

Safeguarding in the Age of AI

Keeping Safeguarding Human, Proactive and Grounded in Lived Experience



A TransforMAtive Roundtable Networking Dinner
Event partner – Smoothwall, 10.06.26, The Ivy Asia, Leeds

Executive Summary

Artificial intelligence is rapidly reshaping education. It offers significant opportunities to reduce workload, improve accessibility, personalise learning and support inclusion. Yet, as AI becomes embedded in school and trust life, safeguarding must remain focused on people rather than systems alone.

This TransforMATive Roundtable, in partnership with Smoothwall, brought together MAT leaders, safeguarding leads, IT directors and education practitioners to explore how schools and trusts can respond to the changing safeguarding landscape. The discussion considered DfE filtering and monitoring expectations, the direction of Keeping Children Safe in Education 2026, mobile-phone-free environments, managed and unmanaged devices, and the growing use of AI tools by staff and pupils.

A clear theme emerged: safeguarding in the age of AI cannot sit with one function. It requires shared ownership between safeguarding, IT, curriculum, governance, data protection, senior leadership and vendors. Participants recognised that technical controls are essential, but they are only effective when paired with human judgement, clear processes, staff confidence and lived understanding of pupils' experiences.

The roundtable also highlighted a sector-wide challenge around language. Terms such as "filtering", "monitoring", "AI policy", "risk", "visibility" and "assurance" are often used differently by safeguarding teams, IT teams, governors and suppliers. This creates gaps in understanding and, potentially, gaps in protection.

The discussion concluded that schools and trusts need to move from reactive compliance towards proactive, collaborative safeguarding. AI should not be treated as a separate initiative, but embedded into safeguarding training, digital strategy, curriculum planning, governance reporting and operational decision-making.



Key Learning



1. Safeguarding must remain human-centred

AI and digital systems can identify risk, block harmful content and provide alerts, but they cannot replace professional curiosity, contextual knowledge or trusted relationships with children and young people.

Participants repeatedly emphasised that safeguarding is not simply about what a system detects. It is about recognising patterns, understanding context, knowing pupils well and responding with clarity. The most effective models combine technical controls with staff who understand both the child and the environment in which risk appears.



2. Filtering and monitoring are not the same thing

A significant part of the conversation focused on the difference between filtering and monitoring. Filtering was understood as preventing access to harmful or inappropriate content. Monitoring was understood as gaining visibility of behaviour, language, keystrokes, searches or activity that may indicate concern.

However, participants noted that the sector often uses these terms inconsistently. Some settings believe they have both filtering and monitoring when, in practice, they may only have one. Others receive alerts but do not fully understand what is captured, what is missed, and who is responsible for reviewing and acting on concerns.

The key learning is that trusts should define these terms clearly for their own context and test whether leaders, DSLs, IT teams and governors share the same understanding.



3. Managed devices do not automatically mean visible devices

The group discussed managed and unmanaged devices, including laptops, iPads, phones, smart watches and emerging wearable technology. A managed device can still contain blind spots, depending on configuration, app use, operating system limitations and whether activity takes place inside a browser or a standalone application.

Leaders need to ask not only “is this device managed?” but also:

- ✓ What can we actually see?
- ✓ What can we not see?
- ✓ Who receives alerts?
- ✓ Who reviews them?

Key Learning

- ✔ What happens when activity occurs off-site?
- ✔ What assumptions are we making?



4. Mobile-phone-free environments may reduce in-school risk, but not remove wider harm

Participants from specialist, AP, mainstream and post-16 settings recognised the value of removing or restricting unmanaged devices during the school day. This can provide a “pause” from online conflict, bullying, anxiety and exploitation. However, the risk does not disappear. Harm may happen before school, after school, on transport, at home or through unmanaged platforms. Schools are still left supporting the consequences.

The group saw value in phone-free approaches, but warned against assuming that they solve online safeguarding. They must sit alongside education, parental engagement, pastoral support and managed access to appropriate technology.



5. Safeguarding and IT must work as translators for each other

One of the strongest themes was the need for translation between professional worlds. Safeguarding teams understand risk, vulnerability, context and lived experience. IT teams understand systems, controls, configuration and technical limitations.

Neither perspective is sufficient alone.

The discussion showed that safeguarding colleagues need confidence to ask technical questions, while IT colleagues need to explain systems in practical, risk-based language. Effective safeguarding depends on both groups being in the room when decisions are made.



6. AI governance needs policy, standards and procedures

Participants recognised that AI policy is changing too quickly to be rewritten every few months. A more sustainable model is to create a stable, overarching policy supported by more regularly updated standards, approved tools, procedures and guidance.

This allows trusts to maintain strategic intent while adapting to new tools, risks and use cases.

A mature model may include:

- ✔ An overarching AI or digital safeguarding policy
- ✔ Approved AI tool lists

Key Learning

- ✓ Clear DPIA and procurement routes
- ✓ Safeguarding review before adoption
- ✓ Training for staff and leaders
- ✓ Regular review of risk and usage
- ✓ Governance-level reporting



7. Governance needs assurance, not just compliance

The roundtable raised concern that boards and governors may receive simple yes/no answers to complex questions. For example, asking “Do we have filtering and monitoring in place?” may provide compliance comfort but little meaningful assurance.

Trustees and governors need to understand the limitations, blind spots and residual risks. They do not need to be technical experts, but they do need enough understanding to ask better questions.

Assurance should include what is working, what is not visible, what is changing, what incidents or themes are emerging, and what action is being taken.

Sector Insights



The sector is moving from school-level autonomy to trust-level assurance

Many MATs are still on a journey from historically autonomous school-level decision-making to more centralised digital governance. This is especially visible in AI adoption, filtering and monitoring, cyber security, data protection and safeguarding systems.

Roundtable participants described the challenge of bringing together different schools, systems, platforms and cultures. Mergers and trust growth add further complexity, particularly where one organisation uses Google and another uses Microsoft, or where different schools have different filtering, MIS or safeguarding systems.

The sector insight is clear: trust-level assurance requires common standards, but implementation must still reflect local context.

Sector Insights



Specialist, AP, mainstream and post-16 settings face different risks

There was strong agreement that safeguarding expectations must be interpreted through setting type. A special school, AP setting, mainstream secondary, rural primary and sixth form college will all experience AI and online risk differently.

Some learners may be more vulnerable to exploitation. Some may rely heavily on assistive technology. Some may require close adult supervision. Post-16 and SEND settings may also face tensions between cyber requirements, mobile phone restrictions and accessibility needs.

A one-size-fits-all model will not work. Trusts need consistent principles with flexible implementation.



Staff confidence is uneven

AI confidence varied widely across the room. Some leaders felt technically confident but cautious about safeguarding implications. Others felt confident in safeguarding but less confident in AI terminology or system capability.

This matters because staff confidence shapes culture. If staff are fearful, they may avoid AI entirely. If they are overconfident, they may use tools without understanding risk. The sector needs balanced, practical training that addresses opportunity, ethics, safeguarding, data protection and real-world classroom use.



Parents and carers remain a critical blind spot

Participants highlighted the gap between school awareness and parental understanding of online risk. Schools may restrict unmanaged devices during the day, but pupils often return to environments where settings, supervision or awareness are limited.

This creates a safeguarding burden for schools and trusts. Parent and carer education must form part of the response, especially where children are vulnerable, anxious, socially isolated or at risk of exploitation.

Sector Insights



Vendors are increasingly part of the safeguarding ecosystem

The discussion reflected a shift in the role of technology providers. Vendors are no longer seen solely as firewall, filtering or compliance suppliers. They are increasingly expected to act as safeguarding partners, helping schools understand risk, configure systems, interpret alerts and share emerging practice.

This partnership role is valuable, but it also requires clarity. Schools and trusts must understand what a product does, what it does not do, what is moderated by humans, what is automated, and where responsibility remains with the education setting.

Future Research Questions



- 1 How do mobile-phone-free environments affect safeguarding, wellbeing, behaviour and attainment across different phases and settings?
- 2 What visibility is lost or gained when pupils move from unmanaged personal devices to school-managed devices?
- 3 How well do governors and trustees understand filtering, monitoring, AI risk and digital safeguarding assurance?
- 4 What is the most effective operating model for collaboration between safeguarding, IT, curriculum, data protection and governance?
- 5 How should MATs define “appropriate” filtering and monitoring across diverse school contexts?
- 6 What should a minimum viable AI governance framework include for schools and trusts?
- 7 How can schools identify AI-related harm when activity takes place in apps, encrypted environments or off-network spaces?
- 8 What training helps staff move from AI anxiety to confident, safe and purposeful use?
- 9 How can parental engagement around AI, online harm and device use be made more effective?
- 10 What role should vendors play in supporting sector-wide learning, shared standards and safeguarding maturity?

Conclusion



The age of AI demands a new level of safeguarding maturity. Schools and trusts cannot rely on systems alone, nor can they expect individual teams to solve complex digital risks in isolation.

The roundtable demonstrated the value of bringing safeguarding, IT, curriculum, governance and vendor perspectives together. The most important outcome was not a single answer, but a shared recognition that safeguarding in the age of AI is collective work.

The challenge now is to turn that collaboration into operating models, assurance processes, training, governance and culture. AI will continue to evolve quickly. The safeguarding response must be equally proactive, but always grounded in people, relationships and the lived experience of children and young people.

