Report on the Fourth Meeting of High-Level Working Group for Privacy and Safety

Prof Andy Phippen, Bournemouth University

Prof Emma Bond, University of Suffolk

Introduction and Overview

The 'High-Level Working Group for Privacy & Safety' aims to advocate for a holistic, person-centred approach to online safeguarding that respects people's rights to online participation and to their privacy.

Convened by Prof Andy Phippen and Prof Emma Bond, the Working Group intends to drive discussions where central concepts such as harm, risk, vulnerability, well-being, and the best interest of the child are addressed in a nuanced and contextual manner to move conversations on from the traditional prohibitive narratives that beset the online harms work. In convening this group, Andy Phippen and Emma Bond, who collectively have 40 years' experience working in this area, are hoping to develop a more inclusive and progressive narrative that moves from "someone needs to stop this" to "what can we all do to make online experiences more inclusive while understanding and reducing harm". Current political narratives generally centre around how platforms can reduce or eliminate harms, with little consideration of other stakeholders that might be better placed to mitigate these risks.

The group brings a multi-stakeholder approach, convening experts from regulators, research institutions, private companies, industry associations, non-profit organisations, and academia to better articulate the challenges of tackling online harms in a right based, empowered manner.

As such, the goals of these sessions are:

- Build a community of stakeholders with a progressive view on tackling online harms.
- Placing a more progressive voice into the public domain with broad stakeholder buy in and a constructive conversation between parties aiming to achieve a common goal mindful of children's rights.
- To develop new approaches that stakeholders might adopt that go beyond technical intervention and prohibitive measures.

Sessions take place under Chatham House rules (although some attendees have consented to being named as attendees). Reporting on each session will be conducted through the publication of a detailed article on the discussions that took place (this being the third report in this series). These documents present the discussion that took place and will result at the end of the first three sessions with a recommendations document that brings together all the discussions that have taken place to articulate what a progressive, holistic, and inclusive approach to tackling online harms looks like. These reports are presented as working documents rather than academic analyses of the events with each output will be made publicly available for free. By placing these reports in the public domain, it is our intention to propose ways we might move conversations on from the current cycle of prohibition and prevention and introduce some new voices into the debates around online harms. The views reported in these documents reflect the feelings of those who contributed to the discussions rather than being a factual exploration of the issues that arose in the meetings, where there was conflict in views this will be represented. As such, the reports present a broad mix of views by progressive thinking in tackling privacy and safety issues in online platforms.

Session 4 - AI for Age-Appropriate Experiences: Balancing Youth Privacy and Safety

The fourth of these discussion sessions took place on the 6th June, 2024 online, breaking the discussion into two group sessions to allow for all contributors to be heard in the online environment. As the discussions with the group develop, so does the policy space and with the assent of the Online Safety Act in the UK, and the subsequent Ofcom consultations, there has also been a focus on AI as both a purveyor of harms as well as a "solution" to tackling AI facilitated harms. In particular, there has been much discussion about the use of AI in age assurance that is viewed, by some in the policy space[1], as the silver bullet to preventing harms and ensuring age-appropriate experiences for young people while others are less confident[2]. However, we should also be mindful that a recent technical report from NIST[3] had suggested that age assurance approaches were perhaps not as accurate as some vendors purport, and cannot be seen as a solution in their own right. We are also mindful, as a group, that all age assurance systems are not equal and their development should remain cognisant on the rights of the child in how they are developed.

Therefore, while it was agreed that AI was something we should explore as a group, we felt that a focus was necessary to pursue issues in sufficient depth, rather than looking at AI in general where the risk would be ending up with a collection of points covering a very broad range of issues but lacking depth. Specifically, the session aimed to unpick what the AI boom means to the current debate around youth privacy and safety, the balancing of the two and whether the current hype around AI matches how it is currently being used and how it might be used in the future. This was not a technical discussion, and the focus was on concerns related to its usefulness and the risks it brings, particularly related to rights, rather than having specific debates around correct terminology or algorithmic functionality. Therefore, in the discussions below we will generally simply refer to Artificial Intelligence based approaches as AI, and not differentiate between machine learning approaches and other forms of AI.

In organising this discussion, we used the following approach. Using the broad scenario:

[1] https://www.ofcom.org.uk/online-safety/protecting-children/implementing-the-online-safety-act-protecting-children/

[2] https://www.biometricupdate.com/202407/highly-effective-age-assurance-poorly-defined-in-ofcom-consult-says-avpa

[3] https://doi.org/10.6028/NIST.IR.8525

Services must assess any risks to children from using their platforms and set appropriate age restrictions, ensuring that child users have age-appropriate experiences and are shielded from harmful content.

We structured the discussion in the following manner:

> *Identify major benefits and limitations/challenges around using AI for age-appropriate experience.*
>
> 1. Understand: AI boom and the policy debate
>
>   a. Trends/Common narratives
>
>   b. Implications for the current debate around youth privacy and safety
>
> 2. Use cases: AI for age-appropriate experience
>
>   a. What are the use cases that can highlight potential and risks
>
> 3. Topics for consideration
>
>   a. Key principles such as risk-based approach, accuracy, proportionality (i.e., data minimization), inclusion, usability
>
>   b. Roles of different players in the ecosystem
>
> 4. Looking ahead: How to encourage use of more/better AI for age appropriate experience and considering the legislative/regulatory
>
>   a. What works well
>
>   b. Open questions that remain (Key considerations/watchouts from privacy & safety perspectives)
>
> 5. Conclude with suggestions to lawmakers and enforcers.

The following provides a summary of the discussions and the key issue that emerged from this. As will the previous groups, while the discussion guide was a useful structure to follow new issues emerged naturally from the conversations.

The AI Boom and the Policy Debate – An Overview

As an overarching theme, the AI boom was discussed as the driver to all the new AI related debates and the associated moral panics. An observation from some members of the groups, who we might observe as older members, reflected upon the fact that, as with several tech policy areas, these issues had arisen before and there seemed to be a persistent unwillingness by policy makers to learn from these historical discussions and instead decide that everything is new, which it is not. AI has become viewed as an essential for emerging tech development and personalized services, but the reality is that it has been used for years, it is only the Generative AI boom that has brought it into public consciousness as a “new” technology. As, given its high media profile, it attracts the interests of policy makers and NGOs they will inevitably raise concerns about abuse. There is a fundamental challenge in that AI is both poorly understood across stakeholders, whose experiences are generally as an end user, and an over promising by vendors looking for investment and sales. This does not make for an informed debate on AI’s capabilities nor how it can be regulated.

Platforms like Meta have implemented sophisticated AI tools to moderate content at scale, addressing the sheer volume of user-generated content for many years. Scalability is a fundamental aspect of the use of AI for platforms supporting billions of users worldwide. However, platforms acknowledge that despite these advancements, challenges remain, such as in accurately identifying nuanced harmful content and balancing moderation with freedom of expression. Nevertheless, AI is not new to social media platforms.

The key concerns in AI policy debates about age-appropriate experiences for young people centre around safety, privacy, and data use (which, of itself, is a privacy concern). Therefore, ensuring the safety and privacy of children should be a driving narrative within the use of AI in development age-appropriate systems which, arguably, requires stringent regulatory scrutiny. However, these is also concern that those who are developing the regulation are not sufficiently appreciative about the history and capabilities of AI, and do not have sufficient technical knowledge to implement an effective regulatory framework.

Regulation should be concerned with whether risk assessments have been conducted and whether testing has been effective prior to launch. However, given the money and prestige in the AI tech world now, there is perhaps pressure to cut corners in a rush to launch. Regulators should be able to see testing and risk assessment approaches, but there are currently no standardized datasets everyone's working on when it comes to testing. A further, well established, challenge is that often it is not known what is going on to inside machine systems that are producing specific outputs. This, of course, makes it very difficult to determine the efficacy of the systems being regulated and there is a history of poor training data resulting is unpredictable outcomes.

Furthermore, obtaining data from individuals under 18 is, quite correctly, difficult under privacy regulation and should require explicit consent. Therefore, if we are calling for AI systems to be used to identify whether an end user is a minor, the effectiveness of AI systems designed to do this might be challenged due to the lack of training data or the complexities in consent to obtain significant volumes. Clear consent requirements and frameworks are necessary for using children's data in training AI models, particularly for content moderation and age verification, but this will impact on efficacy. However, the group was not confident this was well understood by policy makers, as many seemed to be caught up to the "AI is magic and can do anything" hype wave. There was broad agreement among the groups that AI is certainly not magic, it is computer code processing data at scale.

If legislation and regulation is to be effective, it must understand the function of these systems and their flaws. Saying "use AI to determine how old an arbitrary user is" does not necessarily reflect the technical complexity of doing this, particularly when it concerns young people. Policy makers need to appreciate the scale of both computing power and training data volume that is needed to elicit results such as those seen in Gen AI. This is not a simple process that can be conducted by any vendor, those who do it well requite massive computer power and data, which, in turn, introduces questions and ethical considerations regarding how that data is obtained.

For example, there is a view that harmful content can be policed using AI based approaches. However, the subjectivity of this makes it a challenge for a technical approach that needs to be trained on what is, and is not, offensive content. A fundamental ethical question remains regarding who decides what is harmful? Is it government or a regulator? Or is it an expectation placed upon a platform to resolve themselves or whether the decision can ultimately reside with content moderators themselves. In all these cases there is a potential impact on freedom of speech and expression, given offence or harm are not well defined in the same way that, for example, illegal content such as Child Sexual Abuse Material is. With subjective interpretations come risks of over blocking or restricting expression and accusations of erosion of rights.

Transparency and societal understanding of AI's use of data, especially among parents and educators, are also critical issues. Many parents and teachers lack a full understanding of AI (or even any understanding), making it harder to protect and educate children about online safety and perhaps as importantly helping children understand the importance of not giving away their personal data without being aware of why data is being collected.

There is a critical need for comprehensive education around children's privacy, highlighting significant gaps in understanding among parents and educators. These gaps make it challenging to protect and educate children effectively about online safety and privacy. Children's technological capabilities often outstrip those of their teachers, leading to a disconnect that needs addressing. Young People frequently lack awareness of privacy issues and may not realize that their online activities may be harvested for data, underscoring the necessity for better education on privacy and data protection. Improved training for parents and educators as well as better education for young people are essential to bridge these knowledge gaps and to help children use technology responsibly.

Conversely, transparency and honesty in how children's data is used by authorities and companies are vital to building trust and ensuring ethical practices, and there is a role here for regulators in being clear on what transparency looks like and how privacy policies are defined. It is crucial to ensure that children and their guardians understand what data is being collected and how it is used, and opaque policies do not help with this understanding. Ethical and transparent data practices are vital, with clear communication about these practices to foster trust and ensure that data collection aligns with privacy protection standards, and that data that is used is done so with the consent of those whose data it is.

Age-Appropriate Experiences?

As is typical of these discussions, we often begin by deconstruction the premise – in this case, what do we mean by an age-appropriate experience? The discussions highlighted a clear differentiation between the broader use of AI for enabled age-appropriate experiences online, and the use of AI to verify/assure age so that platforms can effectively determine age and implement age-appropriate experiences, and each bring their own challenges.

When considering the use of AI in age estimation/assurance systems, proportionality in data use is addressed through concerns about obtaining and using data from individuals under 18. The difficulty in accessing such data due to strict regulations can impact on the effectiveness of AI systems designed to protect children (i.e. in order to determine whether the presentation of a minor on camera can be verified as a minor requires a lot of training data of similar). The discussions emphasize the need for consent and transparent data practices, balancing the necessity of data for AI system improvement with ethical considerations. This reflects a proportional approach to data use, where data collection and usage are weighed against privacy and ethical implications. However, as we will explore at length through this report, understanding proportional data use requires a level of literacy across the ecosystem which many in the group felt was not in place at the current time

It was suggested that ensuring age-appropriate experiences involves creating online environments that are suitable for children's developmental stages and protecting them from harmful content. However, there are challenges include defining what constitutes age-appropriate content, which can vary widely across cultures and communities, and determining who should set these standards is also complex. This question is not only about technological capabilities but also about societal norms and values. Different stakeholders, including regulators, technology companies, educators, and parents, play a role in defining and enforcing these standards. However, there is often a lack of consensus on what is deemed appropriate or harmful, which complicates the implementation of effective AI solutions.

However, it was clear from those who conduct research in these areas that there is not a consistent body of literature to support "age appropriateness" in such reductionist approaches. While there are clearly developmental phases, it should also be acknowledged that young people are individuals and develop at different ages, so arbitrary age limits to experiences can be problematic.

While there are age limits for certain online experiences (for example access to adult content) which are less contested, the "age 13" limit which permeates a lot of online debates, it was suggested, is poorly understood by policy makers, who believe there is a developmental or safeguarding rationale for this age limit, whereas the reality is that it was driven by privacy debates.

The 13-age limit on social media platforms originates from the Children's Online Privacy Protection Act (COPPA) in the United States, enacted in 1998[4]. COPPA imposed requirements on operators of websites and online services directed at children under 13, including obtaining verifiable parental consent before collecting, using, or disclosing personal information. As compliance with COPPA was already burdensome (as it, quite rightly, necessitates privacy policies, confidentiality measures, and security practices), many social media platforms set their minimum age requirement at 13 to avoid the need for parental consent (which, historically, platforms promoted to young children, such as Club Penguin and Moshi Monsters still conducted).

While the age of 13 was chosen for COPPA and adopted by social media platforms for a number of reasons, it was predominantly as the move to adolescence that marks a developmental milestone where children begin to understand complex social interactions and privacy and the alignment aligns with the transition from primary to secondary education in the US. However, there is no body of literature that supports the now conventional wisdom that 13 exists for child protection reasons or that children of 12 will be at risk in a manner that those who are 13 would not. Therefore, there needs to be more debate around who decides on age limits and age appropriateness, and whether youth voice is sufficiently represented in this discussion. Furthermore, this needs to be better understood in the policy space.

The integration of AI into digital platforms also raises several critical questions regarding youth privacy and safety which are explored in more detail below. AI has the potential to significantly enhance user experiences by personalizing content, providing tailored educational resources, and ensuring safer online environments (while bearing in mind the caveats discussed above). However, it also poses potential risks, such as the misuse of personal data, exposure to harmful content, and the perpetuation of biases.

[4] https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa

It was felt that the core of the policy debate is the challenge of balancing the benefits of AI with the need to protect young users. Policymakers are tasked with creating regulations that safeguard children's privacy and safety without stifling technological innovation which involves setting clear guidelines on data usage, ensuring transparency in AI operations, and implementing robust safeguards to prevent misuse. However, it was felt that policy makers were not sufficiently knowledgeable about the technical capabilities of the systems they are proposing be regulated, and this was a particular issue with AI because of both its technical complexity and current high visibility, where policy makers felt the need to comment in the public domain about its regulation.

Privacy and Safety Concerns

There was much discussion about the fundamental tension between protecting children's privacy and ensuring their safety online. For example, while AI can help monitor and restrict harmful content, it often requires extensive data collection, which poses privacy risks. The role of "safetytech" in the online safety and privacy ecosystem remains one of great debate. While such systems might help to detect abuse and some might argue that the detection of abuse means that young people can be made "safer" the manner in which this is approached raises significant concerns around children's right to privacy (for example, in the case of systems that allow parents to see all of a child's communications on a device). AI has the potential to ramp up these privacy concerns further and introduce a far more industrial scale collection of children's data in order to ensure they are safe.

It was observed that this is an underlying issue often missed from these debates - privacy is not just threatened by the operation of the AI based systems, but the data collected and used to train such systems. Many attendees in the discussions raised concerns around the ethical use of children's data, particularly in training AI models. Balancing these concerns is crucial to develop trustworthy AI systems that protect children without infringing on their privacy rights. There is a need for clear guidelines on data collection, consent, and usage and this is certainly something that remains absent from, for example, school use of children data. While it might be argued, it was suggested, that schools are told that must be compliant with data protection regulation, there is little in national guidance that goes beyond that broad statement and with a need for schools to generate additional revenue, there is clearly a temptation to share data.

A lack of AI literacy among policy makers is writ large once again. Consider the tension between policy makers insisting platforms verify the age of users and that they have to use "highly effective" AV/AA systems to do so. If such systems are going to be highly effective, they need to be trained on significant amounts of images of young people, and the law requires consent to obtain these images. This ethical tension highlights, once again, that technical solutions cannot be perfect and do not replace effective social policy which defines responsibilities for a broad range of stakeholders. It was raised by attendees that not all age estimation and assurance technologies are the same and there are some providers who adopt privacy preserving approaches in their solutions.

Transparency and honesty between different stakeholders in this ecosystem is essential, and we will visit transparency in more detail below. It should not be a contentious statement that children have autonomy and agency – this is defined in well-established rights standards. It was suggested that people who are in positions of power (authorities, companies, schools, etc.) tend to be risk averse rather than protecting children and would rather reduce or prevent participation than enabling it and putting safety measure in place that are both mindful of the young people's rights and also the duties of the stakeholder. In the case of a prohibitive environment the question was raised regarding what specifically protects children in what circumstance? Again, as is typical in these discussions, the crucial importance of knowledgeable young people was brought up, alongside a lack of confidence that current educational approaches might facilitate this. Again, this raises issue beyond technical solutions or the use of emerging technologies including AI solutions in keeping children "safe", this requires a far more nuanced solution than might be implemented in code.

It was also raised that safety solutions can equally be abusive of children's rights. For example, Defend Digital Me has investigated the NSPCC Report Remove campaign[5] and found several concerning examples of privacy abuses. In this system young people could go online via the Report Remove website if they had been a victim of non-consensual intimate image sharing, report it and have it removed. However, to use the system age assurance technology was in place to "assure" that the end user was indeed a minor, and there was evidence to suggest that such age assurance data was retained beyond this assurance process to train AV providers own dataset. Another partner in the project, the Internet Watch Foundation, it was reported, also retained images for further enhancement and training of their models. Yet the rhetoric of the service is one of support and encouraging disclosure, with little clear detail on data/image retention practices. It was suggested that if teenagers come to you and trust you with their data, they don't expect the data to be used for something else. The covert nature of this data collection and lack of transparency undermines the whole purpose of doing it and brings the integrity of such organisations into question.

While this is a single example, it is a clear illustration of how systems purported to protect young people might be abusing their rights to achieve this. As discussed below, there are other approaches which are far more aligned with victim empowerment and privacy preservation.

Regulatory and Legislative Frameworks

It was acknowledged that, at the present time, the regulatory landscape on some of these issues (related to risk and safety) is still very much in their infancy whereas other areas of regulation, related to data protection, are more embedded although can still have potential flaws in this area. It was agreed that effective regulation is essential to ensure that AI systems used by children are safe, reliable, and ethical. However, regulations often lag behind the rapid pace of technological change and as they catch up, the leading edge of tech has moved on again.

By way of example, it was mentioned that the EU and several nation states are now moving to take action against OpenAI for data scraping practices. Following significant regulatory scrutiny, several EU countries, including France, Germany, Ireland, Spain, and Switzerland, have ongoing investigations into OpenAI's compliance with EU data privacy regulations, particularly concerning its data scraping activities used to train AI models like ChatGPT. However, this has taken place "post event", and because of public outcry, rather than horizon scanning on the part of legislator.

The focus on controlling the underlying technology means that the law will always be playing catch up. This might be one reason why Safety by Design approaches tend to be better regarded than some emerging legislative approaches. This approach focusses on utility of the service

[5] https://defenddigitalme.org/wp-content/uploads/2023/06/Response-to-DDM-Email-V5-NSPCC-450415.pdf

being provided and an understanding of the risks that might be associated with the use of a service, rather than a “prohibit any harm” approach which will generally lead to conservative practice by those being regulated which potentially has knock on impact on rights conditions (see above regarding age estimation making use of minor’s data in order to improve performance). Legislation must be adaptable to address new risks as they emerge, providing a framework for ongoing oversight and adjustment and important aspects include setting standards for data protection, ensuring AI systems are tested and audited before deployment, and creating mechanisms for accountability, rather than the more intangible demands of some safety regulation.

Balancing the need for data to train AI systems with the ethical considerations of using children's data is a significant challenge. Policies must ensure that data collection is done transparently and with proper consent. Children’s data is particularly sensitive, and its misuse can have long-lasting consequences.

For example, the General Data Protection Regulation (GDPR) in the European Union mandates strict guidelines for data collection and usage, emphasising transparency and consent. However, enforcement remains a challenge, particularly with cross-border nature of the issues and varying interpretations of the law.

Discussion emphasised that while GDPR provides a robust framework for data protection, there are significant challenges and gaps, particularly when it comes to the use of children's data for AI training and the balance between privacy and the need for large datasets to improve AI systems. One key challenge noted was the tension between the need for children's data to train AI models, such as those used for age verification and content moderation, and the stringent requirements of GDPR that limit the collection and use of such data. This creates a circular problem where AI systems need accurate data to be effective but are hampered by privacy regulations that restrict data collection.

Moreover, the discussion pointed out that there is a lack of standardized datasets and testing protocols, which complicates compliance with GDPR and other regulatory requirements. This lack of standardization can lead to inconsistencies in how AI systems are developed and tested, potentially affecting their reliability and trustworthiness.

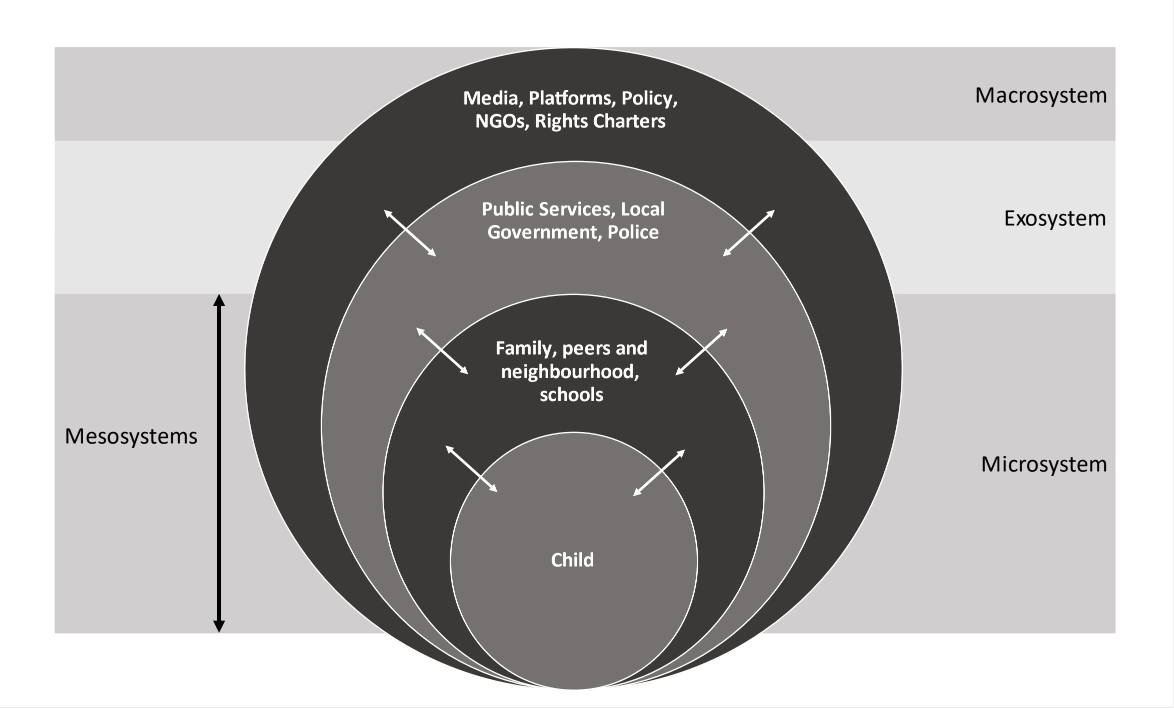
The importance of transparency, clear privacy policies, and the involvement of young people in designing privacy-friendly technologies were emphasised as crucial steps to improving the current situation. And this, in turn, would require a more AI literate ecosystem, including among young people, so that consent would be informed and from a place of knowledge.

AI introduces new challenges to the regulatory environment as it continues to evolve, with calls for more flexible and adaptive approaches to keep pace with technological advancements. For instance, companies should conduct regular audits to identify and mitigate biases in their AI systems, ensuring equitable outcomes for all users. Additionally, transparency reports can provide insights into how AI tools are used and their impact on users. Transparency was an issue that was returned to in many contexts and one we will explore in more detail below. There was a view that this might be one aspect of emerging legislation that might add value to the ecosystem.

The concept of risk-based approaches is implicit in discussions about the need for regulatory scrutiny and testing of AI systems before their launch. The emphasis on ensuring that AI tools are thoroughly tested and validated to prevent rushed launches highlights a risk-based perspective, where the potential risks associated with AI are assessed and mitigated before deployment. Additionally, the need for adaptable and flexible regulatory frameworks to keep up with technological advancements underscores a risk-based approach to AI regulation.

The Role of Different Stakeholders

As is typical with these discussions, different stakeholder positions were both represented, and the importance of collaboration was recognised. Stakeholders, including policymakers, technology companies, educators, and parents, play crucial roles in ensuring that systems making use of AI for age-appropriate experiences are effectively implemented and regulated. Collaboration among these players is essential to create a cohesive and comprehensive approach to youth safety and privacy and each group brings unique perspectives and expertise to the table.


Media, Platforms, Policy, NGOs, Rights Charters
Macrosystem
Public Services, Local Government, Police
Exosystem
Family, peers and neighbourhood, schools
Mesosystems
Microsystem
Child

We returned to the ecosystem model that has permeated all these discussions, an updated representation of which is presented above. The ecosystem model is acknowledged by those who attend the group as a means to holistically understand the importance of stakeholders at different levels around the child and the need for collaboration and communication.

As such, inclusion is highlighted through the emphasis on understanding and addressing the educational gaps among parents, educators, and children regarding AI and data privacy. It was suggested that the significant societal lack of understanding in these areas points to the need for inclusive educational initiatives that can bridge this knowledge gap. Ensuring that all stakeholders, including children, parents, and educators, are informed about AI and its implications fosters an inclusive environment where everyone can engage with and benefit from technological advancements safely. However, this requires, at the macro level, greater AI literacy among those developing the legislation, and there may be a role there for those who make use of AI in the implementation of online services.

Usability was discussed in terms of balancing the removal of harmful content with maintaining a positive user experience. AI's role in making processes like age verification less intrusive and more user-friendly is noted as a potential key benefit but the issues discussed above (and in the limitation section below) were also acknowledged. The emphasis on ensuring that AI systems do not negatively impact user experience while effectively addressing safety concerns underscores the importance of usability in AI design. Moreover, the discussions about trial opportunities and independent audits for AI systems suggest a focus on continuous improvement and user feedback to enhance usability, and the role of safety regulators was discussed, in conjunction with the need for consistency across safety regulators which can be informed by standardized data sets and risk assessment practices.

The challenges of problematic practice by stakeholders were also discussed and the question was raised regarding whether there are others outside of the focus on platform challenge that cannot claim to be making a positive contribution to the improved awareness of the use of AI and its impact on children's privacy. A scenario from the NGOs space was discussed above, and it was also raised that schools remain a challenge and many collect huge volumes of data from children that they are selling to providers, with little awareness by either young people or parents, and with questionable consent in this scenarios (i.e. getting parents to sign a "consent" form at the start of the year that supposedly covers all uses of their children's data by the school, rather than obtaining consent for specific instances). Once again, a common theme in these discussions, problematic practice by stakeholders away from the regulatory gaze, was raised as a concern if achieving age appropriate and rights informed experiences for young people.

Transparency and Accountability

Transparency will be picked up again later in this report when considering "best practice" across the stakeholder space however, it is worth some discussion in a separate section because it is view by the group as such an essential part of a health ecosystem. For example, AI-powered age verification systems can help platforms enforce age restrictions by estimating users' ages based on various data points such as facial recognition and behavioural analysis. This ensures that underage users cannot access content or services that are not suitable for their age group. These systems should offer a more seamless and less intrusive way to verify ages compared to traditional methods.

However, the National Institute of Standards and Technology (NIST) in the US has been testing age estimation algorithms to improve accuracy[6]. This firstly demonstrates the importance of independent and transparent evaluation of these systems, rather than simply saying "use extremely effective age assurance". In this report it was clear, when the data was unpicked, that the systems proposed by some policy makers as the solution for age-appropriate experiences, and therefore one which platforms felt pressure to implement, were not infallible and can be inaccurate to a degree which would be impact on an individual's rights. For example, the levels of error in the systems would suggest many adults might be caught in age assurance systems and be prevented from access services when they are legally entitled to do so. Furthermore, it was pointed out in the discussions, these systems, when geo-located, can

[6] https://www.nist.gov/news-events/news/2024/05/nist-reports-first-results-age-estimation-software-evaluation

be circumvented, and this issue needs to be acknowledged by policy makers, who need to more effectively define what “doing enough” or “highly effective” looks like for platforms.

It was stated that there is a lack of standardized datasets and testing protocols for AI tools, leading to inconsistencies in their performance and reliability. Regulatory bodies such as NIST are working on testing standards, but more is needed to ensure uniformity. Standardized testing can help ensure that AI systems meet consistent safety and quality benchmarks and NIST has become a de facto auditor because of the size of their test dataset, and because no one else is doing anything to this scale. This papers over the cracks of the need for international standards and standardised data sets.

It was agreed that transparency in AI development and deployment is essential to build trust. This involves being clear about how AI systems work, what data they collect, and how they use that data. Accountability mechanisms, such as independent audits and standardized testing, help ensure that AI systems are reliable and ethical and there is a need for AI systems to undergo rigorous testing before being deployed can prevent harm and build public trust in these technologies.

Tech companies should be transparent about how their AI systems work and the data they use, and how it was collected. This builds trust and ensures that users and regulators understand the implications of AI tools, and they should also provide clear and accessible explanations of AI decisions and actions.

For example, companies can publish transparency reports detailing their data practices, content moderation policies, and AI performance metrics, and growing regulation will mean that these become more commonplace. There was hope among the groups that these reports can help users understand how their data is used and the measures in place to protect their privacy. However, once again, this is not something that companies can do on their own, it requires other stakeholder to be aware of transparency reporting and how it might impact upon their own practice. For example, for educators, having evidence of how platforms deal with disclosures and address harms should play a part in harm reduction approaches to education, rather than simply delivering prohibitive messages.

Collaboration between tech companies, regulators, and educators can lead to the development of better AI tools. Initiatives like co-design with young people can ensure that AI systems meet the needs and expectations of their intended users and there was discussion about these sorts of activities taking place in industry. As part of these discussions, it was also acknowledged how much stakeholders can learn from listening to young people rather than telling them what to do. It was also recognised that not enough listening to young people takes place across the ecosystem.

Educational and Awareness

A facet of the ecosystem to which we always return, regardless of use cases being discussed in a working group meeting, is education. The is a consistent view among attendees that education is crucial for *all* stakeholders to understand AI, its benefits, and its risks. Children, parents, and educators need to be informed about online safety, privacy rights, and the ethical use of technology, and policy makers need to either develop their own AI literacy, or listen to industry discussions around what is, and is not, possible with the technology and what risks such approaches bring.

Fundamental to any discussion around effective education is how we better help young people making informed decisions about using online services and understanding the impact across all their rights. There is a need to address the knowledge gap through targeted educational programs can empower children and adults to make informed decisions and use technology responsibly. In an ideal world, digital literacy programmes should be integrated into school curriculums and community initiatives. However, attendees working around the educational aspects of the ecosystem recognise that this is far from the reality for most young people and there is broad recognition that digital literacy is not a priority for many school leaders because it is not a priority for education regulators and inspectors.

There was discussion around how children's technical knowledge tends to be better than that of either parents or education professionals. Therefore, sometimes the need for education is dismissed, through a result of lack of understanding and confidence in delivering this education. However, there was also agreement among those who work in education that young people's knowledge of privacy, why it matters, and their rights, is generally very poor. Which is unsurprising given the lack of education in this area. Online safety education tends to focus on the edge cases and moral panics, such as sending nudes or abuse, and the approach is prohibitive. Far better, it was suggested, that education focusses on understanding risk and harm reduction, and having good education and understanding around privacy would be a crucial facet of this. There would likely be a reduction in young people having data harvested if they had a better understanding of privacy policies and what companies do with their data, but a better understanding of privacy would also align with appreciating consent and rights, and while build resilience more generally.

Empowering young users through education is essential to ensuring their safety and privacy in digital environments. This includes providing comprehensive digital literacy programs, supporting parental involvement, and promoting a culture of responsible AI use. Educating young users about their rights and how to protect their privacy can help them navigate the digital world safely and confidently.

However, it was also discussed that there might be resistance to educating children about their rights might be viewed as problematic by those who wish to control them. It was observed that there is certainly a tension between strict behaviour policies in many UK schools (mandated by the government) and children's knowledge of rights. Discussions can be frustrating because they tend to be adult led and reductionist in scope. This is as much as making sure children do not misbehave and there have been examples where privacy abuses have been used to maintain this. Therefore, there could be some opposition to young people having a great awareness. This point returns, once again, to the need for all stakeholders to be open and transparent. Clearly this is not, as is typical in this ecosystem, something that educators and parents should do independently. Parents and educators need support and resources to understand AI technologies and guide their children effectively. This includes providing training and resources to help parents navigate the digital landscape and support their children's online activities. We discussed the initiatives and support that platforms already provide, and how this is not sometimes recognised by other stakeholders. There is still a view that harms are something that platforms should prevent, rather than recognise it as a multi-stakeholder need with young people at the centre with their needs being listened to. And rights frameworks *should* provide an agreed and, arguably, well understood foundation for addressing these challenges.

Technical Limitations, Challenges and Emerging Best Practice

The value in bringing together technical stakeholders with those who consider social, legal and civil issues is that the discussions can acknowledge technical flaws in achieving desired outcomes in providing age-appropriate experiences for young people. The discussions highlights firstly that accuracy is not the same in all experiences. Those attending from the age assurance sector, reinforced by the NIST report, stated that age assurance for young children is generally more accurate than it is for those who are teenagers – age assurance is not a one size fits all solution. This more nuanced understanding is needed by policy makers rather than assuming the AI can provide complete solutions and it 100% accurate. There are still questions around what "extremely effective" age assurance actually is (and what are the thresholds for accuracy) and policy makers should acknowledge the limitations of systems as well as their potential. This also requires honesty and transparency from vendors about the capabilities of there solutions.

This need for understanding of nuance and complexity extends to the use of AI in making age-appropriate experiences "safe" through tools such as content moderation. AI can be used in systems that analyse social media interactions to identify and flag potential risks, providing an additional layer of protection. AI can also monitor and analyse user behaviour to detect signs of distress or risky behaviour. This can include identifying patterns of cyberbullying, self-harm, or grooming. However, the ethical implications of behavioural profiling, including privacy concerns and the potential for misuse, must be carefully managed, as well as expectation regarding accuracy. There are also certain use cases that would be extremely problematic to implement. The functional demand for machine learning approaches requires training data to match the task. The collection of training data from illegal, child abuse content, which requires severity classification through NCMEC, adds complexity and constraints on its use by law enforcement. However, as recent Online Safety Act debates have shown, that does not mean policy makers will not suggest CSAM image recognition without understanding the functional requirements.

Bullying/abuse is also difficult to detect because the language used within it is so broad – so the language that *might* be used to abuse might also be used in friendly discourse between individuals. Therefore, again, if platforms are expected to detect this abuse, they will build large corpuses of training data to do so, but thresholds will mean that sometimes discourse that is not abusive will be detected. Some contributors mentioned warnings on social media platforms that they felt was clearly not abusive, however it was flagged as such because it follows similar patterns, or used similar words, to content that might be. While this is a debate for society around how many false positives/negatives are acceptable (which requires a far more AI literate population) it once again demonstrates that these systems cannot, at least at the present time, be perfect.

Services such as the Revenge Porn Helpline highlight the challenges in automated responses to abuse. It is clear from their published research that nuance, and context are the most important thing they deal with, every victim has a different experience and requires different support[7]. Such support is rarely successful with automated responses because they require a consistent critical mass of training data to be able to recognise abuse.

[7] https://revengepornhelpline.org.uk/assets/documents/revenge-porn-helpline-report-2023.pdf

Scale and reach also remains an issue – for global platforms there is a requirement to detect abuse in any language, not just majority languages where there would be more training dataA further perennial question is who decides what is harmful? Platforms can use their own data to determine what content has been reported as abusive, and what has been captured by moderators, but given the constant rhetoric by policy makers that "self-regulation doesn't work", will they be deferring to regulators to decide what is and is not abusive and/or harmful, and if so, are they subsequently policed by ministers?

Furthermore, there were questions raised regarding whether content take down and blocking is a "solution" to tackling these issues. It was suggested the building resilience is facilitated by being exposed to harm within a supportive environment, rather than never being exposed to harm.

Fundamentally, as we have discussed in previous reports, technology will only ever be part of the solution in addressing social problems, regardless of whether these issues occur online, and it requires those who use the technology, and those who legislate and regulate the technology, to recognise this.

There was broad agreement that one of the most promising aspects of technology regulation was transparency. The importance of transparency has been discussed above, but it is worth acknowledging that best practice in transparency has the potential to bring the ecosystem together.

If done effectively, transparency will lead to better understanding of what companies currently do, and a greater appreciation of best practice such as co-creation with young people. A significant part of these discussions lies in a broad range of stakeholder learning more about what platforms are doing, and there is a general feeling that companies should be encouraged to be more transparent as this is powerful information to engage the ecosystem and challenges persistent media and political narratives about companies not doing anything.

Meta, as a provider, are currently rolling out services that require both parental and child consent for the installation of parental controls. This is an excellent model because it involves stakeholders in a dialogue, rather than expecting resolution to simply take place by the provider. However, visibility of this approach is not high, and does not fit into the media narratives around platform scapegoating which claims platforms do not care about harms and do nothing to prevent them.

Furthermore, it does not have to be AI that provides the solutions just because AI is de-rigour. For example, StopNCII[8], used in NCMEC's Take It Down service[9] for young people, uses "old" hashing technology to recognise non-consensual image sharing, and it does so in a collaborative, community-based model with stakeholder buy in. It does not use AI for image recognition, the end user hashes their own images on their own device and uploads the hashes, which are shared by platforms where such images might be uploaded. If someone tries to upload a hashed image, it will be captured and prevented for being posted. This approach is far more privacy supportive than approaches such as ReportRemove, discussed above and shows that stakeholders working together can produce more effective solutions that those developed by a single stakeholder.

[8] https://stopncii.org/

[9] https://takeitdown.ncmec.org/

Industry best practice around transparency could potentially lead to changes in the policy landscape. Adoption of safety by design practices and transparency standards means that regulators will have to recognise that there are many in industry who already do a lot of ensure safety and privacy, the narrative of “industry needs to do more” collapses and other stakeholders become more exposed (or the regulator becomes the next scapegoat!).

Conclusions

The fourth session of the High-Level Working Group for Privacy and Safety explored the intricate balance between leveraging AI for age-appropriate online experiences and ensuring youth privacy and safety. The discussions underscored the potential of AI to enhance user experiences through personalized content, educational resources, and safer online environments. However, they also highlighted significant challenges, such as privacy risks from extensive data collection, ethical concerns regarding the use of children's data, and the complexities in accurately implementing age assurance systems.

Participants emphasized the importance of a multi-stakeholder approach involving regulators, technology companies, educators, parents, and children themselves. This inclusive strategy aims to foster a comprehensive understanding and collaborative effort towards creating a safe and empowering digital ecosystem for young users. The necessity of robust regulatory frameworks that keep pace with technological advancements was a recurrent theme, alongside the critical need for transparency and accountability in AI development and deployment.

Educational initiatives emerged as vital for bridging knowledge gaps among parents, educators, and children regarding AI and data privacy. Empowering young users through comprehensive digital literacy programs is essential for helping them navigate the digital world safely and confidently. The group also called for greater AI literacy among policymakers to ensure informed decision-making in the regulation of AI technologies.

The session concluded with a call for continuous improvement and adaptation in regulatory practices, encouraging transparency and ethical data practices. By fostering open communication and collaboration among all stakeholders, and by emphasizing the rights and voices of young people, the group aims to move towards a more inclusive and protective digital environment. This approach seeks not only to mitigate risks but also to harness the potential of AI to create enriching and safe online experiences for children.