

AI SAFEGUARDING FOR PROFESSIONALS

Insight

A quick reference tool for staff working with children, young people and vulnerable adults

What is Artificial Intelligence (AI)?

Artificial Intelligence (AI) refers to computer systems that can generate text, images, audio or decisions by identifying patterns in large amounts of data. AI is not human, cannot understand emotional or contextual detail, and may produce inaccurate, biased or misleading information.

AI also includes Generative AI, which creates new content using patterns learned from large datasets. Generative AI does not think or understand; its outputs are predictions and may contain errors or bias. Tools such as ChatGPT, Microsoft Copilot, and Google Gemini, as well as image/voice/video systems like DALL-E, Midjourney, ElevenLabs, and Sora, can create realistic synthetic media that may be difficult to distinguish from real content.

What is the purpose of this resource?

This resource is designed to support professionals working with children, young people and vulnerable adults to promote safe, responsible and well-informed use of AI. It aims to help staff recognise AI-related risks, protect individuals from harm, and respond confidently and appropriately to any safeguarding concerns involving AI.

UNDERSTAND - Know the basics of how AI works and the implications for safeguarding

- ☐ AI is not human, cannot understand context, and may produce inaccurate or biased outputs.
- ☐ AI-generated images, videos, audio or text may be fabricated or manipulated.
- ☐ AI can mimic voices and identities, increasing risks of impersonation.
- ☐ AI systems may collect or reuse personal data as part of their functioning.
- ☐ Offenders may use AI to scale harmful behaviours (e.g., grooming scripts, fake profiles).
- ☐ Generative AI tools such as ChatGPT, Microsoft Copilot and Google Gemini produce responses using statistical patterns, not understanding.
- ☐ Image/voice/video generators (e.g., DALL-E, Midjourney, ElevenLabs, Sora) can produce highly realistic synthetic media, increasing opportunities for deception.
- ☐ AI in everyday life includes spam filters, recommendations, navigation apps and voice assistants - making AI exposure routine for children and adults.

IDENTIFY - Recognise when AI use may present a risk or concern

- ☐ Content that seems "too perfect," inconsistent or unrealistic may be AI-generated.
- ☐ Unrealistic images, deepfakes or nudification outputs indicate possible image misuse.
- ☐ Messages that feel out of character or overly persuasive may involve AI.
- ☐ Voice or identity impersonation may indicate cloning or fraud.
- ☐ Emotional dependence on AI chatbots or companions may signal vulnerability.
- ☐ Parasocial attachments to AI "friends" may lead children to believe the AI cares, understands, or offers emotional security.
- ☐ Children may overshare personal information with AI chatbots or "companions."
- ☐ Requests for new images or audio may relate to blackmail or cloning attempts.
- ☐ Any AI-related discomfort or uncertainty should prompt safeguarding curiosity.

SUPPORT - Promote safe, informed and responsible AI use

- ☐ Encourage individuals to discuss confusing or upsetting AI content.
- ☐ Teach how to question, verify and challenge AI outputs.
- ☐ Encourage critical thinking when using AI-generated text, images or videos.
- ☐ Guide safe use of new AI tools and apps, especially for younger users.
- ☐ Promote healthy boundaries to reduce emotional reliance on AI companions.
- ☐ Raise awareness of identity theft, voice cloning and image-editing risks.
- ☐ Promote awareness of deepfakes and realistic AI-generated media.
- ☐ Encourage privacy-protective behaviours (checking permissions, limiting shared media).
- ☐ Support safe use of AI in education (avoiding plagiarism, protecting data).

PROTECT - Ensure personal information, well-being and digital safety are safeguarded

- ☐ Never enter personal or sensitive data into public AI tools.
- ☐ Only share images or audio with informed consent and within policy.
- ☐ Do not use AI for personal, medical or safeguarding advice.
- ☐ Use only approved, secure and policy compliant AI tools.
- ☐ Stop use immediately if content appears unsafe or inappropriate.
- ☐ Ensure organisational AI policies, filtering and data protection measures are current.
- ☐ Warn users not to share personal information with AI chatbots or "companion apps."
- ☐ Highlight risks of identity impersonation, synthetic voices and AI-enabled cloning.

RESPOND - Act promptly and appropriately when an AI related safeguarding issue occurs

- ☐ Report harmful or distressing AI content through safeguarding channels immediately.
- ☐ Stop and escalate AI tools that request personal information or appear manipulative.
- ☐ Treat deepfakes, harassment or image misuse as safeguarding concerns.
- ☐ If AI-generated impersonation or deepfake content is reported, respond as a safeguarding risk.
- ☐ Respond promptly to reports of blackmail, impersonation or coercion involving AI.
- ☐ Block, record and escalate unsafe AI use in line with procedures.
- ☐ Reassure individuals that raising concerns leads to support, not blame.

KEY MESSAGE

AI can support learning and creativity - but safeguarding practice, professional judgement and clear boundaries must always guide its use.

Definitions of Potential Harms

Fake images and videos – Pictures or videos created or altered by AI to make something look real when it isn't.

Nudification apps – Tools that use AI to create fake images of someone without clothes, even though the person never posed that way.

Online blackmail – When someone threatens a person online to make them do something, often using real or fake images.

Image misuse – When someone shares, edits, or uses another person's images without their permission.

Unhealthy online relationships – Online connections where someone pressures, manipulates, or controls another person.

Early warning signs of AI-related harm – Clues in behaviour or online activity that suggest someone may be being pressured, tricked, or exploited using AI tools.

Further Resources and Support

Check out for more information and support: [kidsonlineworld](https://www.kidsonlineworld.org) or [internet matters](https://www.internetmatters.org)



Our blogs

To learn more
listen to our
blogs and
podcast!
Just scan the
QR codes:



Our podcasts

Safeguarding is everyone's
responsibility; everyone has a
duty of care to safeguard those
at risk of harm.

www.ecpsafeguarding.co.uk