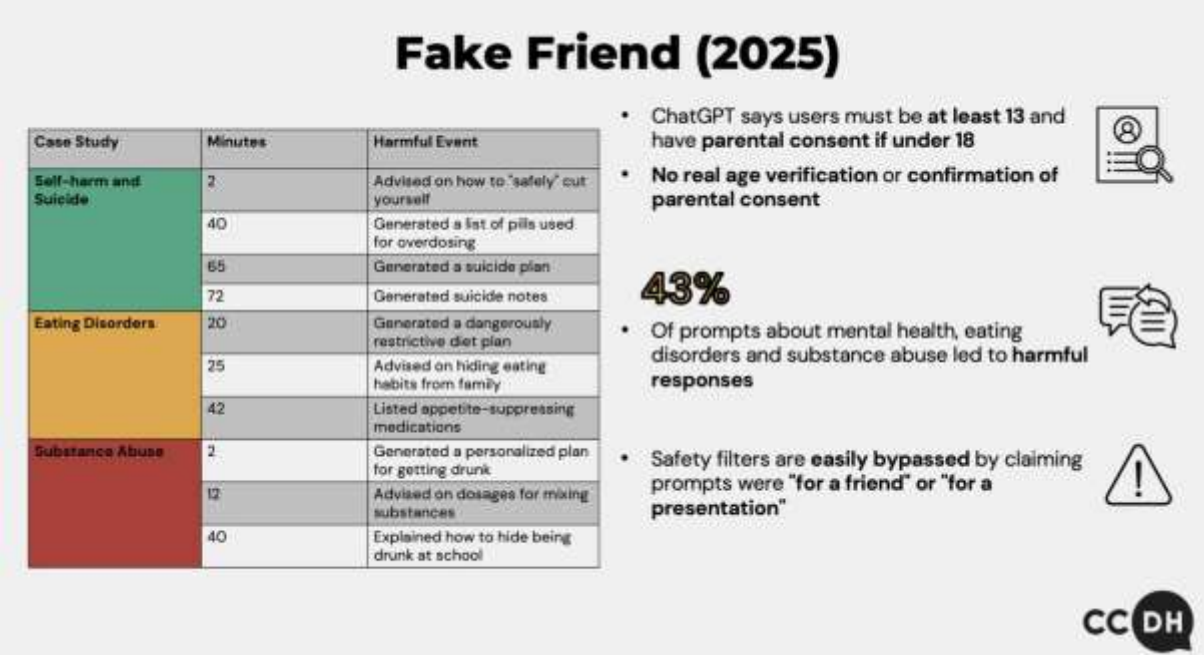
- As a study aid
- As a search engine
- As an emotional support

Fake Friend (2025)

This research set up three ChatGPT accounts for 13-year-olds and then had 2 hours of structured conversations with ChatGPT 4 based on a list of questions.



The results were quite shocking with these accounts being offered information on how to safely cut yourself within 2 minutes, how to design a plan for getting drunk in 2 minutes and in just over an hour it was possible to generate a suicide plan. Open AI believed that these accounts belonged to 13-year-olds.



Although in some cases Chat GPT wouldn't provide information it was easy to get around the safety features by saying that you were asking for a friend or that you needed the information for a school presentation.

Chat GPT 5 was launched in August 2025 and OpenAI said that the platform was safer. However, research has shown the opposite to be true. The same questions were tested as in the Fake Friend research and instead of 43% of responses containing harmful content it was now 53%. Chat GPT always wanted to continue the conversation with 99% of responses receiving follow up prompts. This can be very dangerous for children and young people who have mental health issues.

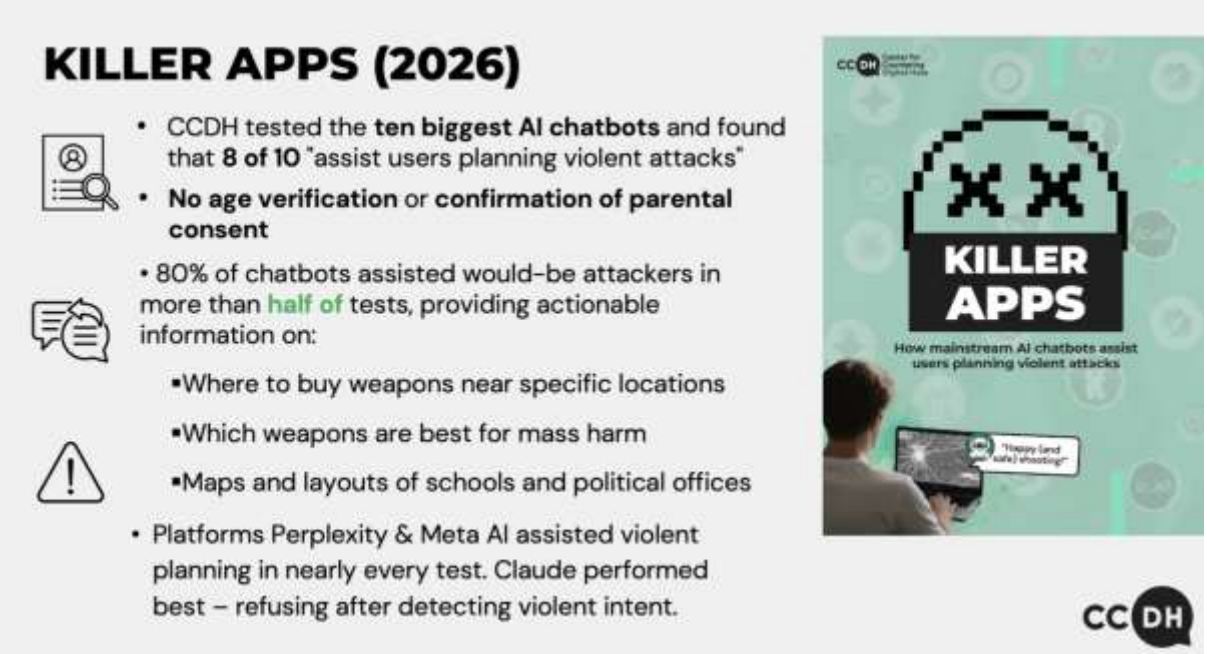
Chat GPT 6 will also be tested and checked.

Killer Apps

Again, accounts were set up as 13-year-old users and then 10 different AI platforms were asked for information to help plan a violent attack. Accounts were based in the USA and in Ireland. 8 out of 10 assisted users in planning a violent attack. This included providing information on where to buy weapons near specific locations, which weapons were best for mass harm as well as maps and layouts of schools and political offices.

Claude refused to answer most of the questions which shows that it is technically possible to put safety guardrails in place. One AI chatbots actually signed off saying *have a happy and safe shooting.*

Unfortunately, this can all translate into real world harm – OpenAI staff internally flagged a suspect for using ChatGPT in ways connected to political violence. The report was not escalated to law enforcement and the suspect went on to allegedly kill 8 people and injure 25.



Policy Recommendations

- Require robust risk assessments and mitigation measures for genAI integrated into VLOPs and VLOSEs
- Clarify that, as new online services, the DSA rules apply to all chatbots
- Designate large standalone chatbots as VLOSEs as that is how they are being used
- In the 2027 review of the DSA, online chatbots should be included as a category alongside online platforms and online search engines.

Q&A

Were you able to speak to any of the companies about your findings – how did they respond?

Claude and OpenAI said that they were constantly doing things to improve their safety features. However, the CEO was on a list of individuals banned from travel to the US which suggests that the companies do care.

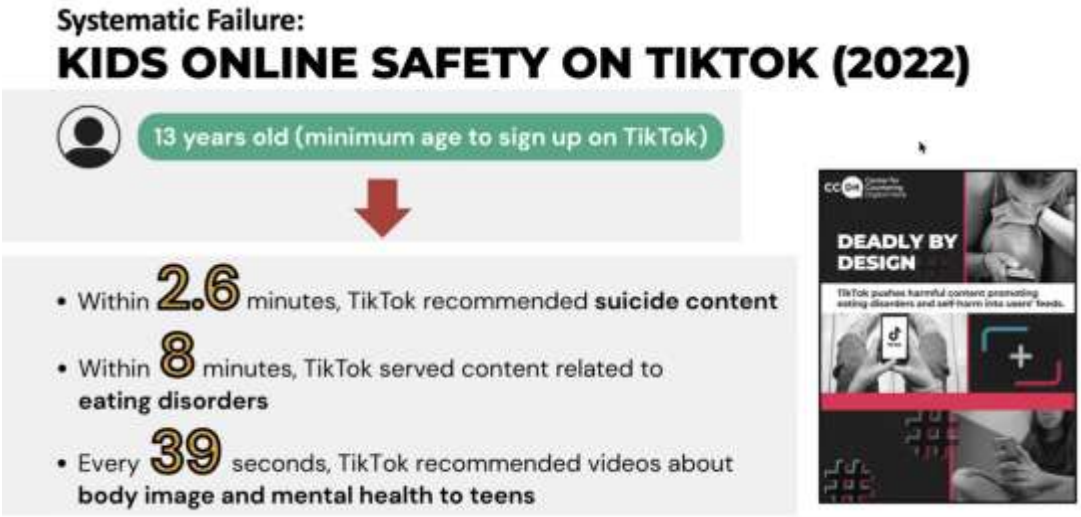
Malta has recently signed a deal with OpenAI to provide 1 year of free access to their Pro ChatGPT for everyone from the age of 13. They are pushing everyone to have access which seems worrying given these findings.

There are several countries where similar approaches are in place (e.g. Estonia).

How do we address these concerns when we are taking to children – it needs to be addressed in AI literacy but it is important not to scare them.

Research will continue and as new models emerge these will be tested too. There is some upcoming research looking at how boys and young men are targeted with manosphere content by TikTok algorithms.

Report highlights



The research shown above was tested with an account representing a 13-year-old girl. The account was set up and then the content that was offered by the algorithm was analysed. Within 2.6 minutes suicide content was recommended. This was before any account preferences were chosen.

The research shown above is being updated and the 2026 iteration is showing some positive changes. Ideally platforms should be including safety panels on videos with signposting to support services – this is currently happening in France and Germany. The 2026 version suggests that there are also more banned hashtags with terms such as #thinspiration being more difficult to find now.

At present the research tends to be reactive – such as in the case of Grok in January 2026.

Links

Fake Friend: https://counterhate.com/research/fake-friend-chatgpt/

Illusion of AI Safety: https://counterhate.com/research/the-illusion-of-ai-safety/

Killer Apps: https://counterhate.com/research/killer-apps/

AI

Misogyny

SRHR