ISSN 2281-5147

A STORY *OF* AND *FOR* CHILDREN: THE LIFECYCLE LOOP OF CHILD RIGHTS-BASED AI

Sara Tibidò, Nadia Spatari, Sara Lilli e Maria Vittoria Zucca[*]

Abstract

This paper traces the lifecycle loop of child rights-based AI - from the initial phase of design through development and deployment - while mapping the ethical and regulatory landscape surrounding AI technologies designed for, accessed by, or impacting children. Building on established frameworks, the study advocates for the implementation of regulatory sandboxes and risk assessment measures to protect children's rights and interests against threats and emerging cyber risks. This research argues for the essential integration of a child rights-based approach at every stage and phase of an AI system's lifecycle, asserting that this leads to the development and deployment of more secure, child-centered systems.

Table of Contents

[*] Sara Tibidò, Ph.D. Student in *Cybersecurity*, University of Bari "Aldo Moro" and IMT School for Advanced Studies Lucca, ORCID: 0009-0004-0646-0558.
Nadia Spatari, Ph.D. Student in *Cybersecurity*, National Inter University Consortium for Informatics (CINI) and IMT School for Advanced Studies Lucca. ORCID: 0009-0009-4935-7135.
Sara Lilli, Ph.D. Student in *Cybersecurity*, Sant'Anna School of Advanced Studies in Pisa and IMT School for Advanced Studies Lucca. ORCID: 0009-0000-2796-3067.
Maria Vittoria Zucca, Ph.D. Student in *Cybersecurity*, Sant'Anna School of Advanced Studies in Pisa and IMT School for Advanced Studies Lucca, ORCID: 0009-0004-0049-9611.
Double blind peer-reviewed contribution.

ISSN 2281-5147

Keywords

Artificial Intelligence - Children's rights - Children-centred AI - Digital Safety - Child Impact Assessment - Regulatory Sandbox

1. Starting the lifecycle loop of child rights-based AI

Once upon a time, there was a doll named Cayla[1], designed to be a friendly playmate for children. But behind her smiling face and sweet voice, she hides the potential of *listening* - and *sharing*. What was meant to be an AI embedded toy became a warning story of how innovation can overlook safety, privacy, and the fundamental rights of

[1]See the articles from BBC, 'German parents told to destroy Cayla dolls over hacking fears', (*BBC News* , 17 February 2017) https://www.bbc.com/news/world-europe-39002142 accessed 06 July 2025; and World Economic Forum (WEF), 'Generation AI: What happens when your child's friend is an AI toy that talks back?' (*World Economic Forum*, 22 May 2018) https://www.weforum.org/stories/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/ accessed 06 July 2025; other relevant cases should also be considered, such as the chatbots Wysa and Woebot, for which reference can be made to the following article: Geoff White, 'Child advice chatbots fail to spot sexual abuse' *BBC* (London, 11 December 2018),https://www.bbc.com/news/technology-46507900 accessed 06 July 2025; and Karen Brown, 'Something Bothering You? Tell It to Woebot. When your therapist is a bot, you can reach it at 2 a.m. But will it really understand your problems?', *The New York Times* (New York, 01 June 2021), https://www.nytimes.com/2021/06/01/health/artificial-intelligence-therapy-woebot.html accessed 06 July 2025.

ISSN 2281-5147

the youngest users. Cayla's conversations have indeed been found vulnerable to hacking, allowing strangers to listen and communicate directly to children.

While significant steps have been undertaken to improve safety and protection from similar situations (for instance, the adoption of *privacy*- and *security*-by-design approaches) different international organizations and associations, like UNICEF'[2] and the Institute of Electrical and Electronics Engineers - IEEE[3], and international non-governmental organizations (NGOs), such as the 5Rights foundation[4]'[5], are calling for stronger, child-specific measures. These measures underscore the importance of integrating children's rights from the outset of the innovation process, ensuring their safety, protection, and participation.

While children should not be excluded from the digital world, as also stated by the UN General Comment No.25[6], they should be protected by the risks (both *old* and *new*) they may face when using digital products or services. To move towards a welcoming, as well as more safe and secure digital ecosystem for children, it is crucial to integrate children's rights - along with safety and security measures - from the very beginning of the innovation process. This approach is particularly important when developing AI systems[7]. Indeed, the interaction between children and AI systems is

[2] UNICEF - V. Dignum, M.Penagos, K.Pigmans and S.Vosloo, 'Policy Guidance on AI for Children (Version 2.0)' (November 2021). https://www.unicef.org/innocenti/reports/policy-guidance-ai-children accessed 12 May 2025.

[3] IEEE Std 2089-2021, 'IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children' (vol., no., pp.1-54, 30 Nov. 2021). DOI: https://doi.org/10.1109/IEEESTD.2021.9627644.

[4] Digital Futures Commission and and 5Rights Foundation, 'Child Rights by Design' (11 March 2023). https://5rightsfoundation.com/resource/child-rights-by-design/ accessed 04 July .2025.

[5] 5Rights Foundation, 'Children & AI Design Code' (March 2025). https://5rightsfoundation.com/children-and-ai-code-of-conduct/ accessed 04 July 2025.

[6] UN Committee on the Rights of the Child, 'General comment No. 25 (2021) on children's rights in relation to the digital environment' (02 March 2021) CRC/C/GC/25.

[7] Acknowledging that there is no internationally shared definition, for the purpose of this paper, we intend an "AI system" as defined by Article 3(1) of the EU AI Act and as further explained by the European Commission (February 2025) in its guidelines on AI systems definition (available online

complex and not limited only to those systems designed *for* children to be the main end users (*e.g.*: AI-enabled toys or systems used in the EdTech field), but also to those systems not meant for them but with which they *interact* in everyday lives contexts (*e.g.:* smart home assistant or recommender systems in social media and streaming platforms), and systems that can directly or indirectly *impact* them (*e.g.:* AI systems used to support decision process of social workers dealing with case of child maltreatment[8])[9]. Attention should also be paid to factors that can influence AI's impact on children, such as socioeconomics, geographic and cultural context and norms, as well as other elements like children's developmental stages related to their physical, cognitive, emotional and psychological capacities.[10]

Accordingly, this story begins far back in the innovation process, from the discovery phase through the design and development phases, and it is grounded in the children's rights as defined by the UN Convention on the Rights of the Child (UNCRC). Indeed, since its adoption by the UN General Assembly in 1989 and its entry into force in September 1990, the UNCRC has become the world's most widely ratified human rights treaty[11]. With its ratification, States are legally bound to respect, protect, and fulfill the rights as outlined in the Convention[12]. Therefore, although the digital environment and new technologies may pose new challenges, the Convention (guided in its implementation in relation to the digital environment by the UN General Comment No.25) can still be considered an authoritative source on children's rights.

at: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application - accessed 14 March 2025).

[8] See, for example, A. Kawakami and V. Sivaraman, and L. Stapleton, and H.F. Cheng, and A. Perer, and Z.S. Wu, and H. Zhu, and K.Holstein, "'Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts' (ACM Designing Interactive Systems Conference, online, 13-17 June 2022).

[9] UNICEF - V. Dignum, M.Penagos, K.Pigmans and S.Vosloo (November 2021).

[10] *Ibidem.*

[11] UNICEF, 'How the Convention on the Rights of the Child works' https://www.unicef.org/child-rights-convention/how-convention-works accessed 12 May 2025.

[12] *Ibidem.*

ISSN 2281-5147

In 2011, the UN Human Rights Council endorsed the "Guided Principles on Business and Human Rights" (UNGPs)[13], implementing the 2008's UN "Protect, Respect and Remedy" framework for business and human rights and recognizing business's responsibility to respect also those rights as enshrined in the UNCRC[14]. The UNGPs '*are applied to the digital context through the UN Human Rights B-Tech Project*'[15] (*e.g.*: the briefing, conducted together with UNICEF and published in 2024, on "Taking a Child Rights-Based Approach to Implementing the UNGPs in the Digital Environment" unpacks core headlines on the implementation of UN principles with a child rights perspective[16]). A year later, in 2012, UNICEF, the UN Global Compact and Save the Children developed the "Children's Rights and Business Principles", a range of actions companies can undertake in different contexts to respect and support children's rights[17]. Although those Principles do not constitute a legally binding document, they are instruments of soft law that have been '*incorporated or referenced in legislation, industry codes of conduct, and market-entry requirements in various sectors of the economy, including the digital sector*'[18].

Unlike such voluntary approaches, the European Union has imposed some legal obligations to online intermediaries and platforms. In particular, first in 2018 with the "Audiovisual Media Services Directive" (AVMSD), coordinating national legislations and setting out responsibilities for media service providers (*e.g.:* protection of users, children in particular, from certain kinds of content or programs and establishment

[13] UN, 'Guided Principles on Business and Human Rights' (01 January 2012) 978-92-1-154201-1.

[14] *Ibidem*.

[15] OECD, 'Shaping a Rights-Oriented Digital Transformation' (28 June 2024), No. 368, OECD Digital Economy Papers (citing). https://www.oecd.org/en/publications/shaping-a-rights-oriented-digital-transformation_86ee84e2-en.html accessed 12 May 2025.

[16] UNICEF and UN Human Rights, 'Taking a Child Rights-Based Approach to Implementing the UNGPs in the Digital Environment' (November 2024) https://www.unicef.org/childrightsandbusiness/reports/b-tech-contribution accessed 05 July 2025.

[17] UNICEF, the UN Global Compact and Save the Children, 'Children's Rights and Business Principles' (2012) https://www.unicef.org/documents/childrens-rights-and-business-principles accessed 12 May 2025.

[18] OECD (28 June 2024), citing.

of age verification systems in video-sharing platforms)[19], and then in 2022, with the "Digital Services Act" (DSA). The DSA, which refers to international standards (including the UNGPs)[20] and aims at regulating online platforms and intermediaries (to be specific: very large online platforms and search engine, online platforms, host services and intermediary services), contains some child-specific provisions (*e.g.*: Article 14 on comprehensible child-friendly explanations of conditions and terms of use, Article 28 on appropriate and proportionate measures to protect children's safety, security and privacy, and Articles 34 and 35 on mandatory annual fundamental rights risks' assessments and mitigation measures).[21]

Designing with children's rights in mind is no simple task, but retrofitting a product to comply with these rights after development can be both difficult and costly.[22] Accordingly, this paper proposes a children's rights-based approach to the entire AI system lifecycle, emphasizing the integration of children's rights, needs, and perspectives - alongside safety, security, and stakeholders inputs - at every phase. The aim is to ensure that systems are well-designed from the outset to be compliant with children's rights standards and obligations, thereby reducing the need for substantial post-deployment corrections. Therefore, the following sections will describe a story of innovation that begin from (i) the legal, policy and technical frameworks shaping the *design* and *development* phases of an AI system for/impacting/accessed by children, passing through (ii) the phases *of testing* and *validation* with the use of regulatory sandboxes, to (iii) the phases of *deployment* and *post-deployment*[23].

[19] *Ibidem*.

[20] *Ibidem*.

[21] OECD, 'Towards digital safety by design for children' (19 June 2024), No. 363, OECD Digital Economy Papers. https://www.oecd.org/en/publications/towards-digital-safety-by-design-for-children_c167b650-en.html accessed 05 July 2025.

[22] Digital Futures Commission and 5Rights (11 March 2023).

[23] For the division of the phases constituting the AI system lifecycle, we recall the work of D. De Silva and D. Alahakoon, 'An Artificial Intelligence Life Cycle: From Conception to Production' (2022) 3(6) Patterns, https://doi.org/10.1016/j.patter.2022.100489 accessed 12 May 2025. Indeed, the Authors consider an AI system's life cycle made of three main phases: "*design*", "*develop*" and "*deploy*", each of them made of different "*stages*". While the Authors do not consider a separate phase for testing and validation, in the "*deploy*" phase it is considered a "*post-deploy*" stage (stage no.16).

ISSN 2281-5147

2. ***Design*** **and** ***develop*****: towards clear and practical child rights-based guidelines for practitioners**

State have the duty, under international human rights law, to protect people in their jurisdiction or/and their territory from human rights abuses, and corporate responsibility to respect human rights exists '*regardless of their size, sector, location, ownership and structure*'[24]. Therefore, States and businesses have different but complementary responsibilities[25]. Accordingly, since the exercise and protection of human rights can be affected by how '*digital technologies are designed, developed and deployed*', it is important to embed human rights in all the phases of an innovation process[26]. However, providing all stakeholders with clear, technically applicable and cross-cutting guidelines is challenging.

Before rights-specific considerations, ethical AI-related challenges have been a central topic of discussion among policy makers, professionals and academics. Indeed, ethical principles and guidelines have been found difficult to be integrated into the engineering process that power AI development: there is a critical gap between these principles, available guidelines and the realities of the engineering practice[27]. Moreover, the accountability gap, in terms of clarity about who should be ought accountable '*for the outcomes of technology use, to whom, and how*'[28], presents a major challenge for engineers (*e.g.:* hierarchies of power in the workplace that may limit their

[24] UN, 'Guided Principles on Business and Human Rights' (01 January 2012), citing.

[25] *Ibidem.*

[26] OECD (28 June 2024), citing.

[27] IEEE SA, 'Report: Addressing Ethical Dilemmas in AI: Listening to Engineers Report' (2021) https://standards.ieee.org/initiatives/autonomous-intelligence-systems/ethical-dilemmas-ai-report/ accessed 05 July 2025.

[28] *Ibidem*, citing.

technical and organizational choices; absence of independent infrastructures to turn to in case of ethical concerns or to report cases of non-compliance)[29].

While various ethical principles have been proposed in relation to the rights of the child and AI systems, their effective implementations and practical applications are still mainly unexplored[30]. Children are different among them and from adults, accordingly AI principles concerning children should not be considered nor treated as a subcategory of other guidelines[31]. Accordingly, Wang *et al.* identify four main '*challenges in translating ethical AI principles into practice for children*'[32]:

1. '*Lack of consideration of the developmental aspect of childhood*'[33]: the vast number of technologies and their various applications make it difficult to provide consistent professional codes and norms for AI applications. Incorporating children introduces a new layer of complexity to this scenario. Their unique needs, diverse age ranges, development stages, backgrounds, physical and psychological traits necessitate special attention;

2. '*Lack of consideration of the role of guardians in childhood*'[34]: parent(s) or legal guardian(s) bear the ethical and legal primary responsibility for the upbringing and development of the child (Article 18 UNCRC) and for the children's provision of appropriate direction and guidance in the exercise of their rights (Article 5 UNCRC). Therefore, the role of parent(s) and legal guardian(s) must be considered and examined, but without falling in the traditional assumption that they possess superior expertise or skills to orient children in the digital landscape;

[29] *Ibidem*.

[30] G.Wang, J. Zhao, M.Van Kleek & N.hadbolt, 'Challenges and opportunities in translating ethical AI principles into practice for children' (2024) Nature Machine Intelligence 6, 265–270 https://doi.org/10.1038/s42256-024-00805-x accessed 04 July 2025.

[31] *Ibidem*.

[32] *Ibidem*, citing.

[33] *Ibidem*, citing.

[34] *Ibidem*, citing.

3. '*Lack of child-centred evaluations considering children's best interests and rights*'[35]: relying solely on quantitative metrics and technical evaluation, while important, can present challenges. Translating ethical AI principles into practice for children requires a more balanced approach between both empirical variables and quantitative measurements, and, in general, a paradigm shift towards a more human-centred approach;

4. '*Lack of a coordinated, cross-sector and cross-disciplinary approach*'[36]: experts from other domains, dealing with analogous issues, often have different vocabularies and methodologies. One of the main challenges lies in their adaptability across different AI principles. Cross-sector and cross-disciplinary collaboration is essential to harmonize and encourage knowledge transfer while avoiding duplicate efforts.[37]

These challenges add other layers of difficulty in integrating children's rights in the design and development of a product or service. Smart toys like Cayla's doll, should not only be secure- and privacy-by-design, but should also *e.g.* take into account children developing language skills, by adopting a child friendly language in accordance of the maturity of the child, while also considering a system of blocking access to content children should not access without adults' supervision. Accordingly the difficulty is not just on how to make the system embedded in the toy technically robust and resilient, but it also concerns dealing with developmental theories, adaptability to different situations (*e.g.:* Is the system capable of adapting content and language according to the child's specificity? and how to make the system able to do that while following the principle of data minimization?), and definitions of concept like "appropriateness" (*e.g.:* What may be considered appropriate for a child of a certain age, maturity and background could not be necessarily considered appropriate for and by another child).

Given all these challenges, engineers and practitioners working on the design and development of AI systems for, accessed by or impacting children, are required to

[35] *Ibidem*, citing.

[36] *Ibidem*, citing.

[37] *Ibidem*.

deal with more than technical problems and solutions. This is for those topics that are indeed '*socio-technical*'[38], meaning that '*social and technical aspects are interwoven in such a way that studying one without due consideration of the other makes for an incomplete investigation and understanding*'[39]. To guide practitioners in diving this scenario, some references are made to existing contributions from academia, industry, international organizations/associations and NGOs.

However, academic contributions on how to design, develop and deploy AI systems compliant with related existing standards and obligations are still few, and mainly summarized as "design implications" at the end of a paper. While literature reviews can offer a valid overview of a topic, few are the works[40] investigating children's rights coverage and inclusion in engineering and computer science' works, and even less are works trying to summarize all these "design implications" in one single and easy to use document. This sum up could be interesting and possibly useful in real life situations, since coming from in-the-field studies, and a service- or product-specific framework can be valuable to achieve precise applicable guidelines.

Nevertheless, industry-partnership projects and international organizations and associations have been mainly focusing on a broader approach, advocating for responsible innovation for children well- being (*e.g.:* LEGO and UNICEF[41]), a child-centered approach to AI system (*e.g.:* UNICEF[42]) and age appropriate services (*e.g.:*

[38] Rashina Hoda., *Qualitative Research with Socio-Technical Grounded Theory. A practical guide to qualitative data analysis and theory development in the digital world* (Springer Charm, 2024), https://doi.org/10.1007/978-3-031-60533-8 citing.

[39] *Ibidem*, citing.

[40] See, for example, G.Wang, J.Zhao, M.Van Kleek, and N.Shadbolt, 'Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children' (CHI - Conference on Human Factors in Computing Systems, New Orleans, LA, April 30 – May 5 2022).

[41] UNICEF and LEGO, 'The Responsible Innovation in Technology for Children (RITEC) Project'. See UNICEF's webpage 'Responsible Innovation in Technology for Children. Project | Digital technology, play and child well-being' (*UNICEF*) https://www.unicef.org/innocenti/projects/responsible-innovation-technology-children accessed 06 July 2025.

[42] UNICEF - V. Dignum, M.Penagos, K.Pigmans and S.Vosloo (November 2021).

ISSN 2281-5147

IEEE[43]). These contributions are one of the most cited when it comes to children and AI.

Contributions coming from (or in collaboration with) businesses and industry are important for their ground on real life scenarios and interests, bridging the gap between academic research and industry actual needs. Integrating a children's rights approach and design for well-being into business strategies can have positive outcomes for both children (their rights, needs and desire with better products) and brands (boosting brand reputation and values, by differentiating themselves from their competitors and within their customers, and attracting possible investors)[44].

The "Responsible Innovation in Technology for Children" (RITEC) project is a collaboration between UNICEF and The LEGO Group, funded by The LEGO Foundation, aiming at investigating how the design of children's digital experiences affects their well-being, and provides guidance on design choices that can promote positive outcomes for children's well-being[45]. From the RITEC project a framework (the final "RITEC-8", updated and published in 2024) and a design toolbox (the "RITEC Design Toolbox") have been developed to provide an '*easy-to-use guidance for designers of digital play*'[46] by including a list of relevant features and examples[47].

The framework developed in the context of this project is called RITEC-8[48] because is grounded in 8 pillars: (i) autonomy (allow children to be in control and make decisions that matter for them and their play); (ii) competence (considering

[43] IEEE, Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children, 2021. DOI: https://doi.org/10.1109/IEEESTD.2021.9627644.

[44] *Ibidem*.

[45] UNICEF, The Business Case for Designing for Children's Well-Being in Digital Play Summary for Executives, 2024. https://www.unicef.org/childrightsandbusiness/reports/business-case-designing-childrens-well-being-digital-play accessed 06 July 2025.

[46] *Ibidem*, citing.

[47] *Ibidem*.

[48] UNICEF, Digital technology, play and child well-being. Responsible innovation in technology for children, 2024. https://www.unicef.org/innocenti/reports/responsible-innovation-technology-children accessed 06 July 2025.

meaningful rewards for progress and allowing children to adjust and improve); (iii) emotions (experience positive as well as more challenging emotions); (iv) relationships (taking into account children's different needs and characteristics, allow them to make new friends and socialize while competing, creating, and/or collaborating with others); (v) creativity (encourage children's curiosity and imagination to invent and experiment); (vi) identities (while playing, allow children to explore and express facets of themselves and of others); (vii) diversity, equity & inclusion (experience intended for different children and needs); and (viii) safety and security (children feel and are kept safe while playing)[49]. The framework is also accompanied by a design toolbox (RDT) with the aim of providing design professionals in the online gaming industry (product, visual, UX, research, but also management levels, and safety professionals) with practical tools for incorporating the RITEC-8 for children's well-being into their design process[50].

UNICEF, before the RITEC Project, has already been focusing on AI systems in its "Policy Guidance on AI for Children"[51]. The document provides nine requirements for child-centered AI, and furnishes a set of '*complementary online resources*' and '*practical implementation tools*'[52]. The guidance is addressed to different stakeholders, from development teams to policymakers, and, while this is important, finding a common both understandable and practical language for all may be challenging. The risk is too high-level guidance, resulting difficult to fully implement into the actual work's duties (*e.g.*: The "transparency" principle does not specify how to explain AI decisions to a child of a certain age or background over a child of another age or background).

Meanwhile, the IEEE, as technical professional organization, elaborated the "Standard for an Age Appropriate Digital Services Framework Based on the 5Rights

[49] *Ibidem.*

[50]UNICEF, 'RITEC Design Toolbox. Designing for children's well-being in digital play' https://www.unicef.org/childrightsandbusiness/workstreams/responsible-technology/online-gaming/ritec-design-toolbox accessed 06 July 2025.

[51] UNICEF - V. Dignum, M.Penagos, K.Pigmans and S.Vosloo (November 2021).

[52] *Ibidem,* citing.

ISSN 2281-5147

Principles for Children"[53] (IEEE 2089-2021)[54]. The IEEE 2089-2021 is practical in its formulation, being developed to be used in '*software engineering and digital services organizations*'[55], including but not limited to those '*providing services and products that engage with children or are likely to be accessed by or engage with children*'[56]. Although its technical nature, the document is informed by the UNCRC and the UN General Comment No.25, and it is based on the principle of the "*best interests*"[57] of the child[58]. The Document is an important attempt to combine a more technical approach with existing policies and regulations on the subject.

NGOs have also attempted '*bridging high-level principles and practical challenges*'[59] by defining what innovators need to know to realise children's rights in their product or service[60]. In 2023, the "5 Rights Foundation" (within the "Digital Future Commission" project) released the "Child Rights By Design": a guidance aiming to provide clear and practical indications to those figures involved in the process of

[53] IEEE Std 2089-2021(2021).

[54] In 2023, the IEEE 2089-2021 has been recognized to serve as the foundation for an *European Committee for Standardization (CEN)/European Committee for Electrotechnical Standardization (CENELEC)* Workshop Agreement (CWA 18016), helping to serve various EU regulations and policies, such as the the DSA and the 'European strategy for a Better Internet for Kids (BIK+)' (see: IEEE SA, 'IEEE 2089™ Provides Foundation for European Reference Document for Children's Protection & Well-being Online' (2023). https://standards.ieee.org/news/ieee-2089-european-reference-document/ accessed 13 May 2025).

[55] IEEE Std 2089-2021 (2021), citing.

[56] *Ibidem,* citing;

[57]The "best interest" principle refers to Article. 3 UNCRC and, according to S. Livingstone et al. (S. Livingstone, N. Cantwell, D.Özkul, G. Shekhawat and B. Kidron, 'The best interests of the child in the digital environment' (March 2024) https://www.digital-futures-for-children.net/our-work/best-interests accessed 14 May 2025), it implies that, when children's rights seem to be in tension or when other parties' interests (such as those of companies or organizations) may conflict with them, to identify "*which rights are to be given precedence*", an independent procedure of "*best interests' determination*" should be designed to avoid "*provide legitimation for whichever right a company may favour*".

[58] IEEE Std 2089-2021 (2021).

[59] Digital Futures Commission and 5Rights (11 March 2023).

[60]*Ibidem*.

ISSN 2281-5147

creation, design, development and deployment of a digital product or services likely to be used by or impacting on children[61]. Grounded on the UNCRC, the guidance calls for a "*by-design*" approach[62], that would mean including children's rights considerations in every phase of an AI system's lifecycle. By collecting inputs from innovators, practitioners, and children, the guide is structured around 11 high-level principles[63] and align with the main crucial phases of an innovation process[64]. Given the peculiar opportunities and challenges AI systems pose, the 5Rights Foundation also published the "Children and AI Design Code. A protocol for the development and use of AI systems that impact children"[65](2025). The Code is composed of distinct stages and developed so as to be applicable in each phase of an AI system's lifecycle[66]. Moreover, it is structured as an '*assessment process*' so that '*non-conformity is identified, evaluated, and mitigated*'[67] and progress are recorded in writing[68]. While recording can help keep track of both progress and risks, the "*requirement checklist*" provided at the end of the Code may be not sufficient to report and elaborate both of them. Here, integrating existing related initiatives can be a valuable asset and can avoid "reinventing the wheel" when other contributions or disciplines have already found a solution (as suggested by Wang et al. when calling for a cross-sector and

[61] *Ibidem*.

[62] As C. Djeffal highlights (in: C.Djeffal, 'Children's Rights by Design and Internet Governance: Revisiting General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment' (2022) 11(6) Laws https://doi.org/10.3390/laws11060084 accessed 05 July 2025), the "*by-design thinking*" has traditionally been applied in the area of privacy, data protection, and security, but it has begun to spread also throughout the legal system. The "*law-by-design norms*" take advantage of "*the law's binding nature and combine it with normative claims that are to be translated into technology*".

[63] 5Rights Foundation's "Child Rights by Design" principles: (i) equity and diversity, (ii) best interests, (iii) consultation, (iv) age appropriate, (v) responsible, (vi) participation, (vii) privacy, (viii) safety, (ix) wellbeing, (x) development, and (xi) agency.

[64] Digital Futures Commission and 5Rights (11 March 2023).

[65] 5Rights Foundation (March 2025).

[66] *Ibidem*.

[67] *Ibidem*, citing.

[68] *Ibidem*.

cross-disciplinary approach). The IEEE 2089-2021[69], for example, foresees the creation of an '*Age Appropriate Register (AAR*)'[70]: a '*medium*'[71], used to document and communicate progressively, and '*handover*'[72] between the competences and responsibilities of the stakeholders involved in one phase to those involved in the subsequent phases[73]. Therefore, the AAR (or a similar tool), can be an important ally in monitoring and ensuring compliance with children's rights (and safety and security standards) throughout the whole AI system's lifecycle.

Whether the use of this or similar tools, in cases such as the doll Cayla, could have been found useful and successful in timely identifying, analysing, and mitigating risks and challenges remains an open question. Further research is needed in order to assess the practical outcomes of applying such frameworks and guidelines, so as to provide effective and actionable indications to practitioners. Retrofitting a product to comply with these rights after development can be equally (if not more) difficult and costly.[74] Accordingly, a child rights approach should be kept as a lighthouse since the pre-deployment phase of an AI system's lifecycle.

3. *Testing* and *validation*: regulatory sandbox environments to ensure safety and compliance

Testing AI systems intended for children within regulatory sandboxes is a crucial step in ensuring the protection of their rights. Children and preadolescents, as particularly vulnerable users, require special consideration from the earliest stages of technology

[69] IEEE Std 2089-2021(2021).

[70] *Ibidem*, citing.

[71] *Ibidem*, citing.

[72] *Ibidem*, citing.

[73] *Ibidem*.

[74] Digital Futures Commission and 5Rights (11 March 2023).

design. It is essential to assess how these systems might affect their privacy, safety, and overall well-being from the outset.

Regulatory sandboxes provide a controlled environment in which innovative digital solutions can be tested, allowing technological development to be balanced with the need for protection. This approach makes it possible to identify and address potential issues before the product is released to the market and its compliance with standard and regulation children's rights by design. Several European States include the use of sandboxes as a means to build a comprehensive legal framework for AI. This trend is supported by the EU, which views regulatory sandboxes as facilitators of innovation and recognizes them as a crucial tool in future regulatory activities concerning AI. A regulatory intervention for the definition of this tool was provided by the AI Act, definitively approved on May 21, 2024, which in Article 57 defines AI sandboxes[75]. Regulatory sandboxes on AI, established by European or national competent authorities, provide a controlled environment to develop and test innovative AI systems before commercial deployment. These activities take place under the direct supervision of authorities to ensure compliance with EU and national regulations. When the systems involve the processing of personal data or fall under other regulated areas, data protection authorities and other relevant bodies must be involved in the sandbox's operation[76]. Regulatory sandboxes can help address these issues by providing regulatory certainty for technology companies and other stakeholders, fostering collaboration and capacity-building with and among regulators, and promoting regulatory clarity and compliance[77].

The use of regulatory sandboxes in Europe to test products aimed at minors is still limited and not yet systematized. However, there are some cases and emerging trends that indicate a growing interest in this area, particularly in relation to financial

[75]EU, 'Artificial Intelligence Act' (2024). Chapter VI: Measures in Support of Innovation. https://artificialintelligenceact.eu/chapter/6/ accessed 12 May 2025.

[76] S. Ranchordas, 'Experimental Regulations for AI: Sandboxes for Morals and Mores' (2021) 1(1) Morals & Machines 86 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3839744 accessed 12 May 2025.

[77] Datasphere Initiative, 'Sandboxes for data: creating spaces for agile solutions across Borders' (2022) https://www.thedatasphere.org/ accessed 12 May 2025.

education for young people, the protection of personal data (including GDPR compliance and age of consent requirements), the responsible use of technology such as AI and digital platforms designed for minors, and the development of secure digital payment solutions for those under the age of 18.

Datasphere initiative[78] has published a case study on regulatory sandboxes, highlighting the inability of current laws and policies to keep pace with rapid technological developments. The study proposes regulatory sandboxes as tools to foster innovation while ensuring effective data governance - particularly when it comes to children's data. The sandbox model described in the study does not allow for temporary suspensions of legal constraints; instead, it promotes innovation within the existing regulatory framework, encouraging solutions that remain compliant with current rules, trends and better oversee foreign products that process children's data within their jurisdictions[79].

The Norwegian Police University College has tested a bot ("PrevBOT") within a regulatory privacy sandbox, aiming to explore the feasibility of developing a tool capable of automatically patrolling the open internet. The goal of this project is to detect and prevent the sexual exploitation of minors by identifying suspicious behavior and grooming attempts in real time. By combining AI-driven language analysis, behavioral profiling, and age estimation technologies, PrevBOT seeks to serve as a proactive digital safeguard, helping law enforcement intervene before harm occurs - while operating within strict privacy and ethical frameworks. PrevBOT is designed to protect minors online by addressing the growing issue of digital grooming. This crime involves adults who use psychological manipulation and digital communication to build trust with children, often with the intent of sexual exploitation. To effectively counter this threat, PrevBOT integrates advanced technologies capable of identifying risky interactions before they escalate. The system is trained to detect grooming language not only in explicit terms but also in the subtle

[78] The "Datasphere Initiative" is a non-profit dedicated to global collaboration on technical and policy solutions for the urgent, multidimensional, and cross-border challenges of data governance (see: https://www.thedatasphere.org//about-us/ accessed 14 May 2025).

[79] UNICEF, 'Regulatory sandboxes . Case study', 2025: https://www.unicef.org/innocenti/media/11091/file/UNICEF-Innocenti-Regulatory-Sandboxes-Case-Study-2025.pdf accessed 14 May 2025.

ISSN 2281-5147

and coded language often used in chats, including slang and emerging online expressions. It can analyze conversation patterns to recognize early signs of inappropriate behavior, even when the language appears innocent. In addition, PrevBOT estimates the age and gender of users based on their writing style and digital behavior. This allows it to identify potentially fake profiles, especially when adults pretend to be minors to gain access to youth-oriented spaces. Recognizing age discrepancies is important for detecting interactions where children may be at risk. The bot also performs sentiment and behavioral analysis by monitoring response times, typing speed, emotional tone, and interaction patterns. This helps identify users who, despite maintaining a calm or friendly appearance, may be displaying signs of persistence, or manipulation - indicators that their intentions might not align with their words. Together, these capabilities enable PrevBOT to provide proactive protection for minors, flagging dangerous behavior early while respecting privacy regulations and promoting safer digital environments for young users[80]. PrevBOT project is still in its early stages, and it will be interesting to see how it manages to strike a balance between the need for freedom and the need for safety. Minors have a right to agency and privacy, but without an adequate level of online protection, they would not be able to fully exercise those rights. Trust is a key element for a project that aims to comply with both current regulations and the principles of ethical and responsible AI. In this regard, emphasizing transparency and actively involving stakeholders throughout the research process provides a strong foundation.

An important experimentation to make in consideration is the case of the UK's ICO Regulatory Sandbox. The United Kingdom's Information Commissioner's Office (ICO) established the ICO Sandbox program in 2019 to support organizations developing innovative data-based products and services, ensuring compliance with privacy regulations. Since 2020, the program has focused particularly on two areas: protecting children's online privacy through the Children's Code and managing the complex sharing of personal data in sensitive sectors such as health, education, finance, and public administration.

[80] The Norwegian Police University College, exit report: PrevBOT (20 September 2024) https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/the-norwegian-police-university-college-exit-report-prevbot/ accessed 14 May 2025.

ISSN 2281-5147

A notable example is the Lookafterme project by FlyingBinary Limited[81], a digital service based on AI designed to support mental health issues such as anorexia and bulimia, including for children from the age of eight. The system monitors online content in real time and alerts users to potentially harmful material, providing integrated clinical support. During its participation in the Sandbox, FlyingBinary ensured full compliance with UK GDPR, the Data Protection Act 2018, and the Children's Code. The company focused particularly on secure and age-appropriate authentication methods for children, the principle of data minimization, and data protection by design. Special attention was given to the protection of health data, considered sensitive, and ensuring that data processing always took place in the best interest of the child, using the "Best Interests Framework", an ICO tool inspired by the UNCRC. The project serves as a replicable model demonstrating how technological innovation and the protection of fundamental rights can be effectively integrated, especially in sensitive fields like health and education.

Lessons learned from various sandbox experiences highlight both their potential and the challenges they pose - especially concerning children's data. Sandboxes can play a crucial role in helping stakeholders balance the benefits of using minors' data with the need to fully safeguard their rights: testing the doll Cayla in such an environment could have helped experts identify those vulnerabilities and issues before its deployment into the market, and possibly avoid children's harm and company's reputational damage. Encouraging tech companies to participate in sandboxes is a key factor in their success. While some sandboxes provide financial support to cover legal, technical, or operational costs, the most valuable incentive is often the regulatory clarity and compliance assurance they offer.

Sandboxes have demonstrated global relevance and potential for cross-border replication. In particular, international sandboxes can enhance regulatory capacity, improve cooperation, foster innovation and compliance, and promote the availability and accessibility of data across jurisdictions and sectors. By engaging directly with emerging technologies - including those developed abroad regulators, especially in countries without a strong domestic tech sector, can stay informed on global trends

[81] Information Commissioner's Office, 'Regulatory sandbox final report: Flyingbinary'(Tech. Rep., 2022). https://ico.org.uk/media2/migrated/4021302/flyingbinary-exit-report-202208.pdf accessed 15 May 2025.

and better oversee foreign products that process children's data within their territory[82].

4. *Deployment* (and *post-deployment*): cyber-threats and risk-driven mitigation

The deployment of AI-based technologies designed for/interacting with/impacting children does not mark the end of the innovation lifecycle but initiates a new phase - one that requires ongoing oversight, responsiveness and ethical commitment. Indeed, ensuring that these systems uphold children's rights over time requires a structured post-deployment framework of assessment, monitoring, and risk mitigation.

interference and, in fact, prove to be particularly vulnerable to a wide range of cyber-threats.[83] Common risks include data breaches that can compromise sensitive personal information (*e.g.*: names, locations and voice recording) or even adversarial attacks that can manipulate system inputs to trigger inappropriate or unsafe outputs, distorting educational content or conversational responses.

As concerns data breaches, particular attention should be paid to the real case of the Smart Toy produced by Fisher-Price[84]. This product represents one of the earliest and most emblematic examples of an Internet-connected smart toy, designed to establish personalized interaction with the child through the use of a rudimentary form of AI[85]. Manufactured by the American company Fisher-Price, a subsidiary of Mattel, the toy was available in three versions - a bear, a monkey, and a panda - and relied on Wi-Fi connectivity and a mobile application managed by parents to oversee its functions. The Smart Toy was capable of gradually learning the child's preferences, customizing

[82] *Ibidem*.

[83] For further reading, S. Shasha et al, 'Playing with Danger: A Taxonomy and Evaluation of Threats to Smart Toys' (2018) 6 IEEE Internet of Things Journal 2986, 2996.

[84] Description of the Fisher-Price Smart Toy Bear, see: http://fisher-price.mattel.com/shop/en-us/fp/smart-toy/smart-toy-bear-dnv31.

[85] For a more in-depth look at the case, refer to: M.C. Gaeta, 'Smart toys and minors' protection in the context of the Internet of everything' (2020) 11(2) Eur J Privacy L & Tech 118.

its content and responses through the use of physical smart cards[86]. However, a technical analysis conducted at the hardware, software and network levels[87] revealed critical vulnerabilities in the system's APIs - the *Application Programming Interfaces* that enable communication between applications and services. These vulnerabilities involved the lack of proper identity verification for message senders, thereby allowing unauthorized third parties to gain access to sensitive personal data, such as the child's name, date of birth, language, activity history, and similar information. More concerning was the demonstrated possibility of modifying or deleting user profiles and even altering the toy's functionality, potentially exposing children to physical and psychological harm. This case highlights how, even in the absence of immediate damage, a cyberattack can deeply compromise a child's private and relational sphere, emphasizing the risks posed by the aggregation of seemingly innocuous data, which can be utilized to construct a detailed and exploitable personal profile.

As for the cyber-risks of manipulation, some smart toys have begun incorporating generative AI systems such as ChatGPT - one notable example is Grok[88]. Grok is a conversational toy designed to engage children through verbal interaction powered by a LLM, and it is among the first toys to feature a voice interface connected to ChatGPT. While the toy's goal is to promote natural dialogue, integrating LLMs into children's products raises significant concerns around safety and control. In Grok's case, researchers conducted an experiment[89] that demonstrated the toy continuously streams audio to external servers without requiring a wake word. It records not only

[86] Fisher-Price described the toy as "*an interactive learning friend with all the brains of a computer, without the screen*", thus emphasising its educational and innovative intent to combine technology and learning in a playful and non-invasive format.

[87] Rapid7, *R7-2015-27 and R7-2015-24: Fisher-Price Smart Toy and HereO GPS Platform Vulnerabilities (FIXED)* (Rapid7 Blog, 2 February 2016), online at: https://community.rapid7.com/community/infosec/blog/2016/02/02/security-vulnerabilitieswithin-fisher-price-smart-toy-hereo-gps-platform).

[88] Shaped like a plush rocket, Grok contains an embedded "voice box" inside a zippered compartment and requires Wi-Fi connection via a companion app. To see the product: Curio Interactive Inc. 2024. Curio - AI Toys, https://heycurio.com/. accessed 05-07-2025.

[89] V. Pavliv, N. Akbari and I Wagner, 'AI-powered smart toys: interactive friends or surveillance devices?' in Proceedings of the 14th International Conference on the Internet of Things (IoT '24, ACM 2025) 172.

intentional commands but also background conversations, including external audio sources or nearby people. This raises privacy concerns, as sensitive information can be captured and transmitted without the user's knowledge. Furthermore, the toy's responses revealed vulnerabilities: although the experiment was not designed to elicit inappropriate content, some replies contained double meanings - for example, "*it's about spirit not size*". This suggests it may be possible to bypass or break out of the system prompt, allowing the toy to produce inappropriate or unsafe statements, representing a child safety risk and a potential avenue for manipulation.

Given the outlined and - not merely theoretical - cyber risks[90] the post-deployment phase must prioritize the implementation of robust cybersecurity safeguards[91].

Article 15 of the AI Act mandates that high-risk systems - including those used in educational and play-based contexts[92]- be developed with a high degree of robustness and cybersecurity, aligned with the state of the art. This includes encryption, anomaly detection and protection against tampering. At a broader level, Article 5(1)(b) of the AI Act explicitly prohibits the use of AI systems that exploit vulnerabilities linked to age, thereby shielding children from manipulative or coercive behaviors.

Nevertheless, ensuring a secure post-deployment environment for children requires more than technical safeguards; it demands ongoing, structured monitoring and accountability throughout the system's lifecycle. As required by Article 71 of the AI

[90] See the BBC News article related to the Cayla doll case: https://www.bbc.com/news/world-europe-39002142 (*BBC,* 2017), accessed 10 May 2025. Consider also that, where children's rights may be compromised, predefined sunsetting or withdrawal protocols should be established to ensure the safe decommissioning of harmful or outdated AI systems.

[91] In this context, it is important to consider that during the negotiations of the AI Act, numerous child rights organizations called for greater attention to the specific needs of children. In particular, they urged the inclusion of educational systems in the list of "*high-risk*" applications, the prohibition of AI practices that exploit vulnerabilities related to age and the development of clear guidelines to ensure transparency and comprehensibility of AI systems for children. While the final text of the AI Act has partially addressed these demands - by, for instance, including educational AI systems in Annex III and banning the use of AI that exploits age-related vulnerabilities - it has fallen short of explicitly recognizing children as a protected group in all provisions and it lacks specific instructions on how to communicate with child users. European Commission, "Commission Seeks Feedback on Guidelines on the Protection of Minors Online under the Digital Services Act" (11 March 2024).

[92] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, OJ L1689/1, Annex III.

ISSN 2281-5147

Act, providers of high-risk AI must implement a post-market monitoring system to collect and assess performance data over time. Rather than a one-off evaluation, this should be seen as a living framework - one integrating technical vigilance with a sustained ethical responsibility to act in the best interests of the child.

Moreover, post-deployment oversight must be equipped to address adversarial threats, such as input manipulation or the covert reprogramming of educational agents for *non*-educational - or harmful - purposes[93]. To mitigate these risks, real-time monitoring systems must be capable of identifying not only technical malfunctions but also indicators of deliberate misuse, unauthorized alterations or manipulation, as these safeguards are essential to ensuring the long-term safety, reliability and trustworthiness of AI systems - provided they are effectively integrated within a continuous risk assessment framework[94].

Central to this evaluation is the integration of the "Child Rights Impact Assessment" (hereinafter, CRIA): a methodology, applied from the design phase, that examines the potential impacts on children of laws, policies, programmes and services, and that can also be applied to assess both the potential and actual effects of AI systems on children's rights[95]. The CRIA process begins with a screening stage to determine whether a policy, service or technology warrants a full assessment. Where significant impacts are identified, a full CRIA follows, starting with an analysis of the proposal's scope and the relevant Articles of the UNCRC. This stage is backed by qualitative and quantitative evidence, including direct consultation feedback with children to ensure their views are considered and to identify recurring themes and priority concerns. The assessment then evaluates general and disproportionate impacts on specific groups of children and outlines corresponding mitigation strategies (*e.g.:* reduction in exposure to harmful content by X%). The process concludes with a set of findings, including

[93] For an in-depth investigation, see B. Guembe et al., 'The Emerging Threat of AI-Driven Cyber Attacks: A Review' (2022) 36(1) Applied Artificial Intelligence 2037254.

[94] NIST AI, Artificial Intelligence Risk Management Framework *(AI RMF 1.0)* (2023); URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[95] *Ex multis*, J. H. and M.A. Stephenson, 'Human Rights Impact Assessment: Review and Practice Guidance for Future Assessments' (2010) Scottish Human Rights Commission Report; L. Payne, 'Child Rights Impact Assessment as a Policy Improvement Tool' in K. Roberts Lyer (ed), Human Rights Monitoring and Implementation (Routledge 2020) 91.

ISSN 2281-5147

recommendations and monitoring mechanisms. Publishing the CRIA enhances transparency and accountability, ensuring that AI systems are developed in a manner that upholds children's rights and delivers long-term, positive outcomes. Alongside this risk assessment approach, periodic impact reports should be mandated for high-risk AI systems, modeled after the "Data Protection Impact Assessments" (DPIAs), but tailored to specifically address child-specific risks, so that developers, providers, regulators and institutional users[96] must share clear, traceable responsibilities for the long-term impacts of AI on children's well-being.

Therefore, post-deployment accountability demands a collective responsibility from multiple stakeholders.[97] Indeed, regulators must define and enforce standards for an ongoing compliance, while civil society, academic and research institutions should serve as "watchdogs" and evaluators of AI's forthcoming impact and industry actors must commit to the long-term stewardship of their technologies. On this point, instruments such as the aforementioned AAR could play a role in ensuring that AI systems consistently meet children's rights and needs. It could serve as a tool for monitoring issues identified in earlier phases and facilitating the transfer of knowledge across different phases of the design and development. This ensures alignment among all stakeholders, enabling ongoing monitoring to maintain compliance throughout the product's lifecycle.

Ultimately, accountability must be understood not merely as a legal or procedural obligation, but as a moral and social responsibility. The best interests of the child, as enshrined in Article 3 UNCRC, can become an enforceable benchmark only if a "post-deployment conscience" is embraced - one that compels designers, developers and even decision-makers to measure AI's success, by its real-world impact on children's rights and well-being.

[96] Such as schools, public agencies and other stakeholders.

[97] T. Merlin, J. Boyd and C. Donovan, 'The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI' (2024) *arXiv preprint* arXiv:2410.04931.

ISSN 2281-5147

5. Closing the lifecycle loop of child Rights-Based AI

And so, this story - one *about* and *for* children - almost comes to an end. It is a narrative where child agency, safety and protection form the hoped-for happy ending. Yet reality proves far more complex. Even when AI systems are designed, developed and deployed in line with children's rights standards, there is no guarantee of their continued compliance in real-world use. Here is where our story begins again, going back to the development phase or even to the design phase, in a never ending, possibly safe and child rights-based loop.

To be fully applicable, the lifecycle loop of child rights-based AI suggested in this work needs to address some limitations:

(i) Existing frameworks (*e.g.*: from UNICEF[98] and IEEE 2089-2021[99]) provide important guidelines to practitioners, but they often miss out on metrics and/or practical implementation tools. These gaps can pose limitations to their applicability, resulting in too high-level recommendations of difficult understanding and/or operationalization for practitioners. At the same time, few academic works, focusing on a specific case or system, rarely offer scalability solutions "*per se*". Consequently, core research priorities are: (i) identifying, evaluating and validating metrics and operational measures specifically for AI systems intended for children, and (ii) integrating these metrics and measures with knowledge from other fields (*e.g.*: development theories). At the same time, practitioners can in the meanwhile refer to valuable already existing materials. To guide the reflection, when creating and building a new service or product for children, practitioners can indeed refer to contributions

[98] UNICEF - V. Dignum, M.Penagos, K.Pigmans and S.Vosloo (November 2021).

[99] IEEE, Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children, (2021).

such as the ones highlighted above in this paper, or others like the 5Rights' "Playful by Design Toolkit"[100] or Save the Children's guide on "Child-Centered Design"[101].

(ii) Regulatory Sandboxes are expected to be created in the EU by 2026[102]. Regulatory Sandboxes can be very effective tools to bridge the gap between technological innovation and slow regulatory adaptation. This gap is particularly evident in sectors such as fintech, AI, blockchain and biotech, where technology is advancing faster than regulators can regulate it. The sector concerning the protection of minors in the use of technology presents serious regulatory gaps, making it difficult to effectively safeguard the rights of young people in the digital environment. The Italian case is an emblematic example. Since the entry into force of the new European Electronic Communications Code[103] (December 2020), a derogation that allowed ICT companies to monitor and report child sexual abuse material online has lapsed. This regulatory gap has had direct and measurable consequences: reports to the competent authorities have decreased by 46% across Europe, negatively impacting prevention and enforcement efforts against child abuse. Furthermore, the "Caivano Decree" (September 2023) [104], in an effort to strengthen child protection, delegated to AGCOM the task of defining technical tools for age verification and secure access to digital content. However, to date, no concrete implementing measures have been

[100] 5Rights Foundation, 'Playful by Design' (2021). https://playfulbydesign.5rightsfoundation.com. Accessed 11 September 2025.

[101] Save the Children Finland, 'Child-Centered Design' (2020). https://resourcecentre.savethechildren.net/document/child-centered-design. Accessed 11 September 2025.

[102]European Parliament and Council. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (AI act) (text with eea relevance), 2024.

[103] Directive (EU) 2018/1972 establishing the European Electronic Communications Code [2018] OJ L321/36; transposed into Italian law by D. lgs., 8 november 2021, n. 207, GURI n.292, 9 December 2021.

[104] Decreto Legge 15 Settembre 2023, n°123 "Misure urgenti di contrasto al disagio giovanile, alla povertà educativa e alla criminalita' minorile, nonche' per la sicurezza dei minori in ambito digitale" https://www.gazzettaufficiale.it/eli/id/2023/11/14/23A06292/sg accessed 06 July 2025.

adopted: only guidelines are in force, which are not legally binding, and actual implementation by operators remains inconsistent. In this context, innovative tools such as AI regulatory sandboxes could represent a strategic opportunity to overcome the regulatory deadlock. Sandboxes offer a regulated yet flexible environment in which to test technologies and solutions (such as age verification systems, AI-based parental control, or the automated detection of illegal content) before their full legal application. The experience of the United Kingdom, for instance, shows how regulatory experimentation can contribute to the development of dedicated legislation. The UK Information Commissioner's Office (ICO) has used sandboxes to develop the principles of the "Children's Code", a legal framework that has since established new standards for the design of digital platforms with a focus on respecting children's rights.

However, so far, there are few examples of attempts to create such environments. A recent paper[105] proposes a regulatory framework for child-friendly AI sandboxes that integrates the EU AI Act with UNICEF guidelines and other international references (UN, OECD, UNESCO). This framework is structured around a multi-stakeholder, modular, and iterative process aimed at ensuring that the development and testing of AI systems respect the rights and well-being of children. Given the international relevance of the topic, interesting new contributions are expected in the near future;

(*iii*) *Zero risk doesn't exist,* cybersecurity threats may still emerge over time. Therefore, it is essential to move beyond voluntary guidelines and soft law (meaning, codes of conduct and non-binding recommendations). To ensure the long-term protection of children's rights in digital environments, companies must be encouraged - and, where necessary, compelled - to take shared responsibility through binding legal frameworks and effective enforcement mechanisms. In the post-deployment phase, proactive regulation is crucial to clearly define the duties and liabilities of AI producers, software developers and platform operators, with enforceable measures such as substantial fines for damages and explicit rights of claim for affected parties (*post*-damage protection). This ongoing accountability should be anchored in systematic monitoring, inspired by the CRIA or comparable methodologies, and guided by

[105] V. Charisi and V. Dignum, "Operationalizing AI Regulatory Sandboxes for Children's Rights and Well-Being" in Human-Centered AI (Chapman and Hall/CRC 2024) 231.

robust indicators. Relevant measures may include: (i) tracking the number and severity of cyber-incidents involving children, (ii) assessing the speed and effectiveness of responses to identified risks, (iii) evaluating the participation of children in post-deployment reviews, (iv) analysing the distribution of impacts across different groups of children in order to detect disproportionate effects, and (v) collecting data on children's own perceptions of safety and well-being when engaging with digital systems. Embedding such evidence-based indicators within regulatory frameworks ensures that accountability extends beyond the design stage, turning compliance into a continuous, transparent and participatory process that protects children's rights throughout the entire life cycle of AI systems.

Also, future efforts should aim to overcome these limitations by developing more effective strategies for engaging children directly - such as through interviews, surveys and focus groups - and by fostering a collaborative approach that integrates diverse professional and academic expertise. This strategy will better position the final AI system to meet security standards and ensure compliance with children's rights and related obligations.

ISSN 2281-5147