Safety-by-Design

Parliamentary roundtable hosted by the Minderoo Centre for Technology and Democracy

13 October 2021 – 14:00-16:00

This workshop examined the challenges and opportunities of mechanisms to ensure online safety, the experiences of diverse user communities, monitoring and interoperability across applications, and experience with the Ofcom regulatory frameworks.

The format of a discussion allowed for an informative exchange of different views to bring to light aspects of the practical implications of the Bill that may otherwise be more difficult for the Committee to appraise. This format highlighted the information gap that still exists between Parliamentarians and the actual impact anticipated by different stakeholders of the proposed legislation. Throughout the discussion, Committee members asked open-ended questions to which further research could help answer.

Recommendations

Below are the key recommendations from the workshop on the concept of 'Safety-by-Design' in the Online Safety Bill.

- **The undefined scope of 'harm' makes any standard for 'Safety-by-Design' nearly impossible to articulate in primary legislation.** The Bill does not provide a definition of harm or guidance on the difference between 'illegal harms' and 'harms'. 'Safety-by-Design' is intended to keep Internet users 'safe' from 'harm,' but unless there is a clear articulation and threshold to be met, this concept will remain rather vague in law. Johnathon Purcell (Match) emphasised that the harm definition needed to recognise the cumulative or aggregate harm that was created by systems as a whole.

- **Powers of the regulator need to be defined clearly**. Much of the discussion concerned the lack of firm divisions of the role of Parliament as the legislator and regulators as responsible for the implementation of 'codes of conduct.'[1] While the participants iterated the need for flexibility in the legal framework, Baroness Kidron was careful to emphasis that this flexibility had to be based in statute.

[1] However, the discussion did not explicitly reference the Delegated Powers Memorandum (<https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985030/Delegated_Powers_Memorandum_Web_Accessible.pdf> accessed 18 October 2021).

- **International standards cannot replace law.** The use of global industry-led standards has many advantages including ensuring that compliance efforts are not duplicated, that the regulatory framework fits the technology, that they can be updated on a need-be-basis, and that they address the need for a global framework to tackle online harm, even if they alone cannot solve the problem. John Havens, from the professional organisation and standards body IEEE offered examples of how the 7000 and 7010 IEEE standards were based on ethical 'value-based' engineering which built esoteric human values into systems. Duschica Naumovska (INHOPE) was also concerned about the jurisdictional boundaries addressed by online harms and therefore wanted a global code.

- The Committee noted that standards could be a useful framework to copy and adapt, and that they had a function as 'soft power' for global consensus building across jurisdictions, but standards alone could not act as a, to use a colloquial phrase, a silver bullet in this instance.

- There is a need to put platform-specific mechanisms in place to ensure companies conduct due diligence to test products for potential harms both before and after launch.

- **Marginal voices must be included in devising codes of conducts and standards.** Several of the participants stressed that 'marginal voices' must be included in the process of devising the regulatory framework. Hilary Watson of the anti-online harms advocacy Glitch was particularly concerned with how the gender-neutral language of the Bill did not reflect how women are more targeted for online abuse, and the negative effect that this has on women's freedom of speech and freedom of participation online. She highlighted that, in that regard, the language of the Bill was out-of-step with the language of the Equality Act[2] and the 'Violence Against Women and Girls' framework.[3]

- **Age banding must be made clearer.** For those aged 0-18, age-banding should be made clearer to ensure there is an age-appropriate design code fit for purpose. The draft bill does not currently account for those under age 3. Professor Aiken (University of East London) discussed the construct of age bands and the use of the term "age groups" in the legislation, in effect ranging from 0 to 18, and argued that evidence-based guidelines would be required to differentiate between vastly deferring age

[2] The Equality Act. 2010 (<https://www.legislation.gov.uk/ukpga/2010/15/contents>) accessed 18 October 2021.

[3] Violence Against Women and Girls Guidance by the Crown Prosecution Services published in November 2019; see also A Framework to Underpin Action to Prevent Violence Against Women published by UN 2015.

groups in the context of the impact of online harm. Moira Patterson (IEEE) used Baroness Kidron's 'age-appropriate digital code' based on five principles for tracing the entire lifecycle of information products and services aimed at children as an example of the utility of standards as a regulatory mechanism in the online space.[4]

- **External access to platforms and data are needed for verification and scrutiny.** Professor Gina Neff repeatedly highlighted the lack of access to data by regulators and researchers which precluded meaningful assessment of adherence to the law (and which would therefore by implication affect the robustness (or lack thereof) of any enforcement mechanisms).

[4] https://5rightsfoundation.com/in-action/blog-age-appropriate-design-code-data-protection-bill.html.

SUMMARY

Parliamentary Roundtable hosted by the Minderoo Centre for Technology and Democracy on Zoom (13 October 2021); Summary by Ann Kristin Glenster, Minderoo Centre Law & Policy Affiliate.

I. Introductory Remarks

The parliamentary select committee workshop held 13 October 2021 with invited participants focused on the concept of 'Safety-by-Design' in the Online Safety Bill. The discussion, chaired by Professor Gina Neff at the Minderoo Centre for Technology and Democracy, was concerned only with this concept as it is used in the context of this specific legislative instrument. Given the broad legislative intent, it would be remiss not to contextualise the concept of 'Safety-by-Design' somewhat in the broader architecture of the Bill and, indeed, the workshop discussion did at times branch into broader themes.

'Safety-by-Design' is not defined in the Bill.[5] It may relate to other concepts found in law or regulation, such as 'Privacy-by-Design,[6] 'data protection by design and default.'[7] It can also be assumed to draw from product safety standards in consumer protection law, and the 'precautionary principle' found in environment law.[8] The ambivalence of the proposed statutory language suggests that the concept may be interpreted in many different ways by different actors, which was evidenced in the workshop.

The workshop was designed as a structured conversation with the participants rather than a formal session where each participant in turn present prepared evidence. The purpose of the session was therefore not to offer a doctrinal dissection of the proposed legal text, but instead offer the Committee some insight into the topics of the Bill; its practical implications and the opportunities it offered companies, organisations when designing systems that would meet the new legal requirements. Thus, the participants were not (and could not be) fully prepared for the turn of the discussion, and their contributions should be seen in light of this format.

[5] However, the government has published non-binding guidance on the subject on 29 June 2021 (https://www.gov.uk/guidance/principles-of-safer-online-platform-design accessed 18 October 2021).
[6] Ann Cavoukian, 'Privacy by Design: The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices" <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf> accessed 26 August 2021.
[7] Article 25 GDPR.
[8] Principle 15 Rio Declaration of Environment and Development 1992.

II. General Observations

First, the conversation reflected the tension between three regulatory powers; parliament as legislators of primary legislation, regulators as authors and overseers of secondary legislation in the form of binding codes of conduct, and external organisations (such as the IEEE or ISO) that develop, often global, industry standards.[9] Much of the discussion concerned the lack of firm divisions of roles between these mandates, but did not refer to the Delegated Powers Memorandum that accompanies the Bill.[10] In short, the participants iterated the need for flexibility in the adaption of any technical standards imposed on them by the Bill. Some participants were concerned that a lack of flexibility would make the Bill too narrow and quickly outdated, thereby making it unable to prevent emerging harms; others were concerned (although they did not necessarily clearly express it as such) that a Bill that was too prescriptive would impose technical legal standards that would be too onerous and restrictive for businesses, and particularly small businesses. Questions by the Committee often returned to the division between primary and secondary legislation, while the workshop participants often dwelled in the intersection between global standards and regulatory codes of conduct.

As the notes below will bear witness to, a considerable amount of the conversation was dedicated to advocacy for the introduction of international standards. The use of global industry-led standards has many advantages including ensuring that compliance efforts are not duplicated, that the regulatory framework fits the technology, that they can be updated on a need-be-basis, and that they address (although do not solve) the need for a global framework to tackle online harm. However, international standards cannot replace law which is enacted by a democratically accountable elected institution. Thus, the advocacy of standards could also be seen as a refusal to meet the Bill's objective as articulated in § 5 of the Delegated Powers Memorandum to '(…) end the era of self-regulation.' As such, it is worth reiterating here The Committee's repeated insistence that any code of conduct or standard adopted had to have a statutory base.

III. Undefined Harm

The first issue that arises is invariably the lack of specification in the Bill of what is meant by 'harm,' and especially the neglect to provide a legal distinction between 'illegal harms' and 'harms' (presumably by suppressing the latter, this category would also by logic become a form of illegal harm). The Committee queried whether the Bill needed to have a specified harm definition, but none of the participants offered a direct answer to this question. In that regard, it is worth bearing in mind that 'Safety-by-Design' is intended to keep Internet users

[9] Institute of Electrical and Electronics Engineers ('IEEE'); International Organization for Standards ('ISO').
[10] Delegated Powers Memorandum (<https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985030/Delegated_Powers_Memorandum_Web_Accessible.pdf> accessed 18 October 2021).

'safe' from 'harm,' but unless there is a clear articulation and threshold to be met, this concept will remain vague in law.

The Committee believed the standard of harm should be based on a collective, known standard and not be left to the subjective interpretation of individual companies. While a sensible suggestion which would place a manageable ringfence around the concept of harm, it is difficult to see how this approach would work without further specification of when and how harm would come in under the collective definition, especially given the repeated theme in the discussion of how harms continue to emerge with new technological and social trends. In response to a question from the Committee, Professor Mary Aiken offered evidence of harm from her research and confirmed that the cumulative effect of exposure to content produced greater negative impact. Similarly, in response to a question regarding 'pushed content' and shared technology from the Committee, Johnathon Purcell (Match) emphasised that the harm definition needed to recognise the cumulative or aggregate harm that was created by systems as a whole and multiple comments, and not just single platforms and services.

Professor Aiken was adamant that the Bill was needed, but she found the lack of definition of harm and e.g., the differentiation of children's ages (or lack thereof) difficult. She also queried how 'cumulative' or 'aggregated' harms would be identified and prevented. Aiken raised the broader issue of how legislation historically was focused on data and not the individual, and how this perspective should change as the Bill concerned the overall identity of human beings, not just pieces of data. She referenced her Ofcom report on harm where she had made a taxonomy demonstrating that harm was hierarchal and multifactorial, and mired in moral complexities.[11] Similarly, the Committee queried the definition of harm and noted how it is not specified in the Bill. He wondered what the harm threshold should be for a regulator to pursue (which leads to questions of resources), and also whether the specification should be in primary or secondary legislation, and if left to secondary legislation, whether that would give the regulator 'too much discretion.'

Hilary Watson (Glitch) was particularly concerned with how the gender-neutral language of the Bill did not reflect how women are more targeted for online abuse, and the negative effect this fact has on women's freedom of speech. She highlighted that, in that regard, the language

[11] Research on the Protection of Minors: A Literature Review and Interconnected Frameworks. Implications for VSP Regulation and Beyond. 2021. https://www.ofcom.org.uk/__data/assets/pdf_file/0023/216491/uel-report-protection-of-minors.pdf

of the Bill was out-of-step with the language of the Equality Act[12] and the 'Violence Against Women and Girls' framework.[13]

The Committee queried whether the precautionary principle from environment law could be used in this context. The precautionary principle mandates that companies have a duty to foresee and prevent any environmental harm that may arise from their activities. Indeed, Tracey Breeden (Match) praised the Bill for being proactive to 'Safety-by-Design'.

IV. Flexible and Futureproof Standards

The second issue and core of the workshop discussion centred on the potential for the use of flexible (and thus 'futureproofed') standards in the regulatory framework of codes of conduct to specify the harm and legal obligations on companies to prevent such harm through the adoption of 'Safety-by-Design.'

Emphasising the need to think holistically, Moira Patterson (IEEE) used 'age-appropriate digital code' based on five principles for tracing the entire lifecycle of information products and services aimed at children as an example of the utility of standards as a regulatory mechanism in the online space.[14]

John Havens (IEEE) offered examples of how the 7000 and 7010 IEEE standards were based on ethical 'value-based' engineering which built 'esoteric human values' into systems. Havens emphasised that the IEEE strive not only to prevent harm, but to encourage human and environmental flourishing. The Committee queried how the 7010 IEEE standard could regulate algorithms when algorithms 'do so much' and whether the standard could be adapted for the area of the proposed Bill. Havens reiterated the need for other benchmark measurements than GDP such as plants, people and profits, and the concept of 'responsible innovation.'

Duschica Naumovska (INHOPE) was also concerned about the jurisdictional boundaries and therefore wanted a global code. She iterated that it was important not to duplicate efforts. The point about not duplicating standards is important for market actors as they will otherwise have to duplicate their technical efforts and costs and may also have to have compliance policies to resolve 'conflict of laws.'

[12] The Equality Act. 2010 (<https://www.legislation.gov.uk/ukpga/2010/15/contents>) accessed 18 October 2021.
[13] Violence Against Women and Girls Guidance by the Crown Prosecution Services published in November 2019; see also A Framework to Underpin Action to Prevent Violence Against Women published by UN 2015.
[14] https://5rightsfoundation.com/in-action/blog-age-appropriate-design-code-data-protection-bill.html.

The Committee referenced the UK's BSI Standard for Robots and Robotic Devices and asked how the verification process would work in this case, especially in relation to age.[15] Havens responded that the BSI was 'the first standard on human ethics,' and the IEEE works globally, so its standards are complementary to national frameworks. Havers further offered information on the IEEE's forthcoming AI standards and how it wants to introduce a labelling system for consumers, whereby companies will be able to demonstrate their attempts to act more responsibly. There was also a question regarding how a labelling system could incentivise companies when customers had no other options (e.g., Facebook). Michael Tunks (Internet Watch Foundation) noted that human oversight of the standards and codes of conduct is important.

Professor Gina Neff asked if the relevance of these standards was to illustrate how standards could be transposed from one legal framework to another. Patterson confirmed this point and noted that broad standards could branch out into more specialised (and thus narrow) standards. Professor Neff queried how standards could allow for the planning (and presumably prevention) of emerging unknown risks. In response, Havens emphasised the importance of consensus building across (academic) disciplines and the understanding that the issues at hand are socio-technical issues where disciplines may operate with different definitions for same or similar terminology.

The Committee emphasised that standards needed to be based in statute. She would like to know what a code of conduct produced under the Bill would look like. She noted that standards could be a useful framework to copy and adapt, and that they had a function as 'soft power' for global consensus building across jurisdictions, but standards alone could not act as a silver bullet in this instance. Breeden and Naumovska both emphasised how important it is that codes are flexible and thus can be rapidly updated, are malleable (and thus futureproofed). The Bill should therefore not be overly prescriptive, but the requirements should be in the codes of conduct. A question regarding what would trigger a revision of the codes of conduct was raised.

Related to the issue of standardisation is the question of how far individuals shall be responsible for their own exposure to risk. This point is linked to the Committee's question regarding the role of media literacy. Breeden pointed out that there was a difference between the technical conceptualisation and implementation of 'Safety-by-Design' and users' expectation of safety. She advocated giving users tools to decide their own exposure to risk, which was refuted by the Committee who stressed that making individuals responsible for their exposure to harm would be wrong, especially in relation to children. Watson noted that often users were given these tools, but that they were not sufficiently communicated. She

[15] https://shop.bsigroup.com/products/robots-and-robotic-devices-guide-to-the-ethical-design-and-application-of-robots-and-robotic-systems?pid=0000000000.

reiterated the importance of ensuring that legislation was not too specific in order to be adaptable to cover emerging trends. She further highlighted the need for 'awareness-raising' campaigns to educate the public on the difference between, for example, political suppression of speech and abuse of women online.

Tunks requested more transparency regarding how different marginal voices were consulted, especially children, when the standards and codes of conduct were drawn up. Havens and

Breeden also iterated the need for 'marginalised' representation in the devising of standards. To illustrate his point, Havens pointed to the use of children's mental health and New Zealand's artificial intelligence ('AI') technology roadmap.[16] Indeed, Abrahams asked which marginal groups should be consulted.

V. Enforcement and Transparency

The third issue that came to the fore in the workshop by its notable *absence* was enforcement. Breeden began the conversation by addressing the legal obligations the Bill places on companies and organisations for conducting internal risk assessments. She specifically asked what the standards were for the associated due diligence and testing requirements. She noted the need for clarity in this area as products were continuously developed ('rolled out').

Abrahams asked a general question regarding the effectiveness of the enforcement mechanisms as they are currently articulated in the Bill. Notably, none of the participants picked up this baton, possibly because, at least in the case of those who are members of industry, favour flexible standards rather than more rigorous enforcement provisions. Professor Aiken did address the question head on by bringing it back to the definition of harm—she prodded how the harm would be evidenced in a manner that would be used in courts and how causality would be proved.

The Committee asked about the role of the Regulator. He contextualised his question by noting how the definition of vulnerable adults had change in the domain of Reality TV, or the difficulty a regulator would have in identifying the harm in relation to girls being exposed to anorexic images online.[17] Tunks asked what would trigger enforcement action by the Regulators and how the accuracy of children's ages would be verified. The Committee queried how AI could be verified and overseen, and how to ensure that the legislation would be fit for the future.

[16] Artificial Intelligence in Health in New Zealand (https://aiforum.org.nz/wp-content/uploads/2019/10/AI-For-Health-in-New-Zealand.pdf) accessed 18 October 2021.
[17] (In an aside, Collins also noted the power struggle between ICO and Ofcom.)

Related to the issue of enforcement is the issue of transparency. Professor Neff repeatedly highlighted the lack of access to data by regulators and researchers which precluded meaningful assessment of adherence to the law (and which would therefore by implication affect the robustness (or lack thereof) of any enforcement mechanisms). Breeden emphasised that a mandate to disclose 'too much information' would give potential malicious actors the information they needed to cause harm. While an important point, it is important that this contribution is contextualised in a wider regulatory framework.

As such, it is notable that the discussants throughout positioned the technology companies as 'neutral.' The Bill also adopts this position as it regulates harm that derives from 'user-to-user' activity, and not activity by the companies. Thus, the role of the companies is to create 'Safety-by-Design' systems that promote the safety objectives of the Bill, but not to be labelled as potential 'malicious actors' themselves.

The discussion (and arguments for and against) the requirements on companies to disclose data and algorithms have been considered thoroughly elsewhere so will not be revisited here other than to say that this is an expected position for a market actor to take.[18] Purcell advocated the use of 'transparency reports' as he believed they would incentivise companies 'to get their houses in order' before exposing themselves to external scrutiny.

There was a short mention of how the Bill could diminish companies' ability to encrypt their services, thereby making the action on them invisible to regulators,[19] but this topic was not explored further. Finally, Professor Gina Neff queried how emerging risks would be identified. Summarising the Bill, Paul Gaskell (DCMS) iterated that the Bill was intended to be flexible and technologically neutral. The question of 'Safety-by-Design' was a question of the duty of care (including risk assessments) that would be imposed on platforms and how they would discharge that duty. He noted that companies would need to be able to verify the age of children users and the priorities of different harms would be presented in secondary legislation.

VI. Optimisation

Several of the participants asked the question regarding what systems were optimised for. While a valid policy question, it is one that must be treated carefully in the legislative context as it implies that the lawmakers are not attempting to impose a *negative* obligation on companies to prevent something from occurring, but rather impose a *positive* obligation for

[18] E.g., Katarina Foss-Solbrekk and Ann Kristin Glenster, 'The Intersection of Data Protection Rights and Trade Secret Privileges in "Algorithmic Transparency"' in EU Data Protection Handbook (Ronald Leenes and Eleni Kosta, eds) (Edward Elgar 2021) (forthcoming).
[19] Referenced Facebook encryption of WhatsApp.

something to occur, which would lead to a wider debate about what that positive result should be and the implications that would have for the proliferation of different actors, products, spaces, content, and opinions online. Answering the question 'what are we optimising for' as a positive legal obligation could potentially lead to a restriction of the online space. A comment was offered that system designers should be asked to explain their 'intentions,' but unless the intentions can be measured against a legally enforceable standard, asking for a self-declared intention (that cannot be tested or to which the designers cannot be held to account) seems rather meaningless in a strict legal context.

VII. Minderoo Centre's Comments

Finally, we would like to take this opportunity to offer a few comments to contextualise some of the topics that and comments of the discussion. First, we would like to offer some caution in the adoption of global standards by NGO or other organisations as they are not beholden to an electorate, and which thereby could therefore be interpreted as an 'outsourcing' of the democratic function. This caution also extends to 'outsourcing' the selection and inclusion of 'marginal voices' to organisations with no legal accountability to the population of the UK.

Second, while immediately appealing as a legal means to prevent online harm, the precautionary principle could be problematic as it would incentivise companies to be (perhaps overly) cautious and thereby lead to the restriction of democratically desirable free speech online. In other words, it is reasonable to assume that the precautionary principle in this context could lead to the censorship of the common online space determined and administered by corporations based on their financial motivations (i.e., minimisation of exposure to the risk of liability penalties). Thus, overall, the use of the precautionary principle, combined with the lack of a definition of harm, could make technology companies acting from a motive of profits responsible for policing the scope of 'freedom of speech' online.

It is also worth bearing in mind that 'flexibility' undermines accountability. Enforcing a legal standard which is flexible is near impossible. Thus, the final word of caution is that the reliance on standards as the framework for legally binding 'codes of conduct' is problematic, as these often become 'aspirational targets' rather than enforceable legal rules. For example, the language of 'human and environmental flourishing' and 'responsible innovation' may be highly appropriate in a policy context, but is not suited for law-making, especially when the benchmarks for these positive concepts are set by organisations and individuals outside the jurisdiction of the democratically elected law-making institution. Promoting human and environmental flourishing may be important goals, but not the objectives of the Online Safety Bill which is concerned with the prevention of online harm. Without further elucidation these concepts will be empty in law and may therefore be susceptible to rulemaking by regulators and others without the transparency of process that is appropriate and necessary for the legitimacy of the law.

Workshop attendees:

- Damian Collins MP
- Debbie Abrahams MP
- Darren Jones MP
- Dean Russell MP
- Suzanne Webb MP
- Baroness Kidron
- Lord Clement-Jones
- Lord Stevenson
- Mary Aiken, Professor of Forensic Cyberpsychology, University of East London
- Hilary Watson, Policy and Campaigns Manager - Glitch
- Dushica Naumovska, Programme Manager INHOPE
- John Havens, Director, Emerging Technologies & Strategic Development, IEEE Standards Association
- Moira Patterson, IEEE Standards Association
- Michael Tunks, Senior Policy and Public Affairs Manager, Internet Watch Foundation
- Tracey Breeden, VP, Head of Safety and Social Advocacy, Match Group
- Johnathon Purcell, Director, Trust and Safety Operations at Match Group
- Professor Gina Neff, Executive Director, Minderoo Centre for Technology and Democracy
- Ann Kristin Glenster, Law & Policy Affiliate, Minderoo Centre for Technology and Democracy
- Irene Galandra Cooper, Project Administrator, Minderoo Centre for Technology and Democracy
- Jeremy Hughes, Communications Co-ordinator, Minderoo Centre for Technology and Democracy
- DCMS Bill team - Becky Woods, Paul Gaskell
- Ofcom - Amy Jordan, Chloe Grant, Alex Galloway
- Jacquie Hughes – Committee specialist adviser
- David Slater – Committee Clerk
- Andrea Dowsett – Committee Clerk
- Pierre Andrews, Office of Damian Collins MP
- Lucy Dargahi, House of Lords
- Bruce Sinclair –HoL Committee team
- Jillian Duke – House of Commons team.