ONLINE SAFETY ACT NETWORK

SAFETY BY DESIGN

This paper by Professor Lorna Woods, Professor of Internet Law at the University of Essex and expert adviser to the OSA Network, considers the concept of safety by design in the Online Safety Act (OSA). It covers:

- *What the OSA says*
- *Current thinking on "by design" obligations*
- *The specifics of other "by design" approaches, including "Privacy by Design" and "Security by Design"*
- *Existing "Safety by Design" principles*
- *How these approaches and principles apply to the interpretation of "safety by design" in the OSA*

The Online Safety Act: section 1 (3)

Section 1(3) Online Safety Act states that the duties in the Act seek to secure that services regulated are:

> (a) safe by design, and
>
> (b) designed and operated in such a way that
>
> (i) a higher standard of protection is provided for children than for adults;
>
> (ii) users' rights to freedom of expression and privacy are protected, and
>
> (iii) transparency and accountability are provided for in relation to those services.

Section 1(3) does not expand the scope of the duties but indicates how they may be operationalised. Although there is more detail as to the service providers' duties and Ofcom's obligations in the Act, those provisions should be read through the framing provided here; s1 OSA provides not just context for the rest of the Act but also a pretty clear statement of purpose for it.

Lord Parkinson, the Government Minister who took the Online Safety Bill through the House of Lords, set out the purpose of (then) clause 1, which the Government introduced as an amendment following cross-party pressure for such an introductory statement.

> "My Lords, this is not just a content Bill. The Government have always been clear that the way in which a service is designed and operated, including its features and functionalities, can have a significant impact on the risk of harm to a user. That is why the Bill already explicitly requires providers to ensure their services are safe by design and to address the risks that arise from features and functionalities. The Government have recognised the concerns which noble Lords have voiced throughout our scrutiny of the Bill, and those which predated the scrutiny of it. We have tabled a number of amendments to make it even more explicit that these elements are covered by the Bill. We have tabled the new introductory Clause 1, which makes it clear that duties on providers are aimed at ensuring that services are safe by design. It also highlights that obligations on services extend to the design and operation of the service. These obligations ensure that the consideration of risks associated with the business model of a service is a fundamental aspect of the Bill. …
>
> The Bill will require services to take a safety by design approach to the design and operation of their services. We have always been clear that this will be crucial to compliance with the legislation. The new introductory Clause 1 makes this explicit as an overarching objective of the Bill. The introductory clause does not introduce any new concepts; it is an accurate summary of the key provisions and objectives of the Bill and, to that end, the framework and introductory statement are entirely compatible." (Hansard 19 July 2023 col 2419)

It is worth noting here that current statutory interpretation focuses on Parliamentary intention, as can be seen in *R (Quintavalle) v Secretary of State for Health*:

> "The court's task, within the permissible bounds of interpretation, is to give effect to Parliament's purpose. So the controversial provisions should be read in the context of the statute as a whole, and the statute as a whole should be read in the historical context of the situation which led to its enactment".

The purpose set out in section 1 – that services are "safe by design" – therefore gives a benchmark against which services' respective performance of their duties should be understood; effectively this is minimum content for those duties.

The Act does not, however, tell us what "safe by design" means. We may find some understanding of this approach by looking at "by design" obligations in other contexts – specifically those related to information and communication technologies.

Why a "By Design" Approach?

This approach is based on the recognition that information technologies are created and that choices have been made about how. Moreover, these technical decisions affect the affordances of the technology - that is the properties of the technology - and the ways it may be used. While none of this render users totally passive or inert - and, indeed, they may use services inventively - the fact remains that technology is not neutral. Users cannot go beyond the technical possibilities of the service; moreover, they may be nudged to use the service in particular ways. Further, different groups of people may perceive affordances differently – or not perceive that they exist at all. As Constanza-Chock noted, design can "systematically privilege [..] some kinds of people over others" (Design Justice, p 37).[1] For some groups, affordances may lead to a particular risk of harm while for others, that risk is lower or non-existent. For designers, there may be a risk of privilege hazard – that is by not recognising their own position and its benefits, they may not consider others are not similarly protected.

This approach also recognises that, given the scale of use, it remains more difficult to deal with instances of bad behaviour individually and after the fact. For example, there might be less mis- and disinformation (potentially caught as illegal content or content harmful to children in the OSA) if there were no financial incentives to create and share it, and if the content prioritisation tools valued an orientation to accuracy rather than sensationalism. At the moment, content moderation seems to be in tension with the design features that are influencing the creation of content in the first place, making moderation a harder job. So, a "by design" approach is a necessary precondition for ensuring that other *ex post* responses have a chance of success.

While a "by design" approach is important, it is not sufficient on its own; there will be a need to keep reviewing design choices and updating them, as well as perhaps considering *ex post* measures to deal with residual issues that cannot be designed out, even if the incidence of such issues has been reduced.

Designing for safety (or some other societal value) does not equate to techno-solutionism (or techno-optimism); the reliance on a "magic box" to solve society's woes or provide a quick fix. Rather, what it acknowledges is that each technology may have weaknesses and disadvantages, as well as benefits. Further, the design may embody the social values and interests of its creators. A product (or some of its features) may be part of the problem. The objective of "safety by design" is – like product safety – to reduce the tendency of a given feature or service to create or exacerbate such issues.

[1] The obvious example of this is people with disabilities (see eg https://cdn.disabilityinnovation.com/uploads/documents/Inclusive-Design-Standards.pdf?v=1572970889). Also relevant re gender (see eg here Wachter-Boettcher, S. 2017. Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech. 1st. New York, NY: W. W. Norton & Company), and race.

"By Design" Approaches

The design of a product reflects the issues the developer has thought about and the issues and values the developer has prioritised. For example, there has been increased emphasis on human-centred design (which does not just focus on an identified user group but considers a range of people). Of course, developers could seek to take into account a range of different values. Some in the design community have talked about value sensitive design (and methodologies for achieving this) – but the difficulty with this approach is that it is open to the designers to adopt from a wide range of values (and values are not necessarily positive or ethical). Hildebrandt has argued that design should reflect the law, and specifically human rights. A "by design" approach seeks to ensure products respect the law and/or protect the value identified; it shifts responsibility for the protection of that value from the user to the manufacturer or service provider.

There are a range of "by design" approaches that could operate as a model for "safety by design": "privacy by design" (PbD) and "security by design" (SbD) are well-established in the information and communication technologies sectors. The re-deployment of these approaches to address other values can also be seen. For example, some have suggested the need for "transparency by design", especially in the context of AI. "Inclusivity by design" was included in Lord Holmes' Bill on AI regulation (cl 2(1)(c)(ii)) and in the recent joint statement by the UK and US governments on protecting minors online, where it was recognised alongside safety and privacy by design.

Significantly, however, the values to be protected in these examples are (or are proposed to be) mandated by law: the precise scope of those values might be the subject of debate but the fact that they should be protected is not – and they cannot be traded away against convenience or profitability.

Privacy by Design

PbD is perhaps the longest established, and one around which there is significant consensus. It is described as: "a process for embedding good privacy practices into the design specifications of technologies, business practices and physical infrastructures. This means building privacy into the design specifications and architecture of new systems and processes". Ann Cavoukian, responsible for the introduction of PbD noted: "Privacy must be embedded into every standard, protocol and process that touches our lives". One of the key principles was that:

> "Privacy by Design is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality."

Rather than waiting for problems to manifest and then deal with them, the developer of the service should seek to understand the risks and unintended side-effects and seek to mitigate them before the harm is caused.

PbD set out 7 principles to follow:

1. Proactive not Reactive; Preventative not Remedial

2. Privacy as the Default

3. Privacy Embedded into Design

4. Full Functionality – Positive-Sum, not Zero-Sum

5. End-to-End Security – Lifecycle Protection

6. Visibility and Transparency

7. Respect for User Privacy

Note that the word "proactive" has developed some unfortunate connotations in the context of internet regulation, as it has been linked primarily to the use of upload filters. Proactive here is linked to the design of the service and is largely content neutral. It can be contrasted with a remediation approach which tries to repair things or compensate after the harm has occurred (and which might focus on content types). PbD in this sense seems to have similarities with the idea of "inherent safety" in other industrial sectors – so the objective is to avoid hazards rather than control them. Principle 1, above, suggests that users cannot be made responsible for their own privacy (or by extension) safety; reliance on user empowerment tools alone would not constitute a "by design" approach. Privacy Enhancing Tools (PETs) have been defined as

> "Software and hardware solutions, i.e. systems encompassing technical processes, methods or knowledge to achieve specific privacy or data protection functionality or to protect against risks of privacy of an individual or a group of natural persons."

There is some uncertainty as to the scope of PETs, and some authors take a broader approach than others; the definition adopted by the ICO focuses on a single issue rather than the more holistic approach envisaged by PbD, and the requirement for privacy to be respected as part of the underlying architecture. Another issue relating to PETs is that they may be targeted at the users of a product rather than the designers/developers, potentially removing the responsibility for user safety (however widely that is understood) from developers.

There is also a specific requirement for "data protection by design" found in the GDPR. Article 25(1) GDPR provides a general obligation that controllers:

> "both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects."

Article 25(2) adds another more specific requirement:

> "The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed."

This makes clear another point – that PbD also includes privacy by default, reflecting Principle 2 of Cavoukian's 7 principles (there is a question as to whether you can have by default without by design and in terms of Article 25's structure, default follows design). But, technical and economic feasibility is a requirement of Article 25 GDPR. What these each mean in a particular context may be open to some debate – depending on levels of acceptable efficiency and cost and also how those two considerations affect one another.

Security by Design

In cyber security, there seems to be consensus about the importance of SbD but not what it requires. It has been described thus:

> " 'Secure by design' means that technology products are built in a way that reasonably protects against malicious cyber actors successfully gaining access to devices, data, and connected infrastructure."

Security is often seen as relating to confidentiality, integrity and availability. There have been some questions in the security field as to whether the focus on design is enough and that the entire software cycle should be considered (design, development, deployment, management, and retirement). Of course, design issues can impact throughout the entire lifecycle of the product, and have consequences for even the ease with which products can be retired (think for example about early genAI models released without sufficient security features; once released open source, can they be retired?) Issues about design and functionality should be considered in each of these phases, as well as other mechanisms and processes to compensate for the gaps left even with good design.

SbD also has guiding principles. The National Cyber Security Centre has proposed the following:

1. establish the context before designing a system;

2. make the compromising of data or systems difficult for attackers;
3. make disruption of the service difficult;
4. make compromise detection easier; and
5. design to naturally minimise the severity of any compromise.

Account safety is not much discussed in the context of online harms but it is an important element of keeping users safe (for eg victims of domestic violence) or preventing misinformation.

More recently, principles agreed internationally focus more on process and governance issues relating to product development, rather than end-goals, such as specific security standards for products:

1. take ownership of customer security outcomes
2. embrace radical transparency and accountability
3. build organization structures and leadership to achieve these goals.

Safety by Design

There is less consensus around what "safety by design" precisely means, though the abstract principle seems generally accepted. DCMS, prior to the introduction of the OSA, explained Safety by Design as:

> "the process of designing an online platform to reduce the risk of harm to those who use it. Safety by design is preventative. It considers user safety throughout the development of a service, rather than in response to harms that have occurred."

It could be said that this is a form of product safety – the UK product safety rules require that producers and distributors must only place safe products on the market. Safety is determined by considering all characteristics of the product, including the impact on other products and taking into account the nature of the users. Amongst other obligations, producers must test a sample of the products to identify risks and keep a register of complaints.

DCMS identified also a number of principles for safety by design that seem to identify how the general idea should be put into practice:

- Users are not left to manage their own safety – eg
    - makes a user aware when they do something that might harm themselves or others

    - makes it harder for users to upload or share content that is illegal or breaks your terms of service

- Platforms should consider all types of user – eg
    - users with protected characteristics which may make them victims of discrimination
    - users with accessibility needs
    - users with low levels of media literacy
- Users are empowered to make safer decisions
    - understanding reliability of content
    - understanding how online activity is seen by others and how to manage that
- Platforms are designed to keep children safe
    - encourage safe, positive interactions and ensure users can interact free from abuse, bullying and harassment
    - limit access to certain features, functions and content which pose a greater risk of harm to them
    - set safety, security and privacy settings to high by default.

Another approach is to look at safety tech, which could be seen as analogous to PETs. The UK Government defines safety tech as:

> "technologies or solutions to facilitate safer online experiences, and to protect users from harmful content, contact or conduct."

While this definition could overlap with safety by design, it certainly seems to encompass technologies that are not part of the original product design, and which may be provided by third parties. It would seem likely that the concerns about PETs focussing on a single issue might map on to safety tech too.

The Australian e-Safety Commissioner identified three principles, though these seem to have sub-elements:

- Service provider responsibility
    - The burden of safety should never fall solely upon the user.

    - Processes to ensure that known and anticipated harms have been evaluated in the design and provision of an online platform or service.

- User empowerment and autonomy – contains a number of elements including

    - Provide technical measures and tools that adequately allow users to manage their own safety, and that are set to the most secure privacy and safety levels by default.

    - Establish clear protocols and consequences for service violations that serve as meaningful deterrents and reflect the values and expectations of the users.

    - Evaluate all design and function features to ensure that risk factors for all users – particularly for those with distinct characteristics and capabilities –have been mitigated before products or features are released to the public.

- Transparency and accountability

    - Ensure that user safety policies, terms and conditions, community guidelines and processes about user safety are accessible, easy to find, regularly updated and easy to understand.

    - Carry out open engagement with a wide user base, including experts and key stakeholders, on the development, interpretation and application of safety standards and their effectiveness or appropriateness.

    - Publish an annual assessment of reported abuses on the service, alongside the open publication of meaningful analysis of metrics such as abuse data and reports, the effectiveness of moderation efforts and the extent to which community guidelines and terms of service are being satisfied through enforcement metrics.

    - Commit to consistently innovate and invest in safety-enhancing technologies.

This list seems to go broader than a safety by design, to include elements of risk management and transparency. Nonetheless, the principles do pick up on aspects of “by design” – notably the obligation to test before going to market (as per the other “by design” approaches) and the need to have governance and/or oversight over the implementation of a “by design” approach. This perhaps reflects the fact that in this sector there is a strong incentive to ship minimum viable products, rather than safe products, and an assumption that it is ok to use your customers as lab rats.

What is Safe?

There is no separate definition of "safe". We might suppose that it stands in opposition to "harm" and harmful content – though not all OSA duties are described by reference to harm. Notably the illegal content duties do not refer to harm – though it might be supposed that content relating to criminal offences is harmful. Section 1(3) is not limited to Part 3 duties, but includes also Part 5. The age verification obligations on pornography providers are not justified by reference to harm directly, though in the children's duties, pornography is identified as primary priority content. User empowerment tools cover further categories of concern (listed in s 16). Again these are not defined as harmful but rather as regulated user-generated content. Safe should probably be taken to refer to these issues identified by the Act, rather than issues more generally.

Some caution should be exercised about the scope of safety here – it seems unlikely (especially given the emphasis on proportionality in the Act) that safety by design requires a maximal interpretation of safety, or safety at all costs. Indeed, it is questionable whether such a position is actually possible. Where the boundary lies, however, is unclear. A duty of care standard talks of reasonable steps to prevent foreseeable risks. Product safety talks about a "safe" product being one which, under normal or reasonably foreseeable conditions of use, does not present any risk - or only the minimum risks compatible with the product's use and which are acceptable and consistent with a high level of protection for the safety and health of persons. Note that safety is not just safety of users, but also other affected persons. It is also standard practice to focus first on the most severe harms and to accept that these may demand more far-reaching solutions than more minor harms.

Safety by Design and the OSA

This leaves the issue of what "by design" means. The structure of s 1(3)(b) suggests design is different from operation. "By design" is an approach seen in other policy fields relevant to information and communications technologies. They may perhaps be seen as types of "value sensitive design". Significantly, however, the values to be protected are mandated by law; while the precise scope of those values might be the subject of debate, the fact that they should be protected are not – and they cannot be traded away against convenience or profitability.

Looking at existing examples there seems to be no one set starting point – save to say that in all instances the general principle requires further specification to become operationalisable (especially when considering specific measures required). This still leaves some questions as to what is precisely required by safety by design: is it about process, or assessing the end product, or both? It seems that a process is required to ensure that the issues are considered. This can be seen in commentary on value sensitive design, but also PbD and SbD. Safety by design is not

just a process-based obligation, important though appropriate internal processes and governance mechanisms are. Section 1(3) says the objective of the Act is to ensure products are safe by design. This seems to cover both the process (including governance mechanisms) and the end design.

Despite differences between the models, some common underlying principles can be seen: that the value in question is a core objective in the design process, considered in the objectives to be achieved from the start of the process. This approach can be contrasted with the use of a particular technology which even if integrated may be adopted late in the design process. An example of this could be the use of age verification on some services to avoid revisiting other design (and business) choices. Of course, this is not to say age verification should not be used, just that as a sole response to issues it may not reflect the safety by design approach. Safety through exclusion of vulnerable groups is insufficient. There is also a risk that unless safety considerations are included from the start, the discussion turns into a zero sum game where functionalities are seen as opposing safety.

As the PbD principles note, however:

> "Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. Privacy by Design avoids the pretence of false dichotomies, such as privacy vs. security, demonstrating that it is possible, and far more desirable, to have both."

One question here is whether the "by design" duty requires the introduction of extra safety features (we might say to create a "good space" for communication), or just the removal of (or provisions of adequate safeguards against risk of harm in relation to) hazards. This latter approach has similarities to the views of some with regard to SbD: "the expectation that technology is purposely designed, built, tested, and maintained to significantly reduce the number of exploitable flaws before it is introduced to the market for broad use". In practice, there may not always be a clear boundary. In the case of misinformation (admittedly outside the Act unless harmful to children or illegal), the rewards for clickbait farms seem to incentivise the creation of problem content and could be seen as a hazard that should be removed; a reward structure that emphasis public service content could be seen as the latter. Moreover, some functionalities may have important safety benefits for some groups, while constituting a hazard for others (eg anonymity). This is different from assessing safety against other interests – like making money (eg recommender algorithms set to outrage). Even if positive obligations are expected, it is important to recognise that technology cannot be a cure for all ills.

Three types of response to hazards and risks can be seen: hazard reduction; then hazard management; and finally remediation. Hazard reduction rather than hazard management or the provision of remediation should be prioritised, with remediation being the mechanisms of last

resort. This emphasis on prevention rather than cure (and/or remediation), which can also be seen in the Ruggie principles, exists in some of the existing models. The "security by design" approach recognises the importance of product testing and in particular abusability testing or red-teaming; so does the e-Safety Commissioner; these are important for hazard reduction before a product or change to a product is introduced. In particular, they may reveal where the production teams have made assumptions about who the user base is and how the product might be used.

In relation to risk assessments, the "by design" approach seems to imply product testing rather than just hypothesising risks, and also envisages an appropriate response, even if there are not "hard" output measures for what adequate safety looks like. The emphasis on embedding and not relying on users to protect themselves could fall within the scope of the safety duties (eg in the reference to proactive steps); indeed, it could be said that those obligations should be understood in the light of the safety by design obligation in s 1(3).

What is rather more explicit in the safety duties is the focus on filtering and moderation, which may have a design element (ie the tools are made available within the system and designed to work with the system) but seem more *ex post* in the way they work. Potentially the same criticism as that levelled at PETS could be applied here - that such tools only address one aspect of the problem and so, while they may play a role in improving things on their own, they would not be sufficient and do not satisfy the requirement for safety by design.

In this the risk assessment processes seem part of a "safety by design" approach – though the considerations in the risk assessments in the Act go beyond design considerations and include operational matters. Note, however, that there is a temporal dimension to this. We might expect risk assessments to take place during product development as well as on the final product, and when the product is changed. Does safety by design expect monitoring, investigations and response (including redesign as well as ex post safety tech solutions) when unanticipated problems become apparent during use? Similar questions arise in relation to existing products, which may not have been designed with safety (as understood in OSA terms) in mind.

Two responses come to mind. The first is that any changes to the product should be subject to a risk assessment (including testing), with consequent appropriate changes to design; this brings in the possibility of not launching the feature – as happened with TikTok lite under the DSA. The second is that a risk assessment process should be understood as requiring investigation (including testing) of products on an ongoing basis even absent product changes (though the frequency of testing and the structure of that investigation may vary across different services), resulting in effective mitigation, including through redesign.

One final area of uncertainty is whether there are any benchmarks for minimum safety or safety practices. Although it might be said that an unsafe product should not be marketed, it must be

recognised that risk and exposure to hazard might not be equally experienced amongst users; some features that are unsafe for some may be safe for others, leading to a difficult balancing exercise in understanding the base level of safety. Even where there are some safe uses, the risks to some should never be lightly or totally disregarded, and the severity of impact is a factor that should weigh heavily in the balance when deciding these issues as harm prevention lies at the heart of the regime.

October 2024

Contact: hello@onlinesafetyact.net