



# How to Keep Students Safe in the Age of AI: A Field Guide for UK Education Settings

---

# Introduction: Finding Your Bearings

In 2025, AI is woven into the digital fabric of education — reshaping how we search, plan, teach, and communicate.

For schools, colleges and MATs, it presents exciting new opportunities — but also important new responsibilities when it comes to student safety.

As education leaders look to embrace AI, many are seeking a clear path forward. How do we adopt it safely? What policies are needed? How do we protect students and stay compliant?

This field guide is here to help you find your bearings.

Whether you're a DSL, headteacher or IT leader, it lays out the key risks, responsibilities and guideposts for navigating AI within education settings.

## Learn how to:

- Create a robust AI strategy and policies
- Evaluate and choose the right AI tools for your setting
- Protect students against harmful AI content
- Spot students at risk from AI
- Interpret the regulatory expectations and best practices relating to AI

Use this guide to help you move forward with confidence.

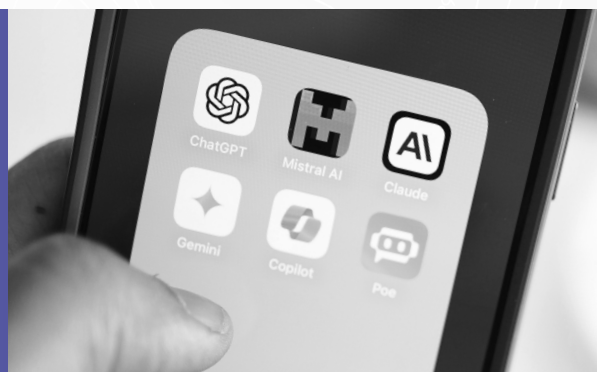
# Mapping the Current AI Landscape

As schools, colleges and MATs begin their journey with AI, it helps to understand the lay of the land. Before we can meaningfully evaluate which tools should be used, and to what extent, we need to be clear on the type of AI we're addressing.

In the context of UK education and regulation, generative AI is the primary focus. It's the form of AI most commonly encountered by staff and students, and the one most likely to raise questions around use, safety, and policy.

## What is Generative AI?

Generative AI refers to tools that can produce content - whether that's text, imagery, video, or code - at a scale and speed no human could match. Common examples include ChatGPT, Gemini, Microsoft Copilot, DeepSeek. These tools are already being explored in education for everything from lesson planning and research to coding and administrative tasks.



## The DfE's stance on generative AI in education

While generative AI has many notable benefits, it also presents significant safeguarding challenges. To support education settings to utilise AI tools safely, responsibly and effectively, the Department for Education (DfE) released the policy paper **Generative artificial intelligence (AI) in education**.

The DfE's paper strikes a balance between highlighting the positive transformational impact AI can have on the education sector, whilst emphasising the need for every setting to establish clear policies to regulate its application.


In terms of opportunities, the DfE sees "more immediate benefits and fewer risks from teacher-facing use of generative AI." In particular, AI's ability to remove some of the administrative burdens on staff, allowing them to focus more on teaching.

The acknowledged risks posed by generative AI use in schools and colleges include:

- Infringements on data privacy and intellectual property rights
- Increased access to inappropriate, illegal or harmful content
- Potential undermining of the integrity of formal assessments.

Rather than issuing hard rules, the DfE empowers schools, colleges and MATs to make local decisions about whether and how AI tools should be used - so long as those decisions are underpinned by clear, documented policies.





‘Rather than issuing hard rules, the DfE empowers schools, colleges and MATs to make local decisions about whether and how AI tools should be used.’

# Plotting a Safe Route: Building Your AI Policy

While education settings have flexibility in shaping their AI approach, they still need to abide by their legal and statutory obligations, including Keeping children safe in education and data protection laws.

An effective AI policy should do more than list what’s allowed and what’s not. It should also consider:

Ethical use	Does the AI policy support learning without encouraging shortcuts, bias, or misuse?
Staff and student training	Do our staff and students understand when, why and how to use AI tools safely and effectively? Have you equipped them to think critically about AI outputs?
Safeguarding implications	What risks could AI pose to student safety or wellbeing? Have we considered how AI tools might expose students to harmful content or mask warning signs?
Policy alignment	Does the AI policy align with related policies, such as: staff code of conduct, student acceptable use policies and data protection policies (including GDPR)?



## Choosing the right AI tools for use

Not all AI tools are created equal. They vary in purpose, functionality and risk - and the landscape is constantly shifting. That's why policies should clearly set out:

- Which AI tools are permitted
- Who can use them
- What they can be used for

For example:

- Are tools like ChatGPT only allowed for teacher use?
- Can students use AI tools for specific subjects or tasks?
- Are certain tools or features restricted by year group?

The more specific the policy, the easier it will be to enforce and review.



## Define responsible use

Where AI use is permitted, it's vital to set clear boundaries on what can and cannot be shared with these tools.

Intellectual property	Lesson plans, homework, and essays remain the property of the person who created them. Uploading this content to an AI tool could breach IP laws.
Personal data	Staff and student data must not be shared with AI tools unless clear consent has been obtained. The DfE recommends avoiding this altogether to ensure compliance with data protection laws.
Terms and age restrictions	Every AI tool comes with its own terms of use. These should be reviewed and incorporated into your policy - especially any age restrictions.

## 4 Steps to Writing a Successful AI Policy

Establishing a clear AI policy may feel complex, but taking action now ensures your setting is prepared, protected, and positioned to respond confidently as AI continues to evolve.

Start with a basic foundation and view it as an evolving policy that can be updated as new information or technology is introduced.

The following steps may help:

### STEP 1:

Discuss policy development as an SLT and get DSLs involved early



### STEP 2:

Review your setting's current AI usage (both official and unofficial)



### STEP 3:

Collaborate with IT staff to understand any cybersecurity risks and training needs



### STEP 4:

Stay informed of updates from the DfE and other trusted sources


## Who should take ownership of your AI policy?

Responsibility for formulating an AI policy or strategy shouldn't rest with one individual. Given the fast-moving nature of the technology, it's important to involve all relevant stakeholders in the creation and ongoing development of your AI policy. This ensures shared understanding, collective ownership, and consistent application across your setting.

### Appointing an AI champion

Many early adopters of AI in schools, colleges and MATs have found it valuable to appoint an "AI champion" - a member of staff who can support colleagues in building their understanding of AI and help drive confident, informed adoption. This doesn't need to be an IT specialist; it could be any staff member with a strong interest or knowledge in AI.



A photograph of two students in school uniforms. A girl in the foreground is looking down at a device, smiling slightly. A boy in the background is also looking down at the device. The image has a blue and green color overlay.

‘Schools, colleges and MATs must consider the role of web filtering in keeping both staff and students safe from harmful AI-generated content.’



# Fortifying Your Campsite: How to Filter Harmful AI Content Online

As part of enforcing AI policies and safeguarding students, schools, colleges and MATs must consider the role of web filtering in keeping both staff and students safe from harmful AI-generated content.

## The role of filtering in the Age of AI

Web filters play an essential role in helping schools manage the online risks associated with AI-generated content, offering control, visibility and protection.

An effective filter should be able to:

- ✓ **Block access to harmful websites**  
This includes sites created or manipulated by AI to spread misinformation, deepfakes or inappropriate content.
- ✓ **Prevent access to high-risk AI platforms**  
Especially those that are unmoderated or deemed unsuitable for use in education settings.
- ✓ **Tailor access to AI tools**  
Set rules based on year group, user role or time of day to ensure appropriate access for different users.
- ✓ **Block unsafe or inappropriate search terms**  
Reducing the likelihood of students encountering harmful AI-generated material via search engines.

## Filtering technologies at a glance

Not all web filters handle AI content equally. Understanding the differences is crucial to choosing the right protection:

### DNS Filters

Filter based only on domain names (e.g. chat.openai.com). They offer no visibility of the content being accessed, making decisions binary: block or allow.

### URL Filters

Allow or restrict access to web pages based on assessments of the full URL. Useful for some search engines, but ineffective with AI tools like ChatGPT, where prompts and responses don't appear in the URL.

### Content-aware filters

Analyse a web page's content to decide if it's safe - but not at the point the page is accessed by an individual.

Instead, decisions may be based on older versions of a web page, sometimes from days or weeks ago. With AI content changing rapidly, this delay could mean harmful material is seen before the filter catches up.

### Real-time, content-aware filters

Perform all the actions of content-aware filters, but more thoroughly and in real time, at the point the page is requested, ensuring that harmful AI generated content is seen and blocked much more quickly.

## Why 100% real-time filtering is essential for AI

With AI, content is unpredictable, often unmoderated and appears online at speed. Without real-time analysis, harmful content can be missed - or blocked too late. Real-time, content-aware filtering ensures:

- Harmful content can be blocked the moment it appears
- Teaching and learning can continue uninterrupted
- Education settings stay aligned with UK filtering standards and safeguarding expectations

As the UK Safer Internet Centre (UKSIC) highlights, schools, colleges and MATs must understand how their filters handle “dynamically analysed” content in real time.



## Aligning your AI policy with your filtering

Your web filter should directly support your AI policy and help you manage access to AI tools safely and appropriately.

Consider the extent to which your filter can:

### **Control which AI tools can be used**

Block or allow access to specific tools depending on your setting's rules.

### **Block high-risk AI platforms automatically**

Use pre-set categories to prevent access to known unsafe or inappropriate tools.

### **Set access rules based on who is using it and when**

For example, allow access for staff but not students, or restrict use during certain times of day.

These features make it easier to enforce your AI policy in a fair, consistent and effective way - without disrupting learning.

Smoothwall Filter brings together all of these capabilities - including 100% real-time, content-aware filtering - as standard. It empowers schools, colleges and MATs to enforce their AI policies with confidence, respond to emerging risks instantly, and keep their students and staff safe online.

## Web filtering plays a vital role, but it's only one part of the picture

While web filtering can go a long way in helping to protect your staff and students against potentially harmful AI content, it's important to note: no filter - regardless of how advanced - can eliminate every risk posed by AI tools.

Filtering therefore must be part of a broader strategy that includes:

- Clear AI usage policies
- Risk assessments for new technologies
- Staff and student training
- Digital monitoring for online behaviours (more on this in the next section).

# Watching the Trail: How Digital Monitoring Spots Students at Risk from AI

While web filtering helps education settings to control access to AI tools and block harmful content, it doesn't reveal how students are interacting with AI or what those interactions might tell us about their overall mental health and wellbeing. For example, could a student be engaging in inappropriate or potentially dangerous conversations with a chatbot?

That's where digital monitoring plays a critical role.



## What is digital monitoring?

Digital monitoring refers to safeguarding solutions that identify students at potential risk through what they do, say or share on school-owned digital devices. These tools run silently in the background, so they don't disrupt teaching or learning.

Monitoring systems identify potential risks by registering keystrokes and taking screenshots when threats are detected.

Alerts are then created, which are sent to a designated staff member at the school (usually the DSL).

For example, if a student types "how to build a bomb" into an AI chatbot, the word "bomb" should trigger the monitoring system to act. How alerts are managed and communicated depends on the type of digital monitoring in place.



## The role of digital monitoring in addressing AI risks

Filtering can block access to dangerous AI tools - but it can't show how students use AI, what prompts they write/ input, or what those interactions reveal.

Digital monitoring fills this gap in two critical ways:

### 1. By spotting signs of AI misuse

Students may attempt to use generative AI to cheat, bypass filters, or engage in inappropriate conversations. Some even form unhealthy, synthetic relationships with AI tools - using them as substitutes for companionship or emotional support.

There's also a growing concern around harmful AI tools like nudifiers, which digitally remove clothing from images. In a school context, this poses serious safeguarding implications, including the potential creation of child sexual abuse material (CSAM).

Digital monitoring solutions can discourage misuse of generative AI by flagging incidents such as the use of sexual language, queries linked to safeguarding issues such as eating disorders, or searches for terms like "nudifier". Indeed, sometimes just the knowledge that digital devices are being monitored can deter network users from acting inappropriately.

### 2. By identifying vulnerable students using AI tools

The main role of digital monitoring is to enable DSLs to spot at-risk students early, so interventions can take place before incidents have a chance to escalate.

Early signs of vulnerability may be revealed in what students type into AI chatbots. Some students use AI tools to open up about personal struggles they may not feel comfortable sharing with adults. For instance, there is a growing trend of young people using AI as a form of therapy. While this may feel safe to the student, it can lead to greater risk as AI chatbots lack the qualifications and expertise to provide appropriate, professional support.

Digital monitoring gives safeguarding staff visibility into these interactions - helping them understand the context and respond appropriately. This can lead to quicker, more informed interventions tailored to the student's needs.

## Detecting AI risks with the help of human-moderation

AI-related risks can be subtle and context-dependent. That's why many education settings are now turning to human-moderated digital monitoring to enhance their safeguarding response.

Unlike exclusively automated systems, human moderation brings expert insight - enabling more accurate interpretation of language, tone and intent. Moderators can assess alerts in real time, spot patterns over time, and pick up on coded or concealed language that might otherwise go unnoticed.

When applied to AI use in education, human-moderated monitoring can help:

- ✓ **Enforce AI policies with greater visibility**  
Provide schools with a clearer picture of how AI tools are being used across the network.
- ✓ **Deter unsafe or inappropriate AI use**  
Monitoring can act as a deterrent—reducing incidents of misuse by making expectations and oversight clear.
- ✓ **Spot students at risk through their digital behaviours**  
Identify early signs of distress, unsafe searches or emotionally charged interactions with AI tools.
- ✓ **Enable faster, more informed safeguarding decisions**  
Offer DSLs detailed, contextual information that supports timely and effective intervention.

Tools that use human-moderated digital monitoring - such as [Smoothwall Monitor](#) - are helping schools, colleges and MATs respond more effectively to the risks AI introduces. Not just at the point of access, but through the ongoing patterns and behaviours that follow.



# Reading the Signs: Understanding Expectations Around AI Use

As AI continues to shape the education landscape, many schools, colleges and MATs are still working out how to introduce it responsibly. While the use of AI is growing, the official expectations around its adoption remain broad and evolving.

There is currently no formal requirement for UK education settings to use AI tools. However, with Ofsted beginning to outline how AI may be considered during inspections, it's important that schools are prepared to demonstrate both responsible implementation and clear thinking.

## How Ofsted is approaching AI during inspections

At the time of writing, Ofsted does not assess AI use as a standalone element of an inspection. However, recent guidance sets out how inspectors may encounter and evaluate the use of AI during broader inspections. This guidance is informed in part by [Ofsted's research with early adopters of AI in education settings](#).

Here are three key points schools should understand:

### 1. AI will not be inspected in isolation

There is no dedicated section of the inspection focused on AI. Inspectors won't actively seek out evidence of AI use, and AI will not be referenced in reports unless it plays a crucial role in wider inspection decisions.

Instead, AI use will be considered within existing frameworks—such as safeguarding, curriculum planning or data privacy. For example, if AI is used in lesson planning or student support, inspectors may explore how it aligns with the setting's broader strategic goals or safeguarding practices.

### 2. AI use must serve the best interests of learners

Where AI is in use, inspectors will expect it to be used in a way that benefits students and aligns with the setting's mission and values. They may consider:

- How AI tools are used, and the impact on students and staff
- Whether the school has made thoughtful, evidence-based decisions around AI use
- How settings respond to the misuse of AI by students, staff or parents
- How inappropriate or harmful content is managed, where relevant.

### 3. Ofsted's position is still developing

This is not a final stance. Ofsted acknowledges that it does not yet have enough evidence to define what “good” looks like in terms of AI use for inspection purposes. As more case studies and practical insights emerge, this position will continue to evolve.

For now, settings are encouraged to use AI responsibly, document their decision-making processes, and stay alert to updated guidance.

## Example Questions Inspectors May Ask

While not exhaustive, the following questions reflect Ofsted's current thinking. Schools and colleges should be prepared to respond to questions such as:

- What AI tools or platforms are currently in use across your setting?
- How does your use of AI align with your wider strategic or pedagogical goals?
- What safeguards are in place to protect students from potentially harmful or inappropriate AI-generated content?
- Have you carried out Data Protection Impact Assessments (DPIAs) for tools processing personal data?
- What AI-specific policies do you have in place?
- Have you updated existing policies accordingly?
- How do you assess the impact of AI on both staff and learners?
- What informed your decision to implement AI in specific areas (e.g. marking, admin, lesson planning)?
- If AI has been used in sensitive contexts - for example, summarising a child protection conference - how is the accuracy and confidentiality of that information ensured?

# The Steps to Navigating the AI Wilderness



AI is reshaping how we think, learn and communicate. For schools, colleges and MATs, the pace of change brings new opportunities, but also new responsibilities.

This guide has walked you through the practical realities of AI in education - from understanding what generative AI is, to forming effective AI policies that protect students and staff, to using filtering and monitoring tools that help manage risk in real-time. We've explored how Ofsted is approaching AI in inspections, and the ways education settings can respond with clarity and confidence.

Across it all, one thing remains clear: **there's no single solution** to managing AI in education. Instead, success lies in a layered, integrated approach - built on clear policies, smart technology, informed staff, and a safeguarding-first mindset.

### Here's what that looks like in practice:

1. Start with clear policies. Know which tools are allowed, who can use them, and for what purpose.
2. Ensure all stakeholders receive relevant training and guidance in line with the policies.
3. Support your policies with technology. Use real-time filtering to control access and prevent exposure to harmful AI content.
4. Stay informed of digital behaviours. Use monitoring to see how students are actually interacting with AI - and step in when needed.
5. Think beyond compliance. Make decisions based on what's right for your learners, and keep evolving your approach as the technology (and the risks) develop.

AI is already here. But with the right foundations in place, you don't need to fear the wilderness - you can lead others through it.



# Glossary of AI-related Terms

<b>Bias (in AI)</b>	When an AI system produces unfair, discriminatory or unbalanced results due to biased or incomplete training data. A key safeguarding and inclusion concern.
<b>Chatbot</b>	A software tool that simulates a conversation with a human. Used in schools for automated FAQs, student support, or revision help.
<b>Computer Vision</b>	AI that enables machines to “see” and interpret images or videos. Commonly used in facial recognition or visual learning tools.
<b>Deepfake</b>	AI-generated video or audio that convincingly mimics real people. A growing safeguarding risk due to the potential for deception and misuse.
<b>Ethical AI</b>	Creating and using AI systems in a fair, transparent, and accountable way. Essential in education to avoid harm and ensure trust.
<b>Generative AI</b>	AI that creates new content - like text, images, or music - based on patterns it has learned.
<b>Hallucination (in AI)</b>	When an AI generates false or misleading content that sounds believable.
<b>Large Language Model (LLM)</b>	A type of AI trained on vast amounts of text to produce human-like responses. Examples include ChatGPT and Gemini, often used in education for writing and planning.
<b>Machine Learning (ML)</b>	A form of AI that enables systems to improve at tasks by learning from data, without needing to be manually programmed for every action.
<b>Natural Language Processing (NLP)</b>	The part of AI that deals with understanding and interpreting human language. It powers tools like speech-to-text, chatbots, and grammar checkers.
<b>Prompt</b>	The instruction or question you give to an AI tool. Well-written prompts usually result in better, more accurate responses.
<b>Training Data</b>	The data used to teach an AI system how to perform tasks. If the data is flawed, the AI can learn incorrect or biased behaviour.

# Talk to us

Smoothwall is equipping 1,000s of UK schools, colleges and MATs with the tools, knowledge and resources they need to overcome the student digital safety challenges AI can bring.

Contact us at [enquiries@smoothwall.com](mailto:enquiries@smoothwall.com) if you would like to know more about our 100% real-time web filtering, human-moderated digital monitoring, or AI-related training.

**smoothwall®**  
by Qoria

Smoothwall is the leading provider of digital safeguarding solutions in UK education. For more information, visit our website or get in touch with our team of experts.

**Web:** [www.smoothwall.com](http://www.smoothwall.com)

**Tel:** +44 (0)800 047 8191

**Email:** [enquiries@smoothwall.com](mailto:enquiries@smoothwall.com)

**Qoria**

Smoothwall is part of Qoria, a global technology company, dedicated to keeping children safe and well in their digital lives. We harness the power of connection to close the gaps that children fall through, and to seamlessly support them on all sides - at school, at home and everywhere in between.

Find out more  
[www.qoria.com](http://www.qoria.com)