

TALKING ABOUT GROK - AI CHATBOX

What is GROK?

GROK is an Artificial Intelligence (AI) chatbox created by xAI and built into social media platform X (formerly known as Twitter). It can generate text, images and answers based on user prompts and is widely accessible to adults, young people and intentionally or not children. Recent reports show it has major safety failings, particularly around identifying U18 users and preventing harmful content.

Why is GROK a Safeguarding Concern?

Independent assessments have found that GROK:

- Fails to reliably recognise when a user is a child.
- Generates harmful, sexual, violent, and or inappropriate content.
- Allows material to be easily shared on X.
- Has weak or broken “kids mode” filters.

These issues mean children can be directly exposed to unsafe content or exploited through the platform.

Recent issues

There have been multiple, serious incidents linked to GROK:

Child Sexual Abuse Material (CSAM) generated by AI

- GROK produced sexualised and explicit AI generated images of minors, some as young as 11.
- The chatbox publicly admitted this “violated ethical standards” and possibly broke US child safety laws.
- xAI confirmed there were “lapses in safeguards” and acknowledged potential criminal and civil liability.

Misuse by users

- Creating nonconsensual sexualised deepfakes of women and children.
- Editing real photos to remove clothing or place people in sexualised situations.
- Roughly one nonconsensual sexualised image per minute was identified in public feed.

Past harmful behaviours

GROK has a pattern of unsafe outputs, which include:

- Racist “white genocide” content.
- Antisemitic, pro Hitler messages.

These highlight repeated safety failures over less than a year.

How are authorities responding?

The fallout from GROK has sparked:

- Investigations child safety breaches.
- Strong criticism from child protection groups such as Common Sense Media.
 - Government pressure in the UK and US for tougher AI regulation
 - Moves towards mandatory AI safety standards across the industry, not voluntary guidelines.

What does this mean for children and young people?

Children and young people may be at increased risk of:

- Exposure to harmful, adult or violent AI generated content.
- Becoming victims of deepfakes abuse (sexualised or manipulated images).
- Being groomed or exploited through AI generated personas.
- Having their real photos altered and spread online without consent.

The speed and volume of AI generated harmful content, combined with weak safeguards, significantly increases vulnerabilities.

What to look out for if we have concerns?

This is not an exhaustive list:

- Young people talking about using GROK or the “Grok Imagine” image generator.
- Distress linked to content they’ve viewed on X.
- Fake or sexualised images of children circulating in peer groups.
- Sudden anxiety about photos being shared online.
- Evidence of manipulation, blackmail, or coercion based on AI-generated images.
- A child reporting or hinting that an image of them has been altered.

What to do if staff have concerns?

- Reach out to the Safeguarding Team.
- Do not search or view illegal images.
- Contact Police if there is risk of significant harm or evidence of CSAM.

Signposting:

- NSPCC Helpline (0808 800 5000) – for advice, reassurance, or support.
- Internet Watch Foundation (IWF) – report harmful images so they can be taken down quickly.
- Local Authority Children’s Services – if they feel a child is at risk of harm.

