Irish Council for Civil Liberties

1st Floor
34 Usher's Quay
Dublin 8
T: +353 1 912 1640
E: info@iccl.ie
W: www.iccl.ie

Brando Benifei MEP
By email

27 October 2021

Loopholes in the scope of the Commission's proposed AI Act

Dear Mr Benifei,

We write on behalf of the Irish Council for Civil Liberties (ICCL), Ireland's oldest independent human rights monitoring organisation. ICCL monitors and campaigns for all human rights, for everyone.

We write to draw your attention to two important elements that are overlooked in the Commission's proposal:
1. AI systems with indeterminate uses, and AI defined objectives; and
2. the interplay of AI systems.

The exclusion of these elements would create a profound loophole in the AI Act, and put at risk the fundamental rights cited in 3.5 of the Commission's explanatory memorandum.

1. AI systems with indeterminate uses, and AI defined objectives

Article 3(1) defines an AI system as

> "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with"

This wording causes two problems.

First, the words "a given set of … objectives" in Article 3(1) does not encompass AI whose uses are indeterminate. One such system has propagated extremist content,[1] and other risky uses have already emerged.[2] Overlooking AI systems with

[1] Kris McGuffie and Alex Newhouse, "The Radicalization Risks of GPT-3 and Advanced Neural Language Models", Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute, 9 September 2020 (URL: https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf).

[2] For example, "Generative Pre-trained Transformer 3" (GPT-3) is a general-purpose AI that produces human-like text. Note that it is not a recommender system, in the meaning of Article 2(o) of the Commission's proposed Digital Services Act. The authors of GPT-3 themselves note a number of misuses that discriminates based on gender and race. See Section 6.2 in Brown et al. "Language Models are Few-Shot Learners", NeurIPS, 22 July 2020 (URL: https://arxiv.org/abs/2005.14165). Even the less

indeterminate uses is dangerous.

Second, the words "human-defined" in Article 3(1) exclude objectives defined by AI systems autonomously, or by other systems in an AI's environment. Recital 6 acknowledges that "AI systems can be designed to operate with varying levels of autonomy", but this is not reflected in Article 3(1).

For example, although motor vehicles are covered,[3] an AI system in a self-driving car may be excluded from the Regulation's scope in circumstances where the AI system defines its own objectives. Humans define only some objectives (destination), and AI systems define others as the vehicle interacts with its environment (for example, avoid collisions).

Recommendations:

i. The words "for a given set of human-defined objectives" should be removed from Article 3 (1).
ii. AI systems with indeterminate uses should undergo risk assessment by an external body before deployment for all foreseeable uses and misuses, so that a provider cannot choose the lowest risk category as a default.
iii. The foreseeable uses of AI systems with indeterminate uses should be included in the publicly accessible EU database referred to in Article 60.

2. The interplay of AI systems

The Commission's proposal considers AI systems in isolation. But in practice, AI systems interact with each other. Interaction of even low-risk AI systems can cumulatively give rise to high-risk.

Much as regulating an individual computer is different from regulating a network of computers, the Commission's proposal should consider the interplay between multiple AIs.

Recommendations:

i. Article 3 (13), which defines "reasonably foreseeable misuse" should be amended to include the words "and with other AI systems", as follows:

> 'reasonably foreseeable misuse' means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems**, and with other AI systems**;

powerful earlier version of GPT-3 can inadvertently reveal the personal data used to train them. See Carlini et al. "Extracting Training Data from Large Language Models", UsenixSec, 15 June 2021 (URL: https://arxiv.org/abs/2012.07805).

[3] Annex II, Section B (6).

ii. Article 6(1) should include a test for low-risk AI systems that, when combined, create foreseeable high-risks.

iii. Article 61 (2) on post-market monitoring should emphasise analysis of the AI environment (including other devices, software, and other AI systems that will interact with the regulated AI system). The Article 11 (1) and Annex IV technical documentation requirement should include the AI environment, too.

Since these are problems of scope, ICCL believes it is particularly important to fix them. Otherwise, the Regulation will fail to protect fundamental rights.

We would be happy to meet with you and your team to discuss this.

Sincerely,

Dr Johnny Ryan FRHistS
Senior Fellow, ICCL

Dr Kris Shrishak
Technology Fellow, ICCL