CYBERSAFE KIDS

Presentation to the Committee on Children, Equality, Disability, Integration and Youth on the topic of "Engagement on the protection of children in the use of Artificial Intelligence"

We would like to thank the Chair and the members of the Committee for inviting us here today. We welcome the opportunity to talk about this important topic.

About CyberSafeKids

Established in 2015, CyberSafeKids is the only Irish charity dedicated to enhancing online safety for children nationwide. Our mission is to ensure children are safer online and the online world is made safer for children. At our core is an education & research programme for primary and post-primary schools, providing expert guidance to pupils aged 8-16 and teachers/parents. We also publish trends and usage data on an annual basis, which helps to paint a picture of what children are actually doing online, the levels of access they have and the areas of vulnerability. Our education programme has reached 65,000 children, 15,000 parents and educators directly across Ireland.

Introduction

We want to begin by acknowledging the hugely important role the internet plays in all of our lives and to recognise it as a very beneficial resource for children for learning, creating, socialising and for entertainment purposes. In 2021, the UN Convention on the Rights of the Child formally adopted General Comment 25, which recognised children's rights in a digital environment to be the same as their rights offline, including the right to participate, the right to access accurate information, the right not to be exploited and the right to be protected from harm.

Whilst the internet brings us opportunities that we couldn't have imagined twenty years ago it also brings risks, particularly for children. The internet was not designed with children in mind: these are environments that many adults struggle to understand and to manage effectively, let alone children and young people.

Children and AI

Whilst much of the current discussion around Artificial Intelligence (or AI as I will refer to it), focuses on the latest developments in generative AI, it has been around for years and has been actively impacting on children in their use of technology over the past 10 years.

AI is behind machine learning, which drives the algorithmic recommender system, which dominates the feeds across social media. The likes of Facebook, Instagram, Snapchat, X, YouTube and TikTok rely heavily on AI algorithms to rank and recommend content to their users. The main aim is to keep eyes on screen. Whilst social media and gaming companies might argue that it's all interest-driven and designed to ensure that we're getting the best content and targeted ads for us, it can be deeply problematic for children through recommendations of inappropriate content related to self-harm, suicide, pro-anorexia and sexual content. Frances Haugen, the ex-Facebook employee turned whistleblower said Instagram's algorithms can lead to addiction in its young users by creating "little dopamine loops". Children get caught in the crosshairs of the algorithm and sent down rabbit holes, engaging with sometimes frightening or enraging content because as Haugen further stated, "it's easier to inspire people to anger than it is to other emotions".[1]

One mother we recently worked with in relation to her 13-year old daughter said:

> *"As a mother I have huge concerns for our teenage children. Last summer it was brought to my attention that my 13-year daughter had been bullied during First Year and by expressing her sadness in a video posted on TikTok, the app starting flooding her daily feed with images of other sad teenage girls referencing suicide, eating disorders & self-harm. The damage & sadness this has caused my family has been immense as we discovered that my daughter saw self-harm as a release from the pain she was suffering from the bullying through the information this app is openly allowing. Anti-bullying efforts by schools are of no use unless these social media platforms are held responsible for openly sharing all this hugely damaging content with children."*

Cybercriminals seeking to sexually extort online users, including children, are using advanced social engineering tactics to coerce their victims into sharing compromising content. A recent report from Network Contagion Research Institute noted an exponential increase in this type of criminal activity over the past 18-months and further found that Generative AI apps were being used to target minors for exploitation.[2] We know that this is impacting children in this country because we have had calls from families whose children have been affected. One such case involved a teenage boy who thought he was talking to a girl of his own age in a different county. He was persuaded to share intimate images and immediately told in the aftermath that if he didn't pay several thousand Euro that it would be shared in a private Instagram group of his peers and younger siblings. The threat is very real and terrifying and has led, in some cases, to truly tragic consequences.

[1] The Verge, 'Facebook encourages hate speech for profit, says whistleblower', 04 Oct 21, source: https://www.theverge.com/2021/10/3/22707860/facebook-whistleblower-leaked-documents-files-regulation

[2] Raffile, Paul et al, A Digital Pandemic: Uncovering The Role Of 'yahoo Boys' In The Surge Of Social Media-Enabled Financial Sextortion Targeting Minors, Jan 2024, source: https://networkcontagion.us/wp-content/uploads/Yahoo-Boys_1.2.24.pdf

To make matters worse, there are new apps that are facilitating such efforts, including ones that remove clothing from photographs, which bypass the need to put people in compromising positions. The photos can just be taken from social media accounts and then sent to the individual to begin the process of extorting them. Such sophisticated technology is greatly increasing the proliferation and distribution of what the UK's Internet Watch Foundation describes as 'AI-generated child sexual abuse material'.[3] There is a real fear, highlighted in the Internet Watch Foundation's report cited in the footnote, that this technology will evolve to be able to create video content too.

We know from recent headlines around celebrity deepfakes, that the problem is becoming more widespread. Deepfake software can take a person's photos and face-swap them onto pornographic videos, making it appear as if the subject is partaking in sexual acts. Research in this area points that whilst much of the abuse is image-based, such as exploiting broadly shared open sourced content to generate CSAM, it can also be used in grooming and sexual extortion text, which poses significant risks to children.

The rise in AI technology poses risks in terms of peer-on-peer abuse, which has been snowballing into a huge area of risk over the last number of years, according to figures from CARI. Peer-on-peer abuse is already massively increasing, and the courts in Ireland have reported underage access to online pornography as being a major contributing factor in serious crimes. In September 2023, 28 Spanish children aged between 11 and 17 were subjected to peer abuse when their social media images were altered to depict them as nude, and these nude images were then circulated on social media. The reports suggest these images were created and circulated by 11 boys from their school.

Over the past year, we've seen new AI features being rolled out and into the hands of children with little thought to the consequences. Snapchat added its beta 'My AI' feature onto every subscriber's account in March 2023; please bear in mind that 37% of 8-12 year olds in Ireland have a Snapchat account. It was touted as being like a friend of whom you could ask anything. If you read the small print you could see that it was still being tested and may return wrong or misleading information. Further testing by external experts found that it forgot that it was talking to a child very quickly into the conversation and started returning inappropriate information.[4] Only 9 months later, in January 2024, did Snapchat add a parental control to restrict the use of 'My AI'.[5]

[3] Internet Watch Foundation, How AI is being abused to create child sexual abuse imagery, Oct 2023, source: https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf

[4] Fowler, Geoffrey, 'Snapchat tried to make a safe AI. It chats with me about booze and sex', 14 Mar 23, source: . https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/

[5] Davison, Tamara, How to restrict Snapchat's AI chatbot, 11 Jan 2024, source: https://www.standard.co.uk/news/tech/snapchat-my-ai-chatbot-how-restrict-b1131780.html

Children are being treated like guinea pigs in the digital world. This was put succinctly by the author of *The Age of Surveillance Capitalism*'s author Harvard professor Shoshana Zuboff who wrote:

> *"Each day we send our children into this cowardly new world of surveillance economics, like innocent canaries into Big Tech's coal mines. Citizens and lawmakers have stood silent, as hidden systems of tracking, monitoring, and manipulation ravage the private lives of unsuspecting kids and their families, challenging vital democratic principles for the sake of profits and power. This is not the promising digital century that we signed up for".*[6]

Why don't the companies behind these services do more to protect children using them? One simple answer is money. They would need to invest a lot more money to bring about real change and in the meantime, they are making billions of dollars of profit off the back of advertising to children. A recent Harvard study found that collectively in 2022, Meta X, Snapchat and TikTok made $11 billion from advertising to children in the US, $2 billion of which was to children under the age of 12.

Actions needed to mitigate against the risks

We acknowledge that there are no easy solutions and this is further complicated by the fact that we take very different regulatory approaches between the EU and US, with the former more bureaucratic and heavily protective of individual's right to privacy. Here are some suggestions:

1. **Harness the power of AI to better protect children in online spaces** such as relying on age assurance to determine the age of child users. We know that technology companies are able to market to users based on age. Further investment on accuracy could see this technology being used to better safeguard children by, for example:
    - preventing underage users from accessing the platforms; we know from our Trends & Usage data that 84% of 8-12 year olds have their own social media profile in Ireland. [7]
    - better protecting child users on platforms from exposure to harmful content, from targeted advertising and data profiling
2. **Amendments to Legislation:** How closely does existing legislation mitigate these risks? Does existing law include artificially created images? The emergence of deepfake technology means there is no longer a requirement for the perpetrator to possess 'real' intimate images of their victim. Nonconsensual pornographic deepfakes are alarmingly easy to access and create. A report by Sensity AI found that 96 percent

[6] Zuboff, Shoshana, 'Effective regulation begins with unmasking these hidden systems', source: https://5rightsfoundation.com/in-action/effective-regulation-begins-with-unmasking-these-hidden-systems.html

[7] CyberSafeKids Trends & Usage report 2023, source: https://www.cybersafekids.ie/resources/#research

of deepfakes were non-consensual sexual deepfakes, and of those, 99 percent were made of women.[8] The Harassment, Harmful Communications and Related Offences Act 2020 was enacted, in part, to criminalise non consensual intimate image abuse. Section 3 prohibits the *recording, distribution or publishing of an intimate image of another person without that other person's consen*t. The definition of intimate image *in relation to a person, means any visual representation (including any accompanying sound or document) made by any means including any photographic, film, video or digital representation.* Section 3 does not appear to clearly extend to images generated without consent.

Notably the new EU Directive on Child sexual abuse will revise the 2011 directive to include updating the definitions of the crime to include definitions on child sexual abuse material in deepfakes or AI-generated material. Ireland should be leading the charge in this arena, given we are regarded as one of Europe's leading tech hubs. Our legislation needs to match this status.

3: Policy, Regulation and Enforcement:

- Safety by Design is a key criteria in devising technologies that are being accessed by children. We know that technology companies are compliance oriented- and largely speaking, as commercial entities, they will not go beyond basic compliance where legislation does not demand it to do so. How can these powerful concepts of safety by design be included in regulation? We note that An Coimisiun na Mean is currently drafting binding Safety Codes. CyberSafeKids has recommended that definitions be extended to include AI generated images, in our response to the public consultation. We suggest however, that the regulations in this area must also be brought into line with any such definitions.
- Algorithm-based recommender systems should not be allowed to serve content to child users.
- Regulation is only beneficial when it is properly enforced and there needs to be great focus on how to do so.

5. Find fresh perspectives: We need new thinking and the confidence to believe we can make real progress on tough issues - like the Paris 2015 climate agreement has been to make work; there needs to be skin in the game and a financial incentive. We suggest that the Government set-up and fund an R&D lab with representatives from academia, industry and the not-for-profit sector solely to look at working on ways to better protect users in meaningful ways. Ireland can and should be a trailblazer in online child protection given our DPC status, EMEA HQ status and the strides taken to protect children via legislation over the past 20 years. This could include economic incentives that will change the behaviours of

[8] Sensity AI, The State of Deepfakes 2019 Landscape, Threats, and Impact, Source: https://sensity.ai/reports/

tech companies for example, if you collaborate with academics and other stakeholders, they would get some kind of financial reward/grant based on outcomes, not just participation.

Conclusion

To conclude, this leaves you as policy-makers, with an enormous and urgent challenge that is growing at pace. There are no quick fixes but a meaningful solution will involve legislation, regulation, education and innovative approaches. Nothing that we have in place currently is good enough or strong enough to take on this challenge properly but we remain hopeful that this will change, but it needs to happen quickly. None of those digital rights referenced in General Comment 31 are being upheld for children online in the current online environment. They are exposed to a wealth of mis- and dis-information. They are being bombarded with harmful content and this is having a genuine impact on their mental health, despite what Mr. Zuckerberg said in the congressional hearing just two weeks ago at which he was shamed into making an apology to parents who can testify intimately to the tragic consequences for their children of insufficient regulation and oversight. While AI offers children opportunities, if not properly regulated from the outset, we will see similar scenarios playout, where children are unwittingly testing dangerous, unregulated products for the profit of corporations.

Thank you for your time today. We look forward to any questions you might have.

Clare Daly
Alex Cooney