Appropriate Filtering for Education settings

May 2025

Filtering Provider Checklist Reponses

Schools (and registered childcare providers) in England and Wales are required "to ensure children are safe from terrorist and extremist material when accessing the internet in school, including by establishing appropriate levels of filtering". Furthermore, it expects that they "assess the risk of [their] children being drawn into terrorism, including support for extremist ideas that are part of terrorist ideology". There are a number of self review systems (eg www.360safe.org.uk) that will support a school in assessing their wider online safety policy and practice.

The Department for Education's statutory guidance 'Keeping Children Safe in Education' obliges schools and colleges in England to "*ensure appropriate filters and appropriate monitoring systems are in place and regularly review their effectiveness" and they "should be doing all that they reasonably can to limit children's exposure to [Content, Contact, Conduct, Contract] risks from the school's or college's IT system*" however, schools will need to "*be careful that "over blocking" does not lead to unreasonable restrictions as to what children can be taught with regards to online teaching and safeguarding*."

By completing all fields and returning to UK Safer Internet Centre (enquiries@saferinternet.org.uk), the aim of this document is to help filtering providers to illustrate to education settings (including Early years, schools and FE) how their particular technology system(s) meets the national defined 'appropriate filtering standards. Fully completed forms will be hosted on the UK Safer Internet Centre website alongside the definitions

It is important to recognise that no filtering systems can be 100% effective and need to be supported with good teaching and learning practice and effective supervision.

Company / Organisation	TrustLayer
Address	Belvedere House 4.2, Basing View, Basingstoke, RG21 4HG
Contact details	support@trustlayer.co.uk 08452309590
Filtering System	TrustLayer CloudUSS (Web & Cloud Security)
Date of assessment	5 September 2025

System Rating response

Where a supplier is able to confirm that their service fully meets the issue identified in a specific checklist the appropriate self-certification colour for that question is GREEN.	
Where a supplier is not able to confirm that their service fully meets the issue identified in a specific checklist question the appropriate self-certification colour for that question is AMBER.	

.

Illegal Online Content

Filtering providers should ensure that access to illegal content is blocked, specifically that the filtering providers:

Aspect	Rating	Explanation
• Are IWF members		Indirectly, our threat intel partner for URL classification (zvelo) is a long-term member.
• and block access to illegal Child Abuse Images (by actively implementing the IWF URL list), including frequency of URL list update		
• Integrate the 'the police assessed list of unlawful terrorist content, produced on behalf of the Home Office'		
• Confirm that filters for illegal content cannot be disabled by anyone at the school (including any system administrator).		

Describing how, their system manages the following illegal content

Content	Explanatory notes – Content that:	Rating	Explanation
child sexual abuse	Content that depicts or promotes sexual abuse or exploitation of children, which is strictly prohibited and subject to severe legal penalties.		The domain and URL monitoring service contains over 500 web categories and granular sub-categories. Top level categories are listed here: https://help.clouduss.com/product-web-security/web-categories-list Custom lists, keywords, and RegEx filters can also be defined. ML algorithms analyse 'unclassified' or new URLs in real-time (default blocked).
controlling or coercive behaviour	Online actions that involve psychological abuse, manipulation, or intimidation to control another individual, often occurring in domestic contexts.		
extreme sexual violence	Content that graphically depicts acts of severe sexual violence, intended to shock or incite similar behaviour, and is illegal under UK law.		
extreme pornography	Pornographic material portraying acts that threaten a person's life or could result in serious injury, and is deemed obscene and unlawful.		
fraud	Deceptive practices conducted online with the intent to secure unfair or unlawful financial gain, including phishing and scam activities.		
racially or religiously aggravated public order offences	Content that incites hatred or violence against individuals based on race or religion, undermining public safety and cohesion.		

inciting violence	Online material that encourages or glorifies acts of violence, posing significant risks to public safety and order.		
illegal immigration and people smuggling	Content that promotes or facilitates unauthorized entry into a country, including services offering illegal transportation or documentation.		
promoting or facilitating suicide	Material that encourages or assists individuals in committing suicide, posing serious risks to vulnerable populations.		
intimate image abuse	The non-consensual sharing of private sexual images or videos, commonly known as "revenge porn," intended to cause distress or harm.		
selling illegal drugs or weapons	Online activities involving the advertisement or sale of prohibited substances or firearms, contravening legal regulations.		
sexual exploitation	Content that involves taking advantage of individuals sexually for personal gain or profit, including trafficking and forced prostitution.		
Terrorism	Material that promotes, incites, or instructs on terrorist activities, aiming to radicalise individuals or coordinate acts of terror.		

Inappropriate Online Content

Recognising that no filter can guarantee to be 100% effective, providers should both confirm, and describe how, their system manages the following content

Content	Explanatory notes – Content that:	Rating	Explanation
Gambling	Enables gambling		The domain and URL monitoring service contains over 500 web categories and granular sub-categories. Top level categories are listed here: https://help.clouduss.com/product-web-security/web-categories-list Custom lists, keywords, and RegEx filters can also be defined.
Hate speech / Discriminiation	Content that expresses hate or encourages violence towards a person or group based on something such as disability, race, religion, sex, or sexual orientation. Promotes the unjust or prejudicial treatment of people with protected characteristics of the Equality Act 2010		
Harmful content	Content that is bullying, abusive or hateful. Content which depicts or encourages serious violence or		

	injury. Content which encourages dangerous stunts and challenges; including the ingestion, inhalation or exposure to harmful substances.		ML algorithms analyse 'unclassified' or new URLs in real-time (default blocked).
Malware / Hacking	promotes the compromising of systems including anonymous browsing and other filter bypass tools as well as sites hosting malicious content		
Mis / Dis Information	Promotes or spreads false or misleading information intended to deceive, manipulate, or harm, including content undermining trust in factual information or institutions		
Piracy and copyright theft	includes illegal provision of copyrighted material		
Pornography	displays sexual acts or explicit images		
Self Harm and eating disorders	content that encourages, promotes, or provides instructions for self harm, eating disorders or suicide		
Violence Against Women and Girls (VAWG)	Promotes or glorifies violence, abuse, coercion, or harmful stereotypes targeting women and girls, including content that normalises gender-based violence or perpetuates misogyny.		

This list should not be considered an exhaustive list. Please outline how the system manages this content and many other aspects

Highly granular rules and policies can be curated to create powerful filtering on domain, URL, category etc. Unclassified content is analysed autonomously and web classifications updated in real-time. The Cloud Application module extends this to full http/https interception and analysis to track activity inside cloud-based SaaS applications, including AI tools.

Regarding the duration and extent of logfile (Internet history) data retention, providers should outline their retention policy, specifically including the extent to the identification of individuals and the duration to which all data is retained.

Log data is stored for 30 days and is queryable in the cloud-based dashboard (90 days retained for backup and DR purposes offline). This is encrypted at rest with AES-256 in a regional database location of the customer's choosing. Log data can be exported on a regular pre-defined cadence via csv/xls for customer retention. Log Streaming is available for integration with SIEM/SOAR systems.

Providers should be clear how their system does not over block access so it does not lead to unreasonable restrictions

Policies can display Warning pages with customer-defined content instead of Block pages. Any category deemed borderline can be logged separately whereby the user has to click a consent/acknowledgement button to proceed to the site. SaaS applications can be controlled with a Risk rating based on individual actions within the service rather than domain or URL-level. Customers can maintain a bypass list of approved services (i.e. intranets)

Filtering System Features

How does the filtering system meet the following principles:

Principle	Rating	Explanation
• Context appropriate differentiated filtering, based on age, vulnerability and risk of harm – also includes the ability to vary filtering strength appropriate for staff		Policies can be deployed based on numerous attributes including device and user characteristics. Web requests can be overridden on an ad-hoc basis by staff and policies tuned per individual if required.
• Circumvention – the extent and ability to identify and manage technologies and techniques used to circumvent the system, specifically VPN, proxy services, DNS over HTTPS and ECH.		The agent is deployed as an OS service and is tamper-proof. All http/https traffic is inspected by the system driver so protection is still delivered over VPN or proxy services. TLS 1.3 is supported.
• Control – has the ability and ease of use that allows schools to control the filter themselves to permit or deny access to specific content. Any changes to the filter system are logged enabling an audit trail that ensure transparency and that individuals are not able to make unilateral changes		Drag-and-drop based policy engine allows quick and easy customisation of rules and policies. An immutable audit trail is logged for every system admin's activity.
• Contextual Content Filters – in addition to URL or IP based filtering, Schools should understand the extent to which (http and https) content is dynamically analysed as it is streamed to the user and blocked. This would include AI or user generated content, for example, being able to contextually analyse text and dynamically filter the content produced (for example ChatGPT). For schools' strategy or policy that allows the use of AI or user generated content, understanding the technical limitations of the system, such as whether it supports real-time filtering, is important.		All http and https protocol requests and responses are analysed (GET, PUT, POST etc.) in detail and policies can be tailored for content detection. Interactions with AI SaaS applications can be controlled and monitored at an individual action level of granularity (i.e. restrict prompts, file uploads/downloads etc.)
• Deployment – filtering systems can be deployed in a variety (and combination) of ways (eg on device, network level, cloud, DNS). Providers should describe how their systems are deployed alongside any required configurations		Hybrid deployment options exist via device agent, local gateway, cloud gateway and hosted mobile proxy. Agents are compatible with 3rd party MDR solutions and endpoint AV agents. Local gateways can be configured as captive portal or transparent proxy.

• Filtering Policy – the filtering provider publishes a rationale that details their approach to filtering with classification and categorisation as well as how the system addresses over blocking		All categories of destination resource is detailed in the log and the associated classification. These can be manually queried in the portal and disputed if required. Policies can be highly tuned to prevent false positives, and SaaS applications can be managed at an action-level o granularity rather than just allow/block at page/domain level.
• Group / Multi-site Management – the ability for deployment of central policy and central oversight or dashboard		Platform is fully multi-tenant and integrates with Entra/AzureAD, Google Workspace etc. for org/site/department/user management.
• Identification - the filtering system should have the ability to identify users and devices to attribute access (particularly for mobile devices) and allow the application of appropriate configurations and restrictions for individual users. This would ensure safer and more personalised filtering experiences.		Detailed attributes logged via agent, can filter on user, device, browser, IP, and more. Policies can be applied at any level
• Mobile and App content – mobile and app content is often delivered in entirely different mechanisms from that delivered through a traditional web browser. To what extent does the filter system block inappropriate content via mobile and app technologies (beyond typical web browser delivered content). Providers should be clear about the capability of their filtering system to manage content on mobile and web apps and any configuration or component requirements to achieve this		Certain native SaaS applications are supported via an API mode of scanning rather than inline. Only http/https traffic is inspected, however individual OS processes can also be captured. Apps that use SSL pinning cannot be tracked (e.g. WhatsApp)
• Multiple language support – the ability for the system to manage relevant languages		Threat Intel and Classification databases are global
• Remote devices – with many children and staff working remotely, the ability for school owned devices to receive the same or equivalent filtering to that provided in school		Devices with agent installed have same level of protection wherever they are located.
• Reporting mechanism – the ability to report inappropriate content for access or blocking		Mis-categorised sites or disputes can be reported manually in the portal. Reclassification and manual

		inspection is done within 24h, often minutes.
• Reports – the system offers clear granular historical information on the websites users have accessed or attempted to access		Highly granular reports and forensic level analytics are available
• Safe Search – the ability to enforce 'safe search' when using search engines		Safe Search by default.
• Safeguarding case management integration – the ability to integrate with school safeguarding and wellbeing systems to better understand context of activity		Logs can be streamed to 3rd party SIEM systems or simple S3 buckets if required. Other tools supported on ad hoc basis depending on demand

How does your filtering system manage access to Generative AI technologies (e.g. ChatGPT, image generators, writing assistants)?

In your response, please describe whether and how your system identifies, categorises, or blocks Generative AI tools; how access can be controlled based on age, risk, or educational need; any limitations in filtering AI-generated content—particularly where such content is embedded within other platforms or applications; and what support or configuration guidance you offer to schools to help them align with the UK Safer Internet Centre's Appropriate Filtering Definitions and relevant national safeguarding frameworks.

Any interaction with a SaaS application via a web browser can be controlled in a granular fashion. We have a catalogue containing thousands of applications and dozens of AI tools. Apps are fingerprinted and classified via our automated profiling engines and manual verification – this means individual actions within Gen-AI tools can be controlled (such as limiting file uploads or specific prompt types). Each action within a SaaS application is allocated a baseline 'Risk' attribute which is displayed to the system admin and can be overridden based on individual preferences.

Filtering systems are only ever a tool in helping to safeguard children when online and schools have an obligation to "*consider how children may be taught about safeguarding, including online, through teaching and learning opportunities, as part of providing a broad and balanced curriculum*".[1]

Please note below opportunities to support schools (and other settings) in this regard

TrustLayer also offers a Security Awareness Training platform to educate users on safe internet usage, appropriate behaviour and digital safety (i.e. social media and social engineering)

[1] https://www.gov.uk/government/publications/keeping-children-safe-in-education--2

PROVIDER SELF-CERTIFICATION DECLARATION

In order that schools can be confident regarding the accuracy of the self-certification statements, the supplier confirms:

- that their self-certification responses have been fully and accurately completed by a person or persons who are competent in the relevant fields
- that they will update their self-certification responses promptly when changes to the service or its terms and conditions would result in their existing compliance statement no longer being accurate or complete
- that they will provide any additional information or clarification sought as part of the self-certification process
- that if at any time, the UK Safer Internet Centre is of the view that any element or elements of a provider's self-certification responses require independent verification, they will agree to that independent verification, supply all necessary clarification requested, meet the associated verification costs, or withdraw their self-certification submission.

Name	Gareth Lockwood
Position	CTO
Date	5/9/25
Signature