## AI CHATBOTS - CRIME AND POLICING BILL COMMONS BRIEFING

Harms caused by AI chatbots are rapidly increasing. These harms represent human rights violations and require urgent action from the government, yet **the Government's proposed approach to legislation does not cover all the harms that need addressing and takes a narrow approach focused on illegal content.** AI chatbots, like all other online spaces and tools, should not be brought to market until they are safe by design for all users. It is the responsibility of the company to prevent harms occurring in the first place rather than trying to deal with them after the fact. This is the established principle that underpins regulation in all other sectors. During Lords report stage of the Bill, Peers voted overwhelmingly in support of an approach to regulation based on proper risk assessments and safety by design, as well as a more expansive understanding of harm, including addictive and manipulative design.

**Our proposal is endorsed by the 45 organisations and individuals below - whose interests span CSAM, child online safety, VAWG, suicide and self harm, mental health, extremism, online hate and abuse, democratic participation and AI regulation. We urge MPs to join us in calling on the Government to take a more ambitious approach to regulating AI chatbots. More detail for Commons consideration of Lords amendments on 14 April is provided in the briefing below and in our research brief.**

AI Youth
CEASE - Centre to End All Sexual Exploitation
Ripple Suicide Prevention Charity
Clean Up The Internet
Institute for Strategic Dialogue (ISD)
Centre for Protecting Women Online
Fair Vote UK
Internet Watch Foundation
5Rights Foundation
Online Safety Act Network
Antisemitism Policy Trust
Molly Rose Foundation
Center for Countering Digital Hate (CCDH)
Internet Matters
End Violence Against Women Coalition (EVAW)
The Jo Cox Foundation
Parent Zone
Everyone's Invited
NSPCC
SWGfL
Suzy Lamplugh Trust
The Jo Cox Foundation
Samaritans
FlippGen
The Safe AI for Children Alliance
Refuge
Mental Health Foundation
Kick It Out
Demos
Marie Collins Foundation
Coalition to End Gambling Ads
Civic Digits
Global Action Plan
Shout Out UK
Check My Ads
Movember
UCL Gender and Tech Lab
PSHE Association
Wikimedia UK
Chris Ashworth OBE, Founder - Click Zero
Professor Gina Neff, Minderoo Centre for Technology and Democracy

Professor Julia Hornle, Queen Mary University of London
Professor Clare McGlynn
Professor Lisa Sugiura
Adele Zeynep Walton
Andy Briercliffe

***We welcome the opportunity to discuss this issue with you, and will provide further information upon request. Please contact tallulah@onlinesafetyact.net***

**Introduction**

Harm related to AI chatbots has already been widely evidenced at both an individual and societal level, including the detrimental impact on young people's mental health and wellbeing, as well as the use of AI chatbots to generate sexualised content of children, perpetrate violence against women and girls, and create non-consensual deepfake images. Just a fortnight ago, an inquest heard that a chatbot played a role in the death of a 16-year-old boy from Hampshire, advising him on the most successful way to commit suicide on a railway tracks. This has chilling parallels with many of the tragic cases that are currently the subject of legal action in the US. Other societal-wide harms, such as those related to privacy, data and security, are less frequently discussed, but nonetheless concerning.

**These harms represent human rights violations at multiple levels, and require urgent attention from the government. Regulating in a piecemeal fashion now will be a missed opportunity with potentially grave consequences for UK citizens, particularly children, young people and vulnerable adults.**

We were pleased to see Peers from all parties rally behind decisive action to hold tech providers to account for the harm caused to users by AI chatbots at Lords Report stage. Their support demonstrates overwhelming endorsement for a safety by design approach - focusing on risk assessment and mitigation across not just illegal content and content harmful to children but on the particular features and functionalities that make the risk of harm from chatbots so concerning. Many Peers' contributions focused on the need for parity with other industries where stringent product testing and risk assessments are the norm. The amendments return to the Commons on 14th April and we have written to Ministers in five Departments to urge the Government, before then, to work with Baroness Kidron and others to consider the principles raised and return to Parliament with their own proposal for more robust action to ensure AI chatbots do not continue to cause harm. At the time of writing, we have had no responses to these letters.

**The issue**

The start of the year brought with it reports of a tsunami of deepfake abuse on X's image-generating chatbot, Grok, which included millions of sexualised images of women and children. Many of these deepfakes reinforced racist tropes against Black and marginalised women, including requests to lighten skin. Reports recently revealed that Grok has also been used to create explicit and derogatory posts about the Hillsborough and Heysel disasters, as well as the death of footballer Diego Jota. Yet these reports were by no means the first of their kind, and represented a systemic failure to both prevent harm occurring in the first place and act swiftly when it was identified.

Indeed, over the last few years chatbots have been connected to serious harm, including the death of Adam Raine, who was advised by a chatbot on methods to take his own life instead of receiving support

when he needed it the most, as well as Sewell Setzer III, who fell in love with a Character.AI chatbot and subsequently took his own life.

However in its current form, the Government's proposal to bring chatbots into the scope of the Online Safety Act falls far short. The focus on 'illegal content' fails to capture the concerns of civil society, who have identified the anthropomorphic features and functionalities of a chatbot as a unique driver of harm, one that creates emotional dependency which can lead to isolation, depression, psychosis, and in extreme cases, suicide. This means that in Sewell's case, the chatbot would only be restricted from providing him with information on suicide - not from forming the emotional bond that ultimately led him to be cut off from the people around him.

In addition to concerns relating to chatbots set out above, one of the key issues identified by researchers and civil society alike has been the inaccuracy of information, often called "hallucinations". The Government's narrow approach fails to address these societal-level harms, such as threats to the democratic process through disinformation related to elections, and broader threats to the information ecosystem resulting from misinformation and bias present in training data. These threats are not abstract, and are already impacting the democratic process.

**What we need**

We need a rights-based approach to AI chatbot regulation which prevents harm from occurring by requiring robust risk assessments and safety by design. These calls have been made across the globe, including by the Council of Europe's Committee of Ministers as well as experts giving evidence to the Human Rights Joint Committee inquiry into the AI Regulation and Human Rights. This would mean that no AI chatbot provider is allowed to roll a product out onto the market that could be harmful to users. Harm must include those associated with the anthropomorphic features and functionalities, unique to a chatbot, that encourage emotional dependency and manipulation of users.

The Government should ensure that product testing and red teaming are included within their approach as they are essential to the objective of ensuring that products are reasonably safe before they are deployed and to reduce the risks of discrimination and bias in the training data they are built on - so limiting the amount of harm caused. In addition, to address concerns related to "hallucinations" and disinformation, we recommend that the Government includes specific requirements on chatbots to ensure that in certain key contexts end-users are directed to official or responsible sources of information, including the NHS, Samaritans and the Electoral Commission.

**What you can do**

While we welcome the Government's commitment to a regulatory approach, the existing regulatory framework is not equipped to deal with the unique challenges posed by AI chatbots. It must be strengthened to adapt to evolving technologies, and we need to act fast to do so.

**We are calling on MPs to support a more ambitious approach to AI chatbot regulation and to ensure the Government's regulatory approach protects all UK citizens from the risk of reckless tech companies - once again - prioritising profits over safety of their users.**