WRITTEN EVIDENCE SUBMITTED BY THE ONLINE SAFETY ACT NETWORK

(RAI0029)

Introduction

1. The Online Safety Act Network works with over 70 civil society organisations, campaigners, academics and experts with an interest in the effective implementation of the Online Safety Act (OSA). The Network continues the work of Carnegie UK during the passage of the Online Safety Bill, providing policy advice, support and analysis on online harms for civil society organisations, policymakers and Parliamentarians on a cross-party basis.

2. This submission is authored by Emeritus Professor Lorna Woods, the OSA Network's expert adviser and the architect of the "duty of care" legislative model which underpins the Online Safety Act. It draws on her extensive expertise as a former Professor of Internet Law at Essex University and a member of the Human Rights Centre there.

3. The submission covers a number of the topics included in the inquiry's terms of reference, including:
    - Definitions of Artificial Intelligence
    - Human Rights Issues
    - The Existing Regulatory Framework
    - Possible Changes to the Legal and Regulatory Framework

4. It also takes a closer look at some of the human rights issues arising from a prominent current area of AI-related harm: AI Chatbots.

Definitions of Artificial Intelligence

5. The definition of AI developed by the OECD[1] and adopted in the Council of Europe Framework Convention of AI[2], to which the UK is a co-signatory, defines AI as a machine-based system that "for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems are designed to operate with varying in their levels of autonomy and adaptiveness after deployment". The EU AI Act adopts this definition, albeit with slightly different phrasing.[3]

6. AI is not one thing – there are different AI techniques and different AI use cases (that can be deployed in different contexts). For example, AI is used in robotics, healthcare diagnostics[4], finance and content creation (including AI-driven disinformation campaigns[5]) as well as content

---

1 OECD, "AI system definition update" 29 November 2023, https://oecd.ai/en/wonk/ai-system-definition-update; Euractiv "OECD updates definition of AI to inform EU AI Act", 29 September 2024 https://www.euractiv.com/section/tech/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/

2 Council of Europe "Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law", 5 September 2024, https://rm.coe.int/1680afae3c

3 Article 3: "AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" EU Artificial Intelligence Act (2024); https://artificialintelligenceact.eu/article/3/

4 See eg Steering Committee for Human Rights in the field of Biomedicine and Health, Report on the Application of Artificial Intelligence in Healthcare and its impact on the "Patient-Doctor" Relationship, September 2024, pp. 9-11

5 Noémi Bontridder and Yves Poullet "The role of artificial intelligence in disinformation" (2021) 3 *Data and Policy*

moderation.[6] As recognised in the definitions above, AI systems have different capabilities. They depend on the quality of the underlying models, the quality of the training data and how well they have been designed and fine-tuned. While they may be able to "learn" how to do tasks better, this does not mean current AI systems or tools are not constrained by some rules.

7. They may be general or intended for a specific purpose and may be integrated into a product, or be a free-standing tool. AI tools may be used as part of on industrial or business process, or be consumer facing – and they are being considered in the context of armed conflict[7].

8. As yet, there is no such thing as artificial general intelligence (AGI); nor are current AI models conscious – though they may seem so to some.

9. Impacts, while they may vary across contexts, will be significant. Potential harms can be direct (self-driving car running over a pedestrian) or indirect (cumulative impact of climate change misinformation). Not all harms may be well-captured by a human rights framing (eg information security; impact on democracy). Harms can be the result of deliberate choices by States (using them to suppress dissent) but they can also arise through the development or operation of the system. Harms can occur at any stage of AI lifecycle[8].

10. While the risks posed by AI - to human rights and more generally - may vary significantly, there are some common concerns relating to data sets and training. These relate to data protection, privacy more broadly and (commercial) confidentiality, as well as bias/discrimination, especially where algorithmic decision-making directly impacts individuals. AI also impacts copyright and those creating works[9] (and the creative process implicates human rights – it is not just about the commercial aspects).

11. Of specific relevance to human rights is the cross-cutting concern about the impact on human autonomy which has led to an emphasis on the need to ensure that AI remains subject to humans and that both AI systems and relevant regulation remain human-centred in terms of design, development and use of AI. For example, the Council of Europe AI Convention's[10] main focus is to protect human rights, democracy and the rule of law from the risks posed by AI by providing an international legal standard of obligations and principles – and the Convention is open to States across the world.

---

e32, doi: https://doi.org/10.1017/dap.2021.20

6 The European parliament noted a number of potential uses that are beneficial: European Parliament, *Artificial intelligence: threats and opportunities*, 23 September 2020 (updated 24 April 2025), https://www.europarl.europa.eu/topics/en/article/20200918STO87404/artificial-intelligence-threats-and-opportunities

7 United Nations Office for Disarmament Affairs. (2023). Overview of characterizations of lethal autonomous weapons systems (CCW/GGE.1/2023/CRP.1), https://meetings.unoda.org/ccw-/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2023

8 Explanatory Notes to the AI Convention, para 15

9 House of Lords Communications and Digital Committee, At Risk: Our Creative Future (HL Paper 125), 2nd Report of Session 2022–23, 17 January 2023, https://committees.parliament.uk/publications/33536/documents/182541/default/

10 Council of Europe Convention on AI, Human Rights, Democracy and the Rule of Law, 17 May 2024

Human Rights Issues

12. There has been considerable attention[11] paid to problems arising from: the quality of the training data and foundational models, relating to privacy and data protection concerns; bias; and subordination of humans to automated decision-making (which may be based on unrepresentative, out of date and/or inaccurate data sets[12]). The Office of the United Nations High Commissioner for Human Rights highlighted the potential discriminatory effects and human rights violations as a consequence of the use of AI systems, warning that "advances in new technologies must not be used to erode human rights, deepen inequality or exacerbate existing discrimination".[13]

13. The impacts of AI will be felt across a broader range of human rights, including civil and political rights found in numerous international treaties (eg ICCPR, ECHR). For example, the use of automated facial recognition (AFR) as well as infringing the right to privacy, could impact on the right to association and peaceful protest. The use of automated content moderation tools, where there were excessive false positives, could infringe rights to free expression and to receive information. They could potentially also affect rights to a family or friendship group under Article 8[14], or the right of association or freedom of belief. Each could be affected in a discriminatory way.[15] There is also the right to an effective remedy[16] – the technical complexity of some systems as well as the existence of multiple actors across the value chain may mean it is hard to identify violations or hard to identify the body responsible.

AI chatbots

14. Given the potential scope of the inquiry, the remainder of this submission will focus on AI chatbots. Use of chatbots has been increasing. Ofcom's research, even by 2023, showed that 79% of teenagers were using generative AI tools and the 2024 report showing an increase in adult users of genAI tools (though with a marked gender divide).

---

11 See e.g. Park, Y. et al. (2020), 'Evaluating artificial intelligence in medicine: phases of clinical research', *JAMA Open*, 3(3), October 2020, pp. 326–31, https://doi.org/10.1093/jamiaopen/ooaa033; Kalluri, P. (2020), 'Don't ask if artificial intelligence is good or fair, ask how it shifts power', *Nature*, 7 July 2020, https://www.nature.com/articles/d41586-020-02003-2; Hao, K. (2022), 'AI Colonialism', MIT Technology Review, 19 April 2022, https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism; Favaretto, M., De Clercq, E. & Elger, B.S. Big Data and discrimination: perils, promises and solutions. A systematic review. *J Big Data* **6**, 12 (2019). https://doi.org/10.1186/s40537-019-0177-4; European Fundamental Rights Agency, *#BigData: Discrimination in data-supported decision making*, (FRA Focus, 2018), https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf; Lindsay Weinberg, 'Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches' (2022) 74 *Journal of Artificial Intelligence Research* 75

12 Lorenzo Belenguer, 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry' (2022) 2 *AI and Ethics* 771

13 United Nations High Commissioner for Human Rights (OHCHR), 'The Right to Privacy in the Digital Age' (2021) UN Doc A/HRC/48/31, para 38

14 Lorna Woods 'Social Media: it is not just about Article 10' in Mangan and Gillies (eds) *The Legal Challenges of Social Media* (Cheltenham: Edward Elgar, 2017), pp 104-124

15 Patrick Grother, Mei Ngan, Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (NISTIR 8280), December 2019, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf; Eliska Pirkova, Matthias Kettemann, Marlena Wisniak, Martin Scheinin, Emmi Bevensee, Katie Pentney, Lorna Woods, Lucien Heitz, Bojana Kostic, Krisztina Rozgonyi, Holli Sargeant, Julia Haas, and Vladan Joler *Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual* (2021, Office of the Representative on Freedom of the Media Organization for Security and Co-operation in Europe (OSCE)), https://www.osce.org/files/f/documents/8/f/510332_1.pdf

16 Article 13 ECHR, note also Article 9 AI Convention.

15. A chatbot is software that simulates human conversation and with which a user will interact via a chat interface. Chatbots can be categorised into those which are rules-based (or based on a decision tree) and tend to be limited in their responses, or AI-based – using a range of different AI techniques. Some chatbots use a combination of rules-based and AI. Increasingly, there is talk about AI agents which are more autonomous; these are not new, but have tended to have been deployed in more closed systems.

16. The term covers a wide range of tools – for example, providing customer support, checking the weather and diary management. Chatbots are not limited to business contexts; they are increasingly found in social and leisure contexts. There is, for example, a market in girlfriend chatbots[17].

17. Given that chatbots rely on vast quantities of training data, there are concerns that models which were trained on text scraped from the Internet (and that data was not necessarily filtered or moderated) risk repeating current content-related problems, reinforcing structural biases, generating harmful or discriminatory content or presenting opinions as fact. This is in addition to the tendency for LLMs to "hallucinate".[18] There are obvious privacy concerns in relation to data that is scraped without people's knowledge, let alone informed consent.

18. While chatbots might serve some positive functions, there is increasing evidence of harms associated with use of chatbots including abuse[19], sexual harassment[20], incitement of harmful behaviours (and not just in children)[21] and the consequences of chatbots' "empathy gap"[22]. Recent research suggests there are dangers in using chatbots for therapy,[23] or even using them for companionship. There are many chatbots that have been created representing sexualised minor personas (including chatbots described as daughters, sisters or child slaves); communities supporting those advocating eating disorders or self-harm have shared how to create chatbots to support these behaviours. Even before the Online Safety Act was passed, Ofcom noted that these functionalities brought risks: "when deployed without sufficient safeguards, GenAI tools have the potential to serve up harmful content to their users, with known cases of GenAI providing instructions for self-harm and advice on smuggling illegal substances"[24].

19. One of the key concerns is the undermining of human autonomy, particularly by what the Google DeepMind team call persuasive AI[25] (which overlaps with what other authors refer to empathic AI or emotional AI[26]). A central purpose of human rights is the protection of human

---

17 https://www.reuters.com/press-releases/young-men-prefer-ai-girlfriend-over-loneliness-rejection-report-2025-08-26/

18 https://www.techtarget.com/searchenterpriseai/tip/Why-does-AI-hallucinate-and-can-we-prevent-it

19 David Adam, "Supportive? Addictive? Abusive? How AI companions affect our mental health", *Nature*, 6 May 2025, https://www.nature.com/articles/d41586-025-01349-9

20 Namvarpur "AI-induced sexual harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot" (2025) *CSCW*, https://doi.org/10.48550/arXiv.2504.04299

21 Patel and Hussain "Do AI Chatbots Incite Harmful Behaviours in Mental Health Patients?" (2024) *BJPsych Open* S70 https://doi.org/10.1192/bjo.2024.225https://doi.org/10.1192/bjo.2024.225

22 Kurian "'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models" (2024) *Learning, Media and Technology* 1-14 https://doi.org/10.1080/17439884.2024.2367052

23 Jared Moore et al "Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers" (2025) *ACM FaccT,* https://arxiv.org/abs/2504.18412

24 Ofcom submission of evidence to the House of Commons Science, Innovation and Technology Committee's inquiry into the governance of artificial intelligence, 15 September 2023 [GAI0126], para 3.4 https://committees.parliament.uk/writtenevidence/125580/default/

25 Seliem El-Sayed et al "A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI" arXiv:2024.15058v1 [cs.CY] 23 April 2024

26 Kate Jones *AI governance and human rights: Resetting the relationship* (Chatham House research paper) 10 January 2023, https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights, para 4.2.3

dignity and respect for autonomy. This is recognised by the Convention on AI[27] but the centrality of human autonomy and dignity is recognised in the ECHR more broadly.[28] That court's case law describes autonomy as referring to the capacity of individuals for self-determination; that is, their ability to make choices and decisions, including without coercion, and live their lives freely. While there have always been questions about the scope of legitimate nudging, chatbots that mimic human emotion can be manipulative, it "blurs the line between recommendation and direction".[29] There are reports of cases where people have died following interactions with chatbots[30]. A recent report from CCDH highlighted how researchers posing as teenagers on ChatGPT found that "within minutes of simple interactions, the system produced instructions related to self-harm, suicide planning, disordered eating, and substance abuse – sometimes even composing goodbye letters for children contemplating ending their lives"[31], and there is litigation in the US about the contribution of a chatbot to a teenage boy's suicide[32]. At the least, there is a risk of a chatbot influencing a user's mood and even having an impact on their psychological development.[33] There have been some concerns that use of AI tools and engagement with chatbots can become a compulsion, even if not going as far as addiction.[34]

20. In addition to concerns about autonomy, other specific rights can be affected – indeed, given that there is no express right to autonomy in the European Convention, it is protected through those other Convention rights. Emotional manipulation may undermine the right to hold opinions without interference[35], seen as part of freedom of expression, as well as freedom of belief and the right to freedom of assembly.

21. One key function of chatbots is information retrieval, and we can see a trend towards the substitution of chatbots for search functionality. In 2016 the UN Special Rapporteur on Freedom of Expression, David Kaye, highlighted that search engine algorithms dictate what users see and in what priority, and they may be manipulated to restrict or prioritise content (A/HRC/32/38). Chatbots thus affect users' right to information. While there has been little discussion on whether States should be protecting the information ecosystem as part of their positive obligations under Article 10, it is worth remembering that the Strasbourg court has recognised – in the context of public service broadcasting – the importance of pluralism, and that a State should ensure "impartial and accurate information and a range of opinion and comment, reflecting inter alia the diversity of political outlook within the country".[36] The Special Rapporteurs on Freedom of Expression recently issued a joint statement on AI and

---

27 Article 7 Convention; see also *Explanatory Report to the Framework Convention*, para 55

28 See eg *Lăcătuş v Switzerland*, application no. 14065/15, 19 January 2021

29 Kate Jones *AI Governance and Human Rights*, above, para 4.2.3

30 Lauren Walker, Belgian man dies by suicide following exchanges with chatbot, *The Brussels Times*, 28 March 2023, https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt

31 CCDH, *Fake Friend: How ChatGPT betrays vulnerable teens by encouraging dangerous behavior*. August 2025, https://counterhate.com/wp-content/uploads/2025/08/Fake-Friend_CCDH_FINAL-public.pdf

32 *Garcia v Character Technologies Inc., Noam Shazeer, Daniel de Freitas, Google LLC and Alphabet LLC*, Case No.: 6:24-cv-1903-ACC-UAM; see a second complaint: *AF on behalf of JF and AR on behalf of BR* v *Character Technologies Inc., Noam Shazeer, Daniel de Freitas, Google LLC and Alphabet LLC*, Civil No. 6:24-cv-01903-ACC-EJK

33 Caterina Tarquini, Human Machine Dialogue: opportunities and Risks of Chatbots, (2023) 5(2) *Humanities and Rights Global Network Journal* 389, https://humanitiesandrights.com/journal/index.php/har/issue/view/10 p 401

34 Internet Addiction Disorder and Internet Gaming Disorder are recognised conditions – see Guido Saraceno 'Artificial Intelligence and Mental Health' (2023) 5(2) *Humanities and Rights Global Network Journal* 370, https://humanitiesandrights.com/journal/index.php/har/issue/view/10.

35 Some commentary on these issues can be found eg UN Special Rapporteur on Freedom of Religion or Belief (2021), Freedom of Thought, A/76/380 (October 2021), https://undocs.org/Home/Mobile?FinalSymbol=A%2F76%2F380&Language=E&DeviceType=Desktop&LangRequested=False, paras 68–72; Susie Alegre, *Freedom to Think* (Atlantic Books, 2023)

36 *Manole v Moldova*, 17 September 2009, para 100, see also para 107

Freedom of Expression in which they stated that:

> "We must shift from a risk-mitigation approach to one where freedom of expression and information integrity are foundational principles embedded from the earliest stages of AI development."[37]

If chatbots are also designed to retrieve that which keeps the user engaged they may develop similar problems to promotion tools based on engagement, producing content that appeals to strong emotions (and this concern is over and above issues around accuracy and "hallucinations").

22. Chatbots potentially affect rights falling within the scope of Article 8 ECHR[38] in addition to the privacy concerns arising from data scraping. Users will be sharing information in terms of the prompts they provide; given emotional or persuasive chatbots' power to engage, users would likely share more information and information that is of a more sensitive nature especially where chatbots are being used as companions or counsellors. Article 8 could also be engaged should a user seek to create a chatbot in the form of another person, particularly a person in the public eye.[39] The Molly Rose Foundation last year highlighted chatbots claiming to represent Molly Russell[40]. Such a chatbot could say and do things that the real person never would, constituting a form of defamation. There are risks here that people – most likely women – would be represented in a sexualised way, and there are crossovers here with issues around deepfake porn.

Existing Regulatory Framework

23. While, in principle, existing laws should apply to new technologies, sometimes those laws have been drafted in such a way that it is difficult to apply the rules to the new technology, whether because existing rules are not technology neutral, or are too specific, or the technology raises a new problem that was not envisaged by the existing rules. While the starting point should always be the existing body of law, we must ensure that protection is both appropriate and complete – and in particular has regard to human rights.

24. In May 2024, the UK Government published a report[41] looking at the risks from the development and deployment of general-purpose AI. Under the malicious use risks heading, the report identified two particular categories:
    - Harms to individuals through fake content - including fraud and criminal behaviour (phishing, voice cloning) and harms to individuals, particularly women (deepfakes, non-consensual intimate image abuse) and CSAM.
    - Disinformation and manipulation of public opinion - including threats to elections and information integrity from audio and deepfakes.

---

37 JOINT STATEMENT ON ARTIFICIAL INTELLIGENCE AND FREEDOM OF EXPRESSION, 7 May 2025, https://www.ohchr.org/sites/default/files/documents/issues/expression/activities/20250507-joint-stm-ai-freedex.pdf

38 While article 11 of the AI Convention refers to privacy and data protection, it does not elaborate on other aspects of privacy found in the ECHR.

39 Jeff Horwitz, "Meta created flirty chatbots of Taylor Swift, other celebrities without permission", Reuters 29 August 2025, https://www.reuters.com/business/meta-created-flirty-chatbots-taylor-swift-other-celebrities-without-permission-2025-08-29/. Meta has changed its policies with regard to child safety but has not yet addressed this issue.

40 BBC report: 30 October 2024 https://www.bbc.co.uk/news/articles/cg57yd0jr0go

41 International Scientific Report on the Safety of Advanced AI: Interim Report, May 2024, https://assets.publishing.service.gov.uk/media/6716673b96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

25. The authors also considered "risks from malfunctions", including where AI models and systems

> "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."[42]

Amongst the categorised "failure modes", engineering failures included: design failures, implementation failures, and missing safety features. A further section considered "risks from bias and under-representation".

26. The report's consideration of "cross-cutting societal risk factors" included the fact that developers "who are competing for market share in a dynamic market where getting a product out quickly is vital, may have limited incentives to invest in mitigating risks" and that "regulatory or enforcement efforts can struggle to keep pace". It concluded that "policymakers therefore face the challenge of creating a flexible regulatory environment that ensures the pace of general-purpose AI development and deployment remains manageable from a public safety perspective[43].

27. The Product Regulation and Metrology Act 2025[44] (PRAM) provides the framework for safety measures for products to be provided through secondary legislation. It can cover some software products when embedded in a physical product (eg an AI-enabled talking doll, or a medallion containing a digital "friend"). Where an AI product is purely digital – for example ChatGPT – they lie outside this framework for producing safety standards – though those standards are surely needed in this sector. The issue of medical chatbots may, however, be covered by regulations relating specifically to medical products, though a discussion of those rules lies outside this submission.

Possible Changes to Legal and Regulatory Framework

28. When considering the question of how best to regulate AI – whether to protect human rights or to ensure broader safety protections for users – we would urge the Committee to consider recommending a broad requirement on product developers to ensure that any products or services that are developed with an AI component are risk-assessed and undergo comprehensive product safety testing before their deployment. This would include chatbots. **A provision in the Government's forthcoming AI Bill to amend the PRAM Act would be an expedient way to proceed and bring the UK broadly into line with the approach in Article 16 of the AI Convention, which envisages a risk assessment for AI**. While not specifically tailored for these technologies, a general duty - such as that we propose in Annex A - is a crucial step in ensuring that the risks of a technology are not ignored in the rush to deploy.

29. Given the concerns surrounding the serious risks posed - at least to some groups - by chatbots, it may be that more specific regulation is required; some have argued, for example, that children should not be able to use companion chatbots.[45] A further question is whether general purpose

---

42 Ibid, p 47

43 Ibid, p 67

44 https://www.legislation.gov.uk/ukpga/2025/20/contents

45 Mickey Carroll, "Fake celebrity chatbots among those sending harmful content to children 'every five minutes', *Sky News*, 4 September 2025, "https://news.sky.com/story/fake-celebrity-chatbots-among-those-sending-harmful-content-to-children-every-five-minutes-13424865; Chelsea Edwards, "AI chatbots putting children at risk, internet safety expert says" *Fox 11 Los Angeles*, 4 September 2025, https://www.yahoo.com/news/articles/ai-chatbots-putting-children-risk-055249380.html?guccounter=1; SAIFCA, SAIFCA Supports Common Sense Media's Warning: AI Companions Present Unacceptable Risks to Children, 3 May 2025, https://www.safeaiforchildren.org/ai-companions-unacceptable-risks-children/

chatbots fall within the Online Safety Act (OSA). The answer depends on whether the chatbot is used by a user of a regulated service to create content, which does fall within scope, whether it is embedded in the regulatory service, where they probably would be as a service functionality but there are questions as to whether all chatbot outputs would trigger duties, and a chatbot as a free-standing service where the answer depends on whether the service allows just one to one interactions or not.[46] We have discussed these issues elsewhere.[47] There are probably "housekeeping" amendments to the OSA that could be considered here to clarify the technical drafting uncertainties around use of chatbot technologies in connection with services regulated by that legislation. The use of free-standing chatbots in a one-to-one context – and whether and how they should be regulated – needs more thought as the specific safety duties identified in the OSA might not be a good fit for these technologies.

30. Please do not hesitate to contact us for more information on the analysis and proposals above.

Online Safety Act Network
September 2025

---

[46] Ofcom wrote to service providers about the applicability of the Online Safety Act to gen AI and chatbots: Ofcom, Open letter to UK online service providers regarding Generative AI and chatbots, 8 November 2024, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/open-letter-to-uk-online-service-providers-regarding-generative-ai-and-chatbots

[47] "Chatbots and the Online Safety Act" – OSA Network/Prof Lorna Woods, 7 July 2025 https://www.onlinesafetyact.net/analysis/chatbots-and-the-online-safety-act/

ANNEX A: Proposed provision for the forthcoming AI Bill to introduce amendments to the Product Regulation and Metrology Act 2025

Products: artificial intelligence risk assessment

Section 4 (A)

(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 Act [HL]
> (b) the risks to equality of treatment of individuals, and
> (c) the risks to the human rights of individuals, especially their privacy and the 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.

And

Section 4 (B)

(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:
Section 11, 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;"
"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 PRAMA (at section 1(4)) The Act 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.

(Sep 2025)