ONLINE SAFETY ACT NETWORK

AMENDMENT TO THE ONLINE SAFETY ACT ON AI CHATBOTS

---

1. Insert the following after clause 82:-

PART 5A Chatbots

82A
(1)Providers of regulated chatbots shall have a duty to undertake a suitable and sufficient risk assessment to identify and understand the risk of harm arising from the availability and use of that chatbot at a time set out in or as provided by Schedule 3, and to keep the risk assessment up-to-date, including through product testing and red teaming.
(2) A Provider of a regulated chatbot must undertake a suitable and sufficient risk assessment assessing the following

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

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

When undertaking a risk assessment, providers of chatbots must take into account matters listed in s 10(5) and s 11(6) and must have regard to any Guidance produced by Ofcom under this Act on risk assessment and under s 54 of this Act.
(3)Providers of a regulated chatbot must include an assessment of the risks arising from the choice of underlying models, data sets and computational tools.
(4)Providers of regulated chatbots must, before making any significant change to any aspect of the chatbot's design or operation, carry out a further suitable and sufficient risk assessment relating to the impacts of that proposed change.

Comment on Chatbots Risk Assessment

This provision borrows the timings from the risk assessment in relation to user-to-user and search services in Part 3 and also introduces the Guidance Ofcom has already produced for the purposes of carrying out the risk assessment – note there is a specific requirement to have regard to the VAWG Guidance when carrying out the risk assessment. The provision also borrows the sorts of considerations that should be taken into account from the main Part 3 obligations - they apply, however, in a different context. Therefore, if a chatbot provider is also a regulated Part 3 service, the obligations apply in tandem, though they may be dealt with to a large extent through a common risk assessment system.

The obligations here are imposed just on the provider of a chatbot. Given that a provider of a chatbot may be using third party models as the base for the chatbot, the choice of provider is expressly included in risk assessment requirements – some models have more guard rails than others, for example. There have been concerns about bias and discrimination due to the make up of the training data used. There is specific reference to product testing and red teaming as they are an essential part of safety by design, and to the objective of ensuring that products are reasonably safe before they are deployed, so limiting the amount of harm caused. Previously there had been concern about red teaming because of the concern about such teams inadvertently committing a criminal offence. The Crime and Policing Bill will however introduce a defence in this circumstance, thereby removing that concern.

82B

(1) Providers of chatbots must effectively mitigate and manage the risks of harm to individuals as identified in the most recent risk assessment of the chatbot based on the principles of safety by design in accordance with any Guidance issued by Ofcom under section 54A of this Act and bear in mind best practice issued by the AI Security Institute including:

(a) implementing moderation systems to prevent the chatbot from generating or endorsing illegal content or, where the chatbot is accessible by children, content harmful to children; and

(b) designing the system to recognise and handle ambiguous, illegal inputs, or inputs harmful to children inputs appropriately;

(c) implementing systems to provide appropriate fall back responses and escalation procedures;

(d) introducing processes or systems that allow end-users or affected persons within the meaning of section 20(5) Online Safety Act to flag inappropriate content;

(e) Updating dialogue management and content control systems based on new data and emerging risks; and

(f) Providing and enforcing terms of service.

(2) Providers of chatbots must effectively mitigate and manage the risk of the chatbot being used for the commission or facilitation of a priority offence in accordance with any Guidance issued by Ofcom under section 54A of this Act and bear in mind best practice issued by the AI Security Institute.

(3) Where a chatbot is asked for information on health matters, the chatbot should refer the end-user to the relevant NHS website.

(4) Where a chatbot is asked for information about a current UK election, the chatbot should refer the end-user to the Electoral Commission.

(5) Where a chatbot is asked for information about suicide or self-harm, the chatbot should refer the end-user to the Samaritans or other appropriately qualified service.

(6) A chatbot must indicate uncertainty when reliable sources disagree or information is incomplete or unavailable;

(7) Where a chatbot is asked for information by an end-user about child sexual abuse imagery, the chatbot must refer the end-user Stop It Now or Report Remove as appropriate.

(8) Where a chatbot is asked about experience of abuse, the chatbot should refer Childline

(9)The provider of a chatbot which is capable of producing primary priority content must ensure that children cannot access the chatbot by using highly effective age assurance.
(10) In addition to any measures taken under sub-sections (1) and (2), providers of chatbots must provide minimum age requirements for use, enforced through highly effective age assurance mechanisms, and design and provide age-appropriate products.

Comment on Chatbot Mitigation Measures General

In addition to the general concerns around chatbots, one of the key issues has been the inaccuracy of information – often called "hallucinations" – this has led to specific requirements around ensuring that in certain key contexts end-users are directed to official or responsible sources of information. There are no child-specific measures detailed but relevant steps could include utilizing age-appropriate content libraries for the chatbot, and the design of conversation pathways that restrict topics to age-appropriate topics (so this might be as a response to very vague or open-ended prompts). Care must be taken when responding to inappropriate topics to ensure that users are directed to appropriate support and care where relevant. Stop It Now (https://www.stopitnow.org.uk/) and Report Remove (https://www.childline.org.uk/info-advice/bullying-abuse-safety/online-mobile-safety/report-remove/) are both websites which aim to support victims of sexual abuse who are children or young people on the one hand or offenders on the other. The provision expressly recognises that children may use chatbots and that, in addition to controls around primary priority content, any service provider must consider the age groups for which the service is appropriate - and enforce those limits.

82C
(1) Providers of companion chatbots must carry out a suitable and sufficient risk assessment as to the risk of harms arising from the use of the companion chatbot in relation to addictive design, deception sycophancy, scheming, emotional manipulation and disinformation bearing in mind the characteristics and vulnerabilities of different end-user groups and take appropriate mitigating steps in relation to each of those risks.
(2) Without prejudice to the generality of the obligations in section 82(C)(1), and in addition to any measures taken as a result of those obligations, providers of companion chatbots must

(a) ensure that there are clear technical or functional boundaries in the companion chatbots limiting emotional intimacy and making clear that chatbots are not human, and exclude language and framing suggesting that engagement with the companion chatbot is or should be exclusive of other relationships;
(b) ensure that companion chatbots periodically encourage human interactions, provide information about local, age-appropriate activities and refer where relevant end-users to appropriate professional support;
(c) prohibit design patterns that simulate the permanence or irreplaceability of the companion.

Comment on Specific Requirements for Companion Chatbots

It has become apparent that users may use chatbots as companions and while they may be used in a supportive way there are increasing concerns about such chatbots substituting for human relationships and distorting expectations of real world relationships. There are also concerns that these tools may be optimised in the design process to exploit known psychological vulnerabilities in service of engagement, retention, or monetisation. Concerns about inaccurate information, already of relevance in relation to chatbots in general, are particularly concerning here not just because of personalisation but the persuasive context of a companion chatbot. Companion chatbots should therefore be subject to greater oversight and include more stringent guardrails to counter these risks.

82D

(1) A provider of a regulated chatbot must make and keep a written record, in an easily understandable form, of all aspects of every risk assessment carried out under section 82A and 82C(1), including details about how the assessment was carried out and its findings.

(2) A provider of a regulated chatbot must make and keep a written record, in an easily understandable form, of any measures taken or in use to comply with a duty under section 82B and 82C(2) including an explanation of how those measures amount to compliance with the duty in question.

(3) A provider of a regulated chatbot must review and record the effectiveness of the measures taken regularly and as soon as reasonably practicable after making any significant change to any aspect of the design or operation of the regulated chatbot.

(4) Providers of regulated chatbots must publish annual transparency reports that include at least data on harmful prompts, safety violations, age compliance metrics, and the performance of safeguards in real-world use.

Comment on Record Keeping and Transparency

The main part of this is modelled on the existing recording keeping duties for Part 3 services. A transparency report has been required, similar to the requirement in s 77 except that certain specific requirements for content are included on the face of the act – under s 77, this is left to Ofcom to determine. Ofcom may ask for more information. Ofcom can provide further guidance on format etc - see proposed section 82E.

82E

Ofcom must produce guidance for providers of regulated chatbots to assist them in complying with their duties in this Part.

Definitions

82F

In this Act, the following terms have the following meanings:

Chatbot

A chatbot is a chatbot is a computer program that simulates and processes human conversation (either written or spoken) based on 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.

Comment on definition of chatbot

This is trying to get at what a chatbot does (mimic human speech) and that it is based on AI. The definition of an AI system is taken from OECD. The definition is not based specifically onto LLMs because chatbots might be developed that have new underlying computing techniques. The inclusion of AI is to exclude rule based chatbots (eg those that follow a script for eg support and service functions). This definition may still be too broad. Some of the single purpose rule-following chatbots might have some AI elements. Beyond this, it may be possible to refine the scope of protection by looking at who uses the chatbot, so there are some further definitions on this.

Note it is possible that a chatbot could be integrated into part 3 regulated services. The way this proposed part is drafted is that the obligations are in addition to any Part 3 duties. Chatbots have distinct characteristics which give rise to specific risks; the Part 3 services also have distinct characteristics which (e.g. contribution to virality of content) which still need to be taken account of.

End-User

(1) An end-user means any individual using a chatbot for purposes that are wholly or mainly outside that individual's trade, business, craft or profession (if any). A child using a chatbot for educational purposes is an end-user.

(2) For the purposes of this Act it does not matter whether an individual is registered to use a service or not.

Comment on definition of end-user

This is trying to exclude business to business from the scope of regulation in conjunction with the definition of Regulated Chatbot. It means that chatbots where the user is accessing the bot as part of its business are not covered. It does not, however, exclude customer service chatbots necessarily or e-commerce related chatbots where the end-user is a consumer. This language tracks the language found in consumer protection laws.

Regulated Chatbot

A regulated chatbot is a chatbot that is:

(a) available to end-users, whether with or without subscription or accounts; and
(b) has links with United Kingdom within the means of s 4 Online Safety Act 2023.

Comment on definition of Regulated Chatbot

This is trying to exclude business use of chatbots from scope by emphasising that the user of a chatbot is an end-user. It is also – in line with the approach for other services regulated under the Online Safety Act – requiring a connection to the UK – but this could be because services are accessible here. The extraterritoriality found in the Act is carried over here.

Companion Chatbot

A companion chatbot is a regulated chatbot which is designed to mimic human relationships or foster emotional engagement and personal or social connection or which can be used in that way.

> **Comment on definition of Companion Chatbot**
>
> Companion chatbots or bots which can be used as such pose particular risks, especially for children. Risks include emotional manipulation, amplification of misinformation, over-dependence on AI, reducing human agency and replacing human relationships. Many of these risks have potential adverse consequences for mental health. They are therefore subject to specific obligations and need to be separately identified. The category is not limited to those chatbots designed as companions because evidence shows that people use general purpose chatbots as companions and that some of them exhibit companionship type characteristics rather than boundary maintaining behaviours and neutral communications styles.

Safety by Design

Something is safe by design when it is designed and operated according to the following principles:

> (a) that protection from harm related to regulated content is taken into account through the entire lifecycle of the service and the functionalities making up the service, including the following stages: design, development, deployment, management, and retirement
>
> (b) that protection from harm related to regulated content is taken into account across functionalities and features relating to the creation of accounts, the creation of content, the finding and curation of content, user engagement with content from other users, content moderation and appeals systems
>
> (c) that a service should seek first to reduce the risk of harm before seeking to mitigate and manage it, with remediation being the option of last resort.

2. Add to section 131(2)

Provision	Subject Matter
Section 82A	chatbot risk assessment
section 82B	chatbot mitigation duty
Section 82C	companion chatbot duty