appg@iwf.org.uk

APPG on Children's Online Safety

All-Party Parliamentary Group (APPG) on Children's Online Safety

Evidence Session on the positive uses of AI for children online

Monday 9 March 2026, 13:45 – 14:45, Macmillan Room

Parliamentarians: Lord Russell (Chair), Steve Race MP, Lord Carlile, Iqbal Mohamed MP

Secretariat: Bobbie Dennis, Sumaiya Zahoor (Internet Watch Foundation)

Participants: Sarah Castro MBE FRSA (SafeToNet), Hugh Milward (Microsoft), Doniya Soni-Clark (techUK)

1. Welcome

1.1. The Chair, Lord Russell, opened the third oral evidence session of the All-Party Parliamentary Group (APPG) on Children's Online Safety and thanked attendees for joining.

1.2. Lord Russell referenced ongoing discussions in the House of Commons regarding amendments to the Children's Wellbeing and Schools Bill. He stated that parliamentarians across both Houses continued to work together to focus attention on online safety issues affecting children.

1.3. The minutes from the previous meeting were approved.

2. Opening Statements

Sarah Castro MBE FRSA (SafeToNet)

2.1. Sarah introduced SafeToNet as a British technology company focused on protecting children online through preventative safeguarding technologies. She stated that much online safeguarding remains reactive, intervening only after a child has already experienced harm. SafeToNet's work instead seeks to move "upstream" and prevent harm before it occurs.

2.2. She explained that the company has developed technologies designed to prevent the viewing, creation and livestreaming of child sexual abuse material (CSAM). Drawing on more than two decades of work in safeguarding and youth violence prevention, she noted increasing concern regarding children harming other children and highlighted links between children's exposure to online pornography and harmful sexual behaviour.

2.3. Sarah stated that SafeToNet's approach differs from platform-by-platform moderation, which she described as "whack-a-mole", by embedding safeguarding directly into devices themselves. She described a prototype Nokia device developed in partnership with Vodafone. The technology operates on-device, monitoring both screens and cameras in order to prevent the creation or consumption of harmful content.

2.4. She emphasised that artificial intelligence must be used positively and responsibly, concluding that the technologies necessary to protect children already exist and that implementation is now "a policy choice".

Hugh Milward (Microsoft),

2.5. Hugh introduced himself as Vice President of Corporate Affairs at Microsoft and stated that he had engaged extensively on child online safety issues both professionally and personally.

2.6. He said Microsoft's objective is to empower young people to use technology safely while accessing educational, social and economic opportunities online. He argued that society should not have to choose between innovation and safety.

2.7. Hugh emphasised that online spaces can create opportunities for inclusion, aspiration and social mobility, and that children are already growing up in a digital world. He stated that the focus should therefore be on equipping children with the skills and protections necessary to thrive safely online.

2.8. He outlined Microsoft's approach to digital safety, stressing the importance of robust safeguards, responsible product design and collaboration between technology companies, educators, academics, parents and policymakers.

2.9. As an example, he described changes implemented within Xbox safety systems. Microsoft had previously treated all under-18 users as a single category but has now introduced a separate "teen" category informed by research into age-specific behaviours and developmental differences. He argued that a simple binary distinction between "adult" and "child" can be unhelpful and that more nuanced approaches may better support children's wellbeing online.

Doniya Soni-Clark – techUK

2.10. Doniya introduced techUK as a membership organisation representing more than 1100 technology companies, including SMEs, employing over one million people across the United Kingdom (UK).

2.11. She stated that online safety remains a top priority for techUK and referenced the organisation's engagement with the Online Safety Act (OSA). While acknowledging serious risks associated with AI misuse, she stressed there could be no commercial or policy justification for harmful uses of AI, particularly where children are concerned.

2.12. Doniya also highlighted the positive opportunities AI can provide for children, particularly those with disabilities, accessibility needs or learning difficulties. She referenced examples including speech-to-text systems, tools supporting dyslexic students and technologies enabling non-traditional communication. She argued that AI can enable children to excel in ways previously unavailable to them, provided appropriate literacy and safeguarding measures are in place.

3. Main discussion

AI Moderation and Harm Prevention

3.1. SafeToNet was asked what evidence exists regarding the effectiveness of AI moderation systems. She stated that existing systems are "not great" and explained that many major platforms currently rely on hash-matching technologies to identify and remove known illegal content. While millions of pieces of content are reportedly removed, she argued these approaches remain largely reactive.

3.2. She described SafeToNet's development of AI systems trained in collaboration with the Internet Watch Foundation (IWF). Training processes were designed to avoid exposing human reviewers to harmful imagery. The company developed standalone digital tools capable of identifying previously unknown illegal material.

3.3. Sarah stressed that AI systems must be trained legally and ethically. She criticised practices where AI-generated content is allegedly used to train moderation systems, arguing this is both unethical and potentially unlawful.

3.4. She then explained the company's "Mustard Seed" technology, which can operate across Android devices, messaging applications and online platforms. The system analyses content before it is sent or received, including within encrypted environments. She outlined possible applications including preventing coercive livestreaming, terminating unsafe feeds involving children and blocking the creation or sharing of harmful material.

3.5. The application itself is called "HarmBlock". Sarah stated that the company has piloted the technology with HMD, the makers of Nokia devices, and is seeking further manufacturing partnerships internationally.

3.6. Lord Carlile asked whether such technologies could distinguish between legal and illegal but harmful content, including incest pornography. Sarah stated that such capabilities are not currently deployed but that the underlying technology exists.

3.7. Lord Carlile also reflected on wider discussions within the House of Lords regarding whether sufficient political and commercial will exists to protect children effectively from harmful online material.

<u>Safety by Design</u>

3.8. Steve Race MP asked witnesses what "safety by design" means in practice.

3.9. Microsoft described responsible design as a continuous process embedded throughout product development and organisational culture. He stated that companies must constantly review how technologies behave in practice and assess long-term impacts during development.

3.10. He argued that safety-by-design approaches require internal ethical standards, operational review mechanisms and ongoing evaluation. While difficult and resource-intensive to implement, he said such approaches are essential.

3.11. TechUK added that many techUK members have implemented formal safety-by-design codes of practice intended to prevent harmful content from reaching platforms in the first place. However, she cautioned that malicious actors frequently repurpose technologies for harmful ends beyond their intended uses.

3.12. She pointed to the Online Safety Act as an important framework requiring companies to consider safety obligations during product development.

3.13. Safe-To-Net argued safeguarding must be embedded "from top to bottom" within companies. Drawing comparisons with mission-led organisations such as NASA, she stated that protecting children online should become a core organisational purpose guiding every operational decision.

3.14. She added that SafeToNet is expanding its work beyond CSAM into areas including strangulation content, gore and violence detection.

Trust, Regulation and Corporate Responsibility

3.15. Iqbal Mohamed MP questioned whether technology companies are sufficiently incentivised to protect children if products remain highly profitable. He raised concerns regarding the amount companies invest in safeguarding relative to monetisation and referenced concerns that some technology executives restrict their own children's use of digital products.

3.16. He also noted that AI systems are not infallible and can create unintended harms, comparing AI deployment to pharmaceutical testing where products can still produce harmful side effects despite extensive review.

3.17. Microsoft responded that businesses ultimately depend on user trust and satisfaction. He distinguished between harms caused directly by technology itself and harms arising from malicious human behaviour using technology.

3.18. He stated that AI moderation should always include "human in the loop" oversight. Within Xbox, AI systems assist moderators by identifying harmful content and behaviours, but human reviewers continue to make moderation decisions.

3.19. The Chair observed that rapid competition in AI development has sometimes resulted in safeguards being deprioritised in favour of speed to market.

Positive Uses of AI

3.20. The Chair asked witnesses to provide examples of positive uses of AI for children.

3.21. Sarah highlighted AI applications in education and safeguarding technologies embedded directly into devices to prevent exposure to harmful content. However, she argued companies often only act where legislation or market pressure requires them to do so.

3.22. She referred to industrial-scale abuse occurring in countries such as the Philippines and criticised what she described as insufficient action by major technology companies despite the existence of preventative technologies.

3.23. Doniya referenced research including a meta-analysis by Hopkin et al. examining AI-supported educational interventions. She stated that speech-to-text technologies and AI tutoring systems have demonstrated positive impacts, including within Government-supported one-to-one tutoring programmes.

3.24. She also raised concerns regarding children turning to unregulated AI chatbots for emotional support and companionship. While acknowledging risks, she suggested there may be opportunities to develop safeguarded AI systems specifically designed to support young people safely.

Regulation and Enforcement

3.25. Lord Carlile outlined four broad approaches to regulation: voluntary action by industry, industry-led regulation, partnership regulation with statutory backstop powers, and direct statutory regulation. He argued that no regulatory system would be effective without meaningful sanctions, potentially linked to global turnover.

3.26. Doniya described the Online Safety Act as a positive example of partnership between industry and government. She stated that AI-related issues are increasingly being incorporated into the framework and suggested the UK approach could become internationally influential.

3.27. Hugh stated that Microsoft broadly supports stronger statutory regulation and has publicly endorsed AI regulation. However, he argued regulation should focus on use cases rather than regulating AI technology itself. He cited facial recognition as an example of technology with legitimate beneficial uses.

3.28. He also warned that purely voluntary approaches leave excessive room for non-compliance.

Platform Governance and Moderation

3.29. Iqbal Mohamed MP asked about Microsoft's enforcement mechanisms and moderation practices, including controls relating to OneDrive and third-party content on Xbox.

3.30. Hugh confirmed that Microsoft's terms and conditions prohibit harmful uses of its products and that sanctions are applied for breaches.

3.31. Regarding OneDrive, he explained that Microsoft does not proactively monitor private stored content but uses PhotoDNA technology when material is shared or distributed. He stated that this approach appropriately balances privacy and protection.

3.32. On Xbox, he outlined Microsoft's community standards and moderation systems, including sanctions for misconduct. He added that third-party developers operating within Xbox ecosystems must comply with Microsoft's rules and approval processes.

4. Closing reflections

4.1. Lord Russell concluded the session by asking witnesses to reflect on both optimistic and pessimistic future scenarios regarding AI and children's online safety.

4.2. He asked witnesses what positive developments they hoped to see in the coming years, what risks most concerned them, and what policymakers should prioritise in order to prevent future harms.

5. Next steps

5.1. The Chair thanked witnesses and attendees for their detailed contributions.

5.2. It was confirmed that materials and minutes would be circulated to members.

5.3. The Secretariat would share further information regarding future inquiry sessions.

6. AOB

6.1. There was no other business and the meeting was closed.