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	Netsweeper UK
Address	Suite 125-126 Pure Offices, 4100 Park Approach Thorpe Park, Leeds United Kingdom LS15 8GB
Contact details	Nick Levey
Filtering System	Netsweeper
Date of assessment	11/08/2025

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		We have been a member of the IWF since 2006. Netsweeper not only uses the list to identify CSAM content, due to Netsweeper wide usage world-wide when new content is identified we share this with IWF to bring down new content found my Netsweeper.
• and block access to illegal Child Abuse Images (by actively implementing the IWF URL list), including frequency of URL list update		The IWF list is used by Netsweeper and Netsweeper list is updated in real time for our end users ensuring all new content found by the IWF is blocked.
• Integrate the 'the police assessed list of unlawful terrorist content, produced on behalf of the Home Office'		The CTIRU (Counter Terrorism Internet Referral Unit) list is used actively by Netsweeper across all its customers.
• Confirm that filters for illegal content cannot be disabled by anyone at the school (including any system administrator).		Netsweeper locks down all illegal content with no ability to disable.

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.		Netsweeper has its own CSAM (Child sexual Abuse Material) category within the platform which it uses to actively block this harmful content. On top of this Netsweeper pulls URLS directly from IWF and Project Arachnid.
controlling or coercive behaviour	Online actions that involve psychological abuse, manipulation, or intimidation to control another individual, often occurring in domestic contexts.		Netsweeper covers this area with multiple categories to cover content related to Controlling and coercive behaviour.

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.		Netsweeper categorisation allows multiple categorises to be used against URLs meaning that Extreme Sexual Violence would be cover by the following Categorises – Violence, Extreme and Adult content. This granular nature ensures URLs do not get miscategorised due to ridged categorisation. It also ensures that DSLs are aware of the type of content that has been accessed when reporting.
extreme pornography	Pornographic material portraying acts that threaten a person's life or could result in serious injury, and is deemed obscene and unlawful.		Netsweeper categorisation allows multiple categorises to be used against URLs meaning that Extreme Pornography would be cover by the following Categorises – Extreme and Pornography. This granular nature ensure URLs do not get miscategorised due to ridged categorisation. It also ensures that DSLs are aware of the type of content that has been accessed when reporting.
fraud	Deceptive practices conducted online with the intent to secure unfair or unlawful financial gain, including phishing and scam activities.		Netsweeper's Scam Category cover all scam and fraud related content. Netsweeper also works with organisations like Scam advisors to block known Fraudulent activity.
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.		Netsweeper has multiple categories to cover this subject matter including Profanity and hate speech. Netsweeper would label the URL based on its specific type of offence.
inciting violence	Online material that encourages or glorifies acts of violence, posing significant risks to public safety and order.		Netsweeper categorisation allows multiple categorises to be used against URLs meaning that inciting violence would be cover by the following Categorises – Extreme, Hate speech and violence. This granular nature ensure URLs do not get miscategorised due to ridged categorisation. It also

			ensures that DSLs are aware of the type of content that has been accessed when reporting.
illegal immigration and people smuggling	Content that promotes or facilitates unauthorized entry into a country, including services offering illegal transportation or documentation.		Netsweeper Criminal skills category would capture all content related to both illegal immigration and people smuggling
promoting or facilitating suicide	Material that encourages or assists individuals in committing suicide, posing serious risks to vulnerable populations.		Netsweeper Self harm and suicide category covers all content related to this matter. Netsweeper also works with leading Suicide Prevention charity “R;pple” to keep on top of all new risk and word used by users looking at suicide ideation content.
intimate image abuse	The non-consensual sharing of private sexual images or videos, commonly known as "revenge porn," intended to cause distress or harm.		Netsweeper can block sexual content via its Adult and pornography categories. On top of this, methods of distribution like personal mail box, communication tools (outside schools control) or storage tools are covered by Web email, Web storage and web chat.
selling illegal drugs or weapons	Online activities involving the advertisement or sale of prohibited substances or firearms, contravening legal regulations.		Netsweeper’s Substance abuse, marijuana, weapons, criminal skills and sale would categorise this content. As an example, an online storing selling drugs would be categorised as – Substance abuse, Criminal Skills and Sales due to the nature of the content. This level of detail ensure DSL are aware of the content accessed by their end users.
sexual exploitation	Content that involves taking advantage of individuals sexually for personal gain or profit, including trafficking and forced prostitution.		Netsweeper following categories would cover this – Adult, Pornography and criminal skills.
Terrorism	Material that promotes, incites, or instructs on terrorist activities, aiming to radicalise individuals or coordinate acts of terror.		Netsweeper has a dedicated Terrorism category for all terrorism related content.

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		Netsweeper Gambling category identifies all known and illegal gambling websites. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using Netsweeper AI engine. This method ensures one of the most accurate categorisation engine on the market.
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		Netsweeper Hate Speech category identifies content related to hate speech and discrimination content. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using Netsweeper AI engine. This method ensures one of the most accurate categorisation engines on the market.
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.		Netsweeper's *Bullying,* Violence and extreme category identify content related to harmful content. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using

			Netsweeper AI engine. This method ensures one of the most accurate categorisation engine on the market.
Malware / Hacking	promotes the compromising of systems including anonymous browsing and other filter bypass tools as well as sites hosting malicious content		Netsweeper has an extensive malicious content database used by our customers. Covered by Netsweeper categories Malware, Phishing, Virus, Infected Host, and on top this a range of some of the top leading security vendors list that we managed and maintain to safeguard our customers networks.
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		Netsweeper News category can be used to block website that are known to promote mis and dis information. This walled garden approach means that trusted new authorities can be allowed on the school network.
Piracy and copyright theft	includes illegal provision of copyrighted material		Netsweeper Copyright Infringement Category blocks content contains sites that use, provide, or distribute information on copyrighted intellectual property or illicitly copied material which violates the owners' rights. On top of this Netsweeper uses the PIPCU (Police intellectual Property Crime Unit) list which guards against Piracy and copyright offences.
Pornography	displays sexual acts or explicit images		Netsweeper's Pornography category identifies content related to harmful content. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using Netsweeper AI engine. This method ensure one of the most

			accurate categorisation engine on the market.
Self Harm and eating disorders	content that encourages, promotes, or provides instructions for self harm, eating disorders or suicide		Netsweeper's Self harm category identifies content related to harmful content. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using Netsweeper AI engine. This method ensures one of the most accurate categorisation engine on the market.
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.		Netsweeper's Hate Speech, Violence and Criminal Skills categories identify content related to harmful content. Using Netsweeper's real time categorisation engine (used by Netsweeper for over 25 years) Netsweeper analysis the content on the pages (Text, Images, Links to other URLS and downloadable) in 47 different languages using Netsweeper AI engine. This method ensure one of the most accurate categorisation engine on the market.

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

The Netsweeper collective community numbers are over billion devices worldwide. This collective, together with our AI technology and human oversight defines URL classification. Netsweeper publishes classification of filtering and categorises on the Netsweeper website as well as a view in real time of new content categorised.
Netsweeper's core competency is using our patented techniques to categorise every URL that passes through our deployed systems. Netsweeper is both real-time and employs a hierarchy of data (URL-to-category), with our Category Naming Service (CNS) as a global-master database.
If any customer anywhere in the world accesses a URL, that URL is submitted to the local policy server, if that policy server cannot find a category match, it is automatically submitted to the CNS and looked up there. If the CNS already has the category mapping it is immediately returned to the local system and cached there for future use, a policy decision is then made by the policy server.
If neither the local system, nor the CNS has a category match, the URL is submitted to our

"Artificial Intelligence" system that will interrogate the content at-and-around that URL, assess the content, detect if it references or contains malware, and assigns one or more categories to the URL into the CNS and then back to the local system, a policy decision is then made by the policy server.
The CNS allows us to adapt to trending URLs immediately due to its world-wide scope. If the local system hasn't seen a particular URL yet, then CNS probably has. If the URL has been assigned one or more categories, local systems see immediate responses (sub-second). If the URL is truly "new" then the AI will typically process the content within 4-5 seconds. The local policy servers can be configured with techniques to minimise the "new URL" wait.

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.

Netsweeper typically retains logs on our deployment for 12 months, however schools, should they wish can choose to have longer retention.

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

Netsweeper policy manager offers granular controls around categorisation. School leaders can enforce the policies based on the guidance set out by the DFE and the Uk safer internet centre. Netsweeper Realtime categorisation ensures the ever-changing landscape of the internet is kept on top of by Netsweeper reviewing content on a regular basis. This ensures categories change when the content changes. Netsweeper’s ability to label URLs with multiple categories ensures Netsweeper can be more specific on categorisation ensuring schools don’t need to over block content due to the risk associated with non-granular categorisation.

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		Netsweeper allows for differentiated filtering based on different user types. For example, students may receive differentiated filtering based on age. Vulnerable users may have certain categories blocked that are not blocked for other uses
• 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.		Netsweeper used advanced technology to detect and prevent attempts to circumvent the system.
• 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		All changes made to the system is logged and auditable by the school. The schools are able to administer their own content filtering.
• 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.		Netsweeper provides on the fly categorisation based on the content and context of text and links that appear on the page. Netsweeper AI category can also be used to safeguard users against more harm generative AI tools.
• Deployment – filtering systems can be deployed in a variety (and combination) of ways (eg on device, network level, cloud, DNS). Providers should describe how		Netsweeper has flexible deployment options. Cloud or Onsite. Network Level and/or client lead. Netsweeper filtering

their systems are deployed alongside any required configurations		provides device level filtering both on network and off network across Windows, Chromebooks, Mac, iPad, iPhones and Android device. This ensures the most comprehensive filtering deployment.
● 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		Netsweeper broad categorisation engine ensures flexibility when applying policies ensure school traffic isn't over blocked.
● Group / Multi-site Management – the ability for deployment of central policy and central oversight or dashboard		Netsweeper multi tenanted platform offers schools and trusts the ability to deliver granular policy management both at a per user level and per school basis.
● 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.		Users can be identified by various methodologies including but not limited to Entra (formally Azure), Local AD, google directory. Netsweeper supports hybrid and cross directory services (School that have both Entra and Google tenancy).
● 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		Netsweeper is capable of filtering all device based content including content delivered via apps.
● Multiple language support – the ability for the system to manage relevant languages		Netsweeper performs dynamic filtering in 47 languages

• 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		Netsweeper can deliver identical filtering to the school's devices onsite and offsite at a device level (not just browser based) across Windows, Chromebooks, Mac, iPad, iPhone and Android device
• Reporting mechanism – the ability to report inappropriate content for access or blocking		Netsweeper's reporting mechanism allows DSL the ability to report on inappropriate content accessed or blocked both on and off network.
• Reports – the system offers clear granular historical information on the websites users have accessed or attempted to access		Netsweeper's reporter can provide detailed reports on with the who, what, when and where. Netsweeper templated reports and custom reports can be set up to proactively send known safeguarding threats to DSL in near real time.
• Safe Search – the ability to enforce 'safe search' when using search engines		Safe search can be enforced across all well-known search engines and lesser-known ones.
• Safeguarding case management integration – the ability to integrate with school safeguarding and wellbeing systems to better understand context of activity		Netsweeper integrates with case management solutions either via direct API to the platform or CSV upload.

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.

Netsweeper's AI categorisation engine can help identify generative AI technology. Netsweeper would suggest using the walled garden approach to safeguarding users from generative AI. As an example schools should only use trust tools like Co-Pilot inside a school's tenancy, thus blocking all over technologies. On top of this Netsweeper onGuard

identifies safeguarding incident in real time and are reviewed by a UK based team of monitors. These can be flagged directly to the DSL and SLT of the safeguarding incident for them to deal with.

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

Ripple – Suicide prevention charity – www.ripplesuicideprevention.com

[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	Nick Levey
Position	Regional Director
Date	10/9/2025
Signature	