TRUSTWORTHY AI AND ONLINE CHILD PROTECTION : AN INTERDISCIPLINARY ANALYSIS

Mónika MERCZ[*]

ABSTRACT: *Artificial intelligence is increasingly shaping the digital environments in which children learn, communicate, and socialize. While AI-driven technologies create valuable opportunities for education, creativity, and online safety, they also expose children to risks such as harmful content, digital addiction, online grooming, and deepfake exploitation. This study examines the sociological, legal, and economic dimensions of AI-driven child protection through an interdisciplinary qualitative approach. It analyzes the impact of AI on children's emotional development and online experiences, compares regulatory approaches in the United States and the European Union, and evaluates the role of trustworthy AI in content moderation and online safety.*
The article argues that effective child protection cannot rely solely on restrictive legislation. Instead, responsible AI governance requires a combination of legal safeguards, ethical corporate responsibility, digital literacy, transparency, and meaningful human oversight. Ultimately, the study concludes that AI should be understood as a dual-use technology capable of both harm and protection, whose societal impact depends largely on the ethical and regulatory frameworks governing its development and deployment.

KEYWORDS : *AI ; children's rights ; child protection ; U.S. ; European Union ; DSA ; AI Act*
JEL Classification : *K24, O33, I28*

DOI: **10.62838/cjjc-2024-0082**

1. INTRODUCTION

Technology plays an increasingly significant role in childhood development, presenting both opportunities and risks. On the one hand, studies indicate that excessive screen time may negatively affect children's social skills (UNESCO 2023), contribute to shorter attention spans, and impair emotional regulation (American Psychological Association 2023). Moreover, AI-driven digital content is often specifically designed to maximize engagement, which can foster unhealthy patterns of digital dependency and reduce meaningful real-world social interaction. On the other hand, the responsible

[*] PhD student, Károli Gáspár University of the Reformed Church Doctoral School of Law and Political Sciences

integration of artificial intelligence can substantially enhance education and creativity. Virtual reality (VR) technologies, for example, can transform history lessons into immersive experiences, while AI-assisted storytelling platforms encourage children to explore creativity and develop narrative skills (American University 2023). The central challenge, therefore, lies in achieving an appropriate balance: utilizing AI as a tool for educational enrichment without allowing it to dominate childhood experiences. At the same time, recent tragedies and public controversies surrounding the uninformed development and use of AI systems highlight the importance of exercising caution regarding the timing, context, and methods of AI implementation in educational settings (SINGER 2025).

Recent constitutional and ethical scholarship on artificial intelligence further emphasizes that children are increasingly growing up in environments where AI systems function not merely as tools, but as emotionally interactive entities. Scholars examining the phenomenon of "deadbots" - AI systems capable of replicating deceased individuals through previously collected digital data - warn that emotionally immersive AI technologies may fundamentally reshape children's understanding of grief, attachment, and authentic human relationships. The normalization of emotionally responsive AI companions may weaken emotional resilience and blur the boundaries between reality and artificial interaction, particularly among younger users whose emotional and social development is still ongoing (HOLLANEK 2024).

Concerns surrounding children's emotional reliance on AI are reinforced by recent empirical findings. According to a 2025 report by Internet Matters, 12% of children reported using AI chatbots because they felt they had no one else to talk to, with the figure rising to 23% among vulnerable children. Additionally, 35% of children stated that interacting with chatbots feels similar to talking to a friend, increasing to 50% among vulnerable groups. The report also found that 58% of children considered using chatbots more effective than conducting independent information searches, while 40% expressed no concerns about following chatbot advice at all (Internet Matters 2025).

The broader ethical concerns raised by these developments are closely connected to the concept of human dignity. Human dignity is widely regarded as inviolable and represents an inherent worth shared equally by all persons (DE BAETS 2023). As a foundational principle of international human rights law - particularly within the European legal tradition, which has deep roots in Christian thought - dignity occupies a central role in the protection of individual rights. While debates surrounding life after death remain philosophical and theological in nature, the principle of dignity should not cease upon death. On the contrary, preserving the dignity of deceased individuals remains essential, particularly because they are no longer able to protect their own interests or reputation (MUDERS 2020). Attitudes toward this issue differ considerably between legal traditions. Within the common law tradition, defamatory statements concerning deceased individuals generally do not provide grounds for legal action by relatives or organizations tasked with protecting the deceased's reputation. In contrast, the continental legal tradition has recognized broader protections. Notably, in its landmark *Mephisto* decision (2007, 1 BvR 1783/05), the *Bundesverfassungsgericht* recognized posthumous personality rights grounded in the principle of human dignity.

The development of deadbots capable of replicating an individual's personality requires access to substantial amounts of personal data, including highly sensitive information. Ideally, the use of a deceased person's data for AI training purposes should be based on that individual's prior consent in order to ensure the greatest possible protection of their dignity. However, important debates remain regarding the scope and limits of such consent. Even where consent has been granted, ethical concerns persist, particularly where AI-generated replicas or deepfakes may damage the deceased person's honor, reputation, or dignity, thereby causing emotional harm to surviving loved ones. This is particularly concerning when it comes to vulnerable populations, such as children (SATELL 2024: 15–24).

Against this background, the present study examines how artificial intelligence can be effectively integrated into online child protection efforts while simultaneously addressing concerns relating to privacy, environmental sustainability, and practical implementation challenges. The study hypothesizes that integrating AI within the economic, sociological, and legal frameworks of content moderation can significantly improve the protection of children from harmful online content while complementing - rather than replacing - existing regulatory measures.

This research adopts an interdisciplinary qualitative methodology to examine the role of artificial intelligence in online child protection from economic, sociological, and legal perspectives. The analysis is primarily based on doctrinal legal research, literature review, and policy analysis, supplemented by a comparative examination of regulatory approaches in the United States and the European Union.

The legal analysis focuses on both existing and emerging regulatory frameworks concerning artificial intelligence, online platforms, data protection, and child safety. The United States prioritizes maintaining global leadership in artificial intelligence, particularly in response to growing competition from China (KAYAALP et al. 2025). Simultaneously, the U.S. government's increasing investment in AI services provided by OpenAI (FIELD 2025) signals a broader integration of AI technologies into public administration, potentially transforming child protection mechanisms, including foster care systems and online safety initiatives. In contrast, the European Union approaches these issues with a considerably stronger emphasis on data protection and privacy concerns (DEWITTE 2025). At the same time, several lawsuits in the United States seek to challenge the policies of the current historically pro-AI administration (CAMERON 2025: 5). In this article, particular attention is devoted to legal instruments such as the GDPR, the Digital Services Act (DSA), the Digital Markets Act (DMA), COPPA, FERPA, and relevant state-level initiatives in the United States. The study also examines the differing regulatory philosophies of the EU and the U.S., particularly regarding the balance between innovation, data protection, and online safety.

From a sociological perspective, the article evaluates the societal effects of AI-driven online environments on children, including issues related to digital addiction, algorithmic content recommendation systems, online grooming, misinformation, emotional development, and AI literacy. The analysis relies on secondary sources, including academic scholarship, institutional reports, policy publications, and media coverage addressing the psychological and social consequences of AI-mediated interactions. The economic dimension of the research examines the rapidly expanding market for AI-based content moderation technologies and the broader economic incentives surrounding the

development of trustworthy AI systems. This includes assessing the costs and benefits of automated moderation tools, the role of AI in supporting platform compliance, and the relationship between child safety, consumer trust, and technological competitiveness.

The study adopts a comparative and analytical approach rather than an empirical one. It does not involve primary data collection, interviews, or quantitative statistical analysis. Instead, the research synthesizes existing academic literature, legal materials, institutional reports, and contemporary case studies in order to identify key trends, regulatory challenges, and policy considerations surrounding AI-based child protection mechanisms.

The overarching objective of this methodology is to provide a comprehensive understanding of how AI governance frameworks can simultaneously promote innovation, support economic development, and ensure the protection of children within digital environments.

2. AI-DRIVEN THREATS AND OPPORTUNITIES IN CHILD PROTECTION

The rapid expansion of digital platforms and AI-driven online environments has fundamentally transformed the way children interact, learn, and socialize. While technological innovation offers unprecedented educational and communicative opportunities, it simultaneously exposes children to harmful, exploitative, and manipulative online content. Policymakers worldwide therefore face the difficult challenge of balancing innovation and economic growth with the protection of minors in increasingly AI-mediated digital spaces.

2.1. Sociological implications of AI on children

Artificial intelligence has become deeply integrated into everyday social interactions, particularly through social media platforms, gaming environments, and algorithmically personalized content. However, the complexity and opacity of AI systems - often described through the “black box” phenomenon - contribute to widespread societal uncertainty and mistrust regarding how these technologies operate and influence behavior (BALTA et al. 2019: 7–11). Combined with the influence of social media, AI is reshaping how individuals work, communicate, and perceive reality at an unprecedented pace.

Platforms such as Instagram and TikTok rely heavily on AI-driven recommendation systems to personalize user experiences, curate content feeds, and maximize engagement. While these systems are commercially effective, children are particularly vulnerable to the spread of harmful or misleading content, which often circulates more rapidly than accurate scientific information (VOSOUGHI 2018). Research has shown that Instagram’s recommendation systems are more likely to expose younger users to inappropriate and sexualized content than adults (HOWRITZ 2024). Concerns surrounding these platforms have intensified in recent years. In late 2024, a bipartisan coalition of fourteen U.S. attorneys general filed lawsuits against TikTok, alleging that the platform intentionally contributes to addictive behavior among minors and negatively impacts their mental health (LEVENSON 2024). Although TikTok briefly faced restrictions in the United States, including temporary removal from app stores, the platform ultimately remained available following political negotiations and executive intervention (VERGE 2025). Despite

growing controversy, TikTok continues to be one of the most widely used social media platforms among users aged 12 to 17 (HATIM 2023).

European regulators have likewise increased scrutiny of social media platforms. The European Commission launched formal proceedings against TikTok under the Digital Services Act (DSA), citing concerns regarding insufficient age verification measures and the exposure of minors to harmful content, with potential fines reaching $368 million (BROWN 2024). Nevertheless, the experiences of both the United States and the European Union demonstrate that platform bans alone are unlikely to effectively address the broader societal harms associated with AI-driven social media ecosystems.

Beyond content recommendation systems, AI has significantly transformed children's interpersonal relationships and emotional development. Digital platforms enable children to connect with peers globally through social media, gaming environments, and virtual communities (NEUGNOT – LAURENTY 2024: 11–15). However, these developments also create new risks. AI-generated deepfake imagery, including non-consensual sexually explicit content created through systems such as Grok, has emerged as a serious issue contributing to gender-based violence and creating environments of fear within schools (NAZAKAT – MALIK 2024: 26-42). In 2025 alone, approximately 1.2 million young people reported that their images had been manipulated into explicit deepfake material. As a result, UNICEF publicly declared that AI-generated sexual deepfakes involving minors should be treated with the same urgency as offline abuse (United Nations 2026).

Children may also struggle to distinguish between authentic human relationships and interactions with AI-powered avatars or chatbots. Because these systems are increasingly designed to imitate emotional responsiveness and human communication patterns, children may form emotional attachments to AI entities or rely on them for social support. Such relationships may ultimately contribute to emotional isolation, loneliness, and difficulties in developing genuine interpersonal connections.

AI technologies are additionally being exploited by online predators to facilitate grooming and manipulation. In response, law enforcement agencies have begun utilizing AI-driven tools to combat these threats. Authorities in several jurisdictions have deployed chatbots designed to mimic children in online environments in order to identify predators and gather evidence (VAN DER HOF 2019). Similarly, the Internet Watch Foundation in the United Kingdom has developed preventive chatbot systems intended to intervene before offenders target real children. Private companies have also contributed technological solutions. Microsoft's Project Artemis, for example, uses predictive AI systems to analyze conversations in real time and identify potential grooming behavior (Europol 2024).

Beyond immediate safety concerns, AI systems increasingly shape children's identity formation, self-perception, and worldview. Algorithmically curated content continuously reinforces particular values, beauty standards, and behavioral patterns, often encouraging social comparison and exacerbating mental health challenges among young users. Moreover, AI-driven profiling and behavioral prediction systems raise broader ethical concerns regarding surveillance, manipulation, and autonomy. Children's personal data may be leveraged for targeted advertising, ideological influence, or commercial exploitation at a stage when they are still developing an understanding of privacy, consent, and personal freedoms.

For these reasons, AI literacy must become a fundamental component of modern education. Children growing up in AI-mediated societies will require the ability to distinguish between human-created and AI-generated content, identify trustworthy sources of information, recognize hallucinations produced by AI systems, and use AI tools responsibly in educational and professional contexts. Importantly, current debates increasingly recognize that technical literacy alone is insufficient. Children must also understand how recommendation algorithms influence their behavior, how misinformation spreads online, and how emotionally manipulative systems may shape perceptions of reality. Consequently, scholars and policymakers increasingly argue that education systems should place equal emphasis on preserving uniquely human skills - including empathy, creativity, interpersonal communication, emotional intelligence, and critical thinking.

At the same time, completely excluding children from AI technologies may ultimately disadvantage them in adulthood, particularly as AI becomes deeply integrated into labor markets, education, and everyday life. The challenge, therefore, is not to eliminate children's interaction with AI, but to ensure that such interaction occurs within environments that continue to prioritize meaningful human relationships, peer interaction, and emotional development (VUICIC 2025).

2.2. Legal and regulatory responses

Governments have increasingly sought to address the risks associated with AI-driven online environments through legislative and regulatory measures. In the United States, several states have adopted child-focused social media regulations. New York enacted the Stop Addictive Feeds Exploitation (SAFE) for Kids Act, requiring platforms to limit algorithmically curated feeds and restrict late-night notifications for minors. Other states, including Arkansas, Louisiana, and Florida, introduced legislation establishing minimum age requirements or parental consent obligations for social media usage by minors. However, several of these initiatives have faced constitutional challenges, particularly regarding potential conflicts with minors' First Amendment rights.

In contrast, the European Union has pursued a broader regulatory strategy focused on platform accountability, data protection, and systemic risk mitigation. The Artificial Intelligence Act (AI Act) establishes risk-based obligations for AI systems, including enhanced transparency, safety standards, and human oversight requirements for high-risk applications. Complementing this framework, the Digital Services Act (DSA) imposes obligations on large online platforms to mitigate risks affecting minors, including harmful recommendation systems and addictive platform design. The Digital Markets Act (DMA) further limits the exploitation of users' personal data for targeted advertising purposes, especially concerning children. These measures build upon the protections already provided by the General Data Protection Regulation (GDPR), which grants enhanced safeguards for children's personal data. The influence of EU regulation extends beyond Europe itself. Through the so-called "Brussels Effect," many jurisdictions and companies worldwide have adopted standards modeled after the GDPR and emerging European AI regulations (IAPP 2025).

Despite these legislative efforts, practical enforcement challenges remain significant. Social media platforms often lack effective age and identity verification systems. The scale of the issue is illustrated by the fact that Facebook reportedly deleted 3.5 times more fake

profiles than the entire global population, demonstrating the ease with which users can create anonymous or multiple accounts (MERTON-MCCANN 2024). Consequently, legal regulation alone cannot fully eliminate risks associated with online manipulation, misinformation, scams, and harmful content exposure.

Existing legal frameworks nevertheless provide important protections. In the United States, FERPA safeguards educational records, COPPA requires parental consent for the collection of children's data under thirteen, and the Federal Trade Commission (FTC) has increasingly relied on Section 5 of the FTC Act in cases involving AI systems and data governance practices.

A broader philosophical divide remains visible between the European and American approaches to AI governance. While the European Union generally emphasizes precaution, data protection, and risk mitigation, the United States more frequently frames AI as an economic and strategic opportunity. Although innovation remains essential, AI governance requires a nuanced approach that acknowledges both opportunity and risk. AI systems remain imperfect: they may hallucinate, incorrectly flag content, or produce discriminatory outcomes, which makes meaningful human oversight indispensable.

From a legal perspective, the divergence between the European and American approaches to AI governance reflects deeper constitutional and philosophical differences regarding the relationship between the state, the market, and fundamental rights. The European Union generally adopts a precautionary and rights-based model rooted in the protection of human dignity, privacy, and consumer rights. This approach treats online platforms and AI systems as entities capable of creating systemic societal risks that justify extensive regulatory oversight. Consequently, the EU has focused on ex ante regulation through comprehensive legislative instruments such as the AI Act, GDPR, DSA, and DMA, which impose obligations before harm occurs. In this framework, child protection is treated as a collective societal responsibility requiring strong platform accountability, transparency obligations, and restrictions on exploitative data practices. By contrast, the United States has traditionally favored a more market-oriented and innovation-driven approach grounded in constitutional protections of free expression and economic liberty. American lawmakers and courts are generally more cautious about regulatory interventions that may interfere with First Amendment rights or hinder technological development. As a result, U.S. regulation often emerges through fragmented state-level initiatives, sector-specific legislation, litigation, and post hoc enforcement mechanisms rather than through a unified federal framework. This approach arguably provides greater flexibility for innovation and economic competitiveness, particularly in the rapidly evolving AI sector, but may simultaneously create regulatory inconsistencies and weaker preventative safeguards for children (BRADFORD 2024).

Determining which model is ultimately more effective remains complex. The European model offers stronger preventative protections and clearer legal obligations for technology companies, particularly regarding harmful recommendation systems, data exploitation, and platform accountability. The "Brussels Effect" demonstrates the practical influence of this framework, as multinational companies frequently adopt EU standards globally in order to simplify compliance operations. However, critics argue that excessive regulation may slow innovation, increase compliance costs, and discourage investment in emerging technologies. Smaller companies and startups may face disproportionate burdens when attempting to comply with highly technical regulatory requirements, potentially

consolidating the dominance of large technology corporations already capable of absorbing such costs.

The American approach, on the other hand, may better support technological innovation, economic growth, and rapid AI development. The comparatively flexible regulatory environment encourages experimentation and private sector investment, allowing companies to develop new AI applications at a faster pace. Nevertheless, the absence of a comprehensive federal AI framework may leave significant gaps in child protection, particularly regarding algorithmic transparency, platform accountability, and age verification standards. Furthermore, reliance on litigation and reactive enforcement may fail to adequately address harms before they occur.

In practice, neither approach appears sufficient on its own. The borderless nature of digital platforms and AI systems creates significant enforcement difficulties for purely national or regional regulatory models. Harmful content, online grooming, deepfake exploitation, and AI-driven manipulation frequently transcend jurisdictional boundaries, while technology companies often operate simultaneously across dozens of legal systems. Consequently, the development of some form of international coordination or common governance framework may become increasingly necessary.

One possible solution could involve the establishment of an international AI governance body operating similarly to existing international organizations in fields such as aviation, trade, or public health. Such an institution could develop minimum global standards concerning child safety, algorithmic transparency, age verification practices, AI ethics, and platform accountability. Alternatively, states could negotiate a multilateral international agreement specifically focused on trustworthy AI and online child protection. A harmonized framework could improve cross-border cooperation, facilitate information-sharing between regulators and law enforcement agencies, and reduce the ability of harmful actors to exploit regulatory loopholes between jurisdictions.

However, substantial obstacles complicate the creation of a unified international framework. States differ significantly in their constitutional traditions, political priorities, economic interests, and cultural understandings of privacy, free speech, and government oversight. Geopolitical competition surrounding AI development further reduces the likelihood of comprehensive global consensus. Many governments may also resist international oversight mechanisms that could limit national sovereignty or domestic economic competitiveness (YILMAZ 2026).

Practical enforcement challenges would likewise remain considerable. Even if common standards were adopted, ensuring compliance by multinational technology companies operating across multiple jurisdictions would require robust monitoring and enforcement mechanisms. Developing countries may additionally lack the institutional capacity or technological resources necessary to implement advanced AI oversight systems effectively. Furthermore, rapid technological development risks rendering international agreements outdated before meaningful implementation can occur.

For these reasons, the most realistic short-term solution may not be the creation of a fully centralized global regulatory authority, but rather the gradual development of interoperable international standards and cooperative governance mechanisms. Increased collaboration between states, regional organizations, technology companies, academic institutions, and civil society actors could help establish shared principles for trustworthy AI while preserving sufficient flexibility for innovation and national legal differences. In

the long term, effective child protection in AI-mediated digital environments will likely depend on achieving a careful balance between innovation, human rights, democratic accountability, and international cooperation.

2.3. Economic dimensions of AI-driven child protection

Artificial intelligence also carries substantial economic implications in the context of online child protection. AI-powered content moderation has rapidly developed into a major global industry, with the automated moderation market projected to grow from $1.03 billion in 2024 to $1.24 billion in 2025, representing a compound annual growth rate of 20.5% (The Business Research Company 2025).

Beyond moderation itself, AI technologies generate new economic opportunities across digital platforms. Social media, content creation, and online entertainment increasingly rely on AI-assisted tools, enabling the emergence of new professions and revenue streams for creators, influencers, educators, and digital entrepreneurs. AI may also support broader child welfare systems. Recent proposals concerning foster care reform suggest that predictive AI systems could assist social workers in identifying patterns associated with neglect, instability, or heightened vulnerability, thereby improving placement stability and resource allocation. Nevertheless, scholars caution that such systems must remain transparent, explainable, and subject to human oversight to avoid discriminatory outcomes.

AI additionally offers economic benefits by supporting regulatory compliance and reducing operational costs. Automated systems can assist companies in complying with evolving child protection and privacy laws while significantly reducing the psychological burden placed on human moderators exposed to harmful material (Cambridge Consultants 2019). Furthermore, AI-driven educational technologies and adaptive learning systems may contribute to long-term social mobility by improving educational access for disadvantaged children and reducing structural inequalities.

However, these opportunities also highlight the importance of developing explainable and trustworthy AI systems. The complexity of AI technologies often makes it difficult even for developers themselves to fully understand how systems generate certain outputs. This lack of transparency complicates both regulatory enforcement and public trust. AI technologies can simultaneously be used to combat child exploitation and facilitate harmful activities such as grooming, misinformation, or the creation of CSAM material (FARID 2024). Consequently, AI must be understood as a fundamentally dual-use technology capable of serving both beneficial and harmful purposes.

Ultimately, meaningful child protection in digital environments cannot rely exclusively on legislative intervention. Effective governance also requires internal accountability within technology companies themselves, including robust ethical guidelines, monitoring procedures, and human-centered development practices. Trustworthy AI systems are not only socially desirable but economically advantageous, as consumers are increasingly unlikely to adopt technologies perceived as unsafe or harmful. Long-term innovation therefore depends not merely on technological advancement, but on the ability to develop AI systems that remain reliable, transparent, and aligned with human well-being.

3. HOW AI CAN BE USED TO PROTECT CHILDREN ONLINE

Although artificial intelligence presents significant risks for children in digital environments, it also possesses substantial potential as a tool for online child protection when developed and implemented responsibly. The effectiveness of AI in this context depends not only on technological sophistication, but also on the establishment of trustworthy governance frameworks, ethical safeguards, and meaningful human oversight.

A central requirement for responsible AI development is the creation of internal accountability mechanisms within technology companies themselves. Developers should implement AI ethics committees capable of reviewing products throughout every stage of development and deployment, while regularly conducting risk assessments to identify potential harms, including bias, misinformation, privacy violations, or discriminatory outcomes. Transparency is equally essential. AI systems affecting children should remain explainable to users, regulators, and external oversight bodies, as transparency both strengthens public trust and promotes long-term economic sustainability. Audit trails documenting AI decision-making processes may further assist regulators and independent experts in verifying legal compliance and ethical standards.

At the same time, governments can encourage the development of trustworthy AI through incentive-based regulatory approaches rather than relying exclusively on restrictive legislation. Financial incentives, such as tax benefits, grants, or public-private partnerships, may encourage companies to prioritize ethical AI development and invest in safer technologies. Such measures could also support the emergence of self-regulatory initiatives, including voluntary safety commitments, trustworthy AI certifications, or industry-wide best practices comparable to existing cybersecurity or environmental standards.

Combining internal corporate accountability with external pressure from governments, civil society organizations, consumers, and independent watchdog institutions may therefore create more effective AI safety standards while preserving technological innovation. This balance is particularly important in the context of children, who remain among the most vulnerable users of digital technologies. If AI systems are perceived as safe, transparent, and ethically reliable, their full protective potential can be more effectively utilized.

One of the most significant advantages of AI lies in its capacity for proactive risk detection. Traditional moderation systems frequently operate reactively, removing harmful content only after damage has already occurred. AI-driven systems, however, can identify behavioral patterns associated with grooming, cyberbullying, harassment, exploitation, or self-harm before these situations escalate. Machine learning tools may analyze suspicious communication patterns, repeated contact attempts, abnormal engagement behaviors, or concerning keyword combinations in real time, enabling moderators or law enforcement agencies to intervene more rapidly and effectively.

Similarly, AI technologies can significantly strengthen efforts to combat child sexual abuse material (CSAM) and online predatory behavior. Advanced AI systems are capable of identifying illegal content, detecting grooming patterns, and assisting investigators in dismantling exploitation networks with far greater efficiency than traditional moderation methods alone. Cross-platform cooperation could further enhance these efforts. AI systems capable of securely sharing hashes of illegal content, behavioral indicators, or threat

patterns across platforms may improve coordinated responses against online predators operating simultaneously across social media services, gaming environments, and encrypted communication channels. Nevertheless, such measures must remain proportionate and subject to strict legal oversight in order to prevent excessive surveillance and protect fundamental rights (HAKAMI – TAZEL 2024).

AI also offers opportunities to create safer and more development-oriented online experiences for children. Recommendation algorithms on social media and video-sharing platforms could be designed to prioritize educational, age-appropriate, and socially beneficial content rather than maximizing engagement through sensationalist or harmful material. In educational contexts, AI-driven personalization systems may adapt learning environments to children's individual needs and pace of development while simultaneously identifying signs of emotional distress, cyberbullying, social withdrawal, or harmful online exposure. AI-assisted educational software could therefore support earlier intervention by teachers, parents, or guardians when children appear to be at risk (LEE – KWON 2024).

Another important area of innovation concerns age verification and online identity management. Existing child protection frameworks often struggle to verify users' ages effectively without creating significant privacy concerns. AI-based age estimation technologies may provide less invasive alternatives to traditional identification systems if implemented transparently, proportionately, and with appropriate safeguards. Such technologies could assist platforms in restricting children's access to age-inappropriate services or harmful content while minimizing unnecessary data collection. At the same time, human oversight remains indispensable because AI systems continue to produce false positives, biased outcomes, or inaccurate assessments resulting from technical limitations or flawed training data. Overreliance on automated moderation may lead to unjustified censorship, privacy violations, or the marginalization of vulnerable groups. Consequently, many experts advocate for hybrid moderation systems that combine the efficiency of AI with meaningful human accountability and review mechanisms (KÖHLER-DAUNER 2025).

The long-term success of AI-driven child protection also depends heavily on digital literacy and public education. Children, parents, and educators must be equipped not only with technical knowledge, but also with the critical thinking skills necessary to navigate increasingly AI-mediated digital environments safely. Understanding how recommendation algorithms function, how deepfakes are created, or how manipulative online behaviors emerge has become essential for responsible internet use. Accordingly, AI literacy should become a central pillar of educational curricula and public awareness strategies.

Beyond technical education, children must continue to develop uniquely human capabilities that AI cannot replace, including empathy, interpersonal communication, creativity, emotional intelligence, and social resilience. The objective should therefore not be to isolate children from artificial intelligence entirely, but rather to ensure that AI complements - rather than replaces - healthy human interaction and emotional development.

International cooperation will likewise remain essential in addressing AI-related risks affecting children. Online platforms and AI systems operate across national borders, while legal and regulatory systems remain fragmented. Without greater international

coordination, harmful actors may exploit inconsistencies between jurisdictions and evade enforcement mechanisms. Effective child protection in digital environments therefore requires cooperation among states, international organizations, technology companies, researchers, and civil society actors in order to establish common standards for trustworthy AI, platform accountability, and online child safety.

Ultimately, artificial intelligence should not be viewed solely as a source of risk, but as a dual-use technology capable of both harm and protection. Its societal impact will depend not on the technology itself, but on the ethical, legal, economic, and institutional frameworks within which it is developed and deployed. Properly governed AI systems have the potential to become some of the most effective tools available for protecting children in increasingly complex digital environments.

4. CONCLUSION

Artificial intelligence is rapidly transforming the digital environments in which children learn, communicate, socialize, and develop their identities. As this study has demonstrated, AI simultaneously creates unprecedented opportunities and serious risks for child protection online. The growing influence of algorithmically curated content, emotionally responsive AI systems, deepfakes, online grooming, and AI-assisted manipulation raises significant sociological, ethical, legal, and economic concerns. At the same time, AI technologies also possess considerable potential to improve online safety, detect harmful behavior proactively, support educational development, and strengthen child protection mechanisms across digital platforms.

The central challenge identified throughout this research is therefore not whether AI should be integrated into child protection efforts, but rather how such integration can occur responsibly. The findings suggest that purely prohibitionist or overly restrictive approaches are unlikely to provide effective long-term solutions. Children are growing up in increasingly AI-mediated societies and excluding them entirely from these technologies would likely create new educational and social disadvantages. Instead, the objective should be to develop governance frameworks that ensure AI remains human-centered, transparent, accountable, and supportive of children's well-being.

The comparative analysis of the United States and the European Union further illustrates the tension between innovation and regulation that currently defines global AI governance. While the European Union emphasizes precaution, data protection, and platform accountability through instruments such as the AI Act, the DSA, DMA, and GDPR, the United States more frequently frames AI as a strategic and economic opportunity. Both approaches offer valuable insights, yet neither alone provides a complete solution. Excessive deregulation may expose children to serious harm, while overregulation risks hindering innovation, technological competitiveness, and the development of beneficial AI applications. A balanced regulatory model is therefore necessary - one capable of protecting fundamental rights without preventing technological progress.

This study also highlights that effective child protection cannot rely solely on legislative intervention. Technology companies themselves must assume greater responsibility for the development and deployment of AI systems. Internal ethical oversight, transparency measures, explainable AI systems, regular risk assessments, and

meaningful human supervision are essential components of trustworthy AI governance. In particular, human oversight remains indispensable because AI systems continue to produce biased outcomes, false positives, and inaccurate assessments that may disproportionately affect vulnerable individuals, including children.

At the societal level, digital literacy emerges as one of the most important long-term protective mechanisms. Children, parents, educators, and policymakers must develop a deeper understanding of how AI systems function, how recommendation algorithms influence behavior, how misinformation spreads, and how manipulative digital practices operate. However, technical competence alone is insufficient. As AI increasingly imitates emotional interaction and social behavior, preserving uniquely human qualities - including empathy, emotional intelligence, critical thinking, creativity, and authentic interpersonal relationships - become equally important.

Economically, the study demonstrates that trustworthy AI and child safety should not be viewed as obstacles to innovation, but rather as preconditions for sustainable technological development. Public trust increasingly influences the success of digital platforms and AI systems. Technologies perceived as unsafe, exploitative, or psychologically harmful are unlikely to maintain long-term societal legitimacy. Consequently, investing in ethical, transparent, and child-centered AI systems may simultaneously promote innovation, consumer confidence, and economic competitiveness.

Finally, because AI systems and digital platforms operate globally, meaningful child protection will ultimately require stronger international cooperation. Regulatory fragmentation creates opportunities for harmful actors to exploit jurisdictional gaps, while inconsistent standards undermine accountability efforts. Future governance frameworks must therefore involve coordinated cooperation among states, international organizations, technology companies, researchers, educators, and civil society actors in order to establish common principles for trustworthy AI and online child safety.

Artificial intelligence should ultimately be understood as a dual-use technology capable of both harm and protection. Its long-term societal impact will depend less on the technology itself than on the ethical, legal, and institutional choices that shape its development and implementation. If governed responsibly, AI has the potential not only to mitigate online harms affecting children, but also to contribute to safer, more inclusive, and more supportive digital environments for future generations.

REFERENCES:

AMERICAN PSYCHOLOGICAL ASSOCIATION (2023): Are Our Attention Spans Shrinking? With Gloria Mark, PhD. Speaking of Psychology, Podcast. https://www.apa.org/news/podcasts/speaking-of-psychology/attention-spans

AMERICAN UNIVERSITY (2023): Benefits of Virtual Reality in Education. American University School of Education Blog, 2023. https://soeonline.american.edu/blog/benefits-of-virtual-reality-in-education/

BALTA, Dian – KUHN, Peter – SELLAMI, Mahdi – KALOGEROPOULOS, Anastasios – KRCMAR, Helmut (2019): Blackbox AI: What is in the Box? https://www.researchgate.net/publication/341042329_Blackbox_AI_What_is_in_the_Box

DE BAETS, Antoon. (2023): The Posthumous Dignity of Dead Persons, Anthropology of Violent Death: Theoretical Foundations for Forensic Humanitarian Action, edited by Roberto C. Parra and Douglas H. Ubelaker, Wiley, 15-37. https://doi.org/10.1002/9781119806394.ch2

BRADFORD, Anu (2024): The False Choice Between Digital Regulation and Innovation. Northwestern University Law Review, Vol. 119, No. 2, 377-453. Columbia Law Scholarship Archive PDF

BROWN, William (2024): European Regulators Hand TikTok $368 Million Fine for Failing to Protect Kids' Privacy. PBS NewsHour, https://www.pbs.org/newshour/world/european-regulators-hand-tiktok-368-million-fine-for-failing-to-protect-kids-privacy

CAMBRIDGE CONSULTANTS (2019): Use of AI in Online Content Moderation. Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/other/cambridge-consultants-ai-content-moderation.pdf?v=324081

CAMERON, Chris (2025): Unions Sue to Block Musk's Access to Treasury Payment Data. The New York Times, https://www.nytimes.com/2025/02/03/us/politics/elon-musk-treasury-payment-data.html

DEWITTE, Pierre (2025): AI Meets the GDPR: Navigating the Impact of Data Protection on AI Systems. In: Smuha, N. A. (szerk.): The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence. Cambridge, Cambridge University Press, 133-157.

EUROPEAN COMMISSION (2024): Artificial Intelligence: Commission Proposes New Rules to Make AI Trustworthy and Human-Centered. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4161

EUROPOL (2024): Policing in an AI-Driven World. Police Chief Online, April 24, 2024.

FARID, Hany (2024): Detecting CSA Online - A Technical and Historical Primer for Policy Makers. The Canadian Centre for Child Protection. https://protectchildren.ca/en/resources-research/hany-farid-photodna/

FIELD, Hayden (2025): OpenAI Launches ChatGPT Gov for U.S. Government Agencies. CNBC, https://www.cnbc.com/2025/01/28/openai-launches-chatgpt-gov-for-us-government-agencies.html

HAKAMI, Ammar – TAZEL, Ravi (2024): The Ethics of AI in Content Moderation: Balancing Privacy, Free Speech, and Algorithmic Control, DOI:10.13140/RG.2.2.19529.97121

HATIM, Abdul (2023): TikTok and Its Impact on Children's Health. Journal of Health, https://www.journalofhealth.co.nz/wp-content/uploads/2023/04/DHH_TikTok_Hatim.pdf

HOLLANEK, T. - NOWACZYK-BASIŃSKA, K. (2024): Griefbots, Deadbots, Postmortem Avatars: on Responsible Applications of Generative AI in the Digital Afterlife Industry. Philos. Technol. 37, 63 https://doi.org/10.1007/s13347-024-00744-w

HORWITZ, Jeff (2024): Instagram Recommends Sexual Videos to Accounts for 13-Year-Olds, Tests Show. The Wall Street Journal, https://www.wsj.com/tech/instagram-recommends-sexual-videos-to-accounts-for-13-year-olds-tests-show-b6123c65

IAPP (2025): Global AI Law and Policy Tracker. https://iapp.org/resources/article/global-ai-legislation-tracker/

INTERNET MATTERS (2025): Me, myself & AI: Understanding and safeguarding children's use of AI chatbots. https://www.internetmatters.org/wp-content/uploads/2025/07/Me-Myself-AI-Report.pdf

KAYAALP, M. E. – PRILL, R. – SEZGIN, E. A. – CONG, T. – KRÓLIKOWSKA, A. – HIRSCHMANN, M. T. (2025): DeepSeek versus ChatGPT: Multimodal Artificial Intelligence Revolutionizing Scientific Discovery. From Language Editing to Autonomous Content Generation – Redefining Innovation in Research and Practice. Knee Surgery, Sports Traumatology, Arthroscopy. DOI: 10.1002/ksa.12628.

KÖHLER-DAUNER F. et al. (2025): Digital child protection in social networks: age verification and age-tiered regulation in Europe. Child Adolesc Psychiatry Ment Health. 2025 Dec 23;19(1):143. doi: 10.1186/s13034-025-01016-x. PMID: 41437378; PMCID: PMC12729629.

LEE, Sang Joon – KWON, Kyungbin (2024): A systematic review of AI education in K-12 classrooms from 2018 to 2023: Topics, strategies, and learning outcomes, Computers and Education: Artificial Intelligence Volume 6, https://doi.org/10.1016/j.caeai.2024.100211

LEVENSON, Eric (2024): 14 States Sue TikTok Over Children's Mental Health. CNN, https://www.cnn.com/2024/10/08/tech/tiktok-sued-14-states-childrens-mental-health/index.html

MERTON-MCCANN, Alex (2024): How To Spot A Fake Facebook Account. McAfee. https://www.mcafee.com/blogs/family-safety/spot-fake-facebook-account/

MUDERS, Sebastian (2020): Human Dignity: Final, Inherent, Absolute? Rivista di Estetica, vol. 75, 84-103. OpenEdition Journals, https://doi.org/10.4000/estetica.7319

NAZAKAT, Tooba – MALIK, Faiza Eiman (2024): Empowering Justice through AI: Addressing Technology-Facilitated Gender-Based Violence with Advanced Solutions, Journal of Law & Social Studies (JLSS) Volume 7, Issue 1, 26-42.

NCMEC (2024): Generative AI. https://www.missingkids.org/theissues/generative-ai

NEUGNOT, Mathilde – LAURENTY, Olga Muss (2024): The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts, arXiv.org, DOI:10.48550/arXiv.2405.19275

SATELL, Greg (2024): AI Anxiety Is on the Rise – Here's How to Manage It. Scientific American, https://www.scientificamerican.com/article/ai-anxiety-is-on-the-rise-heres-how-to-manage-it/

SHIBUYA, Fumiko et al (2023): Teachers' conflicts in implementing comprehensive sexuality education: a qualitative systematic review and meta-synthesis, Tropical Medicine and Health (2023) 51:18, https://doi.org/10.1186/s41182-023-00508-w

SINGER, Natasha (2024): School Employee Arrested After Racist Deepfake Recording of Principal Spreads. The New York Times, https://www.nytimes.com/2024/04/25/technology/deepfake-recording-principal-arrest.html

THE BUSINESS RESEARCH COMPANY (2025): Automated Content Moderation Global Market Report, https://www.thebusinessresearchcompany.com/report/automated-content-moderation-global-market-report

THE DRAGHI REPORT: A Competitiveness Strategy for Europe (2024). European Commission. Forrás:https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en#paragraph_47059

UNESCO (2023): Technology in Education: A Tool on Whose Terms? Global Education Monitoring Report. Paris, UNESCO. https://gem-report-2023.unesco.org/technology-in-education/

UNICEF (2026): 'Deepfake abuse is abuse', Statement by UNICEF on AI-generated sexualised images of children, https://www.unicef.org/press-releases/deepfake-abuse-is-abuse

VAN DER HOF, Simon – GEORGIEVA, Ilina – SCHERMER, Bart – KOOPS, Bert-Jaap (2019): Sweetie 2.0 – Using Artificial Intelligence to Fight Webcam Child Sex Tourism. The Hague, T.M.C. Asser Press. DOI: 10.1007/978-94-6265-288-0.

VERGE (2025): TikTok Ban: All the News on the App's Shutdown and Return in the US. The Verge, https://www.theverge.com/23651507/tiktok-ban-us-news

VOSOUGHI, Soroush – ROY, Deb – ARAL, Sinan (2018): The Spread of True and False News Online. Science, 359(6380), 1146-1151. DOI: 10.1126/science.aap9559.

YILMAZ, Cüneyt (2026): The Geopolitics of Artificial Intelligence: Power, Regulation, and Global Governance, FIU Journal of Social Sciences and Humanities 2(3):39-70 DOI:10.29329/ufusobed.2026.1410.2