Generative AI and the safeguarding risks (KCSIE 2026) Summary

What Generative AI Is

Generative AI creates new content (text, images, audio, video, code) based on user prompts. It can produce highly realistic content that appears genuine — including images or videos of children or staff that **never existed**. AI is now a **day-to-day safeguarding responsibility** under KCSIE 2026.

Key Safeguarding Risks

Image-Based Sexual Abuse (IBSA) AI can generate: • sexualised deepfakes • nudification (removal of clothing) • synthetic sexual imagery • sexualised images of fictional minors • sexualised images of staff All of these are treated as **image-based sexual abuse**, even if the child or adult never posed for the image. Creating or sharing AI sexual imagery of anyone under 18 is a **criminal offence**	**Child Sexual Abuse Material (CSAM)** AI can create: • synthetic CSAM • realistic images of children known to perpetrators • immersive simulations of abuse These are treated **as real CSAM in law**.
Grooming & Online Sexual Abuse AI chatbots can: • simulate grooming • mimic adult behaviours • manipulate children into sharing personal data • escalate into sexual exploitation	**Coercion, Blackmail & Sextortion** AI can be used to: • threaten children with deepfakes • extort money or sexual content • impersonate trusted adults • create fake evidence
Radicalisation & Extremism (Prevent Duty) AI can generate: • extremist propaganda • conspiracy narratives • political manipulation • simulated extremist grooming This is a **Prevent concern**.	**Fraud, Impersonation & Identity Theft** AI can: • clone voices • replicate faces • impersonate staff or pupils • generate fake school messages • create phishing emails
Emotional Manipulation & Mental Health Harm AI can: • encourage self-harm • simulate relationships • isolate children • reinforce negative self-image	**Bias & Discrimination** AI may generate: • racist, homophobic, misogynistic or ableist content • harmful stereotypes • prejudiced narratives
SEND Vulnerability Children with SEND may be: • more trusting of AI • more vulnerable to manipulation • less able to identify deepfakes	**Chatbots and AI companions** • Harmful advice on self-harm, suicide, eating disorders and substances • Missed warning signs and poor mental health support • Disclosures made to a bot instead of a trusted adult • Emotional dependency and withdrawal from real relationships • Sycophancy (agreeing with the user) • Over-trust in inaccurate answers • Believing the chatbot is a real person • Age-inappropriate content and weak age checks • Chatbots built into apps children already use • Sharing personal data • Manipulative or engagement-driven design • Greater risk for vulnerable children

Exposure to Inappropriate Content

Children may access:

- pornography
- violence
- sexual chatbots
- extremist content
- harmful mental-health content

AI platforms often lack age verification.

Misinformation, Disinformation & Fraud

AI can generate:

- plausible false information
- conspiracy theories
- extremist narratives
- fraudulent messages
- impersonation of peers or adults

All AI content must be **fact-checked**.

Impact on Learning

Children may use AI to:

- complete work
- generate answers
- bypass independent thinking

This can mislead teachers and limit cognitive development.
Staff must complete a **risk assessment** and obtain **Head approval** before using AI with pupils.

Privacy & Data Protection

AI systems may:

- collect personal data
- track behaviour
- analyse interactions

The Trust maintains an AI Risk Assessment and Data Protection Impact Assessment.

Filtering & Monitoring

Schools must complete **annual filtering and monitoring reviews** to detect:

- nudification sites
- deepfake generators
- harmful AI tools
- evasion techniques

Staff Actions

If you become aware of harmful AI use:

1. **Speak to a DSL immediately**.
2. **Do not investigate** or ask leading questions.
3. **Record concerns on CPOMS**, using the child's exact words.
4. **Do not confront the pupil or alleged perpetrator**.
5. **Do not attempt to analyse or reverse-engineer the AI content**.

Experimentation with harmful AI tools is a **safeguarding incident**, even if it was a "joke".

DSL Responsibilities

DSLs will:

- assess risk
- consider links to other vulnerabilities (SEND, exploitation, mental health)
- determine whether the incident is IBSA, HSB, CSE, CCE, domestic abuse or Prevent
- liaise with Police, Family Help and Prevent
- ensure safe information sharing
- monitor patterns
- ensure filtering systems are updated
- ensure staff training is current

Bourne Alliance
Multi Academy Trust

AI and the risks relating to safeguarding & GDPR policy

Review: Autumn 2027

This policy forms part of the Trust Safeguarding and Child Protection policy and reflects the latest KCSIE guidance. Keeping children safe in education 2026 (comes into force from 1 September 2026) ***All staff must read Part 1 as a minimum.***

KCSIE 2026 explicitly links **AI-generated and digitally manipulated content** to safeguarding risks including coercion, child-on-child abuse, harmful sexual behaviour (HSB), CSE, CCE, domestic abuse, radicalisation, fraud and image-based sexual abuse (IBSA).

AI is now a **statutory, day-to-day safeguarding responsibility**.

What Generative AI Is

Generative AI is a form of artificial intelligence that creates original content (text, images, audio, video, code) in response to user prompts. It uses deep learning models trained on vast datasets to identify patterns, understand natural language and generate realistic content.

AI-generated content can be:

- edited
- refined
- personalised
- manipulated until it appears entirely authentic.

This creates significant safeguarding risks.

Key Safeguarding Risks (KCSIE 2026)

KCSIE 2026 identifies the following AI-related risks:

Image-Based Sexual Abuse (IBSA)

AI can generate:

- sexualised deepfakes
- nudification (removal of clothing)
- synthetic sexual imagery
- sexualised images of fictional minors
- sexualised images of staff

All of these are treated as **image-based sexual abuse**, even if:

- the child never posed for the image
- the image is entirely fake
- the child is fictional or synthetic

Creating or sharing AI-generated sexual imagery of anyone under 18 is a **criminal offence** under the Online Safety Act and Sexual Offences Act.

Child Sexual Abuse Material (CSAM)

AI can generate:

- synthetic CSAM
- realistic images of children known to perpetrators
- fictional minors

- immersive simulations of abuse

These images are treated **as real CSAM in law**.

Grooming and Online Sexual Abuse

AI chatbots can:

- simulate grooming
- mimic adult behaviours
- manipulate children into sharing personal data
- escalate into sexual exploitation

Coercion, Blackmail and Sextortion

AI can be used to:

- threaten children with deepfakes
- extort money or sexual content
- impersonate trusted adults
- create fake evidence

Domestic Abuse and Coercive Control

AI can be used within domestic abuse to:

- monitor
- track
- threaten
- manipulate
- generate sexualised or humiliating imagery

Radicalisation and Extremism (Prevent Duty)

AI can generate:

- extremist propaganda
- conspiracy narratives
- political manipulation
- radicalising content
- simulated extremist grooming

This is a **Prevent concern** requiring DSL involvement.

Fraud, Impersonation and Identity Theft

AI can:

- clone voices
- replicate faces
- impersonate staff or pupils
- generate fake school communications
- create phishing emails
- simulate genuine systems

Emotional Manipulation and Mental Health Harm

AI can:

- encourage self-harm
- simulate relationships
- isolate children
- reinforce negative self-image
- generate harmful mental-health content

Bias, Discrimination and Protected Characteristics

AI may generate:

- racist content
- homophobic content
- misogynistic content
- ableist content
- stereotypes and prejudice

This can harm children and escalate into bullying or hate incidents.

SEND Specific Vulnerability

Children with SEND may be:

- more trusting of AI
- more vulnerable to manipulation
- less able to identify deepfakes
- more susceptible to misinformation

Violence and Normalisation of Harm

AI can generate:

- violent scenarios
- simulated harm involving known individuals
- content that normalises aggression

Bypassing Filters and Controls

AI can teach children:

- how to bypass filtering
- how to evade monitoring
- how to access harmful sites

This is now explicitly referenced in KCSIE 2026.

Exposure to Inappropriate Content

AI platforms often lack age verification. Children may access:

- pornography
- graphic violence
- sexual content
- extremist content
- harmful mental-health content
- simulated romantic or sexual chatbots

AI may also embed inappropriate content within otherwise legitimate responses. AI-driven advertising can target children with persuasive commercial content.

Misinformation, Disinformation and Fraud

AI can generate:

- plausible false information
- misleading narratives
- conspiracy theories
- extremist propaganda
- fraudulent messages
- impersonation of peers or adults

This can:

- distort children's understanding
- increase vulnerability to radicalisation
- manipulate children into sharing personal data
- facilitate scams or sextortion

All AI-generated content must be **fact-checked**.

Impact on Learning and Cognitive Development

Children may use AI to:

- complete schoolwork
- generate answers
- bypass independent thinking
- avoid problem-solving

This can:

- mislead teachers about progress
- limit cognitive development
- reduce critical thinking skills

Staff and pupils receive training on safe and appropriate use.

Staff must complete a risk assessment and obtain **written approval from the Head** before using AI with pupils.

Privacy and Data Protection

AI systems may:

- collect personal data
- track behaviour
- analyse interactions
- store information without children's understanding

The Trust maintains:

- an AI risk assessment
- a Data Protection Impact Assessment (DPIA)
- strict controls on platform use

Filtering and Monitoring (KCSIE 2026)

All our schools must undertake **annual filtering and monitoring reviews** to ensure systems detect:

- nudification sites
- deepfake generators
- harmful AI tools
- evasion techniques

Primary Technologies completes this for each school.

Bullying and Harassment Using AI

AI can generate:

- humiliating images
- threatening text
- unflattering or distressing content
- violent simulations
- deepfakes
- nudification

Bullying chatbots can normalise aggression and escalate into offline bullying.

Staff Training

Staff receive ongoing training on:

- AI risks
- GDPR

- safeguarding implications
- recognising AI-enabled abuse
- safe workload-supporting use of AI

New staff training (2026 onwards) includes:

- signs of AI-enabled abuse
- deepfake recognition
- IBSA awareness (Image Based Sexual Abuse)

Staff Actions

If staff become aware of harmful AI use:

1. **Speak to a DSL immediately** (in person wherever possible).
2. **Do not investigate** or ask leading questions.
3. **Record concerns on CPOMS**, using the child's exact words.
4. **Do not confront the pupil or alleged perpetrator**.
5. **Do not attempt to analyse or reverse-engineer the AI content**.
6. **Maintain confidentiality** — share only with DSL/deputy.

Experimentation with harmful AI tools is a **safeguarding incident**, even if:

- the child "didn't mean harm"
- it was a "joke"
- the content is entirely fake

Staff MUST NOT enter pupil or safeguarding data into AI tools. Information such as a piece of writing for comparison to moderation criteria could be used for example, but it cannot have the name of the child on it.

Only approved AI tools can be used. See the AI policy and risk assessment for full details.

DSL Responsibilities

DSLs will:

- complete an immediate risk assessment
- consider links to other vulnerabilities (SEND, exploitation, mental health)
- assess whether the incident is IBSA, HSB, CSE, CCE, domestic abuse or Prevent
- liaise with Police, Family Help, Prevent and other agencies
- ensure safe information sharing
- monitor patterns through caseload reviews
- ensure risk information transfers when a child moves school
- ensure filtering and monitoring systems are updated
- ensure staff training is current
- Ensure that they or staff do not enter any personal or safeguarding data into AI tools not use AI tools to write safeguarding records or to make safeguarding decisions
- Check that any AI tools in use have been approved – in line with the AI policy and RA. Ban any other tool which is not approved.

Linked Policies

- Child Sexual Exploitation (CSE)
- Child Criminal Exploitation (CCE) & County Lines
- Nudes and Semi-Nudes
- Deepfakes
- Domestic Abuse
- Child-on-Child Abuse
- Online Safety
- Child Protection & Safeguarding Policy

Useful resources

- Online safety - Kent Safeguarding Children Multi-Agency Partnership (kscmp.org.uk)
- What do I need to know about ChatGPT? A guide for parents and carers | Childnet
- Using artificial intelligence (AI) safely | Internet Matters
- What is generative AI and how can it impact kids? | Internet Matters
- A guide to artificial intelligence (AI) for parents | Internet Matters
- learning.nspcc.org.uk/media/3341/child-safeguarding-immersive-technologies.pdf
- Data protection in schools - Artificial intelligence (AI) and data protection in schools - Guidance - GOV.UK (www.gov.uk)
- Artificial Intelligence and Online Safety | SWGfL
- Using artificial intelligence (AI) safely | Internet Matters
- https://www.gov.uk/government/publications/generative-ai-product-safety-standards/generative-ai-product-safety-standards
- Generative artificial intelligence (AI) in education - GOV.UK
- Using AI in education settings: support materials - GOV.UK
- Generative AI: product safety expectations - GOV.UK
- Generative AI in education: user research and technical report - GOV.UK
- Generative AI in education: educator and expert views - GOV.UK
- Data protection in schools - Artificial intelligence (AI) and data protection in schools - Guidance - GOV.UK (www.gov.uk)
- Artificial Intelligence and Online Safety | SWGfL
- Using artificial intelligence (AI) safely | Internet Matters
- Integrating AI in Schools: New Policy Template Available | SWGfL
- Ofsted's approach to artificial intelligence (AI) - GOV.UK
- National AI Strategy - GOV.UK
-