CHAPTER 2

The Online Safeguarding Landscape

Abstract The "online safety" policy area is beset with preventative views and an overreliance and expectation for technology to prevent harm. Current online safety policy has a history of prevention that can be seen to have its roots in controlling access to pornography. Legislation that has arisen in the last ten years has similarly adopted a preventative approach, yet young people consistently tell us that "online safety" is either boring or ineffective. Even the term "online safety" is doomed to fail—we cannot ever hope to prevent harm online; however, we can equip young people with the knowledge to understand, and mitigate, risk when online.

Keywords Online safety · Digital resilience · Algorithms · Critical thinking · Stakeholder perspectives

This chapter places the Headstart digital resilience workpackage in the wider context of online safeguarding. This allows us to then explore findings from conversations with young people against the current "leading edge" of policy thinking. Drawing extensively from UK legislation and policy, this chapter presents an analysis of the online harms agenda—a safeguarding approach that endeavours to keep children and young people safe online by preventing them from being exposed to adultist

A. Phippen and L. Street, *Online Resilience and Wellbeing in Young People*, Palgrave Studies in Cyberpsychology,
https://doi.org/10.1007/978-3-030-88634-9_2

definitions of harm, with an expectation that platform providers can proactively tackle any potential harmful activity via algorithmic means. The critical analysis of these approaches, drawing upon previous work in the field, demonstrates the lack of youth voice, and the complete disregard for their rights, in this policy direction.

What Do You Mean by Safe Anyway?

In one of our early discussions, a young man challenged the notion of safety online, which brought the reality home to us:

> What do you mean by safe anyway?

His view, articulated very clearly, was that we cannot ensure someone can go online without being presented with some risks—he spoke about gaming with people one doesn't know and the risk they might be abusive, group chats where someone could say something mean and seeing upsetting content when browsing for other things.

In this young man's view, you could not prevent these things from happening when going online, but you can help young people understand that these things might happen, and help them if they are upset when they do. This was not a view that said we should stop trying to talk about online risk, because of course this is important. However, he was of the view that you cannot prevent online harm and pretending you can does not help young people.

One of the fundamental challenges we will explore within this book, drawing extensively on our discussions with both young people and professionals, is that we start from a position of prevention with the term "online safety". We have used this a number of times already within this book, and the use of quotes is deliberate. A lot of online safeguarding discourse draws analogies from road safety—we have frequently heard comments about how "we teach them how to cross the road safely, and we should do the same for being online". However, this is applying an adultist perspective on safety and attempted to transfer it onto a domain where it is inappropriate.

Let us consider road safety, there are few threats but the main one is serious and can cause a young person serious harm—if they are struck by a car while crossing the road. Therefore, we can put simple rules in place to mitigate this risk. We can tell children to look both ways before

crossing, make sure they have clear view up and down the road, listen out for traffic and make use of the tool available within the road environment (such as pedestrian crossing systems) to further mitigate that risk. The focus is entirely upon the prevention of an accident between child and motor vehicle.

If we compare this with the online world—firstly what are the threats? There are many and range from exposure to upsetting content, abuse by peers, unsolicited sexual contact by predators, non-consensual sharing of indecent images, being hacked and having identity data from being shared and so on. And, in contrast to the road environment, which is well controlled, with established standards (e.g. cars travel on roads, pedestrians travel on pavements) and a stable environment (it would be unusual to wake and discover we had decided that cars should now travel on the opposite side of the road to the day before), new online risks emerge as the digital infrastructure evolves and develops. Unlike the "atomic" world of roads, digital environments have few boundaries other than the ever expanding capacity of networks upon which all online services operate, and the imaginations of the developers who put services and platforms in place for billions of citizens to use, which poses the question:

- *What rules can we put in place to make sure a child is* ***safe****online?*

Preventative Approaches to Online Safety

If we take a broadly accepted definition of safety—that something is free from risk or harm—we are sadly chasing a utopian goal that will never be achieved in the online world. There is a way to ensure a child is safe online—we take their digital devices away from them and make sure they have no means to be online. Therefore, they will not be exposed to the risk that exists there. However, we will undoubtedly also be preventing them from the many positive experiences that can be delivered online. So, we reject disconnection as a viable safety route and instead bring other preventative measures to the online safety conundrum. For example, two popular preventative views are:

- *We wish to stop young people seeing upsetting content. Lets filter content to stop the young people seeing it.*

- *We wish to ensure a young person isn't taking and sending intimate images. They should be told its illegal and they should not do it.*

However, if we begin to unpick these issues, we get ourselves into a further tangle. If we wish to prevent access to "harmful" content through some sort of filtering, we need to understand what we mean by this. The Reporting Harmful Content service,[1] provided by the UK Safer Internet Centre to support young people who have been exposed to harmful content, details harmful content in a number of categories:

- Threats
- Impersonation
- Bullying and harassment
- Self-harm or suicide content
- Online abuse
- Violent content
- Unwanted sexual advances
- Pornography
- Terrorist content
- Child sexual abuse imagery

It is not the intention of this book to now consider each of these forms of content, the capability of algorithms to detect it (this is done in far more detail in Phippen & Brennan, 2019) or further approaches that could be adopted to prevent access. However, we will briefly explore a perennial favourite in the online safeguarding policy world—preventing access to pornographic imagery.

Again, a noble cause—while young people we speak to are generally of the view that they are comfortable with (or at least resilient to) pornographic content, they also invariably have the view that access is something we should prevent for younger children.

This is typical for most, regardless of age—they believe they are fine, but those younger would not be. This is a manifestation of the *third person effort* phenomenon (Davison, 1983), a belief as a subject that they are fine, but others may be affected or harmed more significantly. While the origins of the theory lie in mass media communication, it has

[1] https://reportharmfulcontent.com/. Accessed August 2021.

also been applied to subject matter as diverse as hip-hop lyrics (McLeod et al., 1997), violence on television (Hoffner et al., 2001) and online pornography (Lee & Tamborini, 2005).

In our wider experiences in discussions with young people (e.g. see Phippen, 2016), there are on occasions indications of the negative impact of pornography access on young people. We have met young people with performance and size anxiety which we would hypothesise is related to exposure to pornography. Equally, a lot of young people we have spoken to have stated that they believe it gives unrealistic expectations around sexual activity. While causation is difficult to prove (e.g. see Horvath et al., 2013), there are few that would argue for unrestricted access to pornography for young people.

However, prevention is a challenge, as can be seen from efforts for well over ten years to address this problem. The "solution" over this time has been filtering technologies, which make use of software that can identify pornographic materials and prevent access either through matching website addresses or identifying sexual keywords on a website. Once the filtering algorithm has identified a website is providing pornography content, it will block it. While this is a well-established practice in schools (and a statutory expectation as defined in Keeping Children Safe in Education (Department for Education, 2021), the UK Government document that defined safeguarding expectations on schools in England and Wales), the social/home environment presents some challenge. Overblocking is a fundamental challenge with filtering systems—they will block websites that are not providing access to pornography but instead are using similar sexual keywords—for example sites that might support relationships and sex education. While this is an accepted part of internet access at school (where systems can be modified to "bypass" filters to access educational resources), overblocking in the home environment can be more frustrating, particularly when parents will neither have the time or knowledge to manage their filters at a fine level of detail.

Digital technology is very good at clearly defined rule-based functionality in easily contained system boundaries. Or, to put it another way, data processing, analysis and pattern matching of data. Computers are very good at taking data and analysing it based upon rules defined within the system (e.g. identify words that *might* relate to sexual content). However, they are far less good at interpretation and inference or, to use a current popular term for these sort of systems, intelligence.

By way of an illustrative, albeit trivial, example, let us take the word "cock". This is a term that might be related to a sexual context—it could to male genitalia. Equally, it might refer to an avian animal. If we consider this from the perspective of a filtering system, that might be tasked with ensuring an end user cannot access websites of a sexual nature, that system might be provided with a list of keywords that could indicate sexual content. It would be expected that "cock" would be one of these terms. The filtering system would be very good at pattern matching this string of characters to any mentioned within any given website and would, as a consequence, decide the site contained pornography content and block it. We use the term "decide" advisedly—the algorithm has no capability to make a decision in the way a human might, it is merely responding to rules coded into it by a developer. As such, the algorithm is far less good at determining the actual context of the website—it *might* be about sexual activity; however, it might also be about animal husbandry.

Even with this simple example, we can see how it might struggle to prevent access to all sexual content or, equally, result in *false positives*—blocking innocuous[2] sites that are not "inappropriate" for children to see. Another simple and popular example of this comes from the overblocking of the Northern English town of Scunthorpe (Wikipedia, n.d.).

The flaws in filtering systems, still viewed as the best approach to preventing young people from accessing pornography, have attracted the attention of the United Nations, who have concerns that while being successful at preventing access to some pornography (but providing no barrier to pornography shared on social media or peer-to-peer communication) they might impact significant upon human rights. The "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression" in 2018 (United Nations Human Rights Council, 2018) stated that:

> States and intergovernmental organizations should refrain from establishing laws or arrangements that would require the "proactive" monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship.

[2] We will use the term "innocuous" sites to describe those who have been incorrectly blocked based upon the requirements of the filter (e.g. pornography, gambling, drugs and alcohol) and not "legal", because access to pornography is legal in the UK.

We have, in other work (Phippen, 2016), raised concerns about the *safeguarding dystopia*, where an overreliance on technical solutions and an obsession with prevention leads to a safeguarding environment that fails to support young people and instead impacts negatively on their human rights and freedoms.

This brief review looked at prevention of access to one type of fairly unambiguous "harmful" content—pornography. The challenges for algorithmic intervention increase further when considering even more subjective and ambiguous content, such as violent or terrorist content. We do not plan to elaborate upon further technical interventions here, but we hope the point we are making is clear—preventing access to all of these types of harmful content is perhaps not the most progressive approach to supporting children and young people in their online experiences and it is certainly not a complete or particularly successful approach.

We take the second example, which considers the sending of a self-produced intimate image by a young person using their mobile device. We have, in Chapter 1, explored a response to this type of incident with one school. And while the response is not one we would support, we are sympathetic, because with this scenario, there are further challenges.

By the letter of the law (specifically, section 1 of the Protection of Children Act 1978 [UK Government, 1978]), any young person taking an intimate image of themselves, and sending it to someone else, is breaking the law. Of course, this legislation was never intended to address this scenario—it was developed to protect young people from exploitation in the production of "pornography" (more correctly child abuse imagery—for further detail, see Phippen & Brennan, 2020). However, there is no provision in the legislation to say "if the subject is also the taker of the image and sender of the image, and is a minor, this is not a crime", and it has never been modernised, so the law still stands. We know, from many conversations with young people (explored in more detail throughout this book), that this is the key educational message delivered to them—"don't send nudes, it's illegal". If a minor is subsequently subject to abuse as a result of the image being non-consensually shared further, there is a serious and legitimate safeguarding concern. It would seem, in our experiences, that it is the intention of professionals that the mere mention of the legality of the practice is a preventative tool to eliminate this potential harm and keep young people safe.

We would, in an ideal world, wish for the young person to disclose this non-consensual sharing and be supported by safeguarding professionals.

Indeed, if the victim of non-consensual sharing was an adult, this is exactly what they could do, with protection in law (section 33 of the Criminal Justice and Courts Act 2015 [UK Government, 2015]). However, the fact the young person is breaking the law by taking the image in the first place (under the 1978 legislation) and they have been told this through their school life, are they likely to disclose, or are the more likely to suffer in silence?

The Youth Voice

Young people have a different perspective. When conducting research projects in 2009 and 2021 around teen sexting (Phippen, 2016), one question we posed for our survey respondents was "what can adults do to help if someone is upset by a sexting incident". The three most popular responses were:

- Listen
- Don't judge
- Understand

While the initial work in 2009 was survey based and did not provide us with the opportunity to explore these responses in depth, qualitative discussions in 2012 did. What was clear was that the wish for young people was they wanted to be able to disclose harm and get help, not a telling off or judgemental statements like "You shouldn't have taken those images in the first place". Such attitudes exist in other youth-focussed studies around sexting, such as Emily Setty's highly young person centric work (Setty, 2020). And we still see these wishes with the conversations we draw from the Headstart Kernow work, presented in Chapters 4 and 5.

Online Safety Policy

Online safety has, arguably, existed as a safeguarding requirement in schools for fifteen years, but did not become part of any statutory framework until more approximately nine years ago. The two major changes to this online safety landscape have been the inclusion of online safety as part of the OFSTED, the schools regulator of England, inspection framework

in 2012 (OFSTED, 2013) and its inclusion in the Department for Education's (DfE) Keeping Children Safe in Education statutory guidance since 2015 (UK Government, 2021). If we consider the latest requirements regarding online safety in school settings from the Department for Education, we can see there are requirements around training:

> 14. All staff should receive appropriate safeguarding and child protection training (including online safety) at induction. The training should be regularly updated. In addition, all staff should receive safeguarding and child protection (including online safety) updates (for example, via email, e-bulletins and staff meetings), as required, and at least annually, to provide them with relevant skills and knowledge to safeguard children effectively.
>
> 89. Governing bodies and proprietors should ensure an appropriate senior member of staff, from the school or college leadership team, is appointed to the role of designated safeguarding lead. The designated safeguarding lead should take lead responsibility for safeguarding and child protection (including online safety).
>
> 117. Governing bodies and proprietors should ensure that, as part of the requirement for staff to undergo regular updated safeguarding training, including online safety (paragraph 114) and the requirement to ensure children are taught about safeguarding, including online safety (paragraph 119), that safeguarding training for staff, including online safety training, is integrated, aligned and considered as part of the whole school or college safeguarding approach and wider staff training and curriculum planning.

Management of risk:

> 128. Whilst considering their responsibility to safeguard and promote the welfare of children and provide them with a safe environment in which to learn, governing bodies and proprietors should be doing all that they reasonably can to limit children's exposure to the above risks from the school's or college's IT system. As part of this process, governing bodies and proprietors should ensure their school or college has appropriate filters and monitoring systems in place. Governing bodies and proprietors should consider the age range of their children, the number of children, how often they access the IT system and the proportionality of costs vs risks.
>
> 129. The appropriateness of any filters and monitoring systems are a matter for individual schools and colleges and will be informed in part, by the risk assessment required by the Prevent Duty. The UK Safer Internet Centre has published guidance as to what "appropriate" filtering and monitoring might look like

And curriculum:

> 119. Governing bodies and proprietors should ensure that children are taught about safeguarding, including online safety, and recognise that a one size fits all approach may not be appropriate for all children, and a more personalised or contextualised approach for more vulnerable children, victims of abuse and some SEND children might be needed

However, there is nothing in the document that defines *what* online safety training or curriculum should look like (non-statutory guidance from the DfE on teaching online safety was released in 2019 (Department of Education, 2019). The management of risk centres mainly on ensuring appropriate technology is in place to make sure inappropriate content cannot be viewed, and online activity is monitored with appropriate alerts are in place should abuse occur. Furthermore, while we know that Keeping Children Safe in Education makes it clear that online safety should form part of whole school safeguarding training, we know from other work (SWGfL, 2021) that 40% of schools (in a sample of 12,000 schools) have no training in place.

Further clarification of the view of online safety (and safeguarding) from the policy perspective could be seen in 2018's Online Harms White Paper from the Home Office and Department of Culture, Media and Sport (UK Government, 2018), which defined a large list of potential harms that can occur online, and proposed a legislative framework and expectation on service providers to mitigate harm. In essence, online safety has become a preventative and prohibitive method of ensuring young people are free from harm through a mix of control, filtering and poorly defined education. Yet with poorly defined expectations, we cannot be surprised that young people's views and experiences with online safety can vary immensely and professionals view online harms as something that need to be stopped, rather than mitigated or managed.

Nevertheless, the preventative approach continues and arguably becomes strong. At the time of writing, the UK Government has published their draft Online Safety Bill 2021 (UK Government, 2021) and it is portrayed in the media as the UK Government's crowning glory in making "Britain the safest place to go online in the world". While we will explore the Bill in a little more depth in this book, it is not intended that there will be a detailed examination of all 145 pages of the draft bill.

However, it does illustrate once more national policy level thinking on what online safety looks like. The prevailing view is that the heart of online safety is a "duty of care" for online service providers. It is down to them, and the bill is clear, to make sure citizens in the UK are not exposed to illegal material and what is also referred to as "legal but harmful". It is clear that companies that cannot demonstrate duty of care will be found liable for abuse that happens on their platforms, however complex this negligence might become.

What is particularly unclear is whether the duty of care in the bill being defined as related to a form of negligence as defined in civil law? If so, how might the company be able to demonstrate due diligence or protect itself from vexatious or unsubstantiated claims of harm? It would seem, however, that the government is indeed introducing failure to protect from online harm as a form of negligence for which one might make civil claim. While we anticipate much contested legal debate on liability, given what is actually possible through the tools available to platforms (such as algorithmic detection and reporting tools), it is clear the expectation by government is that harms can be stopped.

Changing the Perspective

If we now return to the conversation that started this book should we really be surprised that this teacher has this view. Given the nature of online safety is one that has been beset with preventative messages, and the "leading edge" of policy thinking is further efforts at prevention, why should a professional not think that prevention is the best approach? Of course they will bring their own lived experiences to their views they develop, especially given the dearth of training available to professionals. Without effective training, or training that perhaps reinforces preventative messages, the gaps in knowledge will be filled with conjecture and existing biases.

We will explore this in more depth in later chapters, but one of our key observations from our work across the project is that professionals will bring their social and family experiences into their professional judgements. Given that we use digital technology in our own social and personal lives, we can bring this into our professional expectations. This is, of course, quite inappropriate for a professional safeguarding judgement, particularly given that most knowledge developed around the use of digital technology is done so in an informal and ad hoc manner, but it

is something have consistently observed through this work. To continue with road analogies, we would not expect one's capability to drive a motor vehicle be a good foundation knowledge to diagnose a serious mechanical issue with a school minibus. Or, to put it another way, our social knowledge is no substitute for professional development.

In conclusion, in this chapter, we have explored the initial goals of the Headstart Kernow project against the broader domain of online safety policy. While the goals of the project were to be inclusive and strongly represent the youth voice, online safeguarding policy, and the views of professionals who work in the domain, is strongly preventative.

While there are good intentions with these views, they immediately create a tension between those wishing to protect and those who might need protecting. This tension is further tightened with a dearth of professional knowledge around online safety being in filled with opinion and conjecture, such that there is a belief that online technology has a negative impact upon young people's wellbeing, without having the empirical evidence to underpin this view. In the following chapter, we develop the source of this tension further, with a reflective exploration by one of the authors of this book (Louisa) in over ten years of experience as a youth worker.

References

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly, 47*(1), 1–15.

Department for Education. (2019). *Teaching online safety in schools.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811796/Teaching_online_safety_in_school.pdf. Accessed August 2021.

Department for Education. (2021). *Keeping children safe in education 2021—Statutory guidance for schools and colleges.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/999348/Keeping_children_safe_in_education_2021.pdf. Accessed July 2021.

Hoffner, C., Plotkin, R. S., Buchanan, M., Anderson, J. D., Kamigaki, S. K., Hubbs, L. A., Kowalczyk, L., Silberg, K., & Pastorek, A. (2001). The third-person effect in perceptions of the influence of television violence. *Journal of Communication*, *51*(2), 283–299.

Horvath, M. A., Alys, L., Massey, K., Pina, A., Scally, M., & Adler, J. R. (2013). *Basically... porn is everywhere: A rapid evidence assessment on the effects that*

access and exposure to pornography has on children and young people. https://eprints.mdx.ac.uk/10692/1/Basica. Accessed August 2021.

Lee, B., & Tamborini, R. (2005). Third-person effect and internet pornography: The influence of collectivism and Internet self-efficacy. *Journal of Communication*, *55*(2), 292–310.

McLeod, D. M., Eveland, W. P., & Nathanson, A. I. (1997). Support for censorship of violent and misogynic rap lyrics: An analysis of the third-person effect. *Communication Research, 24*(2), 153–174.

OFSTED. (2013). *Inspecting eSafety*. https://www.eani.org.uk/sites/default/files/2018-10/OFSTED%20-%20Inspecting%20e-safety.pdf. Accessed August 2021.

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

Phippen, A. (2021). *UK schools online safety policy & practice—Assessment 2021*. https://swgfl.org.uk/assets/documents/uk-schools-online-safety-policy-and-practice-assessment-2021.pdf. Accessed August 2021.

Phippen, A., & Brennan, M. (2019). *Child protection and safeguarding technologies: Appropriate or excessive 'solutions' to social problems?* Routledge.

Phippen, A., & Brennan, M. (2020). *Sexting and revenge pornography: Legislative and social dimensions of a modern digital phenomenon*. Routledge.

Setty, E. (2020). *Risk and harm in youth sexting: Young people's perspectives*. Routledge.

UK Government. (1978). *Section 1 Protection of Children Act 1978*. https://www.legislation.gov.uk/ukpga/1978/37/section/1. Accessed July 2021.

UK Government. (2015). *Section 33, Criminal Courts and Justice Act 2015*. https://www.legislation.gov.uk/ukpga/2015/2/section/33/enacted. Accessed August 2021.

UK Government. (2018). *Online harms white paper*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf. Accessed August 2021.

UK Government. (2021). *Draft online safety bill 2021*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf. Accessed August 2021.

United Nations Human Rights Council. (2018). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement. Accessed August 2021.

Wikipedia. (n.d.). *The Scunthorpe problem*. https://en.wikipedia.org/wiki/Scunthorpe_problem. Accessed August 2021.