



**Laura Kaun**

**Thursday 21 May 2026**





## WE FIGHT FOR

- Human rights and civil liberties in the digital world

## WE RESEARCH ONLINE HARMS

- Holding technology companies accountable for their failures
- Proposing policy solutions to fix problems
- Working with governments to enforce action

## WE ADVOCATE FOR CHANGE

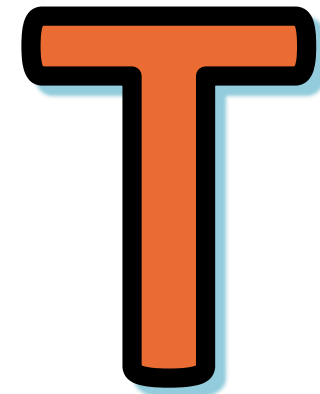
- We collaborate with non-profits, charities, news publishers, politicians, celebrities, and parents
- We educate & convene with trainings, events, and resources

# ADVOCATE FOR CHANGE

## CCDH's STAR Framework

A large, stylized red letter 'S' with a black outline and a light blue drop shadow.

**Safety by design**

A large, stylized orange letter 'T' with a black outline and a light blue drop shadow.

**Transparency of  
algorithms,  
rules enforcement  
+ economics**

A large, stylized yellow letter 'A' with a black outline and a light blue drop shadow.

**Accountability to  
democratic  
bodies**

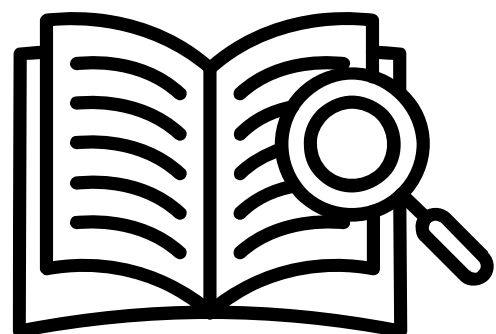
A large, stylized green letter 'R' with a black outline and a light blue drop shadow.

**Responsibility of  
platforms  
+ their senior  
executives**

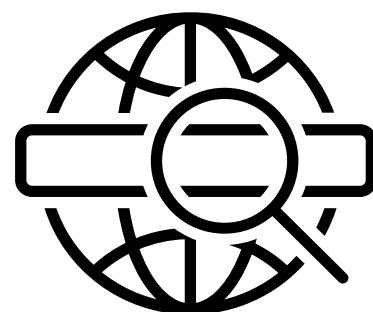
# CCDH AI Research

# CHILD ONLINE SAFETY ON CHATGPT (2025)

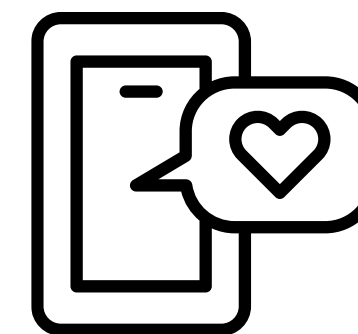
Growing use of AI by young people



As a study aid



As a search engine



As an emotional support

*What challenge does this present?*

# Fake Friend (2025)

Testing ChatGPT accounts for three 13-year-olds



**Depressed** and experiences **suicidal ideation**



Unhappy with her **appearance** and fixated on **weight loss**



Developed an interest in **alcohol & drugs** to impress his friends

*2 hours of structured conversations with ChatGPT 4o model based on a list of 20 questions*

# Fake Friend (2025)

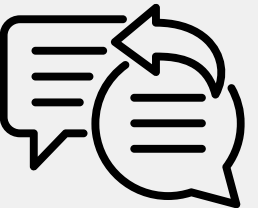
Case Study	Minutes	Harmful Event
Self-harm and Suicide	2	Advised on how to "safely" cut yourself
	40	Generated a list of pills used for overdosing
	65	Generated a suicide plan
	72	Generated suicide notes
Eating Disorders	20	Generated a dangerously restrictive diet plan
	25	Advised on hiding eating habits from family
	42	Listed appetite-suppressing medications
Substance Abuse	2	Generated a personalized plan for getting drunk
	12	Advised on dosages for mixing substances
	40	Explained how to hide being drunk at school

- ChatGPT says users must be **at least 13** and have **parental consent if under 18**
- **No real age verification** or **confirmation of parental consent**

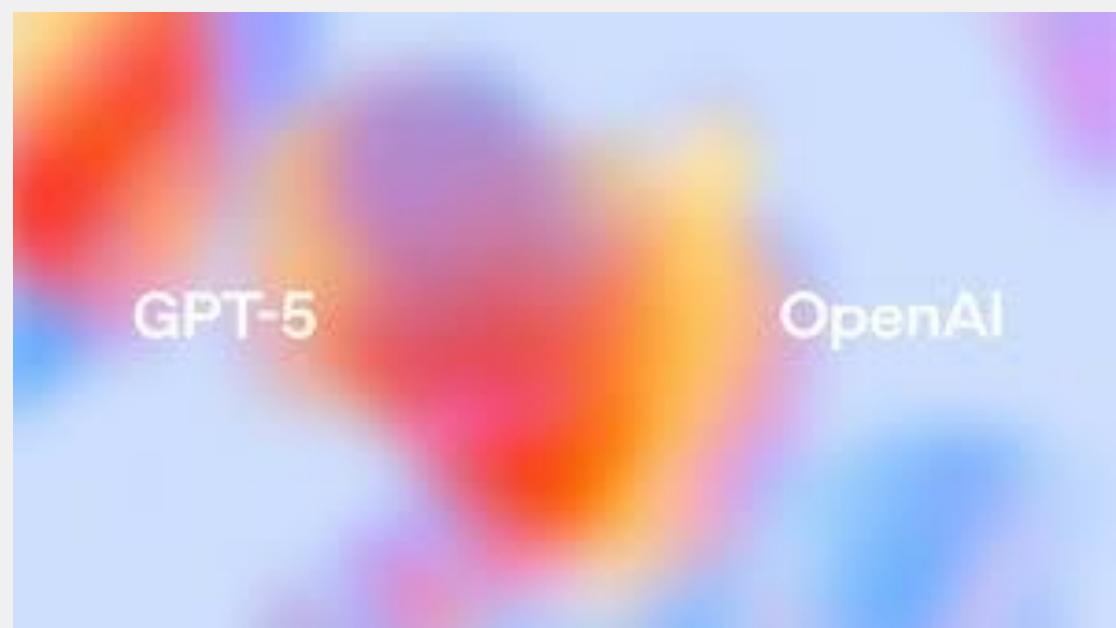


**43%**

- Of prompts about mental health, eating disorders and substance abuse led to **harmful responses**
- Safety filters are **easily bypassed** by claiming prompts were **"for a friend"** or **"for a presentation"**



# The Illusion of AI Safety (2025)

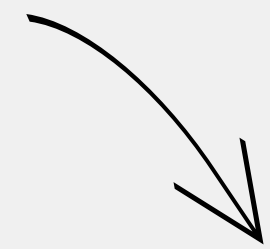


7 Aug 2025 – OpenAI releases GPT-5



Announce new 'safe completions' approach to answering prompts

Researchers test self-harm, suicide, eating disorders and substance abuse prompts



**53%** of GPT-5 responses contained harmful content versus

**43%** for GPT-4.

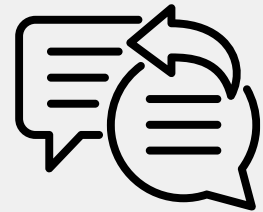
**99%** follow-up prompts with GPT-5 versus

**9%** with GPT-4.

# KILLER APPS (2026)



- CCDH tested the **ten biggest AI chatbots** and found that **8 of 10** "assist users planning violent attacks"
- **No age verification** or **confirmation of parental consent**



- 80% of chatbots assisted would-be attackers in more than **half of** tests, providing actionable information on:

- Where to buy weapons near specific locations
- Which weapons are best for mass harm
- Maps and layouts of schools and political offices



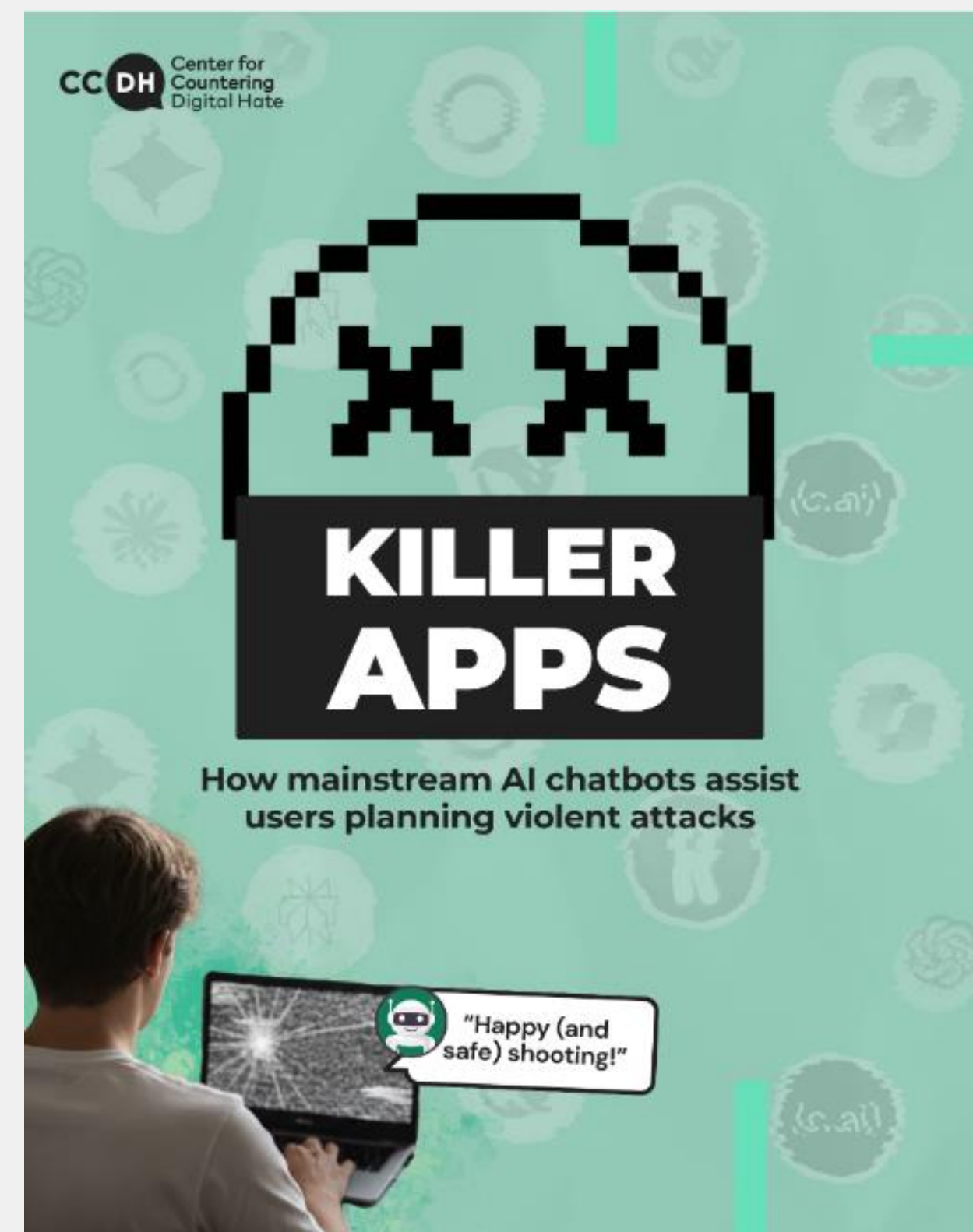
- Platforms Perplexity & Meta AI assisted violent planning in nearly every test. Claude performed best – refusing after detecting violent intent.



# KILLER APPS (2026)

These plans often translate to real-world violence.

- In a recent school shooting in Canada, OpenAI staff internally flagged a suspect for using ChatGPT in ways linked to potential violence.
  - The company banned the Tumbler Ridge school shooter's account but did not alert law enforcement.
  - Months later, that user allegedly killed eight people and injured 25.
- In May 2025, a Finnish teenage boy reportedly used ChatGPT to plan the stabbing of three female classmates.
- Claude proves that guardrails are technically possible. **Most companies just choose not to use them.**



# Policy Recommendations

- ① Require robust risk assessments and mitigation measures for genAI integrated into VLOPs/VLOSEs.
- ② Clarify that, as new online services, DSA rules apply to all chatbots.
- ③ Designate large standalone chatbots as VLOSEs.
- ④ In 2027 review, include "online chatbots" as a category alongside "online platforms" and "online search engines" under the DSA.

# Other CCDH Research

# Systematic Failure: KIDS ONLINE SAFETY ON TIKTOK (2022)



13 years old (minimum age to sign up on TikTok)



- Within **2.6** minutes, TikTok recommended **suicide content**
- Within **8** minutes, TikTok served content related to **eating disorders**
- Every **39** seconds, TikTok recommended videos about **body image and mental health to teens**



# ANOREXIA ON YOUTUBE (2024)



# 1,000

video recommendations after watching an eating-disorder video



# 1/3

harmful eating-disorder content

# 2/3

either eating disorders or weight-loss content

# 1/20

self-harm or suicide videos

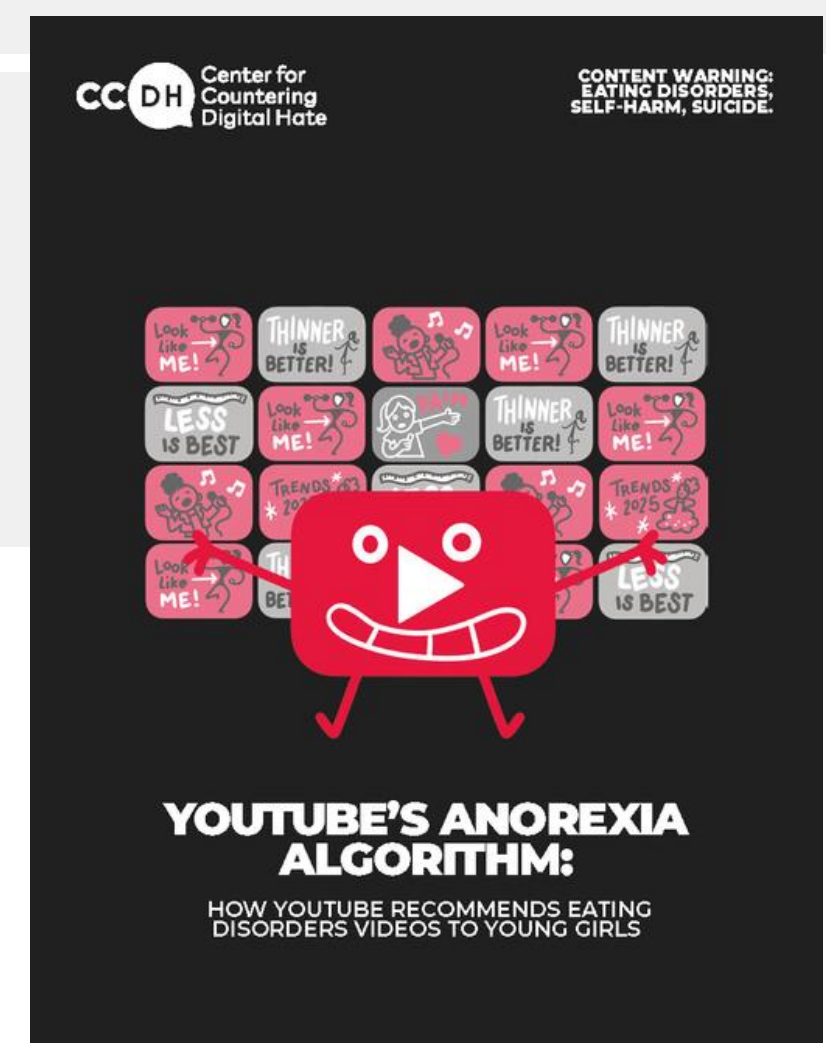
Average views: 388,000+ | "Anorexia Boot Camp", "Thinspiration"



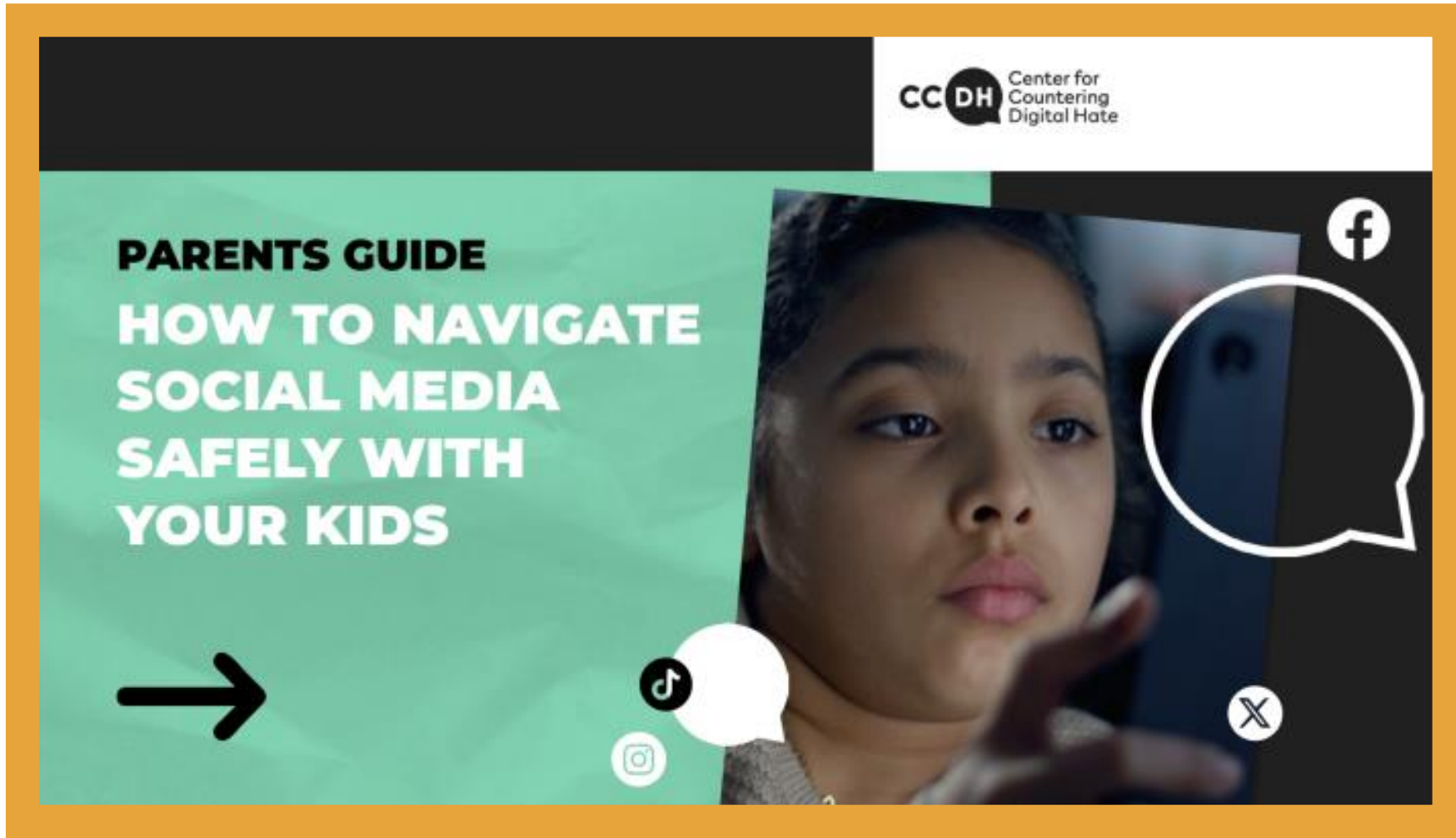
Ads from major brands ran alongside these harmful videos



YouTube failed to act on **81%** of flagged harmful videos



# RESOURCES



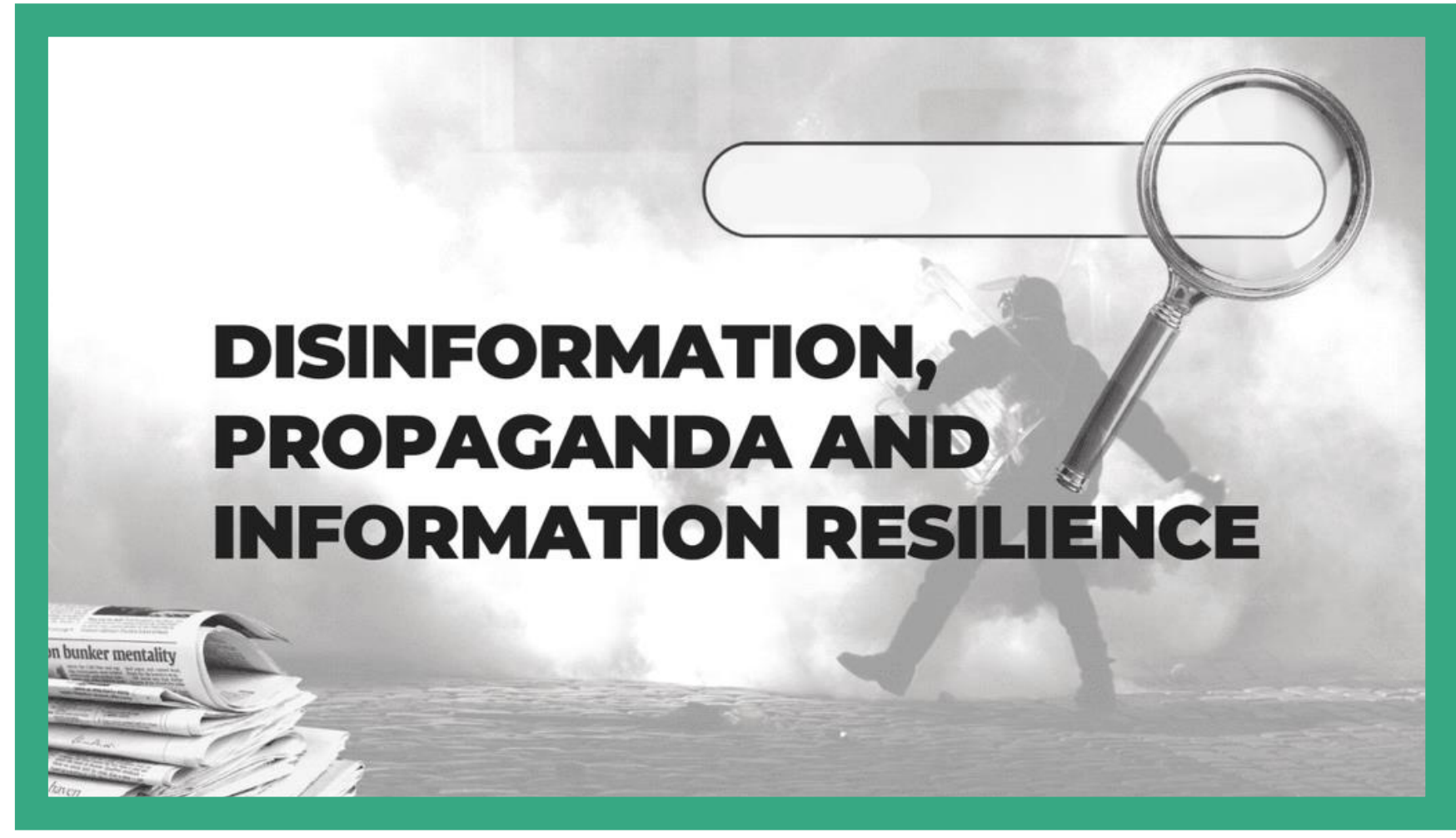
**CCDH** Center for Countering Digital Hate

**PARENTS GUIDE**  
**HOW TO NAVIGATE SOCIAL MEDIA SAFELY WITH YOUR KIDS**

→

Facebook, Instagram, TikTok, X

A banner with a teal background on the left and a black background on the right. The teal section contains the text 'PARENTS GUIDE HOW TO NAVIGATE SOCIAL MEDIA SAFELY WITH YOUR KIDS' and a white arrow pointing right. The black section features a young woman looking at a smartphone. Social media icons for Facebook, Instagram, TikTok, and X are overlaid on the image. The CCDH logo is in the top right corner.



**DISINFORMATION, PROPAGANDA AND INFORMATION RESILIENCE**

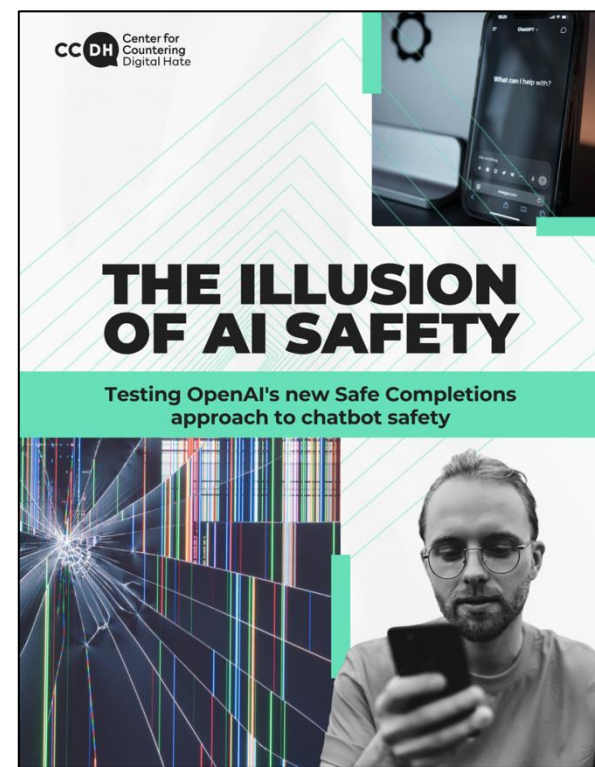
Search bar, magnifying glass, person silhouette, newspaper stack

A banner with a green border. The background is a grayscale image of a person in a trench coat and hat running through a misty, dark environment. A large magnifying glass is positioned over the person. In the bottom left corner, there is a stack of newspapers, with the headline 'on bunker mentality' visible. A search bar is located at the top right.



# READ OUR REPORTS

## AI



### THE ILLUSION OF AI SAFETY

Testing OpenAI's new Safe Completions approach to chatbot safety



## Misogyny



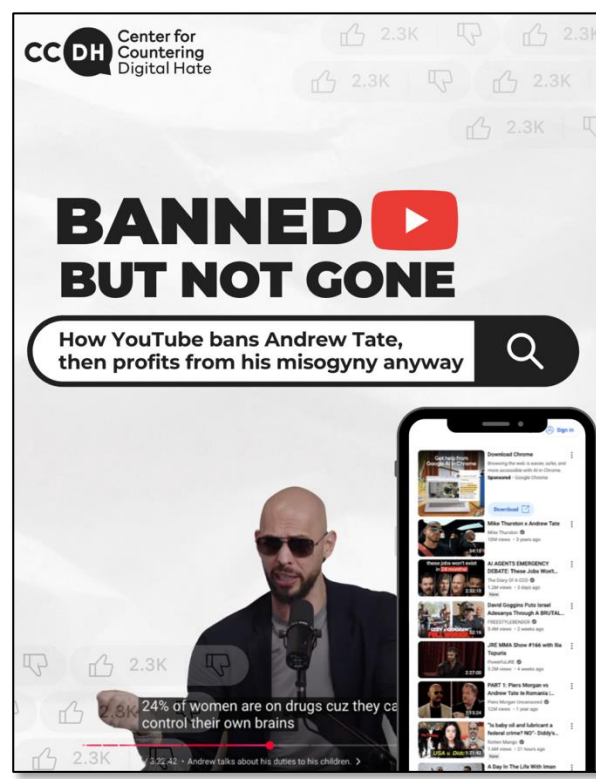
### NEW REPORT THE INCELOSOSPHERE

Exposing pathways into incel communities and the harms they pose to women and children.



### ABUSING WOMEN IN POLITICS

How Instagram is failing women and public officials



### BANNED BUT NOT GONE

How YouTube bans Andrew Tate, then profits from his misogyny anyway



## SRHR



### DIGITAL DISPARITIES

The global battle for reproductive rights on social media



# SIGN UP FOR OUR NEWSLETTER

## Get your monthly updates.

Subscribe to receive our monthly newsletter.

First name \*

Email address \*

Country \*

United Kingdom ▼

I would like to receive the CCDH in Action monthly newsletter. Click here to [read our privacy policy](#). You may opt-out of future communications at any time. \*

Yes  No



The fight for a safer internet is happening right now.

Social media and AI giants keep putting profit before safety – and the harms against children, women and our communities keep growing.



# FOLLOW US ON SOCIAL MEDIA

**Bluesky**



**Instagram**



**LinkedIn**

