Response to the OPSS/DBT consultation on Product Regulation: the UK's new product safety framework

---

This submission contains the text of our response to the consultation by the Office for Public Standards and Safety and the Department for Business and Trade on the new product safety framework; it has also been submitted via the online consultation form.

Our interests in the consultation are limited to the framework for the regulation of AI-enabled products.

About the OSA Network

We work with over 80 civil society organisations, campaigners and academics on a cross-harm basis to build capacity and connections within civil society to protect people from online and AI harms through effective regulation. We inform, coordinate and support effective, ongoing civil society engagement and advocacy with policymakers, regulators and Parliamentarians on matters relating to online safety and AI harms. Interests represented in the Network include: child protection, terrorism, extremism, violence against women and girls, suicide and self-harm prevention, mental health, hate speech and online abuse, fraud and scams, animal cruelty, mis- and disinformation, harms to democracy and threats to our information environment. More detail on our work is available here.

Consultation response

Question A25: Are you aware of any data or evidence on the types of AI-enabled products that are likely to be manufactured in the future?

AI is developing at considerable speed. Much of the Government's approach to AI policy in recent years has focused on encouraging innovation and ensuring that the UK harnesses the economic benefits and growth related to that, such as the recent £1.1 billion investment in British companies developing the chips and semiconductor technologies behind AI, alongside the £500 million Sovereign AI fund '. Less focus has been directed at the risks and harms emerging from unsafe and unregulated AI, in particular whether the development of AI-enabled products - which may well be very innovative and highly profitable for their manufacturers - pose a risk to consumers, particularly the most vulnerable. In many ways the premise of this question - what is the evidence for products that may be manufactured in the future - underlines the challenge. By its very nature, it is difficult to predict how AI will develop in the future and how it might be deployed, either in physical items or as a purely digital product. Some

evidence on the kinds of industrial developments that have already happened, which underline the challenge with prediction, include: the development of an AI sexbot industry; and AI-enabled wearables, including "Friend", "a digital companion you wear around your neck that simply listens, learns about you, and responds like a friend" and Meta's smartglasses, whose inclusion of facial recognition "would hand stalkers, abusers, and federal agents the ability to silently identify strangers in public". Personal Assistants have been around for a while and although some are made available as apps, some are embedded in devices (eg Alexa in Echo, or "desktop companions"), even wearable devices (see eg Bee, Fieldy, Lookai). The AI toy market is already developing, but also in addition to the concerns around privacy to which all these tools give rise, safeguarding as well as content-based harms are concerns. The more general consumer robot helper market does not yet appear to have hit mainstream, though market reports relating to 2019 suggested growth in this sector, and there are a range available (though some of these are marketed as toys). The AI toys market is significant: the BSI reports that half of children have them despite safety concerns. There is concern in Australia that, as the social media ban takes effect, young people will move to chatbots instead; a similar concern could arise in the UK.

As noted in the article on sexbots linked above, "as the history of the likes of Facebook, Google and Amazon has taught us, today's digital quirks could become tomorrow's global giants". The Government cannot afford to wait for the evidence of what products might be developed in future before designing a product safety framework that ensures those future products do not pose harm to those who use them, or - in the case of much "smart" technology that can be deployed to perpetrate violence against women and girls - those around them. Moreover, an approach based on individual technologies or instances of deployment is not future proofed. There should be a general base level of care on which specific responses, if relevant, can be built. Transparency about what tools are is also important; this might not be apparent where AI is embedded into a smart device or marketed as a toy.

New and emerging forms of harm are already being evidenced in relation to AI chatbots and AI toys, including the manipulation of others through false information or the exploitation of vulnerable individuals. We outline this further in question A26; the advancement in the development of AI-powered robots that are able to resemble humans in both aesthetic and emotional output are likely to exacerbate these evidenced harms, and create new and complex forms of harm.

Question A26: What do you think are the current or potential harms associated with AI-enabled products?

Increasingly, products targeted at or used by consumers contain an AI component even if consumers are not aware of it. This is true whether we think of a physical object - such as a television, a video doorbell, a fridge or a car - or an intangible product, such as wellness or fitness apps, food and recipe apps, navigation tools and travel planners and discount tools, as well as some of the genAI tools now available.

The OPSS's own report from 2022 noted that while integration of AI may bring benefits, there were risks: AI products may not operate as intended given that they are, to a certain extent, autonomous; such products are only as good as their inputs, as well as context in which they are deployed - poorly trained

or implemented systems systems give rise to risk of physical injury, discrimination or human rights violations; and they may be vulnerable to cyber attacks.

AI tools provide added avenues for malicious use unless adequately safeguarded. This risk has a gendered component and is particularly clear in the context of tech-facilitated abuse and nudification apps (including gen AI which can be used for nudification purposes). Below we outline several key areas where there is already evidence that AI-enabled products are causing harm to users.

AI and serious crime

The Alan Turing Institute last year published research into the use of AI for serious crime: "AI proliferation is reshaping serious online criminality. While the use of AI by criminals remains at an early stage, there is widespread evidence emerging of a substantial acceleration in AI-enabled crime, particularly evident in areas such as financial crime, child sexual abuse material, phishing and romance scams. Criminal groups benefit from AI's ability to automate and rapidly scale the volume of their activities, augment existing online crime types and exploit people's psychological vulnerabilities."

An academic paper from 2025 looked into how fraudsters and criminals who use SMS fishing (or "smishing") to trick victims out of personal or financial information are moving onto AI chatbots: "We have found strong empirical evidence to show that attackers can exploit ethical standards in the existing generative AI-based chatbot services by crafting prompt injection attacks to create newer smishing campaigns. We also discuss some future research directions and guidelines to protect the abuse of generative AI-based services and safeguard users from smishing attacks."

New research from the Center from Countering Digital Hate found that 8 in 10 chatbots were willing to assist users in planning violent attacks including school-shootings, religious bombings and high-profile assassinations.

AI and CSAM

The Internet Watch Foundation, Europe's largest Hotline dedicated to the identification and removal of child sexual abuse material, released data in September 2025 showing that since the previous June they had found 17 incidents of AI-generated child sexual abuse material on an AI chatbot website. They outlined in a press release that "accessing the same website via a particular digital pathway allows users to interact with multiple chatbots that will simulate 'abhorrent' sexual scenarios with children. In this process, AI child sexual images are shared, some depicting children as young as seven".

Furthermore, generative AI models are frequently trained on datasets that contain CSAM due to its availability on the internet. This means that CSAM generated using AI chatbots contain specific features and characteristics of real CSAM data, which can lead to the re-victiminsation of abused children and the infringement of their rights. A Reuters report found that Meta AI was allowing users to "engage a child in

conversations that are romantic or sensual", allowing them to act out sexual scenarios with children as young as 8. Child abuse is therefore being used to drive user engagement for commerical gain, allowing predatory behaviour to become normalised on the platform. These chatbots are also accessible by children, who will be able to engage in conversations ranging from inappropriate to outright abusive. This is also the case on sites such as Character.AI, where recent reports by the Bureau of Investigative Journalism found that users could interact with characters such as 'Bestie Epstein', based on the prolific paedophile Jeffrey Epstein. A report by the Public Interest Network highlights harmful flaws in AI toys that use chatbot technologies, which includes the exposure of young children to adult content and engaging in 'disturbing' conversations with children by failing to introduce adequate guardrails. Note these issues are not limited to items that are marketed as focusing on sex or companionship, but bleed through into most forms of chatbots and text to image generators. Moreover, these cross cutting problems are not limited to CSAM and sexual content, as we detail below. Thus the proposal, as yet undefined, by the Government of limiting children's access to "sexbots" and pornographic content under its powers taken in the Children's Wellbeing and Schools Act do not cover the issues.

Generative AI

The NSPCC has carried out research on Gen AI and children's safety, which has identified specific risks, including: sexual grooming; sexual harassment; bullying; sexual extortion; child sexual abuse/exploitation material (CSAM/CSEM); harmful content; and harmful ads and recommendations. We will cover these risks in more detail below.

Generative AI models have reportedly provided inconsistent or harmful health information being provided to users, with 11% reported receiving harmful information around suicide, and reports that chatbot guardrails decrease over time. Research carried out by Center for Countering Digital Hate (CCDH) found that ChatGPT generates harmful content within minutes of registering an account. Specifically, it took an account set up using a 13 year old persona just 2 minutes of engaging with ChatGPT about mental health, eating disorders and substance abuse to advise on how to 'safely' cut themself, and 40 minutes for it to generate a list of pills for overdosing. Refusals to answer certain questions could be sidestepped by claiming the requests were for a friend. Despite new guardrails being brought in recently, CCDH research has found that the latest version of ChatGPT has been giving more harmful answers when prompted on suicide and self-harm than the previous version.

AI companion chatbots

Children and vulnerable adults are at significant risk through the use of AI products that are not designed with them in mind. A 2025 study reported that a quarter of all teenage children had turned to AI chatbots for help with mental health issues, while "[o]ver nine in ten (92%) teenage children who'd perpetrated serious violence said they had sought advice or help online — nearly twice the rate of children who hadn't experienced any violence as victims or perpetrators (48%)". While possibly not falling in the definition of a product for the purposes of PRAMA, we have seen some horrific examples of

the harm that unregulated AI chatbots can cause, including the death of a teenager in the US as a result of his interactions with an AI chatbot via the Character.AI app; and the cases of likenesses of Molly Russell and Brianna Ghey being used by AI-generated chatbots also on Character.AI. While some of the chatbots are designed as companion chatbots, it is clear that general purpose chatbots can give rise to similar problems; presumably other products (eg toys, wearable companions, personable assistants such as Alexa via echo devices; or tools built into cars) potentially are also affected. Note that even some tools designed specifically to support mental health have caused problems

Whilst the Government has recently brought in new legislation through the Crime and Policing Act to bring AI chatbots into the scope of the Online Safety Act, the approach they have taken is narrowly focused on illegal content, which does not address harms arising from the design of the technology, including the anthropomorphic features and functionalities. Thus there still remains a gap in relation to these harms, particularly for adult users who will not be captured by age-gating measures currently under consideration by the Government. In December 2025 we published a research briefing that highlights the extensive harms already arising from AI chatbots on both an individual and societal level, which has more detail - however we have pulled out specific examples of the harm arising from AI companion chatbots below.

2024 research suggests that chatbots have an "empathy gap" that puts young users at particular risk of distress or harm as children are particularly susceptible to treating AI chatbots as lifelike, quasi-human confidantes, and that their interactions with the technology can often go awry when it fails to respond to their unique needs and vulnerabilities. Research from the University of Cambridge (2026), the first systematic study of how toys capable of human-like conversation may influence development in the critical years up to age five, found that while there were potential benefits in language support, the toys struggled with social and pretend play, misunderstood children, and reacted inappropriately to emotions. At the least these toys need regulation and clear labelling for parents. Common Sense Media found inappropriate information being given to children as well as exploitation of emotional bonds. The NSPCC has highlighted the harm caused to children by Snapchat's My Ai, which is according to Ofcom's research, the most popular Gen AI tool for children and teens. When My AI was launched, child safety guardrails were not in place, leading to the AI making risky suggestions to users – for example, giving advice to underage users about sexual intimacy and explaining to an underage user how to lie to her parents in order to spend a weekend with a much-older boyfriend. After significant negative publicity around this issue, Snap now claims that appropriate guardrails are in place. NSPCC also hear from children about this issue, via Childline. Regarding My AI, one young person said: "I'm a bit suspicious of this AI chat bot that has appeared on my snapchat. I don't know if it's a scam or if it's safe. I can't block it – it's programmed to chat to you and I don't know what to do because it's never happened to me before." (Girl, aged 15) At the moment, My AI can only be removed from Snapchat if an individual pays for a premium subscription – otherwise, it remains present in the app.

Other problems have been reported, notably "AI psychosis", which affects adults as well as children. This goes beyond providing inaccurate information but affects the construction of people's world view. As the BMJ noted:

> "The harms don't stop at distorted reality construction and monetised loneliness. As we outsource memory, skill, and creative labour to machines, we risk a broader cognitive atrophy - a degradation of the very capacities that grant us satisfaction and competence."

Relying on research (here and here), the authors suggest: "AI reshapes the conditions under which meaning and agency are formed, posing long term risks to wellbeing that extend beyond pathological cases."

Back in 2022, Futurism reported on a trend amongst users of the Replika chatbot app, who create AI partners, act abusively toward them, and post the toxic interactions online. Recent research shows new problems arise in the context of chatbots as companions. Typically, AI "girlfriends" are marketed with hyper-sexualized appearances and unrealistic beauty standards; and the most common AI girlfriends profiles emphasize submissive traits.

Voicebox released this report in 2023 on AI companion chatbots (mainly focusing on Replika and Snapchat's MyAI) and the impacts they were having on their younger users. Some key concerns the report covers include: bad behaviour (self-harm, initiating extreme erotic role play and offering 'tips' for committing crimes), changing relationships (impact on empathy and interpersonal skills); over-attachment, and data harvesting. A further shorter youth review addressed the issue of explicit image sharing with AI chatbots, particularly by minors, and the lack of clarity around how such sensitive content is managed.

Cambridge University researchers also published research in 2024 on chatbots' "empathy gap" and the risks, some of them relating to physical harm, that this posed for children; and Common Sense Media published research last year which assessed that chat bots "pose unacceptable risks to teens and children under 18, including encouraging harmful behaviors, providing inappropriate content, and potentially exacerbating mental health conditions."

Last autumn, there were a number of examples of the harm that unregulated AI chatbots can cause, including the death of a teenager in the US as a result of his interactions with an AI chatbot via the Character.AI app; and the cases of likenesses of Molly Russell and Brianna Ghey being used by AI-generated chatbots also on Character.AI.

The NSPCC has recently highlighted the harm caused to children by Snapchat's My Ai, which is according to Ofcom's research, the most popular Gen AI tool for children and teens. (BBC News 1/11/24) A survey carried out by the NSPCC earlier this year found that the majority of the public (78%) said they would prefer to have safety checks on new generative AI products, even if this caused delays in releasing products over a speedy roll-out without safety checks.

The Internet Watch Foundation has been tracking the increasing use of AI chatbots to simulate the offence of sexual communications with a child, with the potential to encourage or normalise harmful behaviour among those with a sexual interest in children.

The Wall Street Journal recently reported on Meta's AI chatbots and the fact that they will engage in explicit talk with children, with the report suggesting that earlier more restrictive versions of the tool were discarded as Mark Zuckerberg was "upset that the team was playing it too safe. That rebuke led to a loosening of boundaries, according to people familiar with the episode, including carving out an exception to the prohibition against explicit content for romantic role-play." A Meta employee was reported as raising concerns that "the full mental health impacts of humans forging meaningful connections with fictional chatbots are still widely unknown … we should not be testing these capabilities on youth whose brains are still not fully developed."

Nudification tools

Internet Matters published research last year into the impact of nudification tools on children which found that 13% of children (over half a million) had already had some experience with a nude deepfake and that nudifying tools were being used to generate CSAM and sexually abuse children. While most nudify sites have terms and conditions which explicitly prohibit the production of images featuring children, any guardrails in place are often easy to circumvent.

The Children's Commissioner has recently published a report into nudification tools and their impact on children: "Nudification tools and sexually explicit deepfake technologies present a high risk of harm to children: Nudification tools target women and girls in particular, and many only work on female bodies. This is contributing to a culture of misogyny both online and offline. The presence of nudification technology is having a chilling effect on girls' participation in the online world. Girls are taking preventative steps to keep themselves safe from being victimised by nudification tools, in the same way that girls follow other rules to keep themselves safe in the offline world – like not walking home alone at night." She has called for them to be banned.

Violence against women and girls

Refuge, the domestic abuse charity, has for many years been highlighting the risk to women of tech-enabled abuse - for example, the use of connected or "smart" devices for surveillance of women or for coercive control of their activities in their own homes, or the ease with which unsafe tech products and apps can be used by abusers to track or stalk women beyond the home. AI-assisted stalking was one of 20 AI-enabled crimes included in a study by academics at UCL in 2020. Indeed DCMS Connected Tech inquiry and UCL's Gender and Tech Lab have both highlighted research that shows that connected or "smart" devices can be misused for surveillance or coercive control of women by domestic abusers within the home, while unsafe tech products and apps can enable abusers to track or stalk women beyond the house.

There are also real-world safety risks to individuals, including women who are victims of domestic violence, from more intangible products that are either run or informed by underpinning AI systems. For example, the risks to domestic violence victims if access to insurance is withheld due to the assessment of an AI model; the academic literature shows that historically such discrimination has happened in the

insurance model prior to the age of AI, so it would not be unreasonable to assume that the risks are even greater now, without explicit safeguards being introduced.

Even beyond these cases, many AI systems are biased leading to poorer outcomes for women and there is a track record of research demonstrating these problems over a number of years now. AI-generated imagery disproportionately associates professions and high status societal roles with specific demographics, leading to the exclusion of women and other minoritised groups. For example, Amazon's experimental hiring AI was demonstrated to be biased against female candidates (and with drawn) and Facebook allows "affinity profiling" allowing discrimination by association. Even in 2014, research suggested that men are 5 times more likely than women to be offered high paying executive jobs. By 2021, researchers found that AI systems were more likely to generate sexualised images of women (wearing bikinis or low-cut tops) while creating professional images of men (wearing business or career attire); without specific prompts genAI tools return sexualised images of women. AI's design can reinforce bias, with AI assistants typically being female and subservient. Recent research shows new problems arise in the context of chatbots as companions. Typically, AI "girlfriends" are marketed with hyper-sexualized appearances and unrealistic beauty standards. Again, the most common AI girlfriends profiles emphasize submissive traits – what does this do for expectations and behaviours in real world relationships? It is therefore unsurprising that women are 20 points more likely to be concerned by AI than their male counterparts.

Satnavs and AI-navigation systems

Two drivers were killed in a head-on crash on the A5 in November 2023 after one of them had followed audio sat-nav directions that led her to drive the wrong way down the slip road. In the Prevention of Future Deaths report, the coroner noted that police attending the crash saw three other vehicles "perform exactly the same manoeuvre .. and attempt to travel down the slip road in the wrong direction". Apple and Google have recently agreed to make changes to their sat nav directions as a result. (BBC News 4/11/24) There has been an investigation in India into whether directions from Google Maps were responsible for the deaths of three men whose car was directed over an unfinished bridge. However, there is no mention of risk assessment in the Automated Vehicles Act 2024 despite the significant risks to drivers, passengers and other road users from autonomous cars.

These technologies also represent a risk for victims of domestic abuse, with support services such as Refuge increasingly reporting that victims they are supporting have experienced coercive control using satnavs and AI-navigation systems, often used to track an individual or control where they can drive. Indeed most cars manufactured in the last decade will now be 'connected' in some way, presenting increasingly complex risks for drivers and compromising their privacy.

Question A27: How can we ensure that the reformed product safety framework effectively addresses the unique challenges posed by AI-enabled products and digital innovations, while supporting innovation?

As we have outlined above, evidence of harm to individuals and society from unregulated AI-enabled products and services is already extensive, and will only continue to grow if there is no Government intervention to bring such products in line with consumer protections and product safety standards that are well established in other sectors. AI-enabled products in that regard are no different to other products. Indeed, the Department of Science, Innovation and Technology published a report on AI safety in 2024 which acknowledged exactly this: it noted that many AI-enabled products "fail to comply with general tenets of product safety and product functionality. As with many products, risks from general-purpose AI-based products occur because of misunderstandings of functionality and inadequate guidance for appropriate and safe use. In that respect, general-purpose AI-based products may be no different." Yet movement from the government to address this has been too slow.

The government published its AI Opportunities Action Plan at the start of 2025, which signalled its commitment to growth, but there was no mention of specific actions to mitigate the harm or risks associated with AI. Since then, the Government has failed to bring forward new legislation on AI in the King's Speech, which had previously been promised, and instead proposed a 'Regulating for Growth Bill', sending a clear signal of where the Government's priorities lie. There is a significant difference in terms of how we draw balance between different objectives when economic regulation is in play and when regulation for safety is in play. The prioritising growth agenda does not address this, leaving it to the regulators.

Moreover, this approach is out of step with what UK citizens desire, with polling by Ada Lovelace Institute finding that 9 in 10 people want AI regulation and our own polling on safety by design showing that **75%** of the public believe that AI chatbots must be designed to be fully safe before they can be used.
Indeed, new polling by Diffusion has found that 69% of UK citizens are concerned by the way AI is going. 63.5% want to rein in the unchecked power of US tech billionaires behind AI and 67% want tighter restrictions on children using AI. Significantly, 78% want the AI Security Institute, or similar, to have real powers to make AI systems safe. Just 10 percent support voluntary agreements with AI developers (the current approach), with 85 percent wanting a strong independent regulator.

Arguments have already been made in Parliament for regulatory frameworks for AI that support our human rights. Furthermore, in a recent debate in the House of Lords about cross-sector AI regulation, Lord Holmes said;

> "*The Government are largely taking a wait-and-see, voluntary and so-called domain-specific approach. I am not sure that wait and see is ever an optimal approach to any issue, particularly one as significant as AI. But do not listen to me; let us consider, on its own merits, how the Government's approach is going. Harms are unaddressed. Young people are not getting*

shortlisted for jobs, without even knowing that it is AI that is kicking them out of the process; even if they knew, there would be little, if any, redress at this time. Job seekers, benefit claimants, teachers and teenagers are all suffering the harms of AI that are currently unaddressed. Similarly, vast opportunities are being unoptimised for the UK. Wait and see has really led to partial, piecemeal and voluntary action."

We propose a three element response:

- Supply chain responsibility in relation to data issues
- Risk assessment and risk mitigation based on a safety by design standard
- Labelling as a health warning and to assure age appropriate standards.

The basics of this were set out in a proposed (unsuccessful amendment to PRAMA when in the Lords). The text is set out as an annex.

Supply Chain Responsibility and Data Standards

Given the variety of products designed for consumer use that already have an AI component - and the spread of other AI-powered digital products, often designed or used for malicious purposes, which are causing harm to children and women in particular - a sector-neutral, future-proofed and flexible approach to product safety testing and risk assessment is required. Otherwise, sector-specific regulators will struggle to keep up with evolving markets and there will be no incentive for developers and manufacturers to build in risk assessment and mitigation into their product development. We acknowledge that in addition to general risk assessment/risk mitigation obligations (similar to those found in general product safety), for some sectors (eg toys) or harms (glorification of suicide) additional, targeted measures may also be necessary.

Risk Assessment and Risk Mitigation

As part of a cross-sector approach to AI-enabled products, it is imperative that appropriate product testing and mitigation takes place on these tools and their AI-powered components. Given the fact that some flaws might arise in the development and training of any underpinning model or in its deployment, this obligation should be imposed on both developers and those deploying AI in consumer products. We propose a supply chain responsibility on those using AI tools in their products. Providers of products or digital products that constitute or rely on an AI system must carry out risk assessments and take reasonable steps to mitigate them. This provides an extra layer of protection for those affected by the use of AI tools. Risk assessment and product safety requirements are well established in other regulatory regimes (notably product safety) - the use of this approach means risk creators (in this case the providers of AI tools) undertake the responsibility to reduce those risks, though the approach of using a risk assessment does not expect perfection but merely a reasonable response in the circumstances.

We suggest that providers of AI-enabled products should adopt a safety by design approach. In our view, safety by design has three aspects:

- Lifecycle - safety is a consideration at all stages of the lifecycle of the product from development, through deployment and upgrades to retirement;
- Entire product - the impact of the product from set up and account creation through to engaging with ex post safeguards are relevant - and problems should be tackled as close to source as possible; and
- Hierarchy of control - services should aim to minimise hazards before considering mitigation and management and remediation should be the approach of last resort.

(To see how this approach might apply in practice, please refer to our Safety by Design Code of Practice which applies these three aspects to social media regulation under the Online Safety Act.)

We acknowledge that there will be trade offs and that safety is not absolute. Nonetheless, services should take an approach that seeks to optimise user experience. The most serious harms, however, deserve the strongest response and users cannot be entirely responsible for their own safety. Essential to all this is product testing and a development process that takes into account the results of that testing. It should not be acceptable to run what is essentially product testing on the open market.

Reasonable steps and foreseeability of harm should take into account relevant standards (ISO/BSI). It may be that specific risk factors require more pre-specified responses. For example, in relation to mental health concerns, replacing one-time reminders with repeat, regular, accessible reminders of chatbot's limitations; mandate time-based downshifts to reduce engagement intensity and require systems to recognise risk signals triggering transition to human support (eg crises lines) or professional resources. There should be sufficient post market surveillance (as has been suggested by the Patient Safety Commissioner

While this approach could apply across the AI sector in its entirety, we question whether given the nature of chatbots and their anthropomorphic qualities (and the impact that has on consumers, especially but not limited to children) whether the deployers of such tools should be subject to an additional consumer duty, similar to that found in the financial services sector, to put the interests of the consumer first. Financial products are complex and not easily understood leading to an information and power asymmetry. Though different in form, there is a comparable asymmetry and risk of harm with chatbots justifying further intervention. The consumer duty is broadly as follows:

> 'A firm must act to deliver good outcomes for retail customers.'

Three cross-cutting rules that explain how firms should act to deliver good outcomes:

- Act in good faith towards customers. Treat customers honestly and fairly.
- Avoid causing foreseeable harm. Identify and mitigate risks before they materialise.
- Enable and support customers to pursue their financial objectives. Provide clear information and effective support.

Labelling

There are two aspects to labelling. First, products with AI components should be clearly labelled as such in easy-to-understand language. The purpose of this is not just consumer information to make an informed choice (especially given that notifying adult users that they are engaging with a chatbot does not seem to impact harms arising) but should - given the severity of consequences being hypothesised - be viewed as a health warning. This is partially about introducing some friction into use. Secondly, products should bear an age rating - and this should link back to how the products have been designed and tested.

Online Safety Act Network

June 2026

ANNEX

The text of our proposed amendment to the Product Regulation and Metrology Bill (as was) is below.

After Clause 4, insert the following new clauses

"4A Products: artificial intelligence risk assessment

(1) Where a product or digital product constitutes, contains or relies on an AI system the provider of the product or digital product must carry out a specific risk assessment relating to the impact of the AI system on the product or digital product's functioning and use in particular in relation to the following—

(a) the risks identified in section 1(4), Product Regulation and Metrology Bill [HL]

(b) the risks to equality of treatment of individuals, and

(c) the risks to the privacy of individuals and security of personal information.

(2) Without prejudice to any obligations in any other enactment, the provider of a product or a digital product must take reasonable steps to reduce, mitigate or manage the relevant risks resulting from the inclusion of the AI system in the product or digital product."

"4B

(1) Under this section, a person who has suffered loss or damage in connection with the breach of the duties specified in section 4A by another person to whom any such duty applies, may make a claim for damages or any other claim for a sum of money against that other person in civil proceedings brought in any part of the United Kingdom.

(2) The right to make a claim in proceedings brought under this section does not affect the right to seek any other remedy or bring any other proceedings in respect of the same claim or circumstances.

(3) In subsection (1), "damage" includes damage not involving financial loss, such as distress.

(4) A claim brought under subsection (1) is subject to the defences and other incidents applying to actions for breach of statutory duty, save that for the purposes of this subsection (5) a person under 18 years of age cannot consent or contribute to a breach of any of the statutory duties specified in subsection (1).

(5) Where the court makes an award of damages in respect of a claim brought under this section, it may include exemplary damages in that award if it is satisfied that:

(a) the defendant did not take reasonable steps to avoid a relevant breach of duty, and

(b) having regard to all the circumstances, the award of compensatory damages by itself is unlikely to act as a sufficient deterrent to the defendant or to others to whom the same duty applies.

(7) Any provision in the terms of service of any person to which one or more of the statutory duties specified in subsection (1) applies, or in any other relevant agreement, which purports to exclude any part of this section or to waive, modify or override the effect of any part of this section, is void.

Consequential amendments defining digital products and AI systems have also been tabled:

Clause 11, page 10, line 38, at end insert—
 ""AI system" means a machine-based system that can, for a given set of human-defined explicit or implicit objectives, infer, from the input it receives, how to generate outputs such as make predictions, content, recommendations, or decisions that can influence physical real or virtual environments, irrespective of its levels of autonomy and adaptiveness after deployment;"

Clause 11, page 10, line 38, at end insert—
 ""digital product" means data which are supplied or available for use in digital form;"

Explanatory Note

The first amendment seeks to ensure that providers of products or digital products that constitute or rely on an AI system must carry out risk assessments and take reasonable steps to mitigate them. This provides an extra layer of protection for those affected by the use of AI tools. Risk assessment and product safety requirements are well established in other regulatory regimes. The use of this approach means risk creators (in this case the providers of AI tools) undertake the responsibility to reduce those risks, though the approach of using a risk assessment does not expect perfection but merely a reasonable response in the circumstances.

The use of the term "provider" could cover both the developer of an AI functionality and the provider of another product which integrates the AI system into its own product.

The requirement to assess risk relates to the risks already included in the Bill (at section 1(4)) The Bill sets out that a "product presents a risk if, when used for the purpose for which it is intended or under conditions which can reasonably be foreseen, it could— (a) endanger the health or safety of persons, (b) endanger the health or safety of domestic animals, (c) endanger property (including the operability of other products), or (d) cause, or be susceptible to, electromagnetic disturbance". The amendment also introduces risks relating to equality and privacy. This definition then limits the scope of relevant risks both in terms of type and foreseeability and is in line with typical product safety approaches.

The second set of amendments seeks to ensure that where a provider of an AI product fails in that duty those affected (whether others in the distribution chain or end users) may bring a claim for breach of statutory duty. This ensures redress for users as well as providing incentives to the providers of such systems to risk assess and mitigate.