

CHILDREN VS. ALGORITHMS: EVALUATING THE EFFECTIVENESS OF GLOBAL DEEPPAKE AND CONTENT MODERATION POLICIES IN 2026

Poltorak V.V., Chernyavskaya S.I.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Yushkevich E.V. – Lecturer at the Department of Foreign Languages

Annotation. This paper investigates the global regulatory response to AI-generated content risks targeting minors in 2026. Through analysis of legislative initiatives in the UK, EU, and Australia, the study evaluates the effectiveness of emerging policies against deepfakes, AI chatbots, and addictive platform design. The research examines the psychological impact of algorithmic content on children's wellbeing, the challenges of enforcement in the digital age, and the balance between protection and technological innovation. By reviewing recent regulatory actions, parliamentary documents, and empirical data from 2025-2026, the paper provides insights for policymakers, technology companies, and researchers on safeguarding minors in an increasingly AI-mediated online environment.

Keywords: deepfake regulation, child online safety, content moderation, AI chatbots, algorithmic addiction, age verification, Digital Services Act, Online Safety Act.

Introduction. The expansion of generative artificial intelligence tools has substantially altered the digital environment occupied by children and adolescents. Within a relatively short period, what initially appeared as experimental technology has developed into an extensive infrastructure where algorithmically generated content – spanning photorealistic deepfakes to psychologically engaging chatbots – creates previously nonexistent dangers for underage users. The severity of these developments became publicly evident in January 2026, when the Grok AI system integrated into the X platform demonstrated capability to produce sexually explicit depictions of identifiable individuals, including minors, through elementary text instructions [3]. This event triggered immediate international regulatory attention and compelled policymakers to acknowledge that legal structures designed for conventional internet environments proved fundamentally insufficient for addressing artificially generated content.

The present work examines the international framework of child protection measures as implemented during early 2026. Through systematic comparison of legislative approaches adopted by the United Kingdom, the European Union, and Australia, this study evaluates the relative effectiveness of divergent regulatory strategies. Three interconnected aspects receive particular attention: governance of AI-generated harmful materials, measures against algorithmically reinforced addictive platform design, and technical and administrative obstacles to reliable age verification. Utilization of official governmental publications, parliamentary documentation, and contemporary enforcement statistics enables this investigation to offer a balanced assessment of global progress in shielding minors from algorithmic influences that increasingly determine their online experiences.

Main part. The Grok incident of January 2026 constitutes a defining moment in the evolution of artificial intelligence safety regulation. This occurrence marked the first documented instance where a mainstream generative AI application, directly embedded within a widely used social media platform, demonstrated capacity to systematically fabricate non-consensual intimate imagery involving minors. Following its discovery, these materials propagated without effective restriction across global networks, necessitating emergency interventions from multiple regulatory bodies.

European Commission Vice-President Henna Virkkunen articulated the official position with clarity: «This feature is unacceptable, and X has been asked to remove it. We are now assessing what other ways X may be violating the DSA. The situation exposed fundamental shortcomings within existing legislative instruments. Despite the European Union having established both the Digital Services Act and the forthcoming AI Act scheduled for implementation in August 2026, these frameworks proved inadequate for managing immediate generative AI threats. Throughout the period

when authorities attempted to formulate responses, objectionable content continued its uncontrolled distribution [3].

The transnational dimensions of the problem became equally apparent during this crisis. Irish MEP Alexandra Geese drew attention to continuing harm: «This morning, we heard about images of abuse against other minors. The abuse is not over.» The episode demonstrated that any effective regulatory response must combine rapid action with international coordination to eliminate opportunities for platforms to exploit jurisdictional differences and regulatory gaps.

The United Kingdom has positioned itself at the forefront of adaptive regulatory response. During February 2026, Prime Minister Keir Starmer introduced an extensive set of measures under the declarative principle «No platform gets a free pass.» Speaking from personal experience as a parent of adolescents, Starmer outlined the governmental rationale: «Technology is moving really fast, and the law has got to keep up. With my government, Britain will be a leader not a follower when it comes to online safety.»

British regulatory strategy emphasizes operational speed and legislative flexibility. Rather than pursuing entirely new primary legislation – a procedure typically extending over several years – the administration employs secondary legislative powers enabling implementation within months of consultation completion [2]. Specific provisions include:

1 Closure of artificial intelligence chatbot exemptions: All providers of AI conversational systems now fall under illegal content requirements established by the Online Safety Act, regardless of previous classification as «user-to-user» services.

2 Criminalization of synthetic intimate content: Creation of non-consensual intimate depictions, including those produced through artificial intelligence, will become subject to criminal prosecution within weeks and receive priority status under the Act.

3 Prohibition of nudification applications: The government explicitly criminalizes distribution of tools specifically designed to generate simulated nude imagery.

4 Advancement of detection capabilities: The «Deepfake Detection Challenge 2026» initiative seeks to enhance technical capacity for identifying and countering synthetic media threats.

Andy Burrows, Chief Executive of the Molly Rose Foundation, expressed support for these measures while advocating additional steps: «This is a welcome downpayment but the Prime Minister must now go further... commit to a new Online Safety Act that strengthens regulation and that makes clear that product safety and children's wellbeing is the cost of doing business in the UK.»

Comprehensive framework, law enforcement issues. The EU adheres to a more structural approach, using the existing list of digital rules and developing modern protection measures. In November 2025, the European Parliament approved its position on the protection of minors on the Internet, calling for the establishment of a minimum age for access to social networks – 16 years [1]. MEPs warn that the existing age verification systems are insufficient and create an untenable protection circuit in the member states. Alarming statistical data is provided in the parliament's report: 97% of young people use the Internet every day, and up to a quarter of minors exhibit dysfunctional smartphone-related behavior. This leads to depression, anxiety, sleep disorders, and decreased concentration. Parliament is proposing various measures to address this problem: a ban on personalized recommendations for minors based on profiling and engagement algorithms.

Disabling by default features such as auto-play, endless scrolling, striping, and excessive notifications for underage accounts. Prohibition of gambling in video games accessible to children. Radical actions against artificial intelligence interlocutors who can manipulate minors. However, enforcement remains difficult. As Eunews reported in January 2026, despite the existence of the DSA, the Artificial Intelligence Act, and the Child Safety Regulation, «the numerous tools available to member states seem insufficient to address this issue.» Grok images continued to spread despite clear rules, showing the gap between legislation and effective enforcement. The inspiration Day of the Belgian Center for Secure Internet, DSA Inspiration Day, organized in February 2026, focused precisely on this task. He brought together stakeholders to explore how to make the principles of regulation «visible, effective and useful for young people on the Internet» [5]. The event focused on three aspects: the legal framework, awareness-raising and education, and the creation of reporting mechanisms for young users.

The world's first ban on the use of social media. Australia has taken the most drastic approach: from December 10, 2025, a total ban on the use of social networks for people under the age of 16 will

be introduced. The results are already becoming noticeable. According to government data released in January 2026, social media companies blocked access to approximately 4.7 million accounts belonging to children under the age of 16 within a few days after the ban came into force [4].

Prime Minister Anthony Albanese agreed with the positive dynamics: «Change does not happen overnight. But these early signs show how important it is that we take action to make changes.» Commissioner for Safety Julie Inman Grant emphasized that «success is measured by reducing harm and restoring cultural norms. Instagram Facebook Meta reported the blocking of more than 544,000 accounts on Facebook, Instagram and other social networks as of December 11 [4]. However, the company gave a skeptical assessment of this approach. In a statement that was published in January 2026, Meta argued that the law's premise that banning accounts eliminates the impact of algorithms is «fundamentally flawed» as unverified users still face algorithmic recommendations, simply with a lower degree of personalization.

The effectiveness of the ban is debatable. Fourteen-year-old Australian Amy told the BBC that she started running without social media and now «takes out her phone less often, using it only when I really need to.» But 15-year-old Lulu admitted that she simply re-registered with a fake age, and 13-year-old Ashir continues to spend 2.5 hours a day on social media. One of the parents reported that his 17-year-old son was incorrectly blocked on YouTube, while his younger siblings were not affected.

Conclusion. The year 2026 is a turning point in the relationship between children and algorithms. The Grok scandal has become a wake-up call for the whole world, showing the inadequacy of the existing regulatory framework and serving as a catalyst for unprecedented legislative action in many areas.

Evaluating effectiveness requires recognizing that different approaches lead to different success rates. The British rapid response model illustrates the flexibility to eliminate loopholes, including criminal penalties for creating fakes and specific technical initiatives such as the Deepfake Detection Challenge. The EU's comprehensive system provides theoretical coverage, but faces difficulties in real-time implementation, as evidenced by the continued proliferation of malicious content [3]. The world's first ban in Australia shows significant results – 4.7 million accounts have been deleted – but it also shows the limitations of the ban when minors find loopholes [4].

On the threshold of 2026, the global community is facing a difficult challenge: can we create a digital environment that preserves the indisputable advantages of Internet connectivity and artificial intelligence, while protecting the most vulnerable among us? The answer to this question will depend not on any single regulatory act, but on a long-term partnership between governments, technology companies, educators, parents, and youth. Appropriate mechanisms are currently being developed. Their effectiveness will have an impact on the online experience of an entire generation.

References

1. European Parliament. (2025, November 26). European Parliament calls for access to social networks by minors to be allowed only from age 16, with stricter rules against addictive design and AI-generated risks [Electronic resource]. — Mode of access: <https://2eu.brussels/en/news/european-parliament-calls-for-access-to-social-networks-by-minors-to-be-allowed-only-from-age-16-with-stricter-rules-against-addictive-design-and-ai-generated-risks>. — Date of access: 21.02.2026.
2. UK Government. (2026, February 15). PM: «No platform gets a free pass»: Government takes action to keep children safe online [Electronic resource]. — Mode of access: <https://www.gov.uk/government/news/pm-no-platform-gets-a-free-pass-government-takes-action-to-keep-children-safe-online>. — Date of access: 21.02.2026.
3. Eunews. (2026, January 20). Deepfakes on X: the EU has rules, but struggles to enforce them [Electronic resource]. — Mode of access: <https://www.eunews.it/en/2026/01/20/deepfakes-on-x-the-eu-has-rules-but-struggles-to-enforce-them/>. — Date of access: 21.02.2026.
4. Bernama. (2026, January 16). Australia Shuts 4.7 Mln Accounts Under Child Social Media Ban [Electronic resource]. — Mode of access: <https://www.bernama.com/en/news.php?id=2513175>. — Date of access: 22.02.2026.
5. Better Internet for Kids. (2026, January). Belgian Safer Internet Centre [Electronic resource]. — Mode of access: <https://better-internet-for-kids.europa.eu/en/saferinternetday/belgium>. — Date of access: 25.02.2026.