Pelastakaa Lapset

Feedback on the draft guidelines on transparency obligations under the AI Act

Section II - Overview of transparency obligations and horizontal topics

It needs to be clarified, in what manner the transparency obligations will take account of the minimization of biases to ensure non-discrimination and inclusion. Furthermore, clarification is needed on how the training of new AI models is reported and evaluated in a manner that ensures certainty that necessary mitigation efforts to minimize biases have been done and that they are effective.

Section III - Article 50(1): Transparency for interactive AI systems

It's vital that interactive AI systems' impacts on children for short and long term are reported in a manner that allows the data to be used to evaluate the actual effects on the realisation of children's rights. Transparency is vital in how companies report: how they detect harmful content (such as content related to grooming, bullying and exploitation), how they report such content, how they mitigate the risk of such content, and how the mitigation measures have affected the risks. This is the key transparency requirement to be able to evaluate how these AI systems affect the realization of children's rights.

It is central that transparency has strict obligations on how children are notified and informed of the use of interactive AI systems not just once when a child starts to use an interactive AI system but also throughout the interaction.

Section IV - Article 50(2): Marking and detection of AI-generated or manipulated content

Internationally, Save the Children Finland is already seeing an explosive growth in AI-generated child sexual abuse material. For example, in victim identification it is necessary to differentiate AI content from non AI content. Therefore, the guidelines need to have clear obligations on how to report the marking and detection of AI-generated or manipulated content and obligations to report how effective these marking and detections efforts are.

Marking of AI-generated or manipulated content is needed also so that children can differentiate AI and non AI content. This allows children to develop the understanding what is AI content and that it does not necessarily reflect reality.

Section VI - Article 50(4): Labelling of deep fakes and certain text publications

Deep fakes are widely used to spread mis- and disinformation and can be part of hybrid warfare. Deep fake material should be marked as such in a way that it is easy to distinguish from real content, to enable countering their harmful and dangerous use.

Additional information:

Mikko Ahtila, Advisor
Child protection and Finnish Hotline | Save the Children Finland
mikko.ahtila@pelastakaalapset.fi