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

Prof Andy Phippen, Bournemouth University

Prof Emma Bond, University of Suffolk

Introduction

The online harms world is a challenging one, where most parties, we are sure, would agree that it is important that citizens can engage with online platforms and discourses in a manner that mitigates harm and does not expose them to abuse. However, how we achieve this is complex and many stakeholders have conflicting viewpoints. These range from the prohibitive ("Harms occur on platforms, therefore platforms need to stop it") to the progressive ("Harms are caused by people, how do we reduce them by being mindful of people's right to participate free from excessive surveillance?"). There are many views that fit between these two positions. However, it is unquestionable that these perspectives all wish to achieve the same goal – that people, particularly young people, can experience the online world while not being harmed.

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 aims to bring 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.

The Working Group will hold three roundtable discussions starting from April and followed by the sessions in September, and November. Broadly, the three sessions aim to cover three core issues:

Session 1 – Context and Key Concepts/Scene Setting

- Who are the different groups in scope of the discussions?
- Unpacking core concepts such as best interests of the child, harm, safeguarding and privacy.

- The current landscape and the challenges therein – the current regulatory landscape and the "state of the art" in online safeguarding

Session 2 – Interest, Rights and Freedoms

- Digital rights of people and how to respect them while keeping them safe.
- Exploring the relevant vectors for such a balancing exercise; risks, roles, and associated responsibilities (ethical, social, legal and political) as well as capabilities (technical, social, psychological, legal) with respect to privacy and safety

Session 3- Getting the Balance Right – How the Ecosystem Should

- The role of each stakeholder for digital safeguarding and how the whole ecosystem could work together.
- Where are we at with laws and regulations in Europe and what can we do better?
- Is current regulation getting the balance right – what would good regulation look like?

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 first such report) as well as a summary article in an appropriate online space. These documents present the discussion that took place and will result at the end of the 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. 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 these debates.

This report considers the meeting of the High-Level Working Group for Privacy and Safety, which took place online on April 25th, 2023. While the meeting was convened with the organisational support of Meta, no one was funded to attend, and these are not paid events. Everyone gave their time for free.

The first session was intended to be a broad discussion so that attendees could meet each other, present their different perspectives, and provide input based upon the points for discussion. This session was intended to be free form in discussion around the challenges of online harms, with everyone given the opportunity to present their thoughts from their perspectives. Its main goal was to get to know each other, share experiences in the field, and develop an understanding of what this working group is seeking to achieve and the eventual output/end goal. The selection of participants was such that groups who supposedly have competing interests (for example civil society, children's rights groups and platforms) could discuss in a non-confrontational manner in order to explore different perspectives and build an awareness of the current and future landscape that these interests lie within.

Setting the Scene

The online harms world is one where privacy and safety are often placed in tension – particularly when it comes to children and young people. We need to keep young people safe

online, because we know that the impact of online harms can be severe. Therefore, we are told, young people should have to expect to have their privacy eroded – "someone" needs to see what they are doing online, and what they are looking at, to ensure they are safe.

Who this "someone" is remains up for debate – possibly it is parents (although we are also told that parents cannot be expected to "keep up" with technical innovation and therefore cannot be expected to manage their children's safety), but it is more likely that the focus on harm prevention now lies on those who provide the environments in which online interactions, and potentially online harms, take place.

We can see many examples of this in the current policy space. By way of example, we can consider an issue that is currently taking up much policy discussion and media coverage – young people's access to pornography.

The prevention of youth access to pornography has been a political discussion for many years, and already has many abandoned pieces of legislation. It proposes that young people accessing pornography is harmful, and they should be stopped from seeing it. Youth access to pornography is rarely part of the debate – there are very few who would argue they should be allowed access. However, the tensions lie in how to tackle this. There are some, mainly in the policy space, who believe it is the role of providers to check the age of the end user and prevent access to anyone under the age of eighteen. To do this, the end user needs to prove their age. There are some that argue this immediately places some challenges for privacy, as browsing habits of this nature should be private (as they hold sensitive personal data). Others argue that technology exists that can preserve privacy while still providing age verification. Other groups, for example young people, state that they already know the workarounds, such as Virtual Private Networks, which has even led to proposals in parliament to consider the legality of these privacy enhancing technologies. And there are progressive voices that say while prohibition is a utopian goal, it is unrealistic and we need to support young people with both education around what they are seeing, and routes for disclosure should they see something they find upsetting.

Even with this one example, we can see it is complicated. And we can see that an all or nothing approach immediately puts up tensions that are not necessarily needed. A progressive, youth centric approach means that we can provide an environment that allows them to mitigate risks, while being mindful of their rights.

This is certainly not the only scenario where these conversations are taking place – we can see similar conflicts in the debates around end-to-end encryption being a barrier to policing the distribution of child sexual abuse material (CSAM). A privacy enhancing technology, which has been standard in much online communication for many, is once again seen as problematic because, in one use case, it can make detection more challenging.

As a result of use cases emerging that evidence harms, the cyclical political narrative is a simple "we need to stop this", and it is said with the best of intentions. However, the progressive voices say as a comeback "if it was simple to do, don't you think we'd have done that by now?", and highlight that these things are complicated.

A society that affords young people no right to privacy or participation, to ensure they are safe, is not a progressive one. Equally, with the focus in the current policy space almost entirely on platforms ensuring their users are "safe", it does raise questions around what we

mean by safe. Does safe mean free from any risk of harm? Or does it mean the risk of harm has been mitigated or reduced?

Moreover, while platforms undoubtedly have a role to play, how much is "enough"? Platforms adopting conservative content control policies could mitigate a lot of potential harms from abusive comments and content but would also restrict the freedom of expression of many users. By developing "solutions" to prevent abusive comments, without the means to understand context, might they result in unintended consequences for free speech?

It is therefore important to move the conversation on. We have seen these discussions for as long as online interaction has been possible. While the technologies change (in the early 2000s there was much discussion on preventing grooming on platforms such as MSN and cyberbullying on Bebo), the cycle in policy repeats – this thing is occurring online, how do we stop it? Conversely, conversations with young people yield the same calls - regardless that the young people spoken to in the early 2000s are now adults themselves – we need education, we need support, we do not want adults to "freak out" and we need to be heard[1], [2].

Media narratives also present problems that harms are presented as extreme, that online spaces are dangerous, and platforms need to prevent harms. If we consider a recent report by the Internet Watch Foundation[3], this showed they have intercepted more CSAM content than ever. The media, and political, narrative centred upon the concerns around an increase in content (whereas the report could only claim more had been found, because it is impossible to quantify scale of content online) and the lack of interception by platforms where it was hosted (again, in the report it showed that most is not hosted on mainstream social media). A more progressive, evidence-based perspective might be that this excellent organisation who carry out a difficult job of searching for and taking down illegal content have been given more powers for pro-active takedowns are working, and they are receiving more reports from the public than ever. Or, a success for both policy and the wider stakeholder space – informed end users who are aware of the organisation are making use of them.

Which leads to the final point in introducing this topic – we need to remove the emotion and subjective outrage, accept that online harms exist and those who are subject to them need support, and the support should be informed by evidence and data, not opinion. In the drugs harms world, there has been a lot of work in *harm reduction* – accepting drugs harms exist in society and those who have a problematic relationship with drugs need understanding, help and support. In 2010, a wide range of drugs harms specialists published the Vienna Declaration[4], which called for all drugs harms policy to be developed from an evidence-based perspective. In the online harms world, we remain in an environment where the loudest

[1] Phippen, A., & Bond, E. (2023). Policing Teen Sexting: Supporting Children's Rights While Applying the Law. Springer Nature.

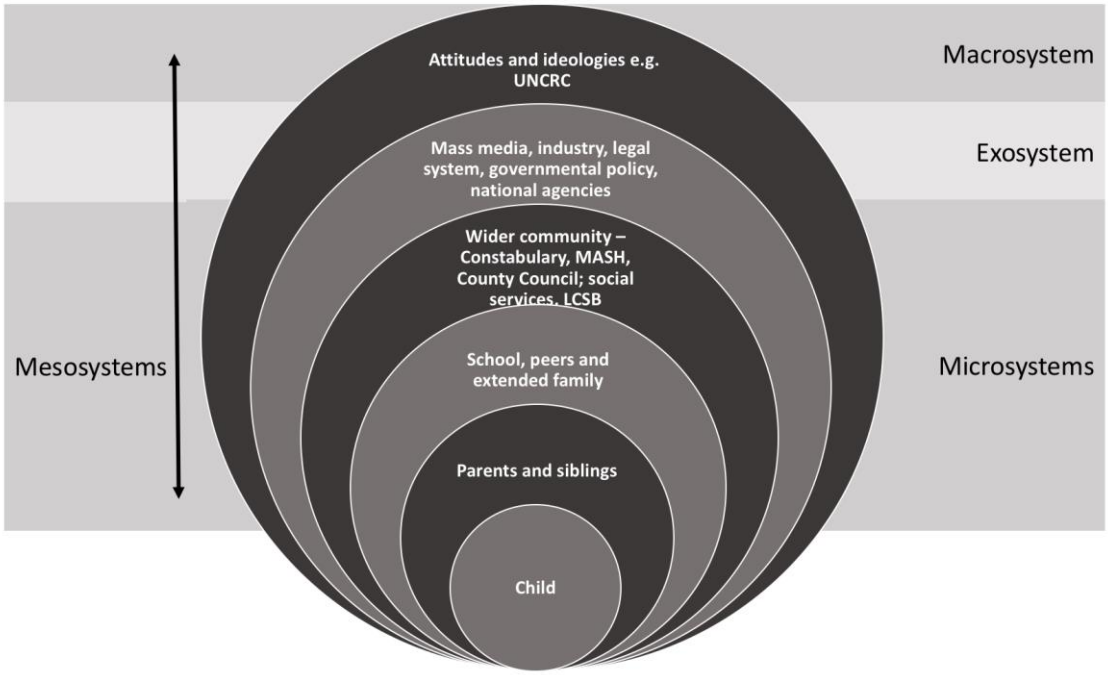
[2] Phippen, A. (2016). Children's online behaviour and safety: Policy and rights challenges. Springer.

[3] https://annualreport2022.iwf.org.uk/

[4] Wood, E., Werb, D., Kazatchkine, M., Kerr, T., Hankins, C., Gorna, R., Nutt, D., Des Jarlais, D., Barré-Sinoussi, F. and Montaner, J., (2010). Vienna Declaration: a call for evidence-based drug policies. The Lancet, 376(9738), pp.310-312.

voices and the biggest outrage tends to drive policy. While the online harms stakeholder space differs from drug harms, the fact remains that prohibition isn't working.

We proposed a stakeholder model that allows all who work in this space to acknowledge both the complexity of online safeguarding and the broad range of stakeholders around the child.



The model is an adaptation of the seminal ecological framework of child development by Brofenbrenner[5]. In his model, Brofenbrenner proposed an ecosystem of interconnections that facilitate the development of the child. There is no one independent entity that ensures positive development of the child, it is the interactions between the child, their immediate environments and wider systems that allows the child to thrive.

By adapting an ecosystem approach for child online safety, we can see the breadth of stakeholder responsibilities for healthy development and safeguarding, and the interdependencies between them. These characteristics will also influence their vulnerability and resilience online. The immediate spheres which influence a child's life and their landscapes of risk include their schoolteachers, peers, local law enforcement, social care, and local authorities as well as everyday physical and virtual environments. All have statutory responsibilities for the welfare and safeguarding of the child both online and offline. Industry, the focus of much policy direction around online harms, has no statutory duty, and exists in the broader ecosystem that provides infrastructure for the child's online experiences and the views of other stakeholders within the microsystems.

Often overlooked and depicted in the model as mesosystems (shown as the arrows in diagram) are the connections or relationships between the microsystems, such as the

[5] Bronfenbrenner, U. (1979). The Ecology of Human Development: Experiments by Nature and Design. Cambridge, Massachusetts: Harvard University Press. (ISBN 0-674-22457-4)

relationship between the parents and the school, but also essentially the relationships between the different stakeholders at different levels of the ecosystem. Conceptualising this as an interconnected system at all levels is important as one stakeholder cannot be isolated and held responsible for providing all the solutions, it must work with others. Consequently, interactions between individuals, groups and social structures are highly significant in improving children's safety and are thus highlighted as being everyone's responsibility.

Also fundamental to this model is the macrosystem that should define the overarching principles surrounding the child in both protectionist and participatory discourses. We, therefore, view the UN Convention on the Rights of the Child (UNCRC)[6] as a vital component of the macrosystem in that it provides extremely clearly defined universal principles, ratified by many governments. If all stakeholders within the systems ensure that the rights of children and young people are paramount in the development of policy, technological, legislative "solutions" and educational responses to keeping them safe online, we would hope that more holistic practices around online safety might be adopted.

This model will be core in the discussions in both this session and those later in the year.

Stakeholder Discussions

The remainder of this report presents the key issues discussed during the session. It is grouped into four key areas which were drawn from the overall scene setting goals:

- Education
- Rights
- Technology as a Solution
- Transparency

As is typical of these discussions, there was some jumping around different themes through the session. The summaries below are not in strict chronological order, but best summarise the discussions around these four key areas.

Education

Education was viewed across attendees as an essential part of keeping young people free from harm online, and there is a need for online harms policy to align far more closely with educational approaches. It is important to understand that, overall, most people (both young people and adults) are using technology for positive, helpful experiences and goals. Harms do occur, but they are in the minority. The "safety" metaphor is not helpful and public health models might provide better approaches to education in this area – young people learn about bodily health, but online health is not a concept that exists to any great degree in the online harms policy space. Instead, the messages are prohibitive and laden with blame. For example, "You shouldn't do this" or "you've been stupid if you are subject to this particular online harm". It was proposed by some attendees that a model that informs users of the risks associated with going online, and how to mitigate them, would be a more progressive and

[6] United Nations (1989). "UN Convention on the Rights of the Child". https://downloads.unicef.org.uk/wpcontent/

uploads/2010/05/UNCRC_united_nations_convention_on_the_rights_of_the_child.pdf

arguably more successful approach that aligns more strongly with what young people are asking for.

There was a clear view among participants that there are most of the messages for young people around technology are not positive, and tend to focus on potential harms. While there is a role for education around where the risks are and the potential harms, they should not swamp all the positive education around technology. Young people who understand both the positive and challenging aspects of technology, which align more realistically with their lived experiences, are far more likely to engage with other stakeholders (if they need to ask for help or disclose harm) and develop resilience. For example, the statutory guidance for Relationship and Sex Education (RSE) in the UK[7] says that opportunities as well as challenges should be covered, but the only curriculum guidance around online issues centres on challenges and legalities (for example accessing pornography or sending a nude is illegal).

The platform centric perspectives around online harm mitigation do little to empower a progressive educational approach, as it restricts the knowledge and understanding to the environment in which the harm takes place, rather than understanding how and why the harm exists. As a result, there is a lack of agency for any actors in this space and there is a need for stakeholders to get those informing policy to understand that the platform is only one element of the harms and to create opportunities to develop better, more holistic educational approaches. We would propose that the stakeholder model above would be a useful tool for this. However, it was observed, education is almost wholly missing from both political debate and policy direction.

It was also acknowledged that education is a challenge in the current political debates because it is both expensive and takes a long time to have an impact. With political cycles usually bound to election schedules, politicians need to show that "their" legislation is having an impact to improve the likelihood of re-election. Stating "we have improved education around online harms and in a few years, we will have a more resilient society" is not likely to achieve this. Furthermore, it can be argued that prohibitive educational messages are "easier" (albeit less effective) to deliver. Far easier to say, "Do not look at pornography, people your age shouldn't" than it is to say, "We'd rather you didn't but if you do see it these are the things you need to understand…".

Similarly, lazy discourse such as "You're addicted to that phone" are unhelpful because they flatten the narratives around online engagement and positive experiences. If one becomes addicted, they can now be excused for their actions. Furthermore, the concept of addiction in this area is still poorly understood yet used frequently by stakeholders around young people's safeguarding. Without clinical diagnosis, should we be using these terms so freely?

Parents as a stakeholder in this space was also frequently raised – there was a general view that public education is worse than statutory education, and there was a serious need for more effective, and responsible, public education in these issues. While there is emerging legislation that places public education expectations on regulators, there is little beyond that

[7] https://www.gov.uk/government/publications/relationships-education-relationships-and-sex-education-rse-and-health-education

they must do “something”. There is certainly little that considers what this “something” might look like and how it could be achieved.

Finally, an educational aspect that some thought was often neglected is that awareness of those training to become the engineers of these platforms around responsible, ethical, and legal practice. It was acknowledged that many will enter into technical disciplines because of their interest in the technology itself, and it was up to those developing their knowledge (for example, universities who provide computer science courses) to ensure they also learn about ethical practice and become thought leaders in tech policy because they both understand the ethical risks and also what technology is, and is not, capable of. It was the view of some in attendance that in the policy space one of the seldom heard voices was the technology community. Sometimes this is because anyone who opposes the latest “big policy ambition” is viewed as a dissenter who does not want further regulation (rather than what is true – their technical knowledge means they know it won’t be successful) but sometimes the technical voices do not speak up. Even in technology providers, those who work in policy areas are rarely from a technical background. A more ethically aware engineer base who understands how to engage in the policy space would help correct this.

Rights

Rights are frequently used in debates around online harms, yet attendees felt that a lot of the time these are not informed by participants who have a holistic view of rights. Rather, they are using rights to get their points across and justify their position. For example, they will use the UN CRC’s Right to Protection from Harm to justify an approach that will place more expectation on platforms to prevent harms without considering a platform centric solution might also impact on other rights such as participation or privacy. There seems, it was observed, a position that rights are used in isolation when it is convenient to make a point. This “rights conflict” can be seen in some of the current policy discussions. While there is traction in the UK Online Safety Bill to make End to End Encryption a problematic element of platforms that needs to be “mitigated”, the UK Information Commissioner’s Office has already stated that this technology is important to protect young people’s privacy rights as much as adults. Furthermore, the Online Safety Bill’s duty of care expectations on platforms potentially present some ethical issues in terms of excessive data collection – should platforms be collecting more data on children to ensure they are safe?

There was a clear view that if rights frameworks such as the CRC were being used in these debates there was a need for a greater knowledge and application of the whole thing, rather than just selective sampling. The UN CRC presents no priority around rights – all rights are equally important. We cannot have policy makers arguing to make young people safe they have to relinquish some of their privacy rights (because we need to make sure we can see they are safe). It is important that rights frameworks are used when considering online harms “solutions”, but they should be used in a holistic manner.

This is challenging for those in the policy space, and those impacted by policy changes, because understanding rights from a holistic perspective is complex, and people want simple solutions to these issues. However, better understanding will lead to more informed debates and solutions. There was a view from some participants that the use of the term “harm” is part of the challenge. Harm is vague and subjective, whereas if we can agree that rights frameworks should be followed (and most countries have ratified the UN CRC), using terms

such as *rights violations* first articulates that something we agree as nations has been violated, and also that it is something that should be dealt with. This is equally true for a use case that results in a violation to the right to protection as one that results in a breach of the right to privacy or participation. This is an excellent way of clarifying the nature of online harm, and harms in the rest of this discussion can be used interchangeably with rights violations.

Another area that garnered much discussion (and agreement) was the importance of youth voice across all stakeholders. While many stakeholders claim to speak for young people, fewer can evidence that they listen to young people. It is important that the best interests of the young person drive these discussions and move away from adultist perspectives around believing we know what is best for young people. Young people's authentic representation is important because we can show from academic literature over many years that the views of young people (better education, better support, knowledgeable adults with safeguarding responsibilities) have not changed. Which would suggest that previous "solutions" have not been effective.

It was suggested that an elimination or prohibition focussed approach is immediately doomed to fail – we cannot "solve" online harms, but we can work together to build better environments to allow young people (and adults) to interact in online spaces with an expectation and understanding around the potential harms that might occur and how they can mitigate them. While technical tools have a part to play in this, it needs to bring in all stakeholders. For example, age verification might be a tool that will make it less likely that a young person will come across pornography by accident, it is not the solution to stop all young people accessing this sort of content. Therefore, we also need education and awareness around this sort of content, rather than hoping we can stop young people seeing it.

The disputes between the polarised debates were once again highlighted as a challenge in this area – there was a need to move away from factions to understand that everyone is working for the same goal and even if there is disagreement with how to achieve this, no one was on the "other side". The challenges around an area such as CSAM regulation highlight this – at the present time in the UK the Online Safety Bill potentially places two regulators – OFCOM and ICO – on a collision course. One regulator will be expecting platforms to show how they can circumvent encryption to be able to identify illegal content, while the other will mandate the need to ensure they cannot examine content that will infringe on net neutrality and end user privacy. Once again, the challenge of progressive policy making was raised – it is easier for policy makers to call on platforms to "do more" than accept this is a complex environment that needs nuance and an understanding of rights and all use cases.

There was, frankly, little confidence that current regulatory proposals were going to do much to improve outcomes for end users, and young people in particular, because of the technocentric approaches being proposed. It was also observed that the discrepancies between large platforms, who would have many new duties under emerging legislation, and smaller ones, who will have no expectations, highlighted how challenging expecting platforms to solve these issues was. It seemed that the legislation acknowledged that smaller platforms did not have the resources to provide effective risk mitigations, which would conflict with governmental goals to encourage technology innovation, should they be expected to fall under the same regulation. Therefore, further fragmenting the stakeholder space between those who have duties, and those who do not.

It was suggested that the High-Level Working Group, with its multi-stakeholder participation (and its general consensus on how these issues are not easily tackled) might have an important role to play in engaging both industry and policy makers in approaching these issues from a rights based perspective for mutual benefit.

Technology As A Solution

Some of the discussion centred on the role technology can play. While it was generally agreed that a technocentric policy approach was not going to be effective (and indeed had been proven to be ineffective in the past) that was not to say that technology did not have a role to play. There was little opposition to the role of age verification systems as part of the ecosystem, and there was much confidence that age verification systems had improved both in terms of performance and privacy preservation in recent times. Certainly, contributions from those in this area reinforced the view that there can be an effective tool and the sector had done a lot of work with young people in developing solutions where cost was no longer a barrier to their use. However, it was still the view, including by those in the sector, that they would not ever be 100% successful in preventing access because there are so many means by which to bypass or simply use cases where they would not manifest (for example a parent leaving themselves signed in on an age verification platform). Certainly, there was a strong view that technology was important, if not viewed as the whole solution. However, there were certainly some examples of best practice across stakeholder groups that showed technology can be used in progressive ways to empower end users. StopNCII[8], a partnership between an online safeguarding charity – SWGfL[9] and Meta – enabled those who have become, or were at risk of becoming, victims of non-consensual intimate image sharing by taking agency for their images and ensuring they are not shared across partner platforms. Using well established hashing technologies to generate unique identifiers the victim can both retain ownership of their images and share these identifiers with platforms to ensure an abuser cannot upload onto these platforms. This is an excellent example of stakeholder working together and delivering on their key expertise, to empower those who might become victims of harm.

However, there were also concerns that, in the rush for technocentric solutions, there was a risk around excessive data collection, and sharing, which is perhaps not discussed in a transparent manner, and privacy infringement perhaps did not have the same visibility as an "online harm" compared to others that are more likely to gain media and political attention. Certainly, the coverage of privacy and issues of data agency and ownership rarely occurred within the existing school curriculum and there was little hope that there was any discussion that this might become the case in the future.

Transparency

The final main topic under discussion extended the conversation around technocentric approaches to challenge platforms on what was possible and to explore an area where there is certainly both potential for value for a wide range of stakeholders and an area where "doing more" was possible. Broadly, we refer to this as "transparency" – how platforms can be more

[8] https://stopncii.org/

[9] https://www.swgfl.org.uk/

open regarding the tools they provide and the processes they must safeguard their communities and how this might be used by stakeholders.

For some attendees this discussion was an opportunity to talk directly to people from a platform (in this case Meta) and challenge what was being done and how they might frame their activities. Certainly, there was concern that platforms might view privacy as something that they owned and could control, whereas some attendees saw privacy, quite correctly, as the right of the individual and something for which they should have ownership. Platforms should provide them with both the knowledge of how their data is used, and what tools the platform might provide to empower the end user to better manage their privacy. There have been cases in the past where platforms have attracted criticism and sanctions from regulators from the collection and processing of children's data.

There was a view that a more transparent perspective on privacy and a reclamation of the language around privacy would help inform a more evidence-based policy direction. While this was valuable for those in civic society, it could also be valuable for platforms – rather than being the focus of criticisms in harm reduction, they could provide detail on the measures they have in place and, perhaps more importantly, show how they are used. There is also a challenge around whether a platform knows a user is a child, given the challenges with age verification and the platforms talked about co-design approaches with young people to develop better age assurance and verification systems that will improve accuracy of age checking to ensure that the systems can be mindful if processing a child's data and therefore enhance practices around both how it is processed and also what safeguards are put in place for both exploitation of data and mitigation of harms.

From the perspective of the platforms, there was a frustration that they do a lot of safety and privacy related activities, and these were not always recognised. In summarising several initiatives and activities, Meta could articulate a holistic approach that considered not just safety but wellbeing and privacy as well. It was interesting to note that a lot of the platform's intervention relies on an informed user base who will make use of the reporting tools provided on their systems. While there are systems in place that can be pro-active in policing the platforms, it is far more effective if abuse and concern is reported so that it might be investigated by either algorithm or human moderator against the publicly available community standards. Some attendees found this particularly interesting given the reluctance of young people with whom they had spoken (and adults with safeguarding responsibilities) who either did not know about the reporting tools or did not use them because they did not think they would be effective. In terms of transparency having more data around reporting, takedowns and sanctions would be extremely valuable in making compelling cases to young people to adopt a more pro-active approach to policing their feeds and make use of the tools that are available.

The interaction between platform users and stakeholders for their safeguarding (predominantly parents) is also something of which platforms are mindful, however, they are also aware that a balance might need to be made – too much intervention or access for a parent might result in a restriction of a child's rights. Platforms are aware of this challenge and that parental controls and supervision that needs further discussion with experts and stakeholder groups.

Other facets of the safety features again required the end user to be honest about their age, or for platforms to have more accurate age assurance systems, so policies around things such

as ad restrictions and strong privacy settings by default could be enacted. Platforms also provide education centres for young people and parents and develop resources for teachers. However, it was acknowledged that these did require parents and young people to seek out these resources and once again highlight the challenges of public education if a stakeholder does not wish to engage with the information that is available.

The challenges of content moderation were also discussed which highlighted the challenges of algorithmic intervention on human communication, and the role human moderation might play at scale. Given earlier discussions around the use of technology as tools, rather than complete solutions, it was no surprise that content moderation is, by its nature, complicated and certainly not perfect, particularly when policing freedom of expression within different geographical locations whose definitions of hate speech will vary and, while AI based solutions continue to improve, they are still going to have large problems with false positives, particularly around identifying satire, sarcasm and other tones that might render a collection of "abusive" phrases an entirely acceptable post.

Both human and algorithmic moderation is subject to third party auditing, codes of conduct and an independent oversight board, the platforms recognised that, as we have discussed above, it is unlikely that moderation will ever be a complete solution without the intervention of the community making use of reporting tools.

It was also acknowledged that while human moderation was more accurate, there was also a duty of care by platforms for those moderators, as they would spend a lot of their working day looking at abusive and harmful content and this can have a serious impact on their wellbeing.

Returning to rights and the intervention of private companies on citizens' rights, it was acknowledged that in some cases this has to be defined by governments and not the platforms themselves. If a platform was too restrictive in their community standards it is unquestionable that policy makers would accuse them of restricting freedom of speech and expression, and, to return to the comments above around the intangibility of harm, it should not be the role of a platform to define what a harm is. Similarly, the transparency and reach of AI based moderation should not be the sole responsibility of platforms, for similar reasons – companies should be expected to work within agreed regulatory structures defined by governments, not simply told to "moderate your content to mitigate risk".

For some attendees the comprehensive nature of the measures put in place by platforms was a surprise (particularly for some who have worked for platforms in the past) and highlighted the need for industry to be more transparent about both the challenges of content moderation and what they are doing to mitigate impacts on safety and privacy. A more open dialogue with stakeholders means the broader safety and privacy ecosystem can better articulate what platforms are doing and "self-silencing" by industry (possibly because of too often being publicly admonished by policy makers and media commentators) helps neither them nor those advocating for progressive change.

Furthermore, some attendees still felt that industry had work to do in grasping privacy as something that belongs to the individual and that privacy by default did not mean anyone outside of the company cannot access the child's data, it meant that no one should be able to. There are still many conversations to be had around these issues, but the openness of platforms in this discussion showed what can be done if stakeholders are less guarded.

Closing

In closing the discussions, which had been wide ranging, what was clear was that there was a great deal of agreement that there is no simple solution to these issues and only through working across stakeholder groups can a better understanding of the issues, and solutions, be achieved. It was also broadly agreed that rights-based approaches are the most effective starting point for these discussions because, firstly, rights frameworks have been ratified by nation states and are far more tangible than the opaque concept of harm. In exploring what might be a starting point for the next session, which will move from context and scene setting to a more focussed exploration of rights and balancing responsibilities with capabilities to begin to frame a realisable progressive approach to tackling online harms, two key issues were raised. Firstly, who decides what is and is not safe for young people, and what is age appropriate? For example, there are many that claim that it is not safe for young people to use social media under the age of thirteen, failing to appreciate that the origins of this age limit lie not in child protection or development, but data protection and consent principles. Therefore, it is important to begin to challenge those who claim authoritative voices and ask for evidence of claims.

And finally, what does a taxonomy of risk and harm look like? It had been mentioned several times throughout the discussions that harm seemed intangible and could not be easily defined or applied. We would propose that defining harms as rights violations is a good starting point for this, in that rights are well defined and cover a broad range of issues and can be applied equitably. This is something that will be developed further at the next session.

Conclusions

This report presents discussions by the High-Level Working Group For Privacy and Safety. It shows a broad stakeholder group committed to achieving a more progressive and evidence-based approach to tackling online harms, underpinned by rights and youth voice.

While the discussions during the session were wide ranging, there are several key issues arising. Firstly, this is complex! In contrast to a lot in the policy area, there was no one who believed this was simply a case of putting some technology in place to identify and prevent harms. Even the concept of harm was challenged – who decides what is and what is not harmful? Far better, it is proposed, to look at harms as rights violations rather than a subjective intangible statement on harm. And rights violations can occur around any established right – there should be a holistic approach to understanding rights and how they are applied in this space, rather than cherry picking specific rights to justify an approach.

There was a large consensus that the key to success is an educated and informed stakeholder group, from young people themselves to those who support them, and those who develop the systems upon which online interactions occur. There was a consistent view, once again, that rights should underpin educational approaches, along with pragmatic views that online harms cannot be prevented, but end users can be informed about potential risks and how to mitigate them.

There was also agreement that regulation should be broader than a duty of care model for industry. While industry is certainly an active stakeholder, there are many others, and current regulatory proposals do little to engage the wider stakeholder group or place expectations on

governments or regulators to define literacy frameworks and education approaches. While technology can be a useful tool, it will never be a solution.

Finally, in terms of what platforms can do more of, rather than focussing on unachievable technical prohibition, it was agreed that greater transparency of process and evidence on how these processes could be used to more effectively inform the stakeholder space which will, in turn, allow more positive education interventions to take place.

The next session will take place in September 2023, and will focus further on rights-based approaches and the balance between responsibilities and capabilities.

In Attendance

Alessandra Tranquilli, We Protect Global Alliance

Andy Lulham, VerifyMy

Andy Phippen, Bournemouth University

David Miles, Meta

DaYoung Too, Meta

Diana Gheorghiu, Child Rights International Network

Eleanor Linsell, We Protect Global Alliance

Emily Setty, University of Surrey

Emma Bond, University of Suffolk

Emma Nottingham, University of Winchester

Iain Corby, Age Verification Association

Janice Richardson, Insight

Jen Persson, Defend Digital Me

Jonny Hunt, University of Bedfordshire

Jonny Shipp, LSE

Julie Dawson, YOTI

Karina Stan, Developers Alliance

Tim Jacquemard, Trilateral Research

Victoria Baines, Greshams College