Written evidence submitted by Professor Andy Phippen, Professor of Digital Rights, Bournemouth University (OSB0121)

This is a response to the Draft Online Safety Bill (Joint Committee)'s call for evidence. While I will address some of the committee questions directly in this submission, I will begin by reflecting more broadly on the aims of this piece of legislation.

My evidence is based upon my experiences researching and engaging in practice around online safeguarding for around 20 years, with a broad stakeholder view and a central focus on youth/victim voice (for example see [1],[2]). I can appreciate that the Bill is addressing a need, and is an attempt to eliminate the piecemeal approach to tackling issues such as harassment, hate speech and abuse online.

At present there are a number of different pieces of legislation that might be able to tackle some of these, but they are often adapted from original intention, built upon thresholds and case law, interpretation by the police and CPS guidance which all together means they can frequently be applied inconsistently and victims are left frustrated or even criminalised themselves (if consider the application of Section 1 of the Protection of Children Act 1978 to the prosecutions of minors for self-produced imagery). We are seeing legislation creaking to keep up and this new legislative approach may mean that a lot of the concerns around user-to-user online abuse can be addressed within a single Act.

The Online Safety Bill feels like a culmination of a (approximately) ten year policy journey that has in its roots the honourable goal of preventing access by minors to pornographic content - a wish few would disagree with but one that is not without its challenges in a globally interconnected system. Along the course of this journey, there have been attempts to use technology to restrict access with some success (such as filtering on public WIFI) and some solutions somewhat inconsistently applied (such as default filtering on home ISP connections - still only used by a minority of households according to OFCOM[3] possibly due to the problematic nature of internet filtering as a content blocking mechanism). Furthermore, there have been some aspects of legislation that have fallen by the way side, such as Part 3 of the Digital Economy Act 2017 – the "age verification" legislation that aimed to force pornography providers to put up age verification gateways for any UK users.

The reason I make this point is to demonstrate that this is not the first attempt to regulate online behaviour via platforms and software and that legislation expecting technology to manage social issues is not without challenge. Whenever considering these legislative attempts to bring technical solutions to safeguarding issues I am reminded of Ranum's Law[4]:

You cannot solve social problems with software.

---

[1] Phippen, A. (2016). "Children's Online Behaviour and Safety: Policy and Rights Challenges". Palgrave.

[2] Phippen, A. and Street, L. (2021). "Online Resilience and Wellbeing in Young People – Representing the Youth Voice". Palgrave MacMillan.

[3] https://www.ofcom.org.uk/__data/assets/pdf_file/0025/217825/children-and-parents-media-use-and-attitudes-report-2020-21.pdf

[4] https://en.wikipedia.org/wiki/Marcus_J._Ranum

However, software can play a role and provide tools that support the stakeholder space. Overall, my thought on the Online Safety Bill is that it is valid attempt to address social problems through platform responsibility and software intervention. It is welcome that the legislation does not attempt to cast every aspect of content moderation and behaviour management in stone and has, instead, appointed a regulator who can make those judgements, based upon the requirements set out in the legislation. This is an innovative and welcome approach, and means the legislation is, to a certain degree, future proof from the evolution of technology and platforms.

The risk assessment and reporting model that is central to the legislation (and platform liability) have the *potential* to be fair and effective – there is nothing fundamentally concerning about providers demonstrating their risk assessments in considering how they support the "safety" of users and providing a level of accountability for those less scrupulous providers who do not see the wellbeing of their platform users as their concern. However, the use of the double quotes around the word safety is deliberate.

The concept of being safe (i.e. free from harm or risk of harm) online is something of a utopian goal and a throwback to poorly comparable concepts such road safety. One cannot make people safe (i.e. free from risk of harm) online, and using this terminology distracts from things we can do, which is to build knowledge and resilience to risk and provide the tools to mitigate potential harms. While some aspects of problematic social policy have steered down more of a harm reduction approach, the discourse of safety around online harms is somewhat problematic.

If we take by way of example, a pornography provider. They wish to be seen to be a responsible platform and ensure all users of its services are adults. While their age verification measures will mean they can verify the age of *many* potential users, it will never be a completely effective system.

It would seem unfair to expect companies to invest in risk assessment and subsequent software tools to implement risk mitigation measures, just to be told they will still be fined as they failed in their duty of care to let in a determined, duplicitous young person who decides to bypass age verification measures with techniques such as the use of a parent's login, a Virtual Private Network or a Tor browser. Age verification in a nation which does not have a consistent means of age ID for all citizens is problematic. If the Government is serious in its view the comprehensive age verification has to be part of the online safeguarding toolkit, and placing expectations on companies to implement a fool proof system, they should at least provide the underpinning infrastructure, which would be something like a national ID card scheme, and all of the privacy concerns and debates that brings.

In addressing a number of the consultation questions specifically:

Will the proposed legislation effectively deliver the policy aim of making the UK the safest place to be online?

Geographically isolated online practices are always hard to achieve, due to the inherent global nature of the technology upon which they run. Geographical boundaries present problems. John Perry Barlow's Declaration of the Independence of Cyberspace[5], written as a reaction to the US' Telecommunications Act 1996, illustrates this point, albeit from perhaps an extreme position. While Barlow's view was that governments should not even try, I feel it is entirely reasonable to bring in legislation that tackles online abuse among their citizens, but a technical focus for this can lead to practical challenges, such as the already mentioned VPNs, Tor browser, or other privacy enhancing features, that will hide a user's geographical location.

A goal of "*making the UK a place where citizens can expect to be provided with useful and relevant tools to mitigate risk online and where providers are expected to have carried out effective risk assessment to consider the potential harms on their platforms for UK citizens*" is a more realistic goal, but is hardly a headline grabber.

Will the proposed legislation help to deliver the policy aim of using digital technologies and services to support the UK's economic growth? Will it support a more inclusive, competitive, and innovative future digital economy?

I do not see much in the bill that will encourage this, unless (and this is of course possible), responsible online service provision becomes a selling point for platforms. Of course, this is a challenge, given that there is rarely anything platform specific that facilitates abuse, abuse is performed by platform users who are probably there because of the access to an audience or peers.

Furthermore, the bill will result in risk assessment measures and the implementation of tools to help end users mitigate potential risks on the service providers platforms that will require the significant investment of resource by the companies. While it is, unquestionably, fair to expect companies to provide end users with the tools to mitigate risk and disclose upset or harms, it will raise the cost of entry to such markets and present challenges for start ups.

The threats of liability and prosecution in the UK if one is to establish a new online service might be enough to move start ups to different countries where liability is less well defined. Depending on the tangibility of proposals being made around duty of care, we might also expect providers to be making far greater use of legal services to both demonstrate duty and challenge rulings against them.

Again, this is a cost that has rarely been considered in technology start ups to date. This is not to say they should not have been, this is just to say it is rarely foremost in the minds of young developers with a "hot" idea looking for investment. Perhaps a side issue but we only need look at computer science curriculum across the higher education sector to see a dearth of teaching around ethical practice and platform liability. And higher education is a key talent pipeline for tech start ups.

[5] https://www.eff.org/cyberspace-independence

Are children effectively protected from harmful activity and content under the measures proposed in the draft Bill?

The measures will undoubtedly mean that providers have to demonstrate they have considered the potential harms their services might provide if used by children. Of course the main route to risk mitigation for platforms is technical tools, whether these are preventative (for example content detection to determine whether a post is appropriate for a minor in terms of sexual or hateful content) and reactive (tools to report, block, mute, etc.). When considering whether children can be "effectively protected" from harmful activity and content again highlights the flaw in using safety, rather than risk mitigation or harm reduction, as the underlying principle. If we wish to ensure children are "effectively protected" we need to understand what causes upset for young people online.

In a survey of over 8000 young people I conducted for the SWGfL[6] (a highly regarded online safety charity) the evidence suggested there is a great deal of different behaviour and content that might upset young people. Some, such as sexual content and peer on peer abuse, is typical, but there were many disclosures around current affairs causing upset (the Manchester Arena bombing and the murder of Lee Rigby are frequently referred to by young people as upsetting online content). If we are to take a position of "can we prevent upset for children online?", no we cannot, because what causes upset for children online is far ranging and individualised.

We must also consider that the paths open for providers are generally technical in nature. In my own extensive work talking with young people it is fair to say the majority do not believe in the efficacy of technical tools to help prevent harm online. So they do not use them.

This is why one of the less discussed aspects of the Bill - Transparency Reporting, which calls on providers to publish details of reports of harmful content/number of takedowns/account suspensions/etc – should be seen as a positive move in empowering users. We know that many users, both adults and minors, will not report abuse because they do not believe anything will be done. Greater transparency around take downs and end user sanctions should give evidence and therefore confidence around user reporting services provided by platforms. And empowering the regulator to compel the less open providers to do so is another positive move.

However, I feel one of the fundamental failings of the Bill lies in the facet of safeguarding that young people have consistently called for in all my time researching this area – better education.

Through the 145 page bill, there are only two mentions of the word "education", in section 103, which relates to media literacy and the role of the regulator in delivering public awareness programmes. This is disappointing given the importance of multi-stakeholder engagement around online safeguarding issues. It is not enough for everyone to demand that platforms tackle online harms, there is also a role in delivering effective education in schools and having safeguarding professional receiving relevant and rigorous training so

[6] https://swgfl.org.uk/assets/documents/what-causes-upset-online.pdf

they can support those at risk of harm and understand how tools provided by platforms might be used more effectively. Platforms can have a role to play in this (for example, providing resources around how to use their safety tools, and information about their transparency reporting), but they cannot do it on their own.

Does the draft Bill make adequate provisions for people who are more likely to experience harm online or who may be more vulnerable to exploitation?

While children are well considered in the, there are no mentions whatsoever in the bill around duty of care toward vulnerable adults. I believe this to be a glaring omission and one I hope the committee can address. I cannot see how adequate provision is being made for those who might be more vulnerable to exploitation and abuse as a result of, for example, a learning difficulty.

Definitions around "legal but harmful" content, as determined in the bill by the intangibly defined "adult of ordinary sensibilities" do little to address the very real concerns of vulnerable adults being at risk of exploitation or accessing content they do not understand is illegal. It would seem we are relying on case law that will follow from the enactment of the bill to determine what legal but harmful actually looks like, and it will be down to a large amount of subjectivity and legal argument. The Mental Capacity Act 2005 gives care teams the powers to remove devices from a vulnerable adult should sufficient concern arise, this should be seen as a last resort and can have serious implications for human rights and deprivation of liberty.

It would be a positive move to see expectations of duty of care more explicitly defined within the Bill so that platforms are expected to consider the needs of vulnerable adults in their risk assessments and provide tools to support them – for example someone being able to disclose both a concern about another user on a platform and also their relationship with that individual. In this case a member of a care team might be able to flag up concern about one of their service users by alerting the platform to an account of concern and the platform could investigate in a non-intrusive manner.

Is the "duty of care" approach in the draft Bill effective?

A present it is fair to say that the duty of care still has a level of intangibility to it. As discussed above, there is little in the bill that defines what "doing enough" looks like. It would be good to see the bill defining more clearly what a good duty of care looks like, rather than having platforms invest considerable resources in risk assessments to still be told "do more". It is fair to acknowledge that this will fall on the shoulders of the regulator when considering whether a company's risk assessment is enough to demonstrate duty of care, but some clear guidance would be valued rather than waiting for the first rulings to be made.

Does the Bill deliver the intention to focus on systems and processes rather than content, and is this an effective approach for moderating content? What role do you see for e.g. safety by design, algorithmic recommendations, minimum standards, default settings?

Technology clearly has a role to play but it can never be a complete solution.

If we take, for example, the blocking of racist content, this is something that a company can do effectively to a certain degree. Anything that uses racist keywords can be easily detected. However, many who wish to post racist discourse online know this and will make sure they are no so explicit in their abuse. Algorithms fail, regularly, to understand context and artificial intelligence still struggles (these approaches will generally be based up a corpus of training data with human identified racist statements, and the algorithm will make an approximate determination whether a newly posted phrase is similar).

Platforms have had a lot of success relying on reporting of racist abuse to then examine it and take it down, rather than automating the whole process. Which is why transparency reporting is to be welcome but technology will not be able to achieve this without an informed, educated, user base.

Are there any contested inclusions, tensions or contradictions in the draft Bill that need to be more carefully considered before the final Bill is put to Parliament?

The fundamental challenge is what "doing enough" looks like, and I've seen little in the draft bill to make this explicit. As discussed above, keeping people free from harm online is almost impossible. Helping them mitigate risk and providing them with the tools for this is more realisable. But providers can only do so much on their own and it needs to be clear how extensive their risk assessment and mitigation should be.

While there is certainly a large role for providers in considering the potential harms that might arise through the use of their platforms, and the need for them to provide users with tools to mitigate risk, technology will not provide solutions to social problem that, while manifesting on platforms, are actually deeper rooted in society. There is a need for a more educated, safeguarding aware workforce, and education curriculum that meets young people's needs to understand online issues and how to get support. While this is possibly not in the scope of the bill, there needs to be an understanding that I think is sometimes missed that, on its own, the Online Safety Bill cannot achieve all of the purported ministerial aims.

Prof Andy Phippen, August 2021.

22 September 2021