Department for Education

Guidance

Generative AI: product safety expectations

Published 22 January 2025

Applies to England

Contents

These expectations outline the capabilities and features that generative artificial intelligence (AI) products and systems should meet to be considered safe for users in educational settings. They are are mainly intended for edtech developers and suppliers to schools and colleges.

Some expectations will need to be met further up the supply chain, but responsibility for assuring this will lie with the systems and tools working directly with schools and colleges.

The expectations are focused on outcomes but do not prescribe specific approaches for meeting those outcomes.

Filtering

> This information is relevant to child-facing products.

Generative AI products used for filtering must effectively and reliably prevent access to harmful and inappropriate content by users. This may be achieved by:

- integrating the highest standards of filtering possible within the product
- using additional filtering solutions that work on top of an AI product

Our expectations

We expect that:

- users are effectively and reliably prevented from generating or accessing harmful and inappropriate content
- filtering standards are maintained effectively throughout the duration of a conversation or interaction with a user
- filtering will be adjusted based on different levels of risk, age, appropriateness and the user's needs - for example users with special educational needs and disabilities (SEND)
- multimodal content is effectively moderated, including detecting and filtering prohibited content across multiple languages, images, common misspellings and abbreviations
- full content moderation capabilities are maintained regardless of the device used, including bring your own device (BYOD) and smartphones when accessing products via an educational institutional account
- content is moderated based on an appropriate contextual understanding of the conversation, ensuring that generated content is sensitive to the context
- filtering should be updated in response to new or emerging types of harmful content

Relevant regulation

Using these expectations could help schools and colleges comply with:

- Keeping children safe in education (particularly parts 1, 2 and 5)
- Filtering and monitoring standards for schools and colleges
- Public Sector Equality Duty: guidance for public authorities

Using these expectations could help edtech developers or suppliers comply with the Online Safety Act 2023 which defines what content is considered to be "content that is harmful to children".

Monitoring and reporting

> This information is relevant to child-facing products.

The generative AI product must maintain robust activity logging procedures. This function may be integrated into an AI tool or be provided by an additional solution working on top of an AI product.

This includes:

- recording input prompts and responses
- analysing performance metrics
- alerting local supervisors when harmful and inappropriate content is accessed or attempted to be accessed

Our expectations

We expect products to:

- identify and alert local supervisors to harmful or inappropriate content being searched for or accessed
- alert and signpost the user to appropriate guidance and support resources when access of prohibited content is attempted (or succeeds)
- generate a real-time user notification in age-appropriate language when harmful or inappropriate content has been blocked, explaining why this has happened
- identify and alert local supervisors of potential safeguarding disclosures made by users
- generate reports and trends on access and attempted access of prohibited content, in a format that non-expert staff can understand and which does not add too much burden on local supervisors

Relevant regulation

Meeting these expectations could help schools and colleges to comply with:

- Keeping children safe in education part 1 - this emphasises the importance of safeguarding responsibilities, including preventing access to harmful content
- the filtering and monitoring standards for schools and colleges

Meeting these expectations could help edtech developers or suppliers comply with the:

- General Data Protection Regulation (GDPR) Article 35 - this requires a Data Protection Impact Assessment for high-risk data processing, which could include monitoring for harmful content
- Information Commissioner's Office (ICO) age appropriate design code - section 11 refers to monitoring, and that children should be clearly told if they are being tracked or monitored - the code can apply to providers of edtech services used in school environments

Security

> This information is relevant to child and teacher-facing products

The generative AI product must be secured against malicious use or exposure to harm. This includes prioritising the technical objectives of:

- reliability
- security
- robustness
- ensuring safe operation under various conditions, including unexpected changes and adversarial attacks

Our expectations

We expect products to:

- offer robust protection against 'jailbreaking' by users trying to access prohibited material
- offer robust measures to prevent unauthorised modifications to the product that could reprogram the product's functionalities
- allow administrators to set different permission levels for different users
- ensure regular bug fixes and updates are promptly implemented
- sufficiently test new versions or models of the product to ensure safety compliance before release
- have robust password protection or authentication methods
- be compatible with the Cyber Security Standards for Schools and Colleges

Relevant regulation

Using these expectations could help schools and colleges comply with the Cyber Security Standards for Schools and Colleges.

This requires that all software used by schools and colleges should:

- be regularly patched with the latest security updates
- have multi-factor authentication to protect accounts with access to sensitive or personal operational data
- ensure that users should only have access to accounts that are relevant to their respective roles

Schools and colleges should also comply with the Computer Misuse Act 1990 which sets out criminal offences related to unauthorised access and modifications to computer material, which could include users reprogramming product functionalities.

Privacy and data protection

> This information is relevant to child and teacher-facing products.

The generative AI product must have a robust approach to data handling and transparency around the processing of personal data, including:

- compliance with all relevant data protection legislation and regulations
- ensuring a lawful basis for data collection

These expectations have been developed to be compatible with these standards, although there are additional expectations focused on generative AI as the technology presents new and distinct risks to users.

Our expectations

We expect products to provide a clear and comprehensive privacy notice which is presented at regular intervals in age-appropriate formats and language with information on:

- the type of data - why and how this is collected, processed, stored and shared by the generative AI system
-

where data will be processed, and whether there are appropriate safeguards in place if this is outside the UK or EU
- the relevant legislative framework that authorises the collection and use of data

We also expect products to:

- conduct Data Protection Impact Assessment (DPIA) during the generative AI tool's development and during the full life cycle of the tool
- allow all parties to fulfil their data controller and processor responsibilities proportionate to the volume, variety and usage of the data they process and without overburdening the other
- comply with all relevant data protection legislation and ICO codes and standards, including the ICO's age appropriate design code if they process personal data
- not collect, store, share or use personal data for any commercial purposes, including further model training and fine-tuning, without confirmation of appropriate lawful basis

Relevant regulation

Meeting these expectations could help edtech developers or suppliers comply with the Data Protection Act 2018 and The UK GDPR . These regulations require data controllers to provide a privacy notice that:

- is written in simple language that can be understood by data subjects, including children
- contains details around data collection, processing, storage, sharing practices, and data subjects' information rights, for example the right to erasure
- complies with the ICO's current position that legitimate interest is likely to be the relevant lawful basis for the processing of personal data in generative AI products used by children

Edtech developers or suppliers should also make sure they are compliant with:

- UK GDPR, which covers DPIAs, the sharing of personal data for training purposes, data controller responsibilities, and organisational data access
- Schedule 1, Part 2, Paragraph 18 of the Data Protection Act 2018, which covers data processing in the context of the safeguarding of children
- ICO's Children's code, which sets out how to ensure that online services appropriately safeguard children's personal data
- ICO's guidance on AI and data protection, which provides broad guidance, including a data protection risk assessment toolkit

ICO is currently conducting a consultation series into generative AI and data protection which may result in further guidance being released.

Intellectual property

> This information is relevant to child and teacher-facing products.

The generative AI product must not store or collect intellectual property created by pupils or the copyright owner for any commercial purposes, such as training or fine tuning of models, unless consented to by the:

- copyright owner
- copyright owner's parent or guardian - where the copyright owner is deemed a minor and therefore unable to consent

Our expectations

We expect that unless there is permission from the copyright owner, inputs and outputs should not be:

- collected
- stored
- shared for any commercial purposes, including (but not limited to) further model training (including fine-tuning), product improvement, and product development

In the case of children that are under the age of 18, it is best practice to obtain permission from the parent or guardian. In the case of teachers, this is likely to be their employer - assuming they created the work in the course of their employment.

Relevant regulation

Meeting these expectations could help edtech developers or suppliers comply with the Copyright, Designs and Patents Act 1988.

This stipulates that a creator of an original work owns the copyright of that work, and has exclusive rights to the use of that work (other than in specific circumstances where copyright exceptions apply).

Design and testing

> This information is relevant to child and teacher-facing products.

The generative AI product must prioritise transparency and children's safety in its design.

In the case of child-facing products, this includes:

- implementing technical and operational mitigations for identified risks
- ensuring child-centred design and operation
- conducting testing with stakeholders, including children, to ensure safety

This may apply to safety features integrated into an AI product, or a separate safety layer.

Our expectations

We expect that:

- sufficient testing with a diverse and realistic range of potential users and use cases is completed
- sufficient testing of new versions or models of the product to ensure safety compliance before release is completed
- the product should consistently perform as intended

Relevant regulation

Meeting these expectations could help schools and colleges comply with The Public Sector Equality Duty (PSED).

Schools and colleges are required to have due regard to the PSED when making decisions and developing policies. This includes the need to eliminate discrimination, harassment, victimisation and other conduct prohibited under the Equality Act 2010.

Meeting these expectations could help edtech developers or suppliers to comply with:

- ICO's Children's code, section 2, which recommends that developers conduct user testing as part of a DPIA to get feedback on children's ability to understand how their data is being used
- the Equality Act 2010, which mandates that services do not discriminate against any of the protected characteristics

- ICO's guidance on AI and data protection, which addresses the need for AI to avoid bias and discrimination, ensuring fairness
- UK GDPR (Articles 13(2)(f) and 14(2)(g), which requires data controllers to provide data subjects with information about the existence of automated decision-making, including:
  - profiling and meaningful information about the logic involved
  - the significance and envisaged consequences of such processing for the data subject
- The General Product Safety Regulations 2005, which states that products must be safe in their normal or reasonably foreseeable usage - find out more from the study on the impact of artificial intelligence on product safety

Governance

> This information is relevant to child and teacher-facing products.

The generative AI product must be operated with accountability. This includes:

- carrying out risk assessments
- instigating formal mechanisms for lodging complaints
- demonstrating that its operations, decision-making processes and data handling practices are understandable and accessible to government agencies and users

Our expectations

We expect that:

- a clear risk assessment will be conducted for the product to assure safety for educational use
- a formal complaints mechanism will be in place, addressing how safety issues with the software can be escalated and resolved in a timely fashion
- policies and processes governing AI safety decisions are made available

Relevant regulation

Meeting these expectations could help edtech developers or suppliers comply with:

- UK GDPR (Articles 77-79) and the Data Protection Act 2018 (s165-166), which cover the right to lodge complaints with a supervisory authority in relation to the processing of personal data
- the ICO's Children's code, which requires the implementation of an accountability programme, including policies to support and demonstrate compliance with data protection legislation

Back to top

Help us improve GOV.UK

To help us improve GOV.UK, we'd like to know more about your visit today. Please fill in this survey (opens in a new tab).

Services and information

Benefits

Births, death, marriages and care

Business and self-employed

Childcare and parenting

Citizenship and living in the UK

Crime, justice and the law

Disabled people

Driving and transport

Education and learning

Employing people

Environment and countryside

Housing and local services

Money and tax

Government activity

Departments

News

Guidance and regulation

Research and statistics

Policy papers and consultations

Transparency

How government works

Get involved

Passports, travel and living abroad

Visas and immigration

Working, jobs and pensions

Help Privacy Cookies Accessibility statement Contact
Terms and conditions Rhestr o Wasanaethau Cymraeg
Government Digital Service

All content is available under the Open Government Licence v3.0, except where otherwise stated

© Crown copyright