

Edited by

Joachim von Braun and Britta Holmberg

Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children



**Proceedings of the Workshop
Casina Pio IV, Vatican City, 21-22 March 2025**



**LIBRERIA
EDITRICE
VATICANA**

Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children

Pontificiae Academiae Scientiarum Scripta Varia 158

The Proceedings of the Workshop on

Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children

21-22 March 2025

Edited by

Joachim von Braun
Britta Holmberg



THE PONTIFICAL
ACADEMY
OF SCIENCES

EX AEDIBVS ACADEMICIS
IN CIVITATE VATICANA • MMXXVI



LIBRERIA
EDITRICE
VATICANA

The Pontifical Academy of Sciences (PAS)
Casina Pio IV, 00120 Vatican City
Tel: +39 0669883195 • Fax: +39 0669885218
Email: pas@pas.va • Website: www.pas.va

The opinions expressed with absolute freedom during the presentation of the papers of this meeting, although published by the PAS, represent only the points of view of the authors and not those of the Academy.

Cover design and illustration: Lorenzo Rumori
Photos and copy-editing: Gabriella Clare Marino

© 2026 – Amministrazione del Patrimonio della Sede Apostolica
© 2026 – Dicastero per la Comunicazione – Libreria Editrice Vaticana
00120 Città del Vaticano
Tel. 06.698.45780
E-mail: commerciale.lev@spc.va
www.libreriaeditricevaticana.va

ISBN: 978-88-266-1083-2
DOI: 10.65398/NCIH3943

Printed by Tipografia Vaticana



“Artificial intelligence is transforming many aspects of our daily lives, including education, entertainment, and the safety of minors. Its use raises important ethical questions, especially concerning the protection of the dignity and wellbeing of children and adolescents, who are particularly vulnerable to manipulation through AI algorithms that shape their decisions and preferences”.

Address of His Holiness Pope Leo XIV to the Participants in the Conference “The Dignity of Children and Adolescents in the Age of Artificial Intelligence”, Clementine Hall, Apostolic Palace, Vatican City, 13 November 2025.









Contents

Risks and Opportunities of AI for Children: Introduction and Concept

Joachim von Braun and Britta Holmberg	11
---	----

Introductory Comments

Cardinal Pietro Parolin	15
-------------------------------	----

Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children

H.M. Queen Silvia of Sweden	18
-----------------------------------	----

SESSION 1. AI AND ITS EFFECT ON CHILDREN'S SAFETY AND DIGNITY

The Beauty and the Threat of the Screen: The Impact of Digital Culture and AI on Children's Development of Language, Reading, and Writing

Maryanne Wolf	23
---------------------	----

The Global Landscape of Child Sexual Exploitation and the Role of AI in Addressing It

Mama Fatima Singhateh	28
-----------------------------	----

Youth Voices in AI Governance: Building a Safe Digital Future for Children

Imane Lakbachi	33
----------------------	----

SESSION 2. RESEARCH AND INNOVATION INSIGHTS: PRESENTATIONS AND A PANEL DISCUSSION

Why Children's and Adolescents' Brains Are Especially Vulnerable

Wolf Singer	43
-------------------	----

Listening to See: Uncovering Dilemmas with Youth and Generative AI

Emily Weinstein and Beck Tench	53
--------------------------------------	----

Should we Ban Social Media in Childhood and Adolescence?

Christian Montag	64
------------------------	----

AI-Generated Child Sexual Exploitation and Abuse Imagery: Research and Actions by Snap Inc.

Jacqueline Beauchere	76
----------------------------	----

Empowering and Safeguarding Youth in the AI Era: Integrating Developmental Neuroscience, Digital Well-Being, and Global Policy Frameworks	
Similoluwa Okunowo	87
 ■ SESSION 3. UNDERSTANDING THE RISKS AND OPPORTUNITIES OF AI FOR CHILDREN	
Safeguarding Children in the Age of AI: A Call to Proactive Action	
Baroness Joanna Shields	93
The Double-Edged Sword of AI: Insights from the Frontlines of Child Sexual Abuse Perpetration Prevention	
Christoffer Rahm	97
Bridging Worlds to Protect Children: Reflections on AI and Cross-Sectoral Collaboration	
Susanne Drakborg	100
AI, Africa, and its Youth: Opportunities and Challenges	
Daya Reddy	106
Panchatantra 2.0: A Culturally Grounded Approach for Children's Digital Rights and Online Safety	
Sagar Vishnoi, Parishrut Jassal & Alisha Butala	114
 ■ SESSION 4. ETHICAL AND RESPONSIBLE AI: ROLES AND RESPONSIBILITIES OF STAKEHOLDERS AND REGULATION	
Rethinking AI for Children: From Technological Potential to Human Responsibility	
Virginia Dignum	125
Protect, Provide, Participate: Children First in the Artificial Intelligence Era	
Josiah Kiarie Kimani	140
Ethical Resources for Helping Artificial Intelligence Benefit the Common Good	
Brian Patrick Green	145
Statement	151
List of Participants	163

RISKS AND OPPORTUNITIES OF AI FOR CHILDREN: INTRODUCTION AND CONCEPT

JOACHIM VON BRAUN

PAS President

BRITTA HOLMBERG

World Childhood Foundation

■ <https://www.doi.org/10.65398/FOGH2829>

The Pontifical Academy of Sciences, in collaboration with the World Childhood Foundation, chaired by Queen Silvia of Sweden, and the Institute of Anthropology of the Pontifical Gregorian University, convened a high-level workshop to address the opportunities and risks AI presents for children. The conference included political leaders, AI expert scientists, youth representatives, United Nations and international organizations, civil society groups, corporate leaders, survivors of AI abuse, and church officials.

The Impact of AI on the Safety and Dignity of Children

Artificial Intelligence (AI) is profoundly transforming the digital experiences of children, presenting both opportunities and risks. The surge in AI innovation brings hope for solving complex problems, enhancing children's digital safety and dignity. However, the rapid development of AI, exacerbated by the pandemic-induced surge in digital usage, has also exposed children to increased online risks such as deepfakes, cyberbullying and sexual exploitation.

While AI-driven tools for analyzing and managing text, images and large datasets can enhance online safety, they also raise significant concerns regarding misuse, privacy violations and algorithmic bias. Ensuring the intentional and ethical deployment of AI is essential to safeguarding children's well-being in an increasingly digital environment, including within social media ecosystems. AI is now deeply embedded in digital infrastructure and integral to the functioning of social platforms. However, AI technologies can be exploited to generate and disseminate child sexual abuse material, facilitate online grooming, and enable human trafficking.

The deployment of AI in surveillance and monitoring, while often intended to enhance safety, can produce unintended consequences. Excessive monitoring risks infringing upon children's fundamental right to privacy, fostering a culture of surveillance that undermines their sense of autonomy and dignity. Moreover, algorithmic bias within AI systems can perpetuate, and in some cases intensify, existing social and structural inequalities.

Despite these challenges, AI has the potential to significantly enhance the opportunities and safety of children in digital environments. One of these opportunities lies in AI's real-time detection and prevention of online threats. Advanced AI algorithms can monitor behavioural patterns across digital platforms, identify potentially harmful interactions, and remove inappropriate or illegal content.

It is imperative to address the associated risks through robust regulatory frameworks, ethical AI development practices, and sustained dialogue among stakeholders. Ensuring that AI technologies are designed and implemented with the best interests of children at their core is essential to harnessing their benefits while mitigating potential harms. As AI continues to evolve, a balanced approach that prioritizes both safety and dignity will be essential for safeguarding children's well-being in the digital age.

In the Global South, where access to traditional education and child protection services may be limited, AI offers the potential to deliver innovative solutions to increase the reach and quality of educational opportunities and essential.

The Conference Organizers

The conference organizers and their invited experts combined related scientific insights with long standing experiences in child protection.

The *Pontifical Academy of Sciences* (PAS) has held various conferences on AI and humanity over the last few years.¹ The goals of PAS include, "Stimulating an interdisciplinary approach to scientific knowledge; Encouraging international interaction; Furthering participation in the benefits of science and technology by the greatest number of people and peoples; Ensuring that science works to advance the human and moral dimension of people; Achieving a role for science which involves the promotion of justice, de-

¹ Power and Limits of Artificial Intelligence. Antonio Battro, Stanislas Dehaene (Eds.). Proceedings of the Workshop 2016, *Scripta Varia* 132, Vatican City, 2017 <https://www.pas.va/content/dam/casinapioiv/pas/pdf-volumi/scripta-varia/sv132pas.pdf>

Robotics, AI and Humanity: Science, Ethics and Policy. von Braun, J., Archer, M.S., Reichberg, G.M., Sánchez Sorondo, M. (Eds.) Springer Nature 2021, Open Access <https://link.springer.com/book/10.1007/978-3-030-54173-6>

Children and Sustainable Development. Ecological Education in a Globalized World, 2015. Springer 2017. https://www.pas.va/en/publications/scripta-varia/sv135_springer.html

Science for Sustainability and Wellbeing in the Anthropocene: Opportunities, Challenges, and AI. Plenary Session | 23-25 September 2024 <https://www.pas.va/en/events/plenary-session.html>

velopment, solidarity, peace, and the resolution of conflict; Fostering interaction between faith and reason and encouraging dialogue between science and spiritual, cultural, philosophical and religious values...”.

With 25 years of pioneering efforts, the *World Childhood Foundation*, founded and spearheaded by Her Majesty Queen Silvia of Sweden, has played a crucial role in preventing child sexual abuse and exploitation. Throughout its history, the Childhood Foundation has consistently championed projects that prioritize children’s safety and dignity, ranging from groundbreaking AI initiatives to innovative community-based programs.

The *Institute of Anthropology. Interdisciplinary Studies on Human Dignity and Care* (IADC) of the Pontifical Gregorian University conducts research and provides education on issues related to dignity and care through an intercultural and interdisciplinary approach, grounded in Christian anthropology. Established in 2012 as the Centre for Child Protection, the Institute contributes to the development of a common model for creating a safer world, and raises awareness of the urgent need to protect minors and vulnerable adults from harm.

Outcomes of the Workshop

The meeting emphasized the shared responsibility of all stakeholders to collaborate effectively within their respective spheres of influence to protect children. Its primary aim was to foster a common understanding of the risks and opportunities that AI presents for child safety and development, and to outline the responsibilities of each stakeholder in advancing a joint commitment to safeguarding children. The key outcomes of the conference include:

1. **Shared Understanding:** An enhanced collective awareness of the current state of AI development, including risks and opportunities related to the safety and dignity of children.
2. **Defined Responsibilities:** A clear delineation of the roles and responsibilities of science, the tech industry, religious communities (including the Catholic Church), researchers, civil society, children and young people, and survivors of abuse.
3. **Joint Commitment:** A joint statement outlining specific, actionable steps each stakeholder may consider to enhance child wellbeing through AI, expand educational opportunities for children and youth, and combat child sexual abuse and exploitation using AI, including regular monitoring and evaluation of progress.

The outcome statement (see below), drawing on insights shared during the workshop, emphasizes five key actions:

- Strengthening international governance and regulatory frameworks to protect children from AI-driven harm.
- Developing AI systems that prioritize children’s safety, privacy, and dignity.
- Promoting the benefits of AI while ensuring transparency, accountability, and fairness.
- Elevating youth voices in AI governance discussions, and in the design and development of AI systems and products.
- Fostering global collaboration to ensure that AI serves, rather than endangers, the intellectual potential, dignity and well-being of every child in the digital age.

AI must be developed responsibly, prioritizing children’s well-being. Therefore, ethical regulations must be further developed and rigorously enforced. In light of the rapid pace of AI innovation, the co-organizers reaffirm their commitment to keeping the theme of AI and child welfare high on their agendas.

INTRODUCTORY COMMENTS

CARDINAL PIETRO PAROLIN

Secretary of State

■ <https://www.doi.org/10.65398/WIFJ5793>

Your Majesty, Queen Silvia of Sweden,
Your Eminence, Card. Turkson,
Your Royal Highness, Princess Madeleine,
Your Excellencies, Members of the Diplomatic Corps,
Distinguished Professor Joachim von Braun
and Members of the Pontifical Academy of Sciences,
Ladies and gentlemen,
Dear friends,

First, I would like to express my cordial greetings to all the participants in this Conference dedicated to the question of Artificial Intelligence and the risks and opportunities that it raises for children in particular. I welcome this meeting and would like to thank the Pontifical Academy of Sciences, its Chancellor, Cardinal Peter Turkson, and its President, Professor Joachim von Braun, for convening it and for gathering experts in child safeguarding from different backgrounds, as well as leaders from tech firms and other industries, members of civil society and of faith-based organisations, philanthropists, youth and indeed survivors of Artificial Intelligence abuse, to explore together this important and challenging issue.

In his address to the G7, last June, Pope Francis insisted that: “*powerful technological progress [...] makes artificial intelligence at the same time an exciting and fearsome tool, and demands a reflection that is up to the challenge it presents*”.¹

Artificial Intelligence systems are extraordinarily fast in their growth and development, and are about to profoundly reshape our societies, and childhood in particular. In the midst of such transformations, it is our responsibility to make sure that this new technology is put to work for the good and safety of each person, especially the most vulnerable, while preserving human dignity and social relationships. As we are on the eve of a

¹ Pope Francis, Address at the G7 Session on Artificial Intelligence. Borgo Egnazia, Puglia, 14 June 2024.

new era, it is worth recalling the Second Vatican Council's pastoral Constitution on the Church in the modern world, which qualifies the norm of human activity as follows: "*that in accord with the divine plan and will, it harmonize with the genuine good of the human race, and that it allow men as individuals and as members of society to pursue their total vocation and fulfil it*".²

Along those lines, the recent Note *Antiqua et Nova*, published by the Dicastery for the Doctrine of the Faith and the Dicastery for Culture and Education, on the relationship between human intelligence and artificial intelligence, reiterates that "*Artificial Intelligence, like any technology, can be part of a conscious and responsible answer to humanity's vocation to the good. However [...], AI must be directed by human intelligence to align with this vocation, ensuring it respects the dignity of the human person*".³

Artificial Intelligence has the potential to enhance children's safety, education, and well-being. At the same time, it may also introduce serious threats, including cyberbullying, privacy violations, Artificial Intelligence-driven addiction, and online exploitation. Preventing risks while seeking to harness the benefits of it is not an easy task. This is not only due to the complexity of the field, but because, as Pope Francis underlined in his message to the 2025 World Economic Forum, "*artificial intelligence raises fundamental questions about ethical responsibility, human safety, and the broader implications of these developments for society*". It requires, therefore, seeking appropriate responses according to specific contexts and to "*the principle of subsidiarity, with individual users, families, civil society, corporations, institutions, governments, and international organizations working at their proper levels to ensure that Artificial Intelligence is directed to the good of all*".⁴

It is critical that governments, tech companies, educators, civil society, and faith-based institutions work together to reflect upon ethical regulations and governance frameworks, as well as data transparency and child-centred policies. Not only is it essential to guarantee children's safety, privacy and respect for their dignity but also to preserve them from Artificial Intelligence-driven harm. It is equally essential to ensure transparency, accountability and fairness in order to make Artificial Intelligence more beneficial for every child.

² *Gaudium et Spes*, n. 35.

³ Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, Note on the Relationship between Artificial Intelligence and Human Intelligence, 14 January 2025, n. 48.

⁴ Pope Francis, Message to the World Economic Forum, 14 January 2025.

Artificial Intelligence's rapid evolution demands urgent and collective action, as well as global collaboration to ensure that this new technology serves, rather than endangers, the dignity and wellbeing of every child in the digital age.

I thank Your Majesty, President von Braun, the Academy and all of you, dear participants, for your contribution and wish you a fruitful meeting.

I thank you all for your kind attention.

RISKS AND OPPORTUNITIES OF AI FOR CHILDREN: A COMMON COMMITMENT FOR SAFEGUARDING CHILDREN

H.M. QUEEN SILVIA OF SWEDEN

■ <https://www.doi.org/10.65398/WOOT4066>

Excellencies,

Ladies and gentlemen,

Childhood. Listen to this word. I hope it stirs up thoughts and memories of play, safety, and care, because that is what childhood should be. A time for safe exploration, growth, learning and development.

The period from birth to the age of 18 is one of the most special phases of human life. So special, in fact, that it is recognised and protected by the most widely ratified international treaty in the world: the Convention on the Rights of the Child.

Every child has rights. And as adults, we bear the responsibility to protect these rights.

Yet as we know, we continue to fail. Too many children are deprived of their fundamental rights to survival, development, protection, and participation. It is unacceptable that, to this day, children are denied access to basic health care, education and the security of loving, responsible adults.

That is also why I founded the World Childhood Foundation 25 years ago, with the mission to combat child sexual abuse – one of the most horrific crimes against children, and one that occurs on an unimaginable scale. Recent global prevalence data shows that one in five girls and one in seven boys experience some form of sexual violence during their childhood.

Child sexual abuse remains a crime surrounded by stigma, guilt and shame. But let us be clear: child sexual abuse is never the fault of the child. As adults and institutions, it is our duty to break the silence, listen actively to what children are telling us, and shift the blame where it belongs – with the perpetrators. Brave survivors are leading the way. Some are here with us today. We stand with you.

Child sexual abuse occurs in every society, every social class, every culture and every religion across the globe. However, some children are more vulnerable than others. Marginalised, such as children with disabilities, children in institutions, and children identifying as lesbian, gay, bisexual, trans-

gender, queer or intersexual (LGBTQI) face a significantly higher risk of abuse. We are here today for all the children who are deprived of their rights.

From the very beginning, the World Childhood Foundation has embraced technology, including artificial intelligence, as a tool for preventing child abuse. It serves as a means to reach and empower vulnerable children and a way to detect and stop perpetrators. But it is a double-edged sword. Technology has also been exploited by abusers to reach children on a massive scale, luring them, exploiting them and spreading abusive material. The accountability of tech providers has been insufficient. Artificial intelligence, undoubtedly, presents new challenges, offering perpetrators new ways to harm children.

However, artificial intelligence can also help us catch up, offering groundbreaking solutions to combat the problem. Technology is not inherently good or evil. We will need AI to fight AI.

The need for action is urgent. This technological revolution is unfolding at lightning speed. We stand on the edge of a paradigm shift that is reshaping the lives and childhoods of future generations. Yet none of us can fully grasp the true implications of generative AI for children growing up today. Now is the time to identify the risks and contribute to solutions to ensure artificial intelligence is used to benefit and protect children, not those who seek to harm them.

Finally, let us recognise that this is not about technology alone. It requires engagement from every field. This is a matter of child rights, science, business, ethics and policy. That is why I am delighted to see such diverse perspectives and expertise in this room. To succeed, we must listen and learn from one another – and act – together.

Thank you!

 SESSION 1.

**AI AND ITS EFFECT ON CHILDREN'S
SAFETY AND DIGNITY**

THE BEAUTY AND THE THREAT OF THE SCREEN: THE IMPACT OF DIGITAL CULTURE AND AI ON CHILDREN'S DEVELOPMENT OF LANGUAGE, READING, AND WRITING

MARYANNE WOLF

University of California, Los Angeles

■ <https://www.doi.org/10.65398/SUHL4822>

“Word-work is sublime...because it is generative, it makes meaning that secures our difference, our human difference. We die. That may be the meaning of life. But we do language. That may be the measure of our lives.”

TONI MORRISON, NOBEL LECTURE, 1993

“Technologies are not mere exterior aids, but also interior transformations of consciousness, and never more than *when they affect the word*.”

WALTER ONG, 2002

The development of oral and written language, by the species and by the child, represents a quintessential achievement in human development. These two inextricably connected aspects of language, one determined by nature and one by nurture, are both potentially propelled and potentially threatened by technological innovation. As described years ago by MIT media scholar Turkle (2011), we do not error because we innovate, we error because we do not ask what the innovation disrupts or diminishes in human lives. Within the overarching context of safeguarding children's development in a digital milieu, this chapter will describe the specific importance of understanding how language, reading, and writing development in childhood are impacted by digital mediums, particularly those reliant on the screen. It is a matter of ultimate concern for our future generations.

To understand how language, particularly written language, can be simultaneously propelled and/or disrupted requires a brief description of reading's development. Based on research in the cognitive neurosciences, this essay begins with the premise that literacy changes the human brain (see Dehaene, Cohen, Morais & Kolinsky, 2015; Wolf, 2018), which changes the individual's life trajectory, and ultimately may alter the future of our species. The second, fundamental premise is that the human brain was never born to read (Wolf, 2007). A semi-miraculous design principle in the human

brain allows it to create new circuits for new behaviors that were never genetically endowed. More specifically, to accommodate human inventions like literacy and mathematics, the brain creates new, plastic circuits by connecting older cognitive and linguistic structures (like vision, language, and conceptual background knowledge). The intrinsic plasticity of this circuit underlies our ability to learn all the very different writing systems in our world. It also means that the circuit will reflect its environmental influences: e.g., how it is taught and what medium is used to read. As the young brain becomes more literate, the first, basic reading circuit becomes elaborated as it adds ever more cognitively complex processes. For this to occur, each of the reading brain's circuit parts needs to become virtually automatic so as to require little attention. In this way, young readers gain additional milliseconds of time to allocate attention to the more sophisticated and more time-consuming cognitive, linguistic, and affective processes that comprise expert, "deep reading."

The key to understanding the potential impact of digital mediums on reading development lies within the allocation of these milliseconds for the deep reading processes. Briefly, these processes include among others: 1) the analogical reasoning ability to couple the reader's background knowledge with new information; 2) inferential/deductive reasoning ; 3) empathy and perspective-taking that enables "passing over" (Dunne, 2006) into the thoughts, feelings, perspectives of others; 4) critical analysis of this information; and 5) insights from contemplation/reflection, described by Thomas Aquinas as the heart of reading (Wolf, 2018). Each of these processes contributes to proficient or expert reading and writing development, as well as to the young readers' intellectual, social-emotional, and ethical development.

The elaboration of the developing reading circuit, however, requires time – both in milliseconds (to allocate attention to deep reading processes) and in years to develop the expert reading brain. Here lies the cerebral rub. The same circuitry that underlies deep reading processes can be *short-circuited* if the reader does not give sufficient time to the formation and/or use of these processes. Extensive research, particularly in eye movement studies, demonstrates that reading on digital screens advantages the quick skimming of information, not the more time-consuming allocation of milliseconds to deep reading processes. As described by UCLA psychologist, Greenfield (2009), "Every medium has its costs and weaknesses... the cost (in the digital medium) seems to be deep processing."

Put more baldly, the digital medium promotes surface or shallow reading, and disadvantages deep reading processes. In an age where we are bombarded with information, skimming might be considered a defense mechanism in the already formed expert reading brain. In the young reading brain circuit, however, it can be an unintended impediment to the development of human thought processes essential for a society. Processes like analogical reasoning, critical analysis, empathy, and reflection provide the foundation for the future growth and development of our citizenry and indeed for democracy itself. They are developed over time through the reciprocal relationships among oral language, reading, and writing, and they can be disrupted, diminished, and fail to develop if they are not nurtured.

We need look no further than the state of literacy for children in the United States. The National Assessment of Educational Progress (NAEP), described as the country's grade card, provides data on the development of reading and math. The most recent NAEP report indicates that only one-third of Grades 4 and 8 students reach the *proficiency* level, while a majority of children of color never reach *basic* levels of reading, much less proficiency.

Although multiple causes account for these results, particularly COVID and inappropriate teaching methods, the amount of digital exposure by children and youth is increasingly cited. Current research on the connection between digital exposure and learning underscores the impact of digital devices on child development in different developmental periods. For example, in the *Journal of the American Medical Association*, Takahashi and colleagues (2023) showed a significant relationship between screen time in one-year-old children and later problems in cognitive and psychosocial processes. Similarly, a larger study in Singapore linked greater, early digital exposure in young children to poorer development of executive processes and lower academic performance in eight-year-old children. A still larger meta-analysis of 50 studies with over 171,000 young adults indicated that when reading the same material on screen or in print, subjects were better at comprehending the material in print (Delgado, Salmeron, et al., 2017;2022). Furthermore, those subjects deemed "digital natives" were even better in print, even though they perceived themselves better on the screen. When probed, they thought they read better on the screen because they were faster. Herein lies the false assumption held by many. Speed is neither illumination nor insight. Indeed, it may detract from both, because the most reflective cognitive capacities take time. When skimming is the *modus operandi* in reading, there is insufficient attention allocated to deep

reading processes, leading to insufficient comprehension and diminished building of the deep reading processes important to child development. The skimming that has become our daily norm gives no one time to think.

Correlational studies do not prove that digital devices caused the findings; neither are they neutral. Rather, together with research from other disciplines, they dictate caution and more rigorous study. As philosopher Ong (2002), the student of Marshall McLuhan advised, digital mediums change the interior processing of words and text. As psychologist Steiner-Adair (2012) early cautioned, digital mediums change the social and psychological well-being of our youth. If we are to ensure that our young are not susceptible to misinformation, disinformation, and to messages from insidious perpetrators, we must safeguard the development of their inferential, analogical, empathic, and critically analytical skills.

This is not a call for binary approaches in which various mediums are conceptualized in opposition. We are enriched by many technological innovations. At the same time, we would be foolhardy to have our young bypass or skip the arduous efforts in acquiring reading and writing skills that build intellectual development. Our society's rapacious desire for efficiency through innovations like ChatGPT and the next "effort-saving" solution should never take the place of *all the very efforts that build the circuits* for our most cherished cognitive, linguistic, and affective processes. We must preserve these processes as we expand other, new processes that are developed with new tools.

That said, my continuing worries about screen reading are far less important than using everything at our disposal to help our children (and ourselves) sustain the desire to learn, read, and be transported by language whatever the medium. Nurturing the *reading life in our children may be our best antidote* to the threats that will persist. We can do this by helping our children learn both the love of being immersed in the reading experience and the importance of thinking more deeply when they read. We must teach our children and ourselves to determine the *purpose* of whatever we are reading. Deep reading processes are rarely needed for email, but they are imperative for important forms of text from contracts and referenda to great literature, all of which reward a quality of attention that goes missing on the screens of our distraction.

Towards these ends, I have proposed what I call the "biliterate brain" as a way to bring about a developmental sequence in learning that allows the parallel development of both deep reading through print and of essential

digital skills like coding and programming through digital media (Wolf, 2018). Throughout this biliteracy development it is both essential and beneficial to us all to give our children lifelong associations: between books and love, as they sit under the crook of our arms while we read to them; and between reading ever more demanding texts and their development of their own generative, *reading sanctuary*, captured for all time by Proust in childhood reveries about reading a book on a summer's day.

We must all be vigilant about preserving the deep reading processes which propel empathy, inferential thinking, and critical analysis, whatever the medium, whether in ourselves and/or future generations. Deep literacy is the foundation for building the reading life, that elusive entity which can transport us at any age to that place where we can think our best thoughts, feel the joys and suffering of others, and know we are not alone. As Toni Morrison wrote, language “may be the measure of our lives”. The reading sanctuary is where we learn to measure it.

References

- Dehaene, S., Cohen, L., Morais, J. & Kolinsky, R. (2015). Illiterate to literate: Behavioral and cerebral changes induced by reading acquisition. *Nature Reviews*, 16, 234-244.
- Delgado, P., Vargas, Ackerman, R., & Salmeron, L. (2018). “Don’t throw away your printed books”: A meta-analysis of the effects of reading media on comprehension. *Educational Research Review* 25, 23-38.
- Dunne, J.S. (2006). *A vision quest*. Notre Dame, IN: University of Notre Dame Press.
- Greenfield, P. (2009). Technology and informal education. *Science*, 323, p. 71.
- Morrison, T. (1993). Nobel Prize for Literature Lecture.
- National Assessment of Educational Progress (2025). Washington D.C.: Department of Education.
- Ong, W. (2002). *Orality and Literacy: The technologizing of the word*: 2nd Ed. New York: Routledge.
- Steiner-Adair, C. (2012). *The Big Disconnect*. New York: HarperCollins.
- Takahashi, I. et al. (2023). Screen time at age 1 and communication and problem-solving developmental delay at 2 and 4 years. *JAMA Pediatrics*, 177: 1039-1046.
- Turkle, S. (2011). *Alone together*. New York: Basic Books.
- Wolf, M. (2007). *Proust and the Squid: The story and science of the reading brain*. New York: Harper Collins.
- Wolf, M. (2018) *Reader, Come Home: The reading brain in a digital world*. New York: HarperCollins.

THE GLOBAL LANDSCAPE OF CHILD SEXUAL EXPLOITATION AND THE ROLE OF AI IN ADDRESSING IT

MAMA FATIMA SINGHATEH

United Nations Special Rapporteur on the Sale,
Sexual Exploitation and Sexual Abuse of Children

■ <https://www.doi.org/10.65398/MZTN9901>

Last year I presented my annual report to the Third Committee of the United Nations General Assembly on the *Existing and emerging sexually exploitative practices against children in the digital environment*,¹ and in that report I highlighted how the misuse of existing and emerging technologies exacerbates and amplifies children's exposure to risks, harms, and various forms of sexual exploitation and sexual abuse. The issues raised herein are derived from said report unless otherwise indicated.

We know that one in three internet users worldwide is estimated to be a child, with a child going online for the first time every half a second.² Children today are spending more time in the digital environment than ever before. Almost 80 per cent of children and young people aged 15 to 24 are driving the force of connectivity globally, compared with 65 per cent of the rest of the population.³ In this context, numerous research and reports have revealed the sustained threat and intensification of manifestations of child sexual abuse and exploitation in the digital environment, both in terms of scale and method.⁴

One must acknowledge that information and communication technologies offer a wide array of opportunities to protect and uphold children's rights, yet their rapid, evolving, and unprecedented capabilities present significant risks to children and expose them to harms caused by the misuse of technologies. This is supported by research which showed that more than 300 million children a year are estimated to be victims of online sexual

¹ Available at <https://docs.un.org/en/A/79/122>

² International Telecommunication Union (ITU), *Guidelines for Industry on Child Online Protection* (Geneva, 2020).

³ ITU, *Measuring Digital Development: Facts and Figures 2023* (Geneva, 2023).

⁴ See A/79/122, available at <https://docs.un.org/en/A/79/122>

abuse and exploitation.⁵ The research also revealed that 12.6 per cent of the world's children have been victims of non-consensual communication as well as non-consensual sharing of and exposure to sexual images and video while another 12.5 per cent have been subjected to online solicitation.⁶

We see every day how existing technology is misused to sexually harass, take and share non-consensual images and videos, generate child sexual abuse materials, sexually extort children, and livestream child sexual abuse. This is amplified through developments in Artificial Intelligence ('AI') which not only replicates existing safety concerns that plague the current digital ecosystem and technological landscape but also exacerbates and facilitates more severe threats against children as well as gives rise to new trends. The misuse of AI has not only significantly facilitated the multiplication of child sexual abuse materials but also driven the emergence of entirely new forms of large-scale abuse facilitated by technology such as grooming, sexual extortion of children and high-definition live-streaming of abuse.

We see how Generative AI, through deepfakes and synthetic images, is increasingly being weaponized to create realistic child sexual abuse content with little to no interaction with real children. These AI models generate materials using image prompts, manipulate real images, or create entirely artificial content that appears indistinguishable from real-life abuse. Perpetrators adapt and customize original child sexual abuse materials into new content, produce and manipulate accessible and benign content of known or unknown children into sexually abusive material, or even create fully AI-generated child sexual abuse materials from scratch. These materials are easily accessible and can be downloaded offline.

This technology enables perpetrators to expand their methods of exploitation, making detection and law enforcement efforts more difficult, while also facilitating anonymity and the evasion of oversight. It is also increasingly difficult for law enforcement to triage, distinguish and identify victims, and identify whether or not there is a real child in danger. This results to situations where investigators spend time and resources pursuing the rescue of children who may turn out to be virtual characters, or waiting for further verification of the quality of the content generated, thereby diverting much needed resources and time away from the rescue of real children.

⁵ Childlight – Global Child Safety Institute, *Into the Light Index of Child Sexual Exploitation and Abuse Globally: 2024 Report* (Edinburgh, 2024).

⁶ *Ibid.*

Despite the forgoing, we have increasingly seen how AI is leveraged to address the risks posed by AI itself and other technologies in the online sexual exploitation of children. Through image matching technologies and advanced algorithms, AI is used to analyse images, videos, and text to detect explicit or exploitative content involving children, content which traditional detection tools are usually unable to detect. It is also used to identify known child sexual abuse materials and flag new content for review. This can dramatically speed up the identification and removal of child sexual abuse materials as well as reduce human moderators from being exposed to sustained extreme content which can have lasting health and wellbeing implications.⁷ Through AI, online platforms can be monitored for suspicious behaviour patterns, such as grooming tactics, inappropriate interactions, or the sharing of sexually exploitative content. It can also be used to assist in identifying victims and perpetrators through the analysis of metadata, facial recognition, and other digital footprints in child sexual abuse materials.

The synthetic nature of AI-generated child sexual abuse materials, however, complicates easy detection, as traditional tools that rely on hashing, like PhotoDNA, to identify known child sexual abuse materials may not recognize new AI-generated materials.⁸ Nevertheless, through capacity building and resource allocation, law enforcement and platforms can adapt their detection systems to identify and block synthetic content, as well as through human expertise ensure that AI systems are accurate to reduce false positives and wrongful accusations or censorship.

Additionally, the sheer scale and complexity of this digital landscape underscore the urgent need for stricter and more robust legal and regulatory frameworks. Laws must be updated to address the creation, distribution and possession of AI-generated child sexual abuse materials, ensuring that both real and synthetic abuse material are criminalized equally under the law.

Collaboration amongst technology companies, law enforcement, and governments is essential to combat child sexual exploitation online, track perpetrators and remove harmful content quickly, especially as such offences are often cross-border in nature.

It is also imperative that the rights of children be mainstreamed within business models, rather than retrofitting safeguards after harm has occurred. Privacy- and safety-by-design approaches would therefore ensure that the

⁷ See <https://www.weprotect.org/thematic/artificial-intelligence-and-gen-ai/>

⁸ *Ibid.*

security and integrity of technological infrastructure are not compromised. Furthermore, strengthening the ethics of due process safeguards, transparency and accountability of the industry makes it possible to not only pinpoint where weaknesses exist in safety practices; identify where discrepancies exist in respect of the extent of their action, performance and presence in jurisdictions; but also determine what innovative and preventive technologies are needed and where law enforcement is required.⁹

Instead of passively relying on warrants, notices and takedown procedures, the burden to report abuse should be placed squarely on technology companies and online service providers. Businesses should be mandated to put in place or improve on systems that proactively and rapidly implement concrete automated and human detection, removal and moderation tools, and to promptly report to the relevant authorities any harmful, exploitative and abusive content of children occurring or amplified through their products and services, including by identifying real-time abuse, and accounts used by perpetrators.

When designing and engineering technologies, greater choices over content moderation and user-controlled moderation settings should be provided. This should include robust age verification and restriction; parental controls; regular reminders and notifications of dangers to users; restricting interactions and making private the follower lists of children; and implementing warning tools and deterrence campaigns to suspected perpetrators.

The challenges posed by the borderless nature of digital crimes can hinder investigative and judicial processes. It would therefore be important to create and strengthening specialized multidisciplinary units with adequate resources both human and financial for effective investigations, as well encourage international cooperation among regulatory bodies, law enforcement, and industry stakeholders.

I want to conclude by emphasising that the evolution of technology and developments in AI come with many benefits and include the protection of children in the digital environment, yet the rapid technological development and change also present opportunities for organized criminal groups and perpetrators to adapt their approach to target their victims.

This raises the question about the effectiveness of current statutory and enforcement regimes, especially given the ease with which criminals can reach children, and child sexual abuse materials can be found in the digi-

⁹ A/HRC/52/61, available at <https://docs.un.org/en/A/HRC/52/61>

tal environment. Without immediate action, effective regulation, product safety, effective content moderation, education, and enhancing the work of law enforcement this phenomenon will only grow.

Children have the right to feel safe in all spaces, wherever they are. A concerted effort is therefore required from all stakeholders to respond to this problem, and technology companies have an important role to play in this regard.

YOUTH VOICES IN AI GOVERNANCE: BUILDING A SAFE DIGITAL FUTURE FOR CHILDREN

IMANE LAKBACHI¹

Digital Transformations for Health Lab²

■ <https://www.doi.org/10.65398/LSIN1835>

This youth statement was prepared in advance of the high-level meeting, *Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children*,³ co-organized by the Pontifical Academy of Sciences,⁴ the World Childhood Foundation,⁵ and the Institute of Anthropology of the Pontifical Gregorian University.⁶

Introduction

Artificial Intelligence (AI) is changing many aspects of our lives, especially how children and young people interact with the digital world. AI has the potential to bring a lot of positive changes – it can improve access to education, help make online spaces safer, and support the overall well-being of children and youth. But at the same time, it can also present serious risks. Issues like data privacy, biased algorithms, and online safety threats can negatively affect the rights, dignity, and future of young people.

Children (under 18 years) and young people (under 30 years) are not just passive users of technology – they are digital natives whose daily lives,

¹ Imane Lakbachi is a youth and gender equality advocate from Morocco; with a strong background in Computer Science and Multimedia, Imane currently serves as the Focal Point of the Digital Transformations for Health Lab (DTH-Lab) to the WHO Youth Council (<https://www.who.int/initiatives/who-youth-engagement/who-youth-council>) and holds the position of Director of Network Engagement at the International Youth Alliance for Family Planning (IYAAP).

² DTH-Lab (<https://dthlab.org/>) is a global consortium of partners working to drive the implementation of The Lancet and Financial Times Commission on Governing Health Futures 2030's recommendations for value-based digital transformations for health, co-created with young people. DTH-Lab operates through a distributive governance model, led by three core partners: Ashoka University (India), DTH-Lab (hosted by the University of Geneva, Switzerland), and PharmAccess (Nigeria).

³ https://www.pas.va/en/events/2025/ai_children.html

⁴ <https://www.pas.va>

⁵ <https://childhood.org/>

⁶ <https://www.unigre.it/en/anthropology/>

education, health, and social interactions are increasingly influenced by AI-driven systems. Yet, policies shaping these technologies are often developed without their input. This statement aims to highlight the main risks and challenges AI poses to children and young people but also focus on the unique opportunities AI can bring to improve those same lives. This is a call to action for all key actors within the digital and AI space to prioritize the best interests of children and youth to create an inclusive digital future.

Key Risks & Challenges

While AI can offer promising avenues for enhancing education, as well as health and well-being, it can present significant risks to children's and young people's safety. It is important for stakeholders to understand the risks that AI can pose in order to effectively build a secure and enriching digital environment for the younger generation. Some of the concerns associated with AI in children's and young people's digital lives include:

- **Online Safety Threats:** AI can be used to create harmful content. One example is AI-generated child sexual abuse material. According to the Internet Watch Foundation (IWF),⁷ there were 245 reports of AI-generated child sexual abuse imagery in 2024, compared with only 51 in 2023 – representing a 380% rise. 193 of these reports involved imagery that was so realistic it had to be treated the same as real photographic child sexual abuse material.
- **Data Privacy Concerns:** A 2024 survey by PR Newswire⁸ revealed that 46% of parents are apprehensive about their children sharing personal data online. In addition, the use of AI-powered surveillance software in schools, intended to monitor students' online activities for safety purposes, has led to unintended exposures of sensitive information. For instance, an investigation by AP News⁹ uncovered that nearly 3,500 unredacted student documents were inadvertently accessible, highlighting substantial cybersecurity risks associated with such monitoring practices.

⁷<https://www.iwf.org.uk/news-media/news/new-ai-child-sexual-abuse-laws-announced-following-iwf-campaign/>

⁸ <https://www.prnewswire.com/news-releases/parents-cautiously-optimistic-on-ai-in-schools-content-safety-and-data-privacy-among-top-worries-302227077.html>

⁹<https://apnews.com/article/ai-school-chromebook-gaggle-goguardian-security-25a3946727397951fd42324139aaf70f>

- **Endless Scrolling and Personalized Notifications:** Social media features like endless scrolling and personalized notifications are designed to maximize user engagement but pose significant threats to the mental health and brain development of children and adolescents. A study from Oxford University¹⁰ found a direct relationship between time spent on social media and mental health issues, with some teenagers using these platforms for up to eight hours daily.
- **AI-Driven Online Grooming:** AI can be misused to impersonate peers or trusted individuals, facilitating online grooming and exploitation of children. The Child Rescue Coalition¹¹ highlights that AI's ability to create realistic personas can be exploited by predators to gain the trust of young users, leading to potential abuse. This misuse of AI technology emphasizes the necessity for vigilant monitoring and education to protect children from such sophisticated threats.
- **Algorithmic Amplification of Harmful Content:** AI algorithms can also promote harmful content to children by optimizing for engagement without adequate safeguards. For example, TikTok's 'For You' feed has been found to risk pushing children and young people towards harmful mental health content. Amnesty International¹² highlights the need for stricter regulations and responsible algorithm design to prevent the dissemination of content that could negatively impact children's well-being.

Addressing these risks requires a collaborative effort among policymakers, tech industry, teachers, and parents to establish ethical guidelines and protective measures in the development, deployment, and governance of AI technologies affecting children.

Opportunities for Positive AI Integration

While AI presents notable challenges, when thoughtfully designed and used with meaningful safeguards, it holds significant potential to enhance the digital safety and well-being of children and youth.

¹⁰ <https://www.thetimes.com/uk/healthcare/article/social-media-is-linked-to-anxiety-in-teenagers-say-researchers-0dczvtgtk?region=global>

¹¹ <https://www.thetimes.com/uk/healthcare/article/social-media-is-linked-to-anxiety-in-teenagers-say-researchers-0dczvtgtk?region=global>

¹² <https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/>

1. **Access to Education in Low-Resource Settings:** AI platforms can provide virtual tutors, language translation, and educational content to children in remote or underserved areas, bridging gaps in educational access. For example, platforms like Wordly¹³ offer live audio translations and captions in multiple languages, facilitating learning for students in diverse linguistic settings.
2. **Personalized and Inclusive Learning:** AI-powered educational tools can customize learning materials to match each child's pace, strengths, and areas for improvement. These technologies can also offer vital support for children with disabilities, enhancing their learning and communication abilities. For instance, innovations such as eye-tracking systems and specialized keyboards offer alternative ways for learners¹⁴ with disabilities to engage with educational content, fostering greater inclusivity and accessibility.
3. **Health Information & Self-Care Guidance:** AI health apps¹⁵ can provide young people with reliable, age-appropriate health information, empowering them to make informed decisions about their physical and mental well-being.
4. **Enhanced Creativity & Innovation:** AI-powered creative tools (e.g., music generators, design platforms, storytelling apps) allow children and adolescents to experiment, create, and innovate easily.

Recognizing the transformative potential of AI to positively shape young people's lives, DTH-Lab's goal is to ensure that these technologies are designed and implemented in ways that truly serve the needs and rights of children and youth. The Lab understands that meaningful youth engagement is key to realizing the benefits of AI, especially in advancing health equity, education, and well-being. That's why DTH-Lab actively partners with young people, empowering them not just as beneficiaries, but as co-creators and decision-makers in shaping ethical, inclusive, and rights-based digital health futures.

By integrating young people with policymakers, technology companies, and other stakeholders, the Lab drives change in four key areas:

¹³ <https://www.wordly.ai/education-translation>

¹⁴ <https://www.jetlearn.com/blog/how-ai-can-help-learning-disabilities>

¹⁵ <https://www.weforum.org/stories/2024/09/ai-diagnostics-health-outcomes/>

1. **Digital First Health Systems:** Co-creating health systems that prioritize digital solutions, ensuring they are designed with and for young people.
2. **Digital and Data Governance:** Advancing value-based governance models to oversee the digital transformation of health, emphasizing transparency and accountability.
3. **Digital Determinants of Health:** Addressing factors within the digital environment that impact health and well-being, such as digital literacy and online safety.
4. **Digital Citizenship for Health:** Empowering young people to meaningfully engage in digital, health, and civic spaces

Through initiatives like the Regional Youth Champions¹⁶ cohort, the #MyHealthFutures Youth Network,¹⁷ and the Research Fellowship,¹⁸ DTH-Lab ensures that youth are not only beneficiaries but also active partners in co-creating health futures.

Call to Action

To ensure that AI is developed and implemented with children's and young people's best interests at heart, we call on both the public and private sectors to take immediate action. Specifically, we recommend prioritizing and investing in the following areas:

- Addressing digital illiteracy to ensure children and young people can use AI safely and effectively.
- Making digital solutions, including AI tools, more accessible to marginalized and underserved communities.
- Empowering young innovators and entrepreneurs to develop child-friendly AI solutions.
- Preserving human interaction and support within AI-driven services, especially in education and health.
- Developing inclusive digital strategies that protect and promote children's rights.

¹⁶ <https://dthlab.org/regional-youth-champions/>

¹⁷ <https://dthlab.org/myhealthfutures/>

¹⁸ <https://dthlab.org/research-fellows/>

- Promoting a culture of continuous innovation coupled with sustainable practices while ensuring ethical AI practices.
- Investing in core infrastructure and addressing the digital determinants of health.
- Empowering educators and the child protection workforce to use AI responsibly.
- Protecting children’s trust by strengthening data privacy and security measures.
- Engaging children, families, and communities through a whole-of-society approach in the design and governance of AI systems.

These recommendations represent just a few of the immediate steps needed to ensure AI technologies are designed and implemented with children’s and young people’s best interests at heart. For a more comprehensive set of recommendations, particularly tailored for governments and policymakers working on digital first health systems, we encourage you to read the Regional Youth Champions (2023-24 cohort) policy brief.¹⁹

Conclusion

Designing AI technologies that safeguard and empower young people requires multisectoral collaboration. The DTH-Lab is dedicated to leading these efforts and calls on all stakeholders to join in shaping a digital future that prioritizes children’s rights and well-being. To access additional resources and explore collaboration opportunities, please visit the DTH-Lab’s website.²⁰

¹⁹ <https://dthlab.org/embracing-digital-first-health-systems/>

²⁰ <https://dthlab.org/>

References

- Internet Watch Foundation. (2024). *IWF Annual Report 2024*. Retrieved from <https://www.iwf.org.uk/report/2024-annual-report%E2%80%8B>
- PR Newswire. Parents Cautiously Optimistic on AI in Schools: Content Safety and Data Privacy Among Top Worries <https://www.prnewswire.com/news-releases/parents-cautiously-optimistic-on-ai-in-schools-content-safety-and-data-privacy-among-top-worries-302227077.html>
- The Times. The link between heavy social media use and teenage anxiety <https://www.thetimes.com/uk/healthcare/article/social-media-is-linked-to-anxiety-in-teenagers-say-researchers-0dczvt-gtk?region=global>
- Lurye, S., & Bryan, C. (2025, March 12). *Schools use AI to monitor kids, hoping to prevent violence. Our investigation found security risks*. Associated Press. Retrieved from <https://apnews.com/article/25a3946727397951fd42324139aaf70f%E2%80%8B>
- Child Rescue Coalition. (n.d.). *The Dark Side of AI: Risks to Children*. Retrieved from <https://childrescuecoalition.org/educations/the-dark-side-of-ai-risks-to-children/>
- Amnesty International. Driven Into the Darkness <https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/>

 SESSION 2.

**RESEARCH AND INNOVATION INSIGHTS:
PRESENTATIONS AND A PANEL DISCUSSION**

WHY CHILDREN'S AND ADOLESCENTS' BRAINS ARE ESPECIALLY VULNERABLE

WOLF SINGER

Ernst Struengmann Institute for Neuroscience in Cooperation
with Max Planck Society, Frankfurt, Germany

■ <https://www.doi.org/10.65398/NTYG4892>

1. Introduction

Human beings are born with extraordinary potential, but realizing that potential depends not only on biology but also on its embedding in the material and cultural world. From the very beginning of life, the brain is shaped by a continuous interplay between genetic instructions and the surrounding environment. While much of the body matures in predictable ways, the brain is unique in that it develops itself through learning. It records experiences not in a notebook, but in its very structure – the intricate, ever-changing web of connections among its billions of neurons.

This process of shaping and reshaping the brain's structure is particularly pronounced during development, which in humans extends until about age 25. This epigenetic shaping of brain architecture allows us to adapt to the environment's specific demands, learn languages, form relationships, and master skills. However, this plasticity also comes with a price: the brain is not only open to enrichment but also vulnerable to harm and deprivation. The more plastic a system, the more its eventual shape depends on the conditions under which it develops. In children and adolescents, the brain is at its most plastic state. It is during these early phases of life that we form the mental models that govern our behavior. The architecture of the brain is not only responsible for perception, memory, and reasoning but also for the formation of moral judgment, social interaction, and emotional self-regulation. It is the substrate of our identity. Crucially, it is not set in stone by our genes. While genetic programs provide a basic scaffold, it is experience – what we see, hear, touch, and do – that determines the final form.

This makes the early years – especially childhood and adolescence – a time of both great opportunity and great risk. Every interaction, every stimulus, and every absence of experience contributes to the shaping of brain architectures and ultimately determines what kind of person a child becomes. It is during these years that fundamental capacities are established: perception, language, emotional regulation, decision-making, moral judgment, and so-

cial competence. These are not traits we simply “grow into.” They emerge from complex patterns of neural activity that gradually stabilize over time.

Yet, the environment in which today’s children grow up is radically different from that of any previous generation. Digital media, smartphones, social networks, and algorithmically curated content play an increasingly central role in how young people experience the world. These technologies are not neutral; they compete for attention, mediate relationships, and shape behavior in subtle but powerful ways. For a brain that is still under construction, this new digital environment can act as a scaffold or a constraint. The critical question we must ask is: are we providing the right kinds of experiences to support healthy brain development?

The challenge is that brain development is largely invisible. A child who spends hours with a screen may seem quiet and content. However, beneath the surface, neural networks are adapting to that experience. They are being tuned to the patterns of stimulation they receive, while other, potentially more valuable pathways may remain underdeveloped. The effects may not be immediately obvious, but their consequences can last a lifetime.

Understanding this process is not just a scientific concern; it is a societal responsibility. If we wish to guide the development of young minds wisely, we must first understand the forces that shape them. This includes not only biology but also technology, culture, and education. Only by recognizing the brain as a system that builds itself through experience can we begin to evaluate which kinds of environments support and which may hinder the growth of capable, resilient, and empathetic individuals.

In the following sections, we will explore what neuroscience has discovered about how the brain develops, why certain periods are especially sensitive, how experience shapes its internal architecture, and what the rise of digital media means for the next generation. This understanding is essential for parents, educators, policymakers, and all who care about the well-being of children in the 21st century.

2. The Functional Architecture of the Brain

The brain does not operate as a collection of isolated parts. It is a network – a dynamic and hierarchically organized system of interconnected regions, each contributing to the integrated whole of perception, action, and thought. This network is what I refer to as the brain’s *functional architecture*, and understanding how it is formed is essential to understanding brain functions and ultimately ourselves.

This architecture is both the product of evolution and the imprint of experience. Genes provide the basic scaffolding, determining the general layout: where different processing modules reside, how they are initially connected, and what kinds of neurons populate them. This inherited structure reflects the accumulated knowledge of our species, honed by millions of years of natural selection. It enables newborns to respond to their environment in meaningful ways from the outset. However, it is not complete; it is a first draft.

What transforms this draft into a functioning adult brain is the experience-dependent refinement of connections. As the child interacts with the world, neural activity shapes the architecture by reinforcing some pathways and weakening others. This process is not random; it follows clear principles: neurons that are activated together tend to connect more strongly – “neurons that fire together wire together.” In this way, the brain becomes a record of experience, a biological memory of the world as it is encountered and understood by the individual.

The cortex, the outer layer of the brain responsible for higher cognitive and executive functions, is a particularly striking example. Despite serving diverse tasks – vision, language, planning, reasoning, and remembering – its regions are constructed according to a shared blueprint. Each area consists of circuits that are canonically structured, but their specific functions are defined by their input and output connections. This means that the same basic module can evolve into a center for face recognition in one region and into a control unit for voluntary movement in another, depending on how it is wired.

Furthermore, these cortical modules are embedded in a dense mesh of reciprocal connections. They continuously communicate with one another, exchanging signals in real time. This reciprocity enables the brain to function not as a collection of discrete processors, but as a coherent system. What we call perception, for instance, is never a simple registration of sensory input. It is an interpretation shaped by prior knowledge stored in the architecture itself.

This brings us to a fundamental insight: all knowledge and all capacity for reasoning or judgment reside in the structure of connections between neurons. There is no separate “seat” of thought apart from the network. Cognition emerges from the architecture, which is shaped by both inheritance and experience.

In practical terms, this means that development is not merely about adding new content to a finished machine. It involves building the machine itself. The child’s brain is under construction, its wiring still flexible, and its

modules still awaiting assignment. This flexibility is a strength – it allows for learning and adaptation – but it also presents a vulnerability. Experiences that are rich, coherent, and emotionally meaningful will shape a brain better equipped to understand and engage with the world. Conversely, experiences that are fragmented, superficial, or disconnected from bodily action may lead to architectures that are less well-tuned for the complexities of real life.

In the sections that follow, I will explore how this architectural development unfolds over time, what makes certain phases particularly sensitive, and why the type of experience, not just its content, matters so deeply for the growing mind.

3. Critical Periods – Windows of Irreversible Change

The development of the brain does not proceed in a smooth, linear fashion. Instead, it unfolds through a sequence of biologically timed windows known as *critical periods*. During these phases, the brain exhibits an exceptional degree of plasticity – an increased capacity to form and refine connections in response to experience. These are the moments when external input has the greatest power to shape internal structure.

Critical periods begin early in life and extend into the third decade, with various functions maturing at different times. Basic sensory systems, such as vision, hearing, and tactile sensation, are the first to develop. This is why infants must be exposed to light and sound from the beginning – without it, the circuits for seeing and hearing fail to form properly. As development progresses, more complex functions emerge: sensory-motor coordination during early childhood, language acquisition, spatial orientation, emotional regulation, and, later, the social and executive abilities that characterize adolescence.

Each of these functions has its own sensitive phase during which the underlying neural circuits are particularly malleable and can be refined, stabilized, or removed. If experience during these phases is lacking, disordered, or misaligned with the brain's expectations, the result can be a permanent loss or an abnormal layout of connections, leading to dramatic consequences for the respective functions. Once the critical period of a particular area of the brain closes, the changes become irreversible. There is no further outgrowth of connections, nor is it possible to physically remove dysfunctional connections. A striking example comes from studies of visual development: if one eye is deprived of input during the critical period, even temporarily, the brain may reorganize itself so that the eye never fully recovers,

despite being structurally intact. The neural connections that would have processed signals from that eye have been reassigned or eliminated. The system has moved on.

These windows of plasticity are marked by intense growth followed by pruning. In early development, the brain overproduces connections, exploring a vast range of possibilities. Later, it eliminates those that are not reinforced by experience. This pruning process is not a loss but rather a refinement – like a sculptor removing what is unnecessary to reveal form. However, once the sculpture is set, it cannot easily be reshaped. This is why we say that the changes that occur during critical periods are, in many cases, irreversible.

Importantly, the sequence of these critical periods follows a progression from lower to higher functions. This indicates that early experiences establish the foundation for later capacities. A child who lacks rich sensory input may struggle later with language. A teenager deprived of meaningful social interaction may find it more difficult to develop empathy or a stable sense of self. These are not merely psychological observations – they reflect the logic of neurodevelopment.

Adolescence, in particular, is a phase of significant reorganization. The brain transitions from an overconnected system with a relatively flat hierarchy to one where central hubs, such as the prefrontal cortex, take control over behavior and decision-making. This transition underlies the development of self-regulation, abstract thinking, and moral reasoning. It is also a time of vulnerability when the balance between different networks is delicate and still in flux. The groundwork laid earlier in life must now be integrated into a coherent system.

Because of this, the quality of experience during critical periods is not just important – it is decisive. The brain cannot simply be “updated” later if these windows are missed or filled with impoverished input. What is built during these years will, to a large extent, define the adult mind.

In the next section, we will examine more closely the types of experiences the brain needs to mature properly and the reasons why passive exposure is insufficient.

4. Experience as the Architect – Why Active Engagement Matters

If the brain is shaped by experience, then the nature of that experience matters deeply. It is not sufficient for the senses to be stimulated. What is important is how the individual engages with the environment – whether passively or actively, superficially or in depth, with or without purpose.

In both the laboratory and daily life, we find repeatedly that only active, self-directed interaction leads to the kind of neural activity that refines and stabilizes the brain's architecture.

This distinction is perhaps best illustrated by a classic experiment in developmental neuroscience: the kitten carousel, introduced by Held and Hein from MIT. In this setup, two young cats are exposed to the same visual environment but in radically different ways. The carousel includes two connected gondolas for the kittens. One kitten can touch the ground and explore the space through its own movements, while the other is passively carried, seeing the same sights but lacking the ability to act. When later tested, only the active kitten develops normal depth perception and visuo-motor coordination. The passive one, despite receiving identical sensory input, fails to build the necessary neural circuits. The difference lies not in what they saw, but in *how* they experienced it.

This principle – active engagement as a prerequisite for learning – is not limited to motor skills or perception; it applies just as much to the development of higher cognitive functions. Attention, motivation, and emotional investment all serve as gates for plasticity, determining whether experience leaves a mark on the brain. A child who passively listens to language will not learn to speak with the same fluency as one who engages in dialogue. A teenager scrolling through images may not form the same emotional insights as one who navigates complex relationships face-to-face.

The mechanisms behind this are now well understood. Synaptic changes, which form the basis of learning and memory, depend on the coordinated firing of neurons. For this to occur effectively, the brain must be in a state of readiness: alert, interested, and actively processing information. Passive exposure does not consistently create the neural conditions required for such change. While it may generate fleeting impressions, it does not build lasting structures.

This has important consequences in the context of digital media. Many forms of screen-based interaction are designed for ease and efficiency. Algorithms anticipate user preferences, reducing the need for exploration. Interfaces flatten complexity. Virtual environments often separate action from consequence and observation from physical feedback.

While there are certainly interactive and educational uses of technology, much of it promotes a mode of engagement that is receptive rather than generative.

Of course, not all passive experiences lack value. We learn from listening, reading, and observing others. However, even in this context, the dis-

inction between active and passive is not absolute. A child who listens to a story and imagines its characters is not passive in the neurological sense. Their brain is engaged, drawing on memory, emotion, and prediction. What matters is the degree of mental activity – the extent to which the experience demands interpretation, response, and connection to the self.

Ultimately, the brain selects the signals it needs for its maturation. It does not indiscriminately absorb all input. It relies on attention to filter what matters and on movement and interaction to test what is learned. This is how the functional architecture becomes adapted to the real world: through a process of inquiry, feedback, and embodied experience.

As we shall see in the next section, adolescence represents a particularly delicate phase in this process. The need for meaningful, engaged experience is never greater – and the risks of misalignment never more serious.

5. Adolescence – A Fragile Phase of Reorganization

Adolescence is often described in psychological terms as a time of identity formation, emotional volatility, and growing independence. From a neurobiological perspective, it encompasses all of these aspects – and more. It represents a phase of radical reorganization in the brain's functional architecture, building on what came before and setting the stage for adult cognition and behavior. However, it is also a period of particular vulnerability when the balance between flexibility and stability becomes delicate.

During this stage, the brain's functional organization undergoes a large-scale restructuring. During childhood, functional networks are organized in a relatively flat manner. In adolescence, this configuration shifts toward a more centralized architecture. Hubs begin to emerge – regions such as the prefrontal cortex assume greater control, integrating information across distant areas and coordinating complex behaviors. This shift enables higher-order functions like impulse control, planning, abstract thinking, and moral reasoning. In a sense, adolescence is when the brain learns to govern itself.

However, this transformation is not smooth. It relies on the timely strengthening of connections, particularly long-range projections between brain regions. These pathways are sensitive to activity-dependent processes and require repeated, meaningful use to consolidate. If these connections are underused or misaligned with the demands of the environment, they may not form as robustly, leaving the system less integrated and more susceptible to dysfunction.

One of the hallmarks of this phase is the transient instability of neural coordination. Research has shown that the synchronization of brain activity – what we call phase locking – temporarily becomes disrupted during early adolescence. This temporary loss of coherence may explain some of the mood swings, impulsivity, and cognitive lapses often associated with this period. It is as though the brain is recalibrating, shedding inefficient patterns and testing new configurations before settling into its mature form.

This reorganization also marks a period when certain neuropsychiatric conditions can begin to manifest. The onset of schizophrenia, for example, typically occurs in late adolescence or early adulthood. While the precise causes remain unclear, the timing suggests a link with the closing stages of neural development – particularly the refinement of long-range connections and the establishment of self-monitoring functions in the prefrontal cortex. It serves as a stark reminder that development does not guarantee successful maturation; instead, it must be supported both biologically and environmentally.

In this context, the role of experience becomes critical once again. Freud called this period “the second chance.” Adolescents need opportunities to test themselves in real-world settings – to negotiate relationships, confront challenges, explore ideals, and deal with consequences. These experiences are not simply shaping social or moral norms deposited in memory; they structure the functional architecture of the brain and contribute to the stabilization of new configurations. Without them, development may stall or deviate.

The risk posed by digital media is that they might substitute for these experiences rather than support them. Virtual interactions often lack the richness and immediacy of face-to-face contact. If the adolescent brain practices control and empathy less frequently, or in oversimplified contexts, the circuits supporting these functions may remain underdeveloped.

To guide this phase well, we must acknowledge its biological foundations. Adolescence is not just a turbulent stage to be endured. It is a fragile and essential process of integration. What is rehearsed during these years is rehearsed for life.

In the final section, we will turn to the questions that arise when the developmental blueprint of the brain is increasingly shaped by digital environments – and what this might mean for future generations.

6. Digital Media and Virtual Environments – The Big Questions

We are experiencing a transformation in the developmental environment. For children and adolescents today, digital media are no longer occasional tools but constant companions. They shape communication, define

play, mediate learning, and increasingly substitute for real-world interaction. In light of what we know about how the brain develops – how it depends on rich, embodied, and actively engaged experience – this shift raises urgent questions.

The developing brain is exquisitely sensitive to the type of input it receives. It constructs its internal models of the world based on experience: from what it hears and sees, from what it touches and moves through, and from what it does. When that experience is passive or abstracted from bodily action, key developmental mechanisms may not be properly engaged. When it is dominated by algorithmically curated content, the opportunity for exploration and self-directed learning can be limited. When social interactions occur largely through screens, emotional feedback may be dulled, nonverbal subtleties lost, and the essential training of empathy and cooperation weakened.

These concerns are not rooted in nostalgia for a pre-digital world; they are grounded in neuroscience. We know that the brain's architecture is shaped by activity and that critical periods exist during which the foundations of perception, reasoning, and social understanding are laid. We understand that these periods are sensitive to deprivation – not only of stimulation but also of meaningful engagement. Additionally, we know that adolescence represents a turning point, when the brain must integrate earlier experiences into a coherent whole. The timing of these processes is not negotiable; what is missed cannot simply be reintroduced later.

The key distinction is not between “digital” and “analog,” but between *passive* and *active*, between *external noise* and *internal structure-building*. A digital environment that encourages exploration, challenges assumptions, fosters dialogue, and supports emotional engagement can effectively promote development. Conversely, one that rewards distraction, minimizes effort, and replaces human presence with artificial feedback may not.

This does not mean that all exposure to digital media is harmful, or that children must be shielded from modern tools. Rather, it emphasizes the need to be intentional in how these tools are integrated into young lives. We must ask: Does this interaction invite curiosity? Does it require agency? Does it connect to the body, to emotion, to others? Is it shaping a brain that is capable of navigating the complexities of the real world?

As a society, we have not yet answered these questions. However, neuroscience has provided us with the tools to explore them effectively. We understand more than ever how the brain develops and why the early years of

life are so profoundly significant. What remains is to apply this knowledge – to recognize that the environment we create for the next generation is not merely a background to their lives. It shapes the architecture of their minds.

Further reading:

The publication cited below is a review published in the Reports of the Ernst Strüngmann Forum that contains references to research papers supporting the claims made in this contribution.

Wolf Singer, “The role of oscillations and synchrony in the development of the nervous system” (2018) in *Emergent Brain Dynamics; Prebirth to Adolescence*. Eds. April A. Benasich and Urs Ribary, The MIT Press, 2018 ISBN 9780262038638. DOI: <https://doi.org/10.7551/mitpress/11957.001.0001>. ISBN electronic: 9780262348591. The chapters of the book can be downloaded from the Home Page of the Forum: esforum.de

LISTENING TO SEE: UNCOVERING DILEMMAS WITH YOUTH AND GENERATIVE AI

EMILY WEINSTEIN, EDD & BECK TENCH, PhD

Center for Digital Thriving at Project Zero, Harvard Graduate School of Education, Cambridge, MA, USA

■ <https://www.doi.org/10.65398/JXKI9561>

Effectively supporting youth requires good insights, not just good intentions. This is admittedly complicated in the context of generative artificial intelligence: it can feel as though there is hardly time to slow down for thoughtful reflection as technologies change at an ever-accelerating pace. Per the International Telecommunication Union (Abdullah, 2025): it took 75 years for the telephone to reach 75 million users; for the radio, it was 38 years; for the television, 13 years; and for the internet, four years. ChatGPT crossed the 100 million mark in only two months.

Youth represent 30% of the global population, and young people are known early adopters of new technologies (5Rights Foundation, 2025). Yet surfacing “good insights” about their experiences can prove challenging because so much about young people’s digital experiences remains invisible to most adults. Technology companies incorporate features and algorithms in ways that are often *unseen*, and young people use technologies in ways that aren’t always *foreseen*. Youth have experiences behind their screens that are routinely out of adult lines of sight. Limited AI literacy compounds the problem: many adults lack the vocabulary to ask the right questions or make sense of the answers. In this essay, we argue that we must listen to see – that is, in order to grasp the new dilemmas that come up for youth as they engage with generative AI, adults in positions of influence (policymakers, NGO leaders, technologists, educators, parents, and researchers) need an ongoing and purposeful practice of listening to young people.

The challenge of seeing not just “what” they do, but “as” they do

The risks of technology use coexist with considerable benefits for youth who often engage with technologies to learn, play, and connect in meaningful ways. Yet the stakes are high. This is evident in cases like the heartbreaking story of 14-year-old Sewell Setzer III who died by suicide after a chat with a bot that encouraged him to “please come home to me,” which was later discovered by adults in his life who had not even known he was even us-

ing this kind of app. Less extreme examples are relevant, too, as teens report using bots to covertly impersonate and evade parents (Common Sense Media, Hopelab, & Center for Digital Thriving, 2024), and as bots offer ready guidance on topics like how to mask the smell of alcohol (Fowler, 2023).

There is an important lesson from social media research that may be relevant as adults turn our collective attention to the range of dilemmas related to generative AI: even when adults see *what* young people do, adults may not see it *as* they do. In the context of social media use, there are at least two reasons for this. First, scholars have for years written about social steganography in teens' social media communication: the practice of "hiding in plain sight" through crafted messages that have different meanings to different audiences (Marwick & boyd, 2014). Song lyrics or an emoji may be "read" one way by those who are part of a teen's intended audience, while their intended meaning is missed or misinterpreted by others.

Second, adults may fail to see a digital interaction *as* young people do because the adults have outgrown developmental sensitivities that fundamentally color adolescents' experiences. Developmental changes during adolescence elevate sensitivity to social information and attunement to potential social evaluation (Somerville, 2013). Practically, we see the influence of these changes as teens describe digital practices like scrutinizing others' Snapscores to see if a friend has been ignoring them while responding to others, or examining Venmo transactions for cues about peer dynamics and exclusion, or worrying about Instagram "follower ratios" – all of which might surprise even regular adult users of these apps (Weinstein & James, 2022).

Developmental sensitivities thus collide with design affordances of technologies in ways that transform adolescent experiences and how they interact with peers (Nesi et al., 2018). There are also good reasons to believe social media might interact with rapid changes during adolescence in behavior, cognition, and neurobiology that increase vulnerabilities to mental health challenges (Orben et al., 2024).

Such developmental considerations are likely to be relevant in the context of *generative AI* use, too. For example, consider a young person who is bringing their most intimate questions to AI at a key developmental moment when their self-concept is changing and taking new shape. They might ask bots questions like: *Is it normal to...? What does it mean if I...? How do I know if...? What should I do if...?*

In short, young people's developmental stages will almost certainly influence the *kinds* of topics and disclosures that they bring to a generative

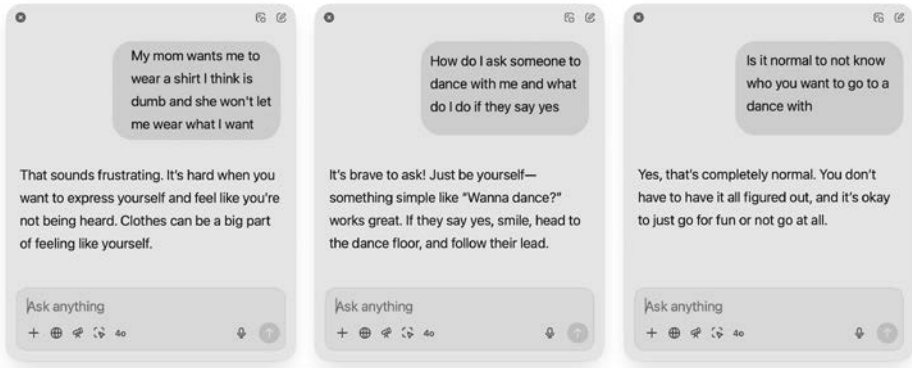


Figure 1. ChatGPT responses to hypothetical questions about a school dance (a common social event in US middle and high school).

AI (the ways they *use* it), their *interpretations* of the responses they get, and their *sensitivity* to these exchanges. Crucially, because adolescents are not a monolith – they have different strengths, contexts, circumstances, vulnerabilities, developmental trajectories – we should also expect their generative AI experiences to play out differently.

The concept of differential susceptibility is a resonant and well-supported principle across media research (Valkenburg & Peter, 2013). The 5 Cs of Media Use from the US Center of Excellence on Social Media and Youth Mental Health (American Academy of Pediatrics, 2024) distills key variables for consideration that may also be relevant to generative AI research and practice, including attention to (1) the child; (2) the content they engage with and their reactions to it; the ways media use is relevant to a child's emotional regulation (captured by the word (3) calm); the concept of displacement or what is (4) crowded out by technology use, and the potentially protective influence of (5) communication with supportive adults. Young people's varied contexts in their offline lives are also well-recognized to shape the meaning and impacts of their online experiences (Ito et al., 2020; Reich et al., 2012).

To summarize: it is a legitimate challenge for adults to “see” young people's experiences with new technologies *as* they do, in ways that elucidate the contours of their interactions. This challenge calls for a practice of listening, by adults who have a dispositional orientation (Perkins et al., 1993) – inclination, sensitivity, and skills – toward curiosity.

Listening to see

Our research team at the Center for Digital Thriving, based at Project Zero at Harvard Graduate School of Education, has spent over a decade refining ways to “see” the dilemmas young people face with emerging technologies. Our goal is to help adults have young people’s backs, rather than looking over their shoulders (Jenkins, 2010), and to support public conversations about teens and technology that are youth-centered. We honor young people’s agency and competence through a design cycle that builds on youth voices to inform both research and resources.

Our design cycle starts and ends with listening to learn. We listen in a way that is inspired by a pedagogy of listening from the Reggio Emilia philosophy (Rinaldi, 2001) – a practice that began in a village in northern Italy and has since gone on to influence early childhood education around the world (Foerch & Iuspa, 2016). Core to the philosophy is a view of the child as competent, capable, and inherently worthy of being listened to – because adults have much to learn from such listening. Reciprocal respect between children and adults underpins Reggio Emilia’s pedagogy and practices, which supports democratic participation both within school walls and across the broader community (Turner & Wilson, 2010).

Just as educators in the Reggio Emilia philosophy ask open-ended questions and then listen with intention and respect, so do we as researchers. We routinely ask young people questions like: What are adults missing that

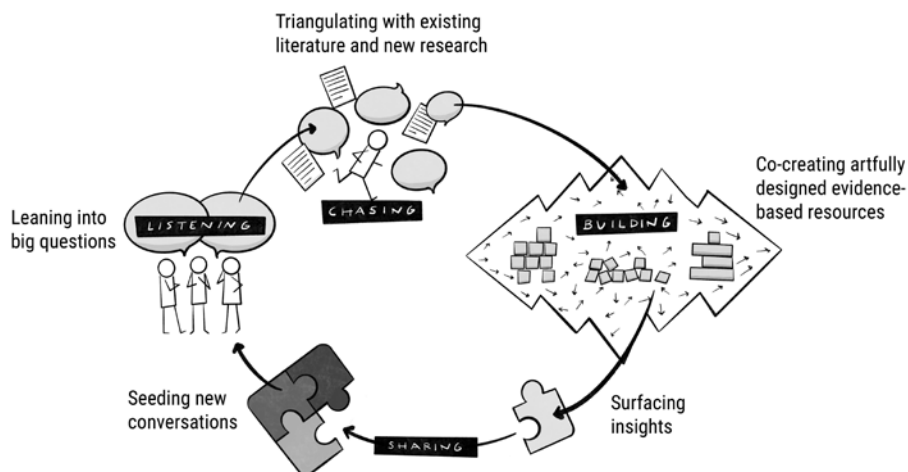


Figure 2. Center for Digital Thriving Design Process.

you most need us to understand? What does it feel like to grow up with a particular app or technology? How do you navigate dilemmas and tensions that are related to tech? We create space for adult researchers to listen with and to teens' thinking about such questions through regular advisory sessions, individual and small group interviews, focus groups, and by conducting surveys and then co-interpreting data with input from teens.

What we hear in young people's responses routinely deepens or shifts our thinking. For example, in past conversations with our teen advisors, some described the pull to stay online when a friend is struggling (Weinstein & James, 2022). They felt torn between disconnecting for self-care – to get rest before a test the next day or to spend time with family – and the urge to be a good friend who answers late-night messages. We saw how these impulses could be pitted against each other in ways that create a pull to their phones that wasn't simply a function of persuasive design or "addiction." As we took the pressures they felt seriously, we also started to appreciate how widespread increases in adolescent mental health struggles (Centers for Disease Control and Prevention, 2024) could confer a hidden toll of friendship on those who are texted for support or who witness signals of peers' distress on social media.

More recently, our research team has widened our focus beyond social media to include generative AI. We collaborated on a national study and collected perspectives from 1,545 US teens on "what adults should know about how teens use artificial intelligence (AI)" (Common Sense Media, Hopelab, & Center for Digital Thriving, 2024). Teens shared responses like:

"We use it for very creative purposes, not just cheating on homework."

"It helps me ask questions without feeling any pressure."

"Some [teens], not all, can use it for getting test answers."

"I use AI to help myself with making myself feel better about being who I am."

"It listens."

In ongoing listening sessions with teens, which include open-ended advisory meetings, structured focus groups, and semi-structured interviews, we are now actively co-interpreting the survey responses and asking young people to share more about their own experiences with generative AI. In one such discussion, a teen shared a story about a family friend who had used ChatGPT to write a card to his wife; the wife was so moved by the card that she cried. The teen's initial reaction was to frame this as "a really cool" use of AI.

Perhaps you have a knee-jerk reaction to the idea of ChatGPT authoring the text for that husband's card; based on our experience sharing this dilemma with adults, we would venture a guess that most adult readers are not so enthusiastic about outsourcing heartfelt messages to an algorithm. As researchers, we strive to lean into curiosity in these moments, and to respond in ways that convey: I believe you, I'm listening; please help me understand. This approach is rooted in a respect for young people's perspectives, and a learned recognition that our own knee-jerk judgments can otherwise shut down conversations with teens. Listening to see requires more than asking good questions; it involves engaging what Elbow (2008) called *the believing game* – listening with openness and trust toward young people's experiences, especially when they differ from your own instincts and assumptions. Engaging the believing game means embracing a willingness to be changed by what you hear. In our conversation with teens about the ChatGPT-authored card, this meant genuinely wanting to understand the young person's positive take, even though it differed from our own. We asked follow-up questions, inviting further reflections from both the teen who shared the story and others in the group. Other teens in the group initially seemed to endorse the idea that this was positive: One shared, “*I don't see that [it's] an ethical problem at all,*” and another elaborated, “*ChatGPT should be used to benefit not only yourself but others also.*”

In the Reggio Emilia philosophy, listening is not only a way to see, but also to be seen; adults “listen to children not only because we can help them but also because they can help us” (Rinaldi, 2001). The researcher facilitating this session (Emily) listened, and then – without suggesting that the teens were “wrong” and she was “right” – also shared she had a differing view:

“I can follow everything that you are saying in your logic about why ... [this] seems like a positive use ... it's so interesting, because – [at the same time] – there is something about it that makes me feel unsettled, almost.”

In response, teens made further efforts to convey their reactions. One teen compared the GPT-authored card to someone buying a ready-made bouquet of flowers. “*If you bought a whole big flower arrangement and you didn't make it yourself, like, it's still a nice effort.*” As the group built upon the bouquet analogy, everyone – including the adult researchers – refined our views and came to appreciate dimensions of other perspectives. The aim of this kind of listening is not for one person to convince another, but rather to engage in ways that surface the logic that underpins different perspectives, which we see as an essential step in our listening cycle.

In subsequent sessions with other groups of teens, we learned that young people vary in their perspectives and draw “lines” in different places related to using AI to help with writing a card for a loved one. Some are comfortable using AI to brainstorm ideas for the card or for wordsmithing, while others object once the tool starts to compose full sentences – or the entire card – for them. As one teen explained, *“Sometimes you struggle getting [thoughts] from your head into words – that’s probably most of the time okay, but asking [ChatGPT] to write an entire card, like, that’s a line.”* The conversations that followed gave us a sense that “the line” was a generative concept, ripe for surfacing data about where young people draw lines in other uses of AI, including with schoolwork. With this insight uncovered, we moved into the next phase of our listening cycle: to introduce a design invitation that makes young people’s thinking about these “lines” visible.

Making thinking visible

Thinking is an internal process that is often invisible, not only to others but also to ourselves (Ritchhart & Perkins, 2008). When we externalize our thoughts – through words written or spoken, images, 3-D models – they become visible artifacts we can reflect and build upon. When visible thinking practices are used in classrooms, they deepen students’ learning, as well as dispositions as good thinkers (Project Zero & Reggio Children, 2001). When sketches are used in design work, these visualizations “convey conceptions of reality” (Tversky, 2002). Design sketches externalize an idea or abstract concept and invite others to iterate and build upon it.

In our research group, we embrace this principle by sketching design prompts that invite teens to make their thinking about technology visible, ideally in ways that deepen awareness of their own thinking, and ours. Following up on the example of the ChatGPT-authored card and “the line” described by one teen, we created several sketches of “the line” as an abstract concept.

We then convened another set of discussion groups with different groups of teens. We told them about the genesis of “the line” idea, and in-

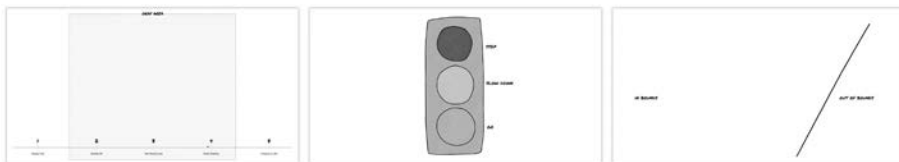


Figure 3. Initial design sketches of “the line.”

vited them to engage with it by choosing a school-related assignment they wanted to explore as a group (e.g., write an essay, create a self-portrait, solve math problems, build an app). After selecting a specific assignment, we invited them to brainstorm all possible ways they might use generative AI to complete it. Notably, at the time of our sessions and of this writing, using AI for homework assignments is considered cheating in many K-12 classrooms in the US. As a result, sharing how to use AI for assignments is sometimes stigmatized and reserved for private conversations and TikTok channels. But the prompt in our sessions intentionally fostered a “judgment-free zone” for students to make such uses visible. We explained that the aim was to surface as many uses as possible, tabling for a moment the question of whether or not those uses were ethical or allowed.

The following example from one group’s brainstorm of the ways they might use AI to help write a 5-page analytical essay on key themes in *The Great Gatsby*.

We then asked each teen to individually map each of these uses using “the line” design invitations above (Fig. 4). The example below (Fig. 5) shows six different teens’ mappings using the “Gray Area” invitation (Tench & Weinstein, 2025). The “Gray Area” sketch, which we came to call “grAIdients,”



Figure 4. Teens brainstorm different ways they might use generative AI in an essay writing assignment.

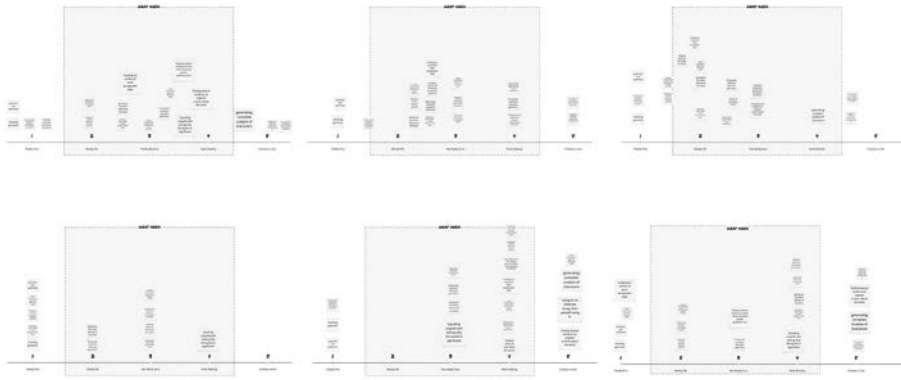


Figure 5. Six different maps from six different teens sorting the same set of activities. Even without reading the details, it is evident that judgments differ from one another.

was a unanimous favorite across groups when we asked teens which sketch showed the most promise for understanding their relationship to “the line.”

The mappings showcased an important insight for our team: teens differ in their assessments of generative AI uses that are acceptable, unacceptable, and in the “gray area” between. This exercise invited further conversation about different uses and their reasoning. Now, we are engaging with educators to explore how the “grAIdients” concept might inform generative AI policies that support essential learning goals, can be narrowed to a specific assignment or scaled to an entire class, and support cross-generational conversation between educators and students through a structure that allows for nuance. This arc – from asking teens what adults are missing, through the ChatGPT-written card and the insight about “the line,” to the design invitation that surfaces data and responds to a real problem in US classrooms – shows how (a) listening to learn and (b) making thinking visible can reveal meaningful insights into the dilemmas young people face. These practices also illuminate potential paths forward, like frameworks that inform classroom policies.

Looking Forward

The rapid development and deployment of generative AI in our society requires new conversations and new literacies to understand both what good and what harm may come from them. Research into young people’s experiences with social media over the last decade has surfaced valuable insights that can illuminate the contours of young people’s engagement with

generative AI, including the ways developmental sensitivities collide with design affordances and the recognition that youth are not a monolith – they bring diverse perspectives, vulnerabilities, and strengths to their experiences with new technologies.

Time and again, we have seen through our own work that while it is challenging to “see” tech as teens do, listening is a key that unlocks insights – and interventions – about the dilemmas they face. As qualitative and youth voice researchers steeped in traditions of making thinking visible and Reggio Emilia’s pedagogy of listening (Project Zero & Reggio Children, 2001; Ritchhart et al., 2011), we benefit from methods and values that encourage a disposition of curiosity. Listening is an essential practice that positions adults to meet this moment alongside young people. Listening helps us to *see* – with clarity and focus – at a time when our technological horizon is shifting faster than our eyes can adjust.

Acknowledgments

Thank you to the Pontifical Academy of Sciences, the World Childhood Foundation, and the Institute of Anthropology of the Pontifical Gregorian University for inviting us to participate in this convening and compilation of essays. We are grateful to Carrie James, Dan Be Kim, and Eduardo Lara for their comments on an earlier version of this essay, to Daniel Wilson and Mara Krachevsky for their time and expertise in Reggio Emilia philosophy and practice, and to Howard Gardner for his ongoing support of our interests in drawing such connections. We are also indebted to the young people who have trusted us with their perspectives and stories. Our work on the projects and listening cycle described in this essay is generously supported by Pivotal Ventures, The Susan Crown Exchange, and Grant 63220 from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of our funders or organizations.

References

- 5Rights Foundation. (2025). *Children & AI design code*. https://5rightsfoundation.com/wp-content/uploads/2025/03/5rights_AI_CODE_DIGITAL.pdf
- Abdullah, N.S. (2025, March). *Protect, provide, participate: Children first in the AI era* [Conference presentation]. *Ethical and Responsible AI: Roles and Responsibilities of Stakeholders and Regulation – Risks and Opportunities of AI for Children*. Pontifical Academy of Sciences, Vatican City.
- American Academy of Pediatrics. (2024). *The 5 Cs of media use*. <https://www.aap.org/en/patient-care/media-and-children/center->

- of-excellence-on-social-media-and-youth-mental-health/5cs-of-media-use/ Centers for Disease Control and Prevention. (2024). *Youth Risk Behavior Survey data summary & trends report: 2013–2023*. US Department of Health and Human Services.
- Common Sense Media, Hopelab, & Center for Digital Thriving. (2024). Teen and young adult perspectives on generative AI: Patterns of use, excitements, and concerns. <https://www.common SenseMedia.org/research/teen-and-young-adult-perspectives-on-generative-ai>
- Elbow, P. (2008). The believing game or methodological believing. *Journal of the Assembly for Expanded Perspectives on Learning*, 14, Article 3.
- Foerch, D.F., & Iuspa, F. (2016). The internationalization of the Reggio Emilia philosophy. *Contrapontos*, 16(2), 321–350.
- Fowler, G.A. (2023, March 14). Snapchat tried to make a safe AI. It chats with me about booze and sex. *The Washington Post*. www.washingtonpost.com/technology/2023/03/14/snapchat-myai
- Ito, M., Odgers, C., Schueller, S., Cabrera, J., Conaway, E., Cross, R., & Hernández, M. (2020). *Social media and youth wellbeing: What we know and where we could go*. Connected Learning Alliance.
- Jenkins, H. (2010, October 18). Raising the digital generation. *Pop Junctions*. https://henryjenkins.org/blog/2010/10/raising_the_digital_generation.html
- Marwick, A.E., & boyd, d. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16(7), 1051–1067.
- Nesi, J., Choukas-Bradley, S., & Prinstein, M.J. (2018). Transformation of adolescent peer relations in the social media context: Part 1 – A theoretical framework and application to dyadic peer relationships. *Clinical Child and Family Psychology Review*, 21, 267–294.
- Orben, A., Meier, A., Dalgleish, T., & Blakemore, S.J. (2024). Mechanisms linking social media use to adolescent mental health vulnerability. *Nature Reviews Psychology*, 3(6), 407–423.
- Perkins, D.N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly* 39(1), 1–21.
- Project Zero, & Reggio Children. (2001). *Making learning visible: Children as individual and group learners*. Reggio Children.
- Reich, S.M., Subrahmanyam, K., & Espinoza, G. (2012). Friending, IMing, and hanging out face-to-face: Overlap in adolescents' online and offline social networks. *Developmental Psychology*, 48(2), 356–368.
- Rinaldi, C. (2001). The pedagogy of listening: Listening perspective from Reggio Emilia. *Innovations in Early Education: The International Reggio Exchange*, 8(4), 1–4.
- Ritchhart, R. & Perkins, D. (2008). Making thinking visible. *Educational Leadership*, 65(5), 57–61.
- Ritchhart, R., Church, M., & Morrison, K. (2011). *Making thinking visible: How to promote engagement, understanding, and independence for all learners*. Jossey-Bass.
- Somerville, L.H. (2013). The teenage brain: Sensitivity to social evaluation. *Current Directions in Psychological Science*, 22(2), 121–127.
- Tench, B. & Weinstein, E. (2025). *GrAIdients 1.0: A framework for making the ethically gray areas of generative AI visible*. Center for Digital Thriving.
- Turner, T., & Wilson, D.G. (2010). Reflections on documentation: A discussion with thought leaders from Reggio Emilia. *Theory Into Practice*, 49(1), 5–13.
- Tversky, B. (2002, March). What do sketches say about thinking? In *Proceedings of the 2002 AAAI Spring Symposium on Sketch Understanding* (pp. 148–151). AAAI.
- Valkenburg, P.M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of Communication*, 63(2), 221–243. <https://doi.org/10.1111/jcom.12024>
- Weinstein, E., & James, C. (2022). *Behind their screens: What teens are facing (and adults are missing)*. MIT Press.

SHOULD WE BAN SOCIAL MEDIA IN CHILDHOOD AND ADOLESCENCE?

CHRISTIAN MONTAG

Centre for Cognitive and Brain Sciences, Institute of Collaborative Innovation, University of Macau, Macao, China;
Department of Psychology, Faculty of Social Sciences, University of Macau, Macao, China;
Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macao, China

■ <https://www.doi.org/10.65398/UNBY3214>

Abstract

Worldwide it is debated if social media should be banned in childhood and adolescence. Recently, the government of Australia even proposed to ban social media until the age of 16. In other countries the age barrier for social media use is often 13. But these age barriers are mostly not enforced.

In the present article, problems around social media use are discussed and it is reflected upon how to best protect young minds from harm on the platforms. Reflections on the appropriate age to use social media are also provided.

Keywords: social media, age, harm, children, adolescence, addiction, eating disorders, cyberbullying, learning

1. Background

At the moment more than five billion people use a social media service (Dixon, 2024) with many of its users being minors. Getting exact numbers on minor-aged users is difficult, because use of social media often is not allowed before the age of 13. Recent data from the US suggest that 49% of 10–12-year-olds from the US use social media (as assessed via self-report of their caretakers).¹ Aside from these numbers, many readers will likely know someone of a younger age navigating platforms such as TikTok, YouTube or Instagram.

In 2024 the book *Anxious Generation* by Jonathan Haidt stimulated a debate on the adverse effects of social media use (Haidt, 2024). In this book, Haidt proposed that underprotection in the online world and overprotection in the offline world might be the cause for the mental health crisis in

¹ https://mottpoll.org/sites/default/files/documents/101821_SocialMedia.pdf

adolescence in the US (Haidt, 2024). Many years of psychiatric research make it very unlikely that it is that simple, because it is well known that psychopathologies share both a nature and a nurture component (Polderman et al., 2015), whereas it has also been put forward that many environmental influences shape our well-being (Biglan et al., 2012).

This said, the literature shows that several problems around social media platforms exist and might also contribute to mental illness. Therefore, it is clear that vulnerable groups such as children need to be protected from harm while being on the platforms and also from harm being a consequence of using these platforms (Montag, Demetrovics, et al., 2024).

Should such protection measures be as radical as proposed by the Australian government aiming to ban social media before the age of 16?² Or are there other better approaches? The present short article provides an overview on the current state of affairs regarding social media use and well-being, and then also reflects upon best policies around regulation of social media in childhood and adolescence.

2. Artificial intelligence and the data business model behind social media

Most social media platforms operate with a surveillance capitalism model (Zuboff, 2015). In short, users leave digital footprints while interacting with the platforms. The digital footprints are studied to better understand what users' interests are and to infer further psychological characteristics of the users. This information can be used for showing personalized ads, which have been mentioned to increase click and buy rates (Matz et al., 2017) and might even change voting intentions (Zarouali et al., 2020). Against the background of this business model, it becomes clear that the industry has a keen interest in prolonging online time mainly for two reasons: First, longer online time means more digital footprints from the users, which can be exploited. Second, longer online time also means that people see more ads on the platforms ("attention economy"). And the social media industry is being paid for providing enterprises with the chance to show their ads on the platforms.

It comes not as a surprise that artificial intelligence (AI) plays an important role on social media platforms (Montag, Yang, et al., 2025): On the one hand, AI supports the study of digital footprints to understand what users are interested in. On the other hand, AI is at the heart of social media's recom-

² <https://www.bbc.com/news/articles/c89vjj0lxx9o>

mendation engine (Guha, 2021), hence supporting the personalization process of the social media feed. Aside from design features such as “Likes” and endless scrolling, the personalization of content can also be seen as a design feature, because “personalization” likely also prolongs online time. Otherwise users would often see non-interesting information in their feeds, which would clearly result in lower online time (Montag et al., 2019). But it is a fine line between filtering content for users according to their (estimated) interests – and by this reducing the cognitive burden of choosing from millions of content pieces – and glueing users to the screens of their smartphones.

3. Can a data business model aiming at prolonging online time be healthy?

In general, the question arises if a business model aiming at prolongation of online time can be healthy. In particular this question is valid, as the current business model also bases on the surveillance of users, hence reducing privacy (Montag & Elhai, 2023b; Montag & Hegelich, 2020). Loss of privacy and prolongation of online time need to be seen very critical in general, but in particular when we discuss use of these products in children and adolescents. Children and adolescents represent very vulnerable user groups with unique psychodevelopmental tasks to fulfill (Montag, Demetrovics, et al., 2025): The classic works of Erik Erikson (Erikson & Erikson, 1998) but also from the neurosciences such as by Jaak Panksepp (Panksepp, 1998) stressed that young kids need sufficient play time to foster motoric skills and social competencies. In elementary school children more and more learn to be part of social groups (outside from the core family) and they start to learn how to regulate their behaviors to focus on school lessons. In adolescence and with the upcoming of puberty, the development of one’s own identity and sexuality become focal points of psychodevelopment. If an industry aims at taking every minute of its users, how can this work in favor of the successful fulfillment of psychodevelopmental tasks of children (Montag, Demetrovics, et al., 2025)? Although social media represents a highly attractive idea allowing users to interact with each other, to establish social capital and to reach out from one to many, the current prevailing business case needs to be seen as problematic in terms of fostering healthy use behavior. Hence, protection of minors also means to rethink the business model behind social media products. In a recent paper we thought about alternatives, which will not be further discussed here (Dhawan et al., 2022).

Complicating matters, it should be mentioned that the debate centering solely around screen time to understand users’ well-being is too constricted.

In the meanwhile, it is well-known that associations between screen time and well-being are very small, but these meta-analytical findings had no focus on children and adolescents (Huang, 2017). In order to shed light on the well-being of users, we need to include information about age, gender, personality, use motives and also on how people are using social media (Montag, Yang, et al., 2021). And for sure younger age is robustly associated with more excessive use of social media (Cheng et al., 2021), likely due to higher self-regulation problems due to the still ongoing neurodevelopmental processes. Therefore, I repeat again that for me an industry maximizing time on the platforms stands in stark contrast with the needed time for the fulfillment of relevant psychodevelopmental tasks of young minds. Unfortunately, to my knowledge empirical research on associations between social media use and fulfillment of such psychodevelopmental tasks in young age is scant.

4. Is excessive social media use of addictive nature?

A highly relevant research question focuses on understanding if excessive social media use is best characterized by addictive behaviors. This is also mirrored in the EU's regulatory Digital Services Act speaking, in its preamble 81, of online platforms, "which may cause addictive behaviours."³ Of note, it is important to say that not everyone spending much time on social media is "addicted", but for sure those who are addicted spend much time on the platforms.

Considering the International Classification of Diseases' 11th edition from the World Health Organization, social media addiction is currently not an officially recognized disorder. Since 2019 Gaming Disorder – including the category of "predominantly online" – has been officially recognized in the manual as a distinct diagnosis (Montag, Schivinski, et al., 2021). At the moment it is discussed if the framework to diagnose Gaming Disorder can be transferred to excessive social media use (Montag & Markett, 2024; Müller et al., 2022). Within this framework, one would speak of disordered social media use, if (1) the users center their everyday life completely around social media (and other formerly important areas of life are neglected), (2) loss of control over one's own social media use is happening and (3) the users of social media keep on with their excessive behaviors although problems arise. For instance, someone scrolls through the feed of social media until late at night, sleeps over in the morning and misses an

³ https://www.eu-digital-services-act.com/Digital_Services_Act_Preamble_81_to_90.html

important exam. After this episode the person does not learn from the bad experience and keeps on with the unhealthy behavior.

Aside from the three described symptoms, functional impairments need to be observed. In other words, the excessive social media behavior must be so grave that it results in significant consequences such as job loss or break up of a romantic relationship. Considering functional impairments is very important because scientists and clinical psychologists should not overpathologize everyday life and need to ensure that diagnosis such as “addiction” are not lightheartedly provided (Billieux et al., 2015).

The reason why we have no diagnosis of social media addiction yet, has to do with the fact that research takes time and several things still need to be answered before such a diagnosis might be valid. Among these questions are: Is the excessive behavior investigated really best captured by an addiction framework, or is it compulsive behavior, or perhaps simply a habit? In particular, empirical evidence from the psychological, psychiatric sciences and neurosciences will be important to finally answer this question (Brand et al., 2022). Of interest, in particular, neuroscientific evidence is surprisingly scarce (Montag et al., 2023) and this is especially true for the investigation of this topic in children and adolescence (but see a recent interesting work: Maza et al., 2023). Aside from finding answers to this study complex, the clinical relevance of the topic also needs to be considered: Do patients reporting addictive behaviors with social media turn up in psychotherapeutical settings? Of note, recent evidence suggests that this is the case (Basenach et al., 2024). Finally, new numbers should be mentioned from the HBSC study (Boniel-Nissim et al., 2024) reporting an increase of problematic (addictive) social media use from 7% (2018) to 11% (2022) in young users. This provides a rough estimate on how common addictive behaviors are in young users – but it needs to be mentioned here that this study relied not on the WHO framework, but a DSM-5 research framework, which in the field of disordered gaming might be associated with inflated prevalences (Pontes et al., 2022).

5. Beyond the addiction debate

Although the debate around the addictive nature of social media is prominent, there are many other topics which need to be discussed when trying to grasp the negative aspects of current use of social media platforms in childhood and adolescence. It is well recognized that social media platforms can scale up the bullying problem (Giumetti & Kowalski, 2022). Hence, when being cyberbullied, many peers will see this, and this can cause additional

psychological distress. Social media use has also been reported to go along with body dissatisfaction and eating disorders (Marks et al., 2020), whereas young female users seem to be stronger afflicted with following thin ideals than young male users (Dane & Bhatia, 2023). Current design of social media platforms can further result in social upward comparison processes (comparing other lives with one's own; and also comparing number of received "Likes") reducing self-esteem (Vogel et al., 2014). And as a recent media report mentioned, gruesome content finds its way into social media feeds, such as the Reels page of Instagram.⁴ A report from 2024 mentioned that 70% of teens have seen real-life violent content on social media.⁵ Another critical area to be mentioned in this section is also poorer sleep, which has been reported to be associated with problematic social media use (Chen et al., 2024); and sleep is critical for healthy development. Such negative aspects for sure led the European Commission in their Digital Services Act to have a strong focus on the protection of minors from harm on so-called VLOPs (very large online platforms) in Europe.⁶

6. A reflection on age barriers

Although much open research questions exist, there is ample evidence that social media use brings problems in adolescents and, in particular, for children. And with the current existing age barriers, children should not be on social media at all.

Further: The age barrier of 13 is often seen as an entry age for social media services but is not based on sufficient empirical evidence. To my knowledge, this age barrier goes back to the Children's Online Privacy Protection Act,⁷ which was put into action to protect children's privacy in the online world before social media existed. So, of course, this age barrier will need to be revisited when we have insights from well-conducted longitudinal studies on well-being and social media use across adolescence. This said, the industry tries to lower the age entry barriers to their services (initiatives

⁴ <https://www.cnn.com/2025/02/27/meta-apologizes-after-instagram-users-see-graphic-and-violent-content.html>

⁵ <https://youthendowmentfund.org.uk/news/70-of-teens-see-real-life-violence-on-social-media-reveals-new-research/>

⁶ <https://digital-strategy.ec.europa.eu/en/news/commission-gathers-good-practices-combat-online-harm-minors>

⁷ <https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa>

such as Instagram Kids, aimed to start at age ten).⁸ Given the clear problems around social media such as body dissatisfaction, social upward comparison processes, confrontation with non-age appropriate content and a business model aiming at the prolongation of online time, from my point of view current social media products do not belong in the hands of children.

Is then the initiative of Australia banning social media until the age of 16 the correct response to the many challenges? I am honestly not sure. Again, no empirical evidence suggests that 16 represents the perfect age barrier, and from my point of view we likely will never have this clear evidence, because the developmental trajectories of young minds are different. The largest problem I see with the age barrier of 16 years has to do with the idea of treating all platforms in the same way (to be seen if Australia handles it this way). Although most social media platforms rely on the data business model, platforms differ in their design, in what they offer users and also their addictive potential (Montag, Wegmann, et al., 2024; Rozgonjuk et al., 2020). Just think of contrasting LinkedIn with TikTok (with this I have no intention of saying that one of the platforms is better).

I am convinced that we need to investigate the platforms separately regarding age entry barriers. Further, 16 might be a bit too late to onboard, as social media is also a relevant communication channel allowing young people to express and discover different views on life and cultures. So what to do?

7. Call for action

In light of the problems around social media, I propose to ensure that children have NO access to social media. This needs to be enforced by clear ID checks when onboarding. Opening up online bank accounts could guide such onboarding processes, and it would not be sufficient if the young users present their IDs, but the caretakers' IDs would also be needed. A third-party data broker could be responsible for the ID checking process, so that social media companies do not get access to sensitive ID data.

Given that childhood (arguably) ends around age 12-13 with the onset of puberty, one could think of allowing some platforms to operate from the age of 13, but again, platforms need to be independently investigated regarding platform design and provided content.

Of importance, parents should test and visit together with their kids the social media product of interest. This is important, because older users

⁸ <https://about.instagram.com/blog/announcements/pausing-instagram-kids>

often do not understand platforms such as TikTok, when they spent their whole life only on Facebook.

Social media use should be accompanied by the chance to speak about experiences on the platforms both in school and families. In particular, this is relevant if young users have bad experiences such as cyberbullying. For further insights into thorny topics such as sexting and grooming, see the book *Behind their screens* (Weinstein & James, 2022).

One further regulation approach could be the ban of smartphones (and the access to social media) in education contexts (Montag & Elhai, 2023a) – perhaps this makes sense even until the 11th grade in schools, in order to create a good social and learning environment. But discussions in this area are also heated around the globe – both in academia and among lay persons. I explicitly mention that a smartphone ban, of course, does not mean that no access to digital content will be provided in schools, but such access to the online world should only happen if interaction with digital content or digital learning methods really support the achievement of a desired learning goal. In the context of the so-called “digital school uniform” we argued recently that schools should establish – aside from smartphone bans – a curriculum preparing young students for a successful digital life (Maas et al., 2024; Montag & Elhai, 2023a). This includes knowledge ABOUT the digital world and the coming AI wave, but this also means ensuring that learning methods in educational settings can be supplemented by digital aspects, if the learning outcomes improve by such an intervention. Of note, this has not been proven everywhere. For instance: A meta-analysis suggests that reading comprehension is achieved in better ways when studying from books versus tablets (Delgado et al., 2018).

Finally, researchers need to be better supported to be able to come up with answers to the pressing questions of our time regarding technology use. Unfortunately, the application programming interfaces that are critical to study the online behavior of users (and also to study what people see in the newsfeeds) are largely closed (Bruns, 2019; Montag, Hall, et al., 2024). Therefore, academics have great problems understanding what is actually happening on social media. Regulation bodies such as the European Commission with their Digital Services Act should enable academics to do their job (Montag, Schulz, et al., 2024). They should also provide researchers with sufficient funds, which could be taken from fines being paid by the industry if not conforming to regulations such as the Digital Services Act (Montag & Becker, 2024).

8. Conclusion

Although no official diagnosis of social media addiction exists at the moment, empirical evidence points to several areas, making social media not a safe product for children to use at the moment. With adolescence, questions on the right age-entry are getting more complicated to be answered, and at the moment I opt for strict age checks starting with the age of 13 for using platforms, but also for the independent investigation of distinct platforms to understand what products might be used at what age level. It should not be forgotten in the discussions that social media represent critical communication/information platforms and adolescents should, at a to-be-defined time point, be allowed to use safe platforms in order to be prepared for their futures as informed citizens of societies.

References

- Basenach, L., Rumpf, H.-J., Dreier, M., Salbach, H., Renneberg, B., Gnauert, O., & Wölfling, K. (2024). Symptoms and functional impairments in patients with Internet Use Disorders participating in an online short-term therapy. *Comprehensive Psychiatry*, 131, 152471. <https://doi.org/10.1016/j.comppsy.2024.152471>
- Biglan, A., Flay, B.R., Embry, D.D., & Sandler, I.N. (2012). The critical role of nurturing environments for promoting human well-being. *American Psychologist*, 67(4), 257–271. <https://doi.org/10.1037/a0026796>
- Billieux, J., Schimmenti, A., Khazaal, Y., Maurage, P., & Heeren, A. (2015). Are we overpathologizing everyday life? A tenable blueprint for behavioral addiction research. *Journal of Behavioral Addictions*, 4(3), 119–123. <https://doi.org/10.1556/2006.4.2015.009>
- Boniell-Nissim, M., Marino, C., Galeotti, T., Blinka, L., Ozolina, K., Craig, W., Lahiti, H., Wong, S.L., Brown, J., Wilson, M., Inchley, J., & Van Den Eijnden, R. (2024). A focus on adolescent social media use and gaming in Europe, central Asia and Canada: Health Behaviour in School-aged Children international report from the 2021/2022 survey. *World Health Organization, Regional Office for Europe*. <https://iris.who.int/handle/10665/378982>
- Brand, M., Rumpf, H.-J., Demetrovics, Z., Müller, A., Stark, R., King, D., Goudriaan, A., Mann, K., Trotzke, P., Fineberg, N., Chamberlain, S., Kraus, S., Wegmann, E., Billieux, J., & Potenza, M. (2022). Which conditions should be considered as disorders in the International Classification of Diseases (ICD-11) designation of “other specified disorders due to addictive behaviors”? *Journal of Behavioral Addictions*, 11, 150–159. <https://doi.org/10.1556/2006.2020.00035>
- Bruns, A. (2019). After the ‘APICalypse’: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22, 1–23. <https://doi.org/10.1080/1369118X.2019.1637447>
- Chen, Y., Li, S., Tian, Y., Li, D., & Yin, H. (2024). Problematic Social Media Use may be Ruining Our Sleep: A Meta-Analysis on the Relationship Between Problematic Social Media Use and Sleep Quality. *International Journal of Mental Health and Addiction*. <https://doi.org/10.1007/s11469-024-01407-9>
- Cheng, C., Lau, Y., Chan, L., & Luk, J.W. (2021). Prevalence of social media addiction across 32 nations: Meta-analysis with

- subgroup analysis of classification schemes and cultural values. *Addictive Behaviors*, 117, 106845. <https://doi.org/10.1016/j.addbeh.2021.106845>
- Dane, A., & Bhatia, K. (2023). The social media diet: A scoping review to investigate the association between social media, body image and eating disorders amongst young people. *PLOS Global Public Health*, 3(3), e0001091. <https://doi.org/10.1371/journal.pgph.0001091>
- Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, 25, 23–38. <https://doi.org/10.1016/j.edurev.2018.09.003>
- Dhawan, S., Hegelich, S., Sindermann, C., & Montag, C. (2022). Re-start social media, but how? *Telematics and Informatics Reports*, 8, 100017. <https://doi.org/10.1016/j.tel.er.2022.100017>
- Dixon, S.J. (2024). *Number of worldwide social network users 2028*. Statista. <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- Erikson, E.H., & Erikson, J.M. (1998). *The Life Cycle Completed (Extended Version)*. W.W. Norton & Company.
- Giumetti, G.W., & Kowalski, R.M. (2022). Cyberbullying via social media and well-being. *Current Opinion in Psychology*, 45, 101314. <https://doi.org/10.1016/j.copsyc.2022.101314>
- Guha, R. (2021). Improving the Performance of an Artificial Intelligence Recommendation Engine with Deep Learning Neural Nets. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–7. <https://doi.org/10.1109/I2CT51068.2021.9417936>
- Haidt, J. (2024). *The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness*. Penguin Press.
- Huang, C. (2017). Time Spent on Social Network Sites and Psychological Well-Being: A Meta-Analysis. *Cyberpsychology, Behavior, and Social Networking*, 20(6), 346–354. https://www.researchgate.net/publication/317628925_Time_Spent_on_Social_Network_Sites_and_Psychological_Well-Being_A_Meta-Analysis
- Maas, R., Montag, C., & Zierer, K. (2024). *Das Digital-Dilemma: Was für die Entwicklung unserer Kinder heute wichtig ist*. Klett.
- Marks, R., Foe, A., & Collett, J. (2020). The pursuit of wellness: Social media, body image and eating disorders. *Children and Youth Services Review*, 119, 105659. <https://doi.org/10.1016/j.childyouth.2020.105659>
- Matz, S.C., Kosinski, M., Nave, G., & Stillwell, D.J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, 114(48), 12714–12719. <https://doi.org/10.1073/pnas.1710966114>
- Maza, M.T., Fox, K.A., Kwon, S.-J., Flannery, J.E., Lindquist, K.A., Prinstein, M.J., & Telzer, E.H. (2023). Association of Habitual Checking Behaviors on Social Media With Longitudinal Functional Brain Development. *JAMA Pediatrics*, 177(2), 160–167. <https://doi.org/10.1001/jamapediatrics.2022.4924>
- Montag, C., & Becker, B. (2024). Use the EU Digital Services Act for more social media research. *Nature*.
- Montag, C., Demetrovics, Z., Elhai, J.D., Grant, D., Koning, I., Rumpf, H.-J., M. Spada, M., Throuvala, M., & van den Eijnden, R. (2024). Problematic social media use in childhood and adolescence. *Addictive Behaviors*, 153, 107980. <https://doi.org/10.1016/j.addbeh.2024.107980>
- Montag, C., Demetrovics, Z., Elhai, J.D., Grant, D., Koning, I., Rumpf, H.-J., Spada, M.M., Throuvala, M., & van den Eijnden, R. (2025). Social Media Use in Childhood and Adolescence: Minimizing Its Adverse Effects Through Corporate So-

- cial Responsibility and European Union Regulations. In D.A. Christakis & L. Hale (Eds.), *Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence* (pp. 477–484). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69362-5_65
- Montag, C., & Elhai, J. (2023a). Do we need a digital school uniform? Arguments for and against a smartphone ban in schools. *Societal Impacts*, 1, 100002. <https://doi.org/10.1016/j.socimp.2023.100002>
- Montag, C., & Elhai, J.D. (2023b). On Social Media Design, (Online-)Time Well-spent and Addictive Behaviors in the Age of Surveillance Capitalism. *Current Addiction Reports*, 10(3), 610–616. <https://doi.org/10.1007/s40429-023-00494-3>
- Montag, C., Hall, B., & Lin, Y.-H. (2024). Is it possible to circumnavigate the APICalypse? On challenges to study mental health in the age of digitalization and AI. *The European Journal of Psychiatry*, 38, 100273. <https://doi.org/10.1016/j.ejpsy.2024.100273>
- Montag, C., & Hegelich, S. (2020). Understanding Detrimental Aspects of Social Media Use: Will the Real Culprits Please Stand Up? *Frontiers in Sociology*, 5. <https://doi.org/10.3389/fsoc.2020.599270>
- Montag, C., Lachmann, B., Herrlich, M., & Zweig, K. (2019). Addictive Features of Social Media/Messenger Platforms and Free-mium Games against the Background of Psychological and Economic Theories. *International Journal of Environmental Research and Public Health*, 16(14), Article 14. <https://doi.org/10.3390/ijerph16142612>
- Montag, C., Marciano, L., Schulz, P.J., & Becker, B. (2023). Unlocking the brain secrets of social media through neuroscience. *Trends in Cognitive Sciences*, 27(12), 1102–1104. <https://doi.org/10.1016/j.tics.2023.09.005>
- Montag, C., & Markett, S. (2024). Depressive inclinations mediate the association between personality (neuroticism/conscientiousness) and TikTok Use Disorder tendencies. *BMC Psychology*, 12(1), 81. <https://doi.org/10.1186/s40359-024-01541-y>
- Montag, C., Schivinski, B., & Pontes, H. (2021). Is the proposed distinction of Gaming Disorder into a predominantly online vs. offline form meaningful? Empirical evidence from a large German speaking gamer sample. *Addictive Behaviors Reports*, 14, 100391. <https://doi.org/10.1016/j.abrep.2021.100391>
- Montag, C., Schulz, P.J., Marciano, L., Roman-Urrestarazu, A., Rumpf, H.-J., & Becker, B. (2024). Safeguarding young users on social media through academic oversight. *Nature Reviews Psychology*, 1–2. <https://doi.org/10.1038/s44159-024-00311-2>
- Montag, C., Wegmann, E., Schmidt, L.D., Klein, L., Rozgonjuk, D., & Rumpf, H.-J. (2024). Insights on associations between the frequency of use of diverse social media products and social networks use disorder tendencies from a German speaking sample. *BMC Research Notes*, 17(1), 189. <https://doi.org/10.1186/s13104-024-06833-5>
- Montag, C., Yang, H., & Elhai, J.D. (2021). On the Psychology of TikTok Use: A First Glimpse From Empirical Findings. *Frontiers in Public Health*, 9. <https://www.frontiersin.org/articles/10.3389/fpubh.2021.641673>
- Montag, C., Yang, H., Wu, A.M.S., Ali, R., & Elhai, J.D. (2025). The role of artificial intelligence in general, and large language models specifically, in better understanding online addictive behaviors. *Annals of the New York Academy of Sciences*.
- Müller, S.M., Wegmann, E., Oelker, A., Stark, R., Müller, A., Montag, C., Wölfling, K., Rumpf, H.-J., & Brand, M. (2022). Assessment of Criteria for Specific Internet-use Disorders (ACSID-11): Introduction of a new screening instrument capturing ICD-11 criteria for gaming disorder and other potential Internet-use disorders. *Journal of Behavioral Addictions*, 11(2), 427–450. <https://doi.org/10.1556/2006.2022.00013>

- Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.
- Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), 702–709. <https://doi.org/10.1038/ng.3285>
- Pontes, H.M., Schivinski, B., Kannen, C., & Montag, C. (2022). The interplay between time spent gaming and disordered gaming: A large-scale world-wide study. *Social Science & Medicine*, 296, 114721. <https://doi.org/10.1016/j.socscimed.2022.114721>
- Rozgonjuk, D., Sindermann, C., Elhai, J.D., & Montag, C. (2020). Comparing smartphone, WhatsApp, Facebook, Instagram, and Snapchat: Which platform elicits the greatest use disorder symptoms? *Cyberpsychology, Behavior, and Social Networking*. <https://doi.org/10.1089/cyber.2020.0156>
- Vogel, E.A., Rose, J.P., Roberts, L.R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture*, 3(4), 206–222. <https://doi.org/10.1037/ppm0000047>
- Weinstein, E., & James, C. (2022). *Behind Their Screens: What Teens Are Facing (and Adults Are Missing)*. MIT Press.
- Zarouali, B., Dobber, T., Pauw, G., & Vreese, C. (2020). Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research*. <https://doi.org/10.1177/0093650220961965>
- Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1), 75–89. <https://doi.org/10.1057/jit.2015.5>

AI-GENERATED CHILD SEXUAL EXPLOITATION AND ABUSE IMAGERY: RESEARCH AND ACTIONS BY SNAP INC.

JACQUELINE BEAUCHERE MBE

Global Head of Platform Safety, Snap Inc.

■ <https://www.doi.org/10.65398/IEEM8679>

1. Introduction

The sexual exploitation and abuse of children is illegal, vile, and, as a topic of polite conversation, often taboo. But these horrific crimes can't be ignored. They need to be discussed in the halls of government, at boardroom tables, and at kitchen tables. Young people need to be attuned to online sexual risks, and adults need to understand the issues so they can assist youth in crisis. As a collective, we need to get comfortable being *uncomfortable*.

These issues are repeatedly discussed and actioned at Snap Inc., the company behind the popular messaging app Snapchat. Indeed, Snap is committed to making its service a hostile environment for illegal or violating activity – including the various forms of child sexual exploitation and abuse and its newest dimension: artificial intelligence-generated sexual imagery of minors.

2. Overview of Snapchat

Snap's mission – to empower people to express themselves, live in the moment, learn about the world, and have fun together – informed Snapchat's fundamental architecture. Adhering to this mission has enabled us to create a platform that reflects human nature and fosters and enhances real friendships. It continues to influence our design processes and principles, our policies and practices, and the resources and tools we provide to our community. Moreover, it undergirds our constant efforts to improve how we address the inherent risks and challenges associated with serving a large online community.

A considerable part of living up to our mission has been building and maintaining trust with our community and partners, as well as child online protection experts, law- and policymakers, and safety organizations around the world. These relationships have been built through our deliberate, consistent decisions to put privacy and safety at the heart of our product design.

Snap has always prioritized the well-being of our community, embracing privacy- and safety-by-design from the app's inception. For us, safety means our community members can maximize their desirable in-app experiences, while minimizing those associated with illegal or potentially harmful Content, Contact, Conduct, and Commercial Activity – what we and others call the “Four Cs.” Every action that we take toward safety at Snap is in furtherance of one or more of our four Safety Goals:

- By design we aim to PROTECT members of our community from the outset
- We want to PREVENT any potential harm from coming to them
- We RESPOND to their needs and concerns, and
- We regularly LEARN and ADAPT, because we know that offenders are doing the same.

2.1 Safety @ Snap

From the production and distribution of contraband imagery to the grooming of children for sexual purposes and financially motivated “sex-tortion,” Snap has been fighting back against the online sexual exploitation and abuse of minors since Snapchat was created. We have adopted and implemented a number of proactive and reactive measures to help mitigate digital risk and reduce potential online harm. Specifically:

- *Research.* Each year, we conduct research about Generation Z teens’ and young adults’ exposure to a range of online risks, including various sexual harms. Although Snap commissions this research, the study covers 13-to-24-year-olds’ experiences across applications, platforms, and services in six countries,¹ with no particular focus on Snapchat. The research is our annual contribution to the evidentiary base across the technology ecosystem, which helps to inform our own policy and product decisions, and we hope benefits others. The latest results, including findings of low knowledge about the illegalities relating to such imagery of minors, will be the focus of this paper.
- *Policies.* Snap has in place aggressive policies that we aim to consistently enforce, prohibiting “any activity that involves sexual exploitation or abuse of a minor, including sharing child sexual exploitation or abuse

¹ The six countries included in the research are Australia, France, Germany, India, the UK, and the U.S.

imagery, grooming, or sexual extortion (sextortion), or the sexualization of children.” When we become aware of such material or activity, we take action at the content, account and, in some instances, the device level. We also report all identified instances of child sexual exploitation and abuse to the U.S. National Center for Missing and Exploited Children’s (NCMEC), as required by federal law, including attempts to engage in such conduct. NCMEC then liaises with domestic and international law enforcement as needed.

► *Detection.* Snap uses established and innovative proactive detection methods to identify illegal imagery or other forms of child sexual exploitation and abuse before it can be reported to us. By way of just a few examples, we use PhotoDNA’s robust hash-matching technology to identify duplicates of known, illegal photos, and we leverage Google’s CSAI Match technology to find duplicates of known, illegal videos. We also employ “signal-based” detection to identify and remove bad actors prior to them having the opportunity to target and victimize others. We participate in all major industry and cross-sector hash- and signal-sharing programs, including the Tech Coalition’s Lantern initiative,² NCMEC’s Take It Down program,³ and Report Remove⁴ from the Internet Watch Foundation and the UK National Society for the Prevention of Cruelty to Children.

► *In-App Tools.* We know that for teens in particular, connecting with both known and unknown individuals, whether in real life or online, presents risk. This is why we have intentionally made it difficult for strangers to identify, much less contact, younger users on Snapchat. In an attempt to guard against unwanted contact from others, we’ve always made available vital tools for blocking accounts and reporting violating content and behaviors to us. We have strengthened the requirements for teens to surface for other users on our platform, and added a series of in-app warnings if teens are contacted by someone who could be suspicious – for example, someone with whom the teen does not share any friends; someone who has been blocked or reported by others; or someone who is from a region where the teen’s social circle isn’t typically located. In addition, our

² See, <https://www.technologycoalition.org/newsroom/announcing-lantern>

³ See, <https://takeitdown.ncmec.org/>

⁴ See, <https://www.iwf.org.uk/our-technology/report-remove/>

Family Center⁵ suite of tools offers parents, caregivers, and other trusted adults insight into who their teens are friends with on Snapchat and who they have been communicating with recently, without revealing the actual content of those communications. This way, our younger users can preserve their privacy and growing independence at a critical stage in their development, and parents can feel empowered to engage in crucial conversations with their teens about staying safe online. At its core, Family Center is about sparking meaningful dialogue between parents and teens about safety and well-being both on and off Snapchat.

- *Collaboration.* Because no one entity or organization can solve these novel and nuanced issues alone, we collaborate with others in industry and across the technology ecosystem. We hold seats on boards, within advisory groups, and steering committees of influential global safety organizations; we partner with industry peers to fight cross-platform abuse; we engage with scores of civil society organizations to share best practices and advocate for safer online spaces for youth; and we support law enforcement agencies in their work to bring offenders to justice. Indeed, Snap is eager to work with individuals and groups that want to make a constructive and productive difference in the lives of young people – on Snapchat and across the technology sphere. In addition, we are often called upon to participate in academic research studies, task forces, and special-project working groups, based on our strong advocacy work. We join these forums with the intent to share, learn, and continually improve.
- *Awareness-Raising and Education.* When it comes to online sexual exploitation and many related issues, raising awareness and maintaining educational efforts are of indispensable importance. We offer resources in-app, on our website,⁶ at our Support Site,⁷ and via our partners and collaborators. Our in-app offerings include four episodes about sexual

⁵ See, https://parents.snapchat.com/parental-controls?utm_source=GoogleSEM&utm_medium=PAIDUA&utm_campaign=SnapchatSafety_MaxClicks_Search_AllDevices&utm_term=Snapchat_For_Parents&utm_content=Parental_Controls&gad_source=1&gclid=CjwKCAjwtdi_BhACEiwA97y8BJLlf3HHmlN4iB7Fw7rmqD-cEArFe7mIPg0frdPeh2zPqnx5KjvxHbxoCPFgQAvD_BwE

⁶ <https://values.snap.com/safety/safety-center>

⁷ See, <https://help.snapchat.com/hc/en-us/articles/7012304746644-How-do-I-stay-safe-on-Snapchat>

risks and potential harms.⁸ We also support others' awareness-raising efforts, with the U.S. Department of Homeland Security's "Know2Protect" campaign⁹ being just one recent example.

2.2 Snap and AI-Generated Sexual Exploitation + Abuse Imagery

As operators of the My AI conversational chatbot and other generative AI features,¹⁰ Snap remains vigilant against both the possible creation, and the re-sharing or distribution, of AI-generated child sexual exploitation and abuse imagery (CSEAI). Snap prohibits use of its generative AI tools to create content that violates our Terms of Service and our Community Guidelines,¹¹ including CSEAI.

My AI and Snap's other generative AI features are built on Large Language Models, like GPT models from OpenAI and Google's Gemini models. The features leverage those models' underlying safety aspects and have a number of additional Snap-bespoke safeguards built in to help ensure content complies with our Community Guidelines. In addition, we conduct safety testing, including regular analysis of My AI's responses. In doing so, a very small percentage of responses have been found to be "non-conforming," meaning they did not adhere to Snap's policies. A common pattern stems from users repeatedly trying to trick My AI into replying with inappropriate content.

It is important to note that when Snap identifies suspected AI-generated CSEAI that was created on another platform and someone is attempting to re-share or distribute the material via Snapchat, Snap treats suspected AI-generated photos and videos as it does "authentic" CSEAI. This means we take action at the content, account and, in many instances, the device level, and report the material and the account to NCMEC as required by U.S. federal law.

⁸ See, <https://www.snapchat.com/p/9e8a90a8-897f-474b-8d65-e29797b3aa40/760067458648064?lang=en-US>

⁹ See, <https://values.snap.com/news/k2p-launch> and <https://www.dhs.gov/know-2protect>

¹⁰ See, <https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it>

¹¹ See this to review Snapchat's Community Guidelines: <https://values.snap.com/policy/policy-community-guidelines>

3. Snap's Digital Well-Being Index

Since 2022, Snap has been conducting research into the digital well-being of teens and young adults in six countries: Australia, France, Germany, India, the UK and the U.S. The research polls 13-to-17-year-olds, 18-to-24-year-olds, and parents of 13-to-19-year-olds about young people's online risk exposure; the relationships they are forging in the digital world; and their level of agreement with 20 sentiment statements. The statements, which span five categories – Positive and Negative emotions, Engagement, Relationships, and Achievement¹² – make up the Digital Well-Being Index (DWBI), a measure of young people's online psychological well-being.

Taking into account, over the preceding three months, all of their online experiences on any device or app – not just Snapchat – respondents are asked to register their level of agreement with each of the 20 statements. For instance, “Generally felt what I did online was valuable and worthwhile,” in the Positive Emotion category, and “Have friends who really listen to me when I have something to say online,” under Relationships.¹³ The DWBI assigns a score of between 0 and 100 to each respondent based on their agreement with the statements. Individual respondent scores then generate DWBI readings per country as well as a six-country average.

In the first two years (2022, 2023), the DWBI across all six countries stood at 62 – a fairly average reading, neither particularly impressive nor especially worrisome. In Year 3 (2024), the DWBI ticked one point higher to 63. This was somewhat noteworthy because the index inched up despite the fact that young people's reported risk exposure also increased – a suggestion of growing resilience on the part of both teens and young adults in 2024.

Below are some additional high-level findings from Year 3:¹⁴

- Eighty percent of 13-to-24-year-olds in the six countries said they experienced at least one online risk in 2024, up nearly five percentage points from the first survey in 2022.

¹² The index leverages the PERNA model, a variation on an established well-being theory [Seligman, M.E.P. (2011) PERMA—A Well-Being Theory], comprising 20 sentiment statements across five categories: Positive Emotion, Engagement, Relationships, Negative Emotion, and Achievement. In PERNA, Negative Emotion replaces Meaning, and Achievement replaces Accomplishment.

¹³ See this link for a list of all 20 DWBI statements: <https://images.ctfassets.net/kw9k15xxztrs/4L5iG6MCaykySxyNyOtyLm/cf7102b4d41f93612f8c2d35645a332c/img2.png>

¹⁴ For full results for years 1 through 3, see this link: <https://values.snap.com/safety/dwbi>

- Deception was common in these risk scenarios with 59% of Generation Z respondents noting they had engaged with someone online who lied about their identity.
- In 2024, more Gen Zers (compared to previous years) said they spoke with someone or sought help after experiencing an online risk. Nearly six in 10 (59%) of 13-to-24-year-olds reported seeking help, up nine percentage points from 2023.
- Just over half (51%) of parents of 13-to-19-year-olds said they actively checked in with their teens about online life, up nine percentage points from Year 2.
- Slightly more parents (45% v. 43% in Year 2) responded that they trust their teens to act responsibly online and don't feel the need to actively monitor them.
- In Year 3, as with the two prior studies, results show that young people with a greater number of "support assets" available to them enjoy stronger digital well-being. Support assets are defined as people at home, at school/work, and in the community to whom the young person can turn for encouragement and support.

3.1 DWBI Deeper-Dive: Sexual Risks

In addition to the annual DWBI research, for the past two years, Snap has delved deeper into some specific risk areas, including polling teens, young adults, and parents of teens about young people's experiences with certain sexual risks. In 2023¹⁵ and 2024,¹⁶ we asked Generation Z about exposure to sextortion – scams that deceive primarily teens and young adults into sharing intimate imagery that can quickly turn to blackmail. And, in 2024, we asked about their involvement with AI-generated imagery, including fake sexual photos and videos.

In 2024, 23% of the 6,004 13-to-24-year-olds surveyed in the six countries said they were victims of sextortion. About half (51%) reported having been lured into certain online situations or having engaged in risky digital behaviors that could have led to sextortion. These included being

¹⁵ See this blog for details: <https://www.weprotect.org/blog/two-thirds-of-gen-z-targeted-for-online-sextortion-new-snap-research/>

¹⁶ See this blog for details: <https://values.snap.com/news/new-sextortion-research-gen-z>

“groomed” (37%), being “catfished” (30%), being hacked (26%), or sharing intimate imagery online (17%).

The online grooming of minors for sexual purposes typically involves an adult befriending a young person, winning their trust through flattery and attention, and then sexualizing the relationship and continuing the abuse. This might include taking photos or videos and even meeting in person. Online catfishing occurs when criminals pretend to be someone they’re not to entice a target into sharing personal information or producing sexual imagery. Hacking usually involves an offender gaining unauthorized access to a target’s devices or online accounts to steal intimate photos or personal information. For the most part, in both the catfishing and hacking scenarios, the videos, photos, or other private information acquired is then used to blackmail the victim into acceding to the perpetrator’s demands in supposed exchange for *not* releasing the compromising imagery to the person’s family and friends.

Voluntarily sharing digital intimate imagery among young people is largely regarded as sexual exploration in the 21st century and that characterization is backed by research.¹⁷ But the practice remains a key risk vector for sextortion and other potential harms stemming from misrepresentation and falsehoods. In 2024, of the 17% of respondents who admitted to sharing or distributing intimate imagery online, 63% said they were lied to by the perpetrator and 58% reported losing control of the material once it was sent. Those under age 18 who admitted to sharing intimate imagery were particularly vulnerable: 76% said they were lied to by the abuser and 66% said they lost control of the imagery.

Meanwhile, Generation Z’s overall involvement with intimate imagery online remained a blindspot for parents in 2024. One in five (21%) parents of teens said they thought their teen had *ever* been involved with sexual imagery online, when in fact, more than a third (36%) had admitted to such connections in just the *last three months* – a 15-percentage-point gap and a gross departure from the accuracy with which parents’ were able to predict their teens’ exposure to, or experiences with, other online risks, like unwanted contact, bullying, harassment, and hate speech. In these other instances, parents said their teens came forward to tell them what happened; they heard about the incident from someone else, like a friend or peer; or

¹⁷ See, <https://www.thorn.org/blog/thorn-research-understanding-sexually-explicit-images-self-produced-by-children/>

they figured it out for themselves. Sexual risks may be a weakness for parents based in no small part on the nature of the risks themselves. Parents may be disinclined to accept that their children are interested and engaged in sexual exploration and discovery, regardless of their age or maturity level. Irrespective of the reasons behind the apparent disconnect, the research findings represent an opportunity to both educate parents and encourage teens to seek guidance from adults about sexual risks as young people appear to be doing with other online issues.

3.2 DWBI Deeper-Dive: AI-Generated Imagery

Finally, the 2024 study also examined Generation Z's involvement with AI-generated imagery, including photos and videos of a sexual nature. Here are some top-level findings about that portion of the research:

- Across teens, young adults, and parents of teenagers aged 13 to 19 (N=9,007) in the six countries, eight in 10 respondents (81%) said they had ever encountered AI-generated imagery of any sort.
- Nearly two-thirds (62%) said they had seen AI-generated imagery in the prior three months.
- Of those who had seen AI-generated imagery of any type, most (54%) had encountered it on social media; smaller percentages had seen the material while browsing the web (22%) or reviewing news articles (19%). Fewer still had received AI-generated imagery of any kind from someone else (9%).
- AI-generated photos (as opposed to videos) were the most common media format seen (83%), and this was the case across all three age demographics. Meanwhile, jokes and memes (53%) and AI-generated photos of celebrities (50%) were the most common AI images seen, while 15% of respondents reported viewing AI-generated imagery of someone they knew in real life.
- As for use of AI-image generation tools, nearly three in 10 (29%) said they had used such tools, the most common age demographic for use being 18-to-24-year-olds at 36%.

Additionally, results show that one-quarter (24%) of those who said they had seen AI-generated photos and videos said they had seen imagery of a sexual nature. Of those, 2% said they believed the imagery included someone believed to be under the age of 18.

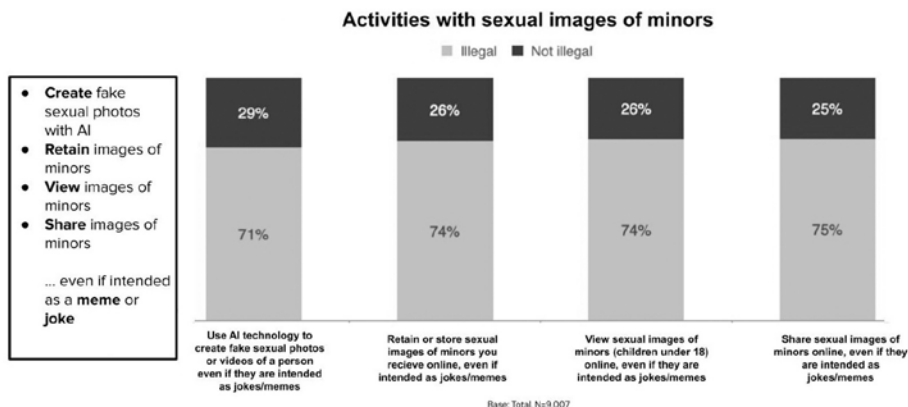


Figure 1. Low knowledge of illegalities around sexual imagery of minors.

Encouragingly, when people saw this type of content, nine out of 10 took some action, ranging from blocking or deleting the content (54%) to speaking to trusted friends or family members (52%). Just 42%, however, said they reported the content to the relevant platform or service, or to a hotline or helpline. (Lower reporting rates on digital safety issues generally follow a broader trend, which is also evidenced in our research.¹⁸)

Even more alarming is that more than 40% of respondents were unclear on the legal obligation for platforms/services to report sexual images of minors, even if such images were intended as jokes or memes. And, while a larger number (70%+) recognized it was illegal to use AI technology to create fake sexual content of another person, or to retain, view, or share sexual images of minors, findings indicate there is considerable work to do to ensure the global general public is aware of the legalities surrounding this content. (See Figure 1 for details)

In the U.S., for example, nearly 40% of respondents said they believe it is legal to use AI technology to create fake sexual images of a person. (See country details in Figure 2 below). In addition, anecdotally, we have heard of a concerning trend from industry colleagues: With the proliferation of

¹⁸ Reporting appears to have gotten a bit of a “bad rap” over the years, as young people have come to normalize exposure to problematic content and conduct online, or equate reporting with snitching or tattletaling. For more see this post and related research: <https://values.snap.com/news/back-to-school-safety>

this type of content, some teenage girls in particular are experiencing “FO-MO” and feeling “left out” if they are *not* featured in AI-manipulated sexual imagery that their peers are inappropriately creating and sharing. This disturbing point further underscores the need to educate and grow awareness of this specific online risk among all stakeholders and for trusted adults and informed peers to play an active role in discouraging this type of behavior.

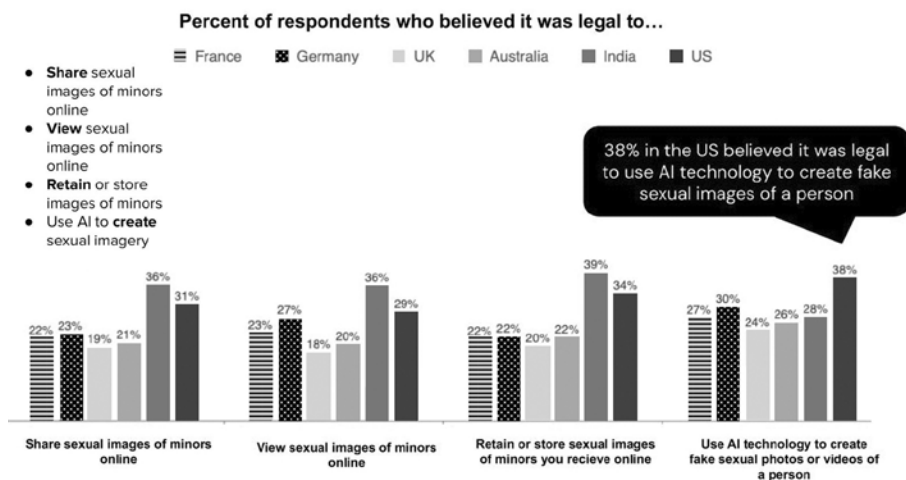


Figure 2. Relatively more respondents in U.S. & India believed it was legal to interact with use sexual imagery of minors online.

4. Conclusion

Child sexual exploitation and abuse online is a reprehensible crime that demands a concerted effort from all sectors of society to eradicate. This pervasive issue necessitates a comprehensive, collaborative approach if we hope to find any viable solutions. At Snap, we are dedicated to playing our part in minimizing risks and reducing potential harm on our platform and across the broader technology ecosystem. We employ a multifaceted strategy to combat child sexual exploitation and abuse, including annual research conducted in six countries. This research helps to inform our own policies and product decisions, and contributes to the broader knowledge base, as we collectively strive to learn, improve, and adapt in our ongoing mission to safeguard young people – and indeed, all individuals – across today’s digital landscape.

EMPOWERING AND SAFEGUARDING YOUTH IN THE AI ERA: INTEGRATING DEVELOPMENTAL NEUROSCIENCE, DIGITAL WELL-BEING, AND GLOBAL POLICY FRAMEWORKS

SIMILOLUWA OKUNOWO

Google DeepMind Scholar, African Institute for Mathematical Sciences (AIMS), South Africa

■ <https://www.doi.org/10.65398/NXFO8751>

Abstract

This paper explores the multifaceted impact of artificial intelligence (AI) on youth, weaving together insights from developmental neuroscience, digital well-being, and global policy frameworks while centering the voices of young people at every stage. It examines the trajectory from brain development to adolescent mental health, digital dilemmas, problematic social media use, AI's dual nature, ethical considerations, and the critical roles of parents and educators. Highlighting both the opportunities and risks posed by AI, this work advocates for robust collaboration among stakeholders – youth, tech developers, policymakers, educators, and researchers – to forge a digital future that empowers young people. The paper concludes with actionable, evidence-based recommendations to ensure AI enhances youth development while safeguarding their well-being.

1. Introduction

Artificial intelligence has become deeply woven into young people's daily lives, from adaptive learning apps and social platforms to mental health chatbots and creative AI tools. By early adolescence, most children navigate AI-driven environments that shape their attention, learning, and social interactions. While these technologies hold promise for personalized education and support, they also pose novel risks – privacy breaches, addictive design, exposure to harmful content, and potential mental-health harms. Centering youth voices, this paper adopts a multidisciplinary lens – integrating developmental neuroscience, psychology, digital media studies, and policy analysis – to explore AI's dual nature in shaping youth development. Our goal is to map current challenges and propose evidence-informed strategies for co-designing AI systems that empower rather than endanger young people.

2. The Developing Brain in a Digital World

2.1 Neuroplasticity and AI-Mediated Learning

Childhood and adolescence are characterized by heightened neuroplasticity, during which experiences exert outsized influence on brain circuits supporting language, executive function, and socio-emotional skills. AI-powered educational platforms can tailor instruction to individual learning profiles, potentially enhancing memory retention and problem-solving abilities.

2.2 Risks of Cognitive Over-Stimulation

Conversely, excessive screen time – driven by AI-optimized notifications and game mechanics – may overstimulate attentional networks, impair impulse control, and reduce opportunities for offline social learning. Balancing interactive digital engagement with real-world experiences is crucial for healthy neural development.

2.3 Co-Designing Developmentally Aligned Tools

Collaborative design processes that bring together neuroscientists, educators, developers, and youth themselves can ensure AI tools respect critical developmental windows. Features such as built-in “brain breaks” and adaptive pacing, informed by empirical benchmarks for attention span, exemplify developmentally sensitive design.

3. Adolescent Mental Health and Digital Dilemmas

3.1 Mental-Health Vulnerabilities

Adolescence is a peak period for the emergence of anxiety, depression, and self-harm behaviors. Online platforms can both exacerbate and alleviate these risks: cyberbullying and social comparison fuel distress, while AI chatbots and peer-moderated forums offer scalable emotional support.

3.2 Problematic Social Media Use

AI-driven recommendation systems amplify engagement at the cost of well-being, trapping users in content loops that heighten anxiety, disrupt sleep, and erode self-esteem. Recognizing patterns of compulsive use and integrating in-app usage reminders are key mitigation strategies.

3.3 Ethical Considerations and Youth Agency

Ensuring AI systems respect young users’ autonomy and emotional safety requires transparency around algorithmic processes and options for

young people to shape their digital experiences. Co-creating ethical guidelines with youth advocates elevates their agency and helps build trust in AI-based support services.

4. Policy and Protection: Addressing Harm and Ensuring Equity

4.1 Gaps in Current Frameworks

Many jurisdictions lack AI-specific protections for children, leaving them vulnerable to data exploitation, biased algorithms, and harmful content. Existing child-rights conventions and privacy laws offer a foundation but require updates to address AI's unique challenges.

4.2 Combating Harmful AI-Generated Content

The rise of AI-generated exploitative imagery demands urgent action: detection technologies, rapid takedown protocols, and survivor support services must be standardized across platforms and enforced through national and international regulations.

4.3 Towards Youth-Centric Policy

Youth councils and participatory policy forums can guide the creation of age-appropriate design codes, data-minimization mandates, and algorithmic impact assessments. Embedding young people in regulatory processes ensures that policies reflect their lived realities and priorities.

5. Conclusion, Recommendations, and Call to Action

Artificial intelligence offers unprecedented opportunities to enrich education, expand access to mental-health resources, and foster creative expression among young people. Yet without purposeful safeguards, these same technologies can compromise cognitive development, trigger psychological distress, and perpetuate inequities. To harness AI's promise while mitigating its perils, we call on stakeholders to:

1. **Co-Design with Youth:** Involve young people at every stage of AI development – from ideation to evaluation – to ensure tools meet their developmental and ethical needs.
2. **Implement Developmentally Informed Features:** Integrate attention-span controls, adaptive pacing, and offline prompts into AI products to support healthy brain development.

3. **Enhance Digital Well-Being:** Mandate transparency in recommendation algorithms, enforce usage reminders, and support research on AI's impact on mental health.
4. **Strengthen Policy Frameworks:** Update child-rights and privacy laws with AI-specific provisions, establish rapid-removal protocols for harmful content, and require algorithmic impact assessments.
5. **Foster Cross-Sector Collaboration:** Unite educators, neuroscientists, mental-health professionals, policymakers, and industry leaders in ongoing dialogues, anchored by youth perspectives.

By uniting these efforts, we can shape an AI-driven digital future that truly empowers young people – nurturing their development, safeguarding their well-being, and amplifying their voices as co-architects of tomorrow's technologies.

 SESSION 3.

**UNDERSTANDING THE RISKS
AND OPPORTUNITIES OF AI FOR CHILDREN**

SAFEGUARDING CHILDREN IN THE AGE OF AI: A CALL TO PROACTIVE ACTION

BARONESS JOANNA SHIELDS, OBE

Founder WeProtect Global Alliance; CEO and Founder Precognition

■ <https://www.doi.org/10.65398/DDAB8263>

1. Introduction

This paper continues the critical conversation on Artificial Intelligence (AI) and its impact on children. It draws from my experience founding WeProtect Global Alliance in 2014 with the goal of uniting governments, law enforcement, and the tech industry to combat online child sexual abuse. Although progress has been made, AI has dramatically reshaped the risks children face online, introducing new, scalable, and harder-to-detect forms of abuse.

My work in the tech industry spans four decades. I have seen the rise of the internet, the shift to mobile, and the emergence of social media – but nothing has matched the speed and scale of AI. Reid Hoffman describes this moment as a Cognitive Industrial Revolution¹ – changing not only what we create but how we think, dream, and invent. AI is driving creativity and enhancing human agency in ways we have never experienced before. But with this extraordinary potential comes profound risk – especially for children.

2. The New Threat Landscape for Children

Artificial Intelligence (AI) is already being weaponised to create synthetic child sexual abuse material (CSAM) at scale, overwhelming detection systems and emboldening offenders. AI-driven chatbots are grooming and manipulating children in real-time, while algorithmic incentives continue to drive harmful content to young audiences, capitalising on outrage and addiction as profitable engagement tools. These harms are not theoretical but are occurring now and with increasing frequency.

To highlight the evolving threat landscape, I commissioned and produced the film *Protect Us*,² to expose the devastating ways in which AI is

¹ Hoffman, R. (2024). *Superagency: What Could Possibly Go Right with Our AI Future*.

² *Protect Us*, WeProtect Global Alliance: <https://www.youtube.com/watch?v=18yPZFpwLFY> and <https://www.weprotect.org/thematic/artificial-intelligence-and-gen-ai/>

already being weaponised against children. The film is based on real-life cases assembled to reflect the patterns of abuse enabled by AI technologies. A domain-specific large language model was employed to analyse these patterns and translate them into a compelling narrative. Although the film is only fifteen minutes long, it delivers a lasting impact. *Protect Us* is available through WeProtect.org and YouTube and serves as a vital tool for raising awareness and driving collective action.

3. The Political Context

The political context we are operating in is just as complex as the technological one, and it would be remiss not to mention the elephant in the room. Political division and increasing geopolitical instability have reshaped the Overton window. For those who may not have heard this term before, it reflects the range of politically acceptable solutions available to policy-makers and regulators at any moment in time.

For the past decade, efforts to protect children online have largely focused on platform responsibility and content moderation. However, this post-harm punitive approach – relying on enforcement after damage has already occurred – was never designed to keep pace with the speed, scale and sophistication of today’s digital landscape. As we have seen, AI-generated content is amplifying these challenges and current systems are struggling to respond effectively. At the same time, enforcing AI and digital regulation is increasingly seen as a hostile act. Vice President JD Vance’s speech at the AI Action Summit in Paris made this clear – he stated that attempts to regulate AI are viewed as an existential threat to American dominance in technology. President Trump has reinforced this stance, making it clear that any effort to regulate AI from outside the U.S. will be seen as a direct attack on American interests.

How do we protect children when the political appetite for regulation is diminishing – and AI development is accelerating?

4. Shifting to a Proactive Model

For too long, we have operated under a post-harm enforcement model – where action is only taken after the damage has been done. That model is no longer viable in the AI era. The scale, speed, and complexity of AI-driven abuse have outstripped our ability to respond after the fact.

Once harmful content is created and shared, it can spread globally in seconds, and removing it entirely is almost impossible. Law enforcement is

overwhelmed, and the platforms responsible for hosting this content often lack the technical or legal capacity to respond quickly enough.

We have reached an inflection point. Content moderation and takedown orders are not enough – by the time action is taken, the harm is already done. The abuse has already occurred, the victims have already suffered, and the psychological scars have already formed.

We need to shift from a reactive model to a proactive one – embedding safety into AI systems from the ground up. Platforms must be incentivised to prioritise safety over engagement and addiction. Developers must build protections into AI systems at the design stage, ensuring that models are hardwired to prevent harm rather than amplify it. Governments need to step up – not by imposing outdated regulatory frameworks, but by setting clear safety standards that balance innovation with accountability.

5. The Incentives Behind AI

There is another pressure at work beneath the surface – one that worries me deeply. Most AI models are currently funded by subscription. However, the costs of developing and running these models at scale are enormous – these systems are computationally intensive and expensive to maintain. To offset these costs, platforms are exploring ad-driven models – turning to the same engagement-based incentives that fuel social media.

If AI responses become shaped by the highest bidder rather than the best answer, the risks will multiply exponentially. We cannot afford AI systems that are designed to manipulate engagement, prioritising profit over truth. Algorithmic manipulation would shift AI from a tool for human empowerment to a tool for exploitation.

Open-source AI is also compounding the challenge. Unlike closed systems, open-source models are freely distributed and can be modified and weaponised by bad actors. Large language models and generative AI tools are already being adapted to create synthetic CSAM, automate grooming, and generate deepfake content at a speed that outpaces existing safeguards. Initiatives like ROOST³ are starting to address these vulnerabilities – but broader action is needed to protect children at scale.

³ ROOST Initiative. (2024). *Addressing Open-Source AI Vulnerabilities*.

6. Conclusion

This is not just about protecting children – it is about shifting the incentives that drive the entire AI ecosystem. AI systems designed for engagement are dangerous. AI systems designed for safety and empowerment are transformative. We need to get this balance right – and we need to act now.

These issues have not been adequately covered during this conference – and they are too important to overlook. Conversations about AI and child protection often focus on individual cases of abuse or platform responsibility, but the deeper challenge lies in the incentives and infrastructure behind these technologies.

AI is evolving faster than regulation, faster than oversight, and faster than our ability to detect and prevent harm. That is why we must confront these challenges head-on.

Let us recognise that this extraordinary moment – the unprecedented pace and potential of AI, combined with the complexity of the political and technological environment – demands new approaches and solutions.

The window has changed – but new doors are open. We must walk through them now, with purpose and resolve.

References

- Hoffman, R. (2024). *Superagency: What Could Possibly Go Right with Our AI Future*.
ROOST Initiative. (2024). *Addressing Open-Source AI Vulnerabilities*.
WeProtect Global Alliance. (2024). *Protect Us* [Film].

THE DOUBLE-EDGED SWORD OF AI: INSIGHTS FROM THE FRONTLINES OF CHILD SEXUAL ABUSE PERPETRATION PREVENTION

CHRISTOFFER RAHM, MD, PhD, DOCENT

Centre for Psychiatry Research, Department of Clinical Neuroscience
Karolinska Institutet & Stockholm Health Care Services, Region Stockholm

■ <https://www.doi.org/10.65398/ALZN2959>

Your Royal Highnesses, esteemed members of the Academy, colleagues, and friends,

Before delving into the risks and potentials of artificial intelligence (AI) in child protection, I wish to share insights from my fieldwork, hoping to convey the urgency I feel regarding the need for immediate action.

As a medical doctor specializing in psychiatry and leading a research group at Karolinska Institutet in Stockholm, Sweden, my team focuses on perpetration prevention in the realm of child sexual abuse (CSA) and exploitation. Collaborating internationally, we develop and scientifically evaluate novel medical and psychotherapeutic interventions – both online and offline – for individuals with a sexual interest in children who are at risk of offending.

Over the past decade, we have developed five interventions. Three have demonstrated significant effectiveness in controlled clinical studies and are currently aiding in the prevention of harm to children daily. A fourth intervention is undergoing evaluation in a randomized controlled trial within Swedish prisons, and we are preparing to launch a program tailored for survivors. We are also collaborating on several others. It is encouraging to witness the rapid global dissemination of our methods.

Notably, we have observed that the interval between initial abusive fantasies and the first hands-on offense typically spans 5–10 years. This window presents a critical opportunity for societal intervention. Evidence suggests that successfully intervening with a single individual can prevent, on average, 20–25 acts of abuse if their interest is in girls, and over 200 acts if their interest is in boys. This does not account for the extensive consumption of child sexual abuse material (CSAM). At the onset of our program, participants report viewing CSAM for 5–7 hours weekly. Post-treatment,

around two thirds cease consumption entirely, while the remainder significantly reduce both the quantity and severity of the material.

This shows that prevention is possible.

When we have the funding to keep our programs open, clients come in hundreds from all over the world. Almost none of them are known to law enforcement.

As a sidenote, before moving on to my observations regarding AI, I would like to comment on yesterday's sessions. While I value the emphasis on future policy development, I urge attention to the current, large-scale challenges we face. Just as nations must bolster defenses against potential future conflicts, they must also allocate resources to ongoing battles. Those of us working directly with potential perpetrators to deter them from offending require substantial support to manage the pressures and advance our efforts. Our work protects children from all kinds of homes. Our interventions are evidence based, cost effective and scalable, we need support to maximise their outreach.

Back to the focus of this panel:

Through my work and patient interactions, I have come to understand that AI is already deeply integrated into perpetrators' methods of child sexual exploitation. The era of hoping that AI would not influence this field has passed. Clients report:

- AI-generated CSAM is employed to create abusive scenarios, incorporating real individuals into synthetic content, thereby tailoring materials to specific preferences.
- AI facilitates scaled grooming behaviors, enabling a single offender to target hundreds of children simultaneously.
- AI is weaponized against law enforcement, inundating detection systems with synthetic material to obstruct investigations and overwhelm investigators.

AI has come to stay, I guess, as an everyday tool in the offender world. Was it worth it, I ask the developers of AI technique.

On our side, we strive to harness AI for protective purposes:

- Machine learning assists in identifying brain activation patterns, shedding light on the mechanisms through which risk-reducing medications operate, thereby informing the development of new treatments.
- AI-powered motivational interviewing interventions are being developed to automate and scale early intervention efforts for at-risk individ-

uals, offering 24/7 availability. By integrating deterrence messaging on platforms hosting CSAM and deploying AI-based chatbots, we encourage individuals seeking such material to connect with clinics within our international network, contributing to abuse prevention.

- At Karolinska, we are developing federated learning methods within secure, sandbox-like environments. These allow private companies to develop and test AI solutions using healthcare data, with the aim of adapting these methods for other sensitive datasets not only from health care but also from law enforcement or social services, all within ethical and legal frameworks.

These examples illustrate the potential of AI for good in our field. However, my primary message is that we are lagging behind.

While offenders have seamlessly incorporated AI into their daily activities, researchers, law enforcement, and public health professionals grapple with bottlenecks – ethical and legal restrictions, compliance with data protection regulations, research ethics hurdles, and limited data access. Too often, discussions about AI’s potential in child protection – be it developing defensive tools to detect and disrupt abuse or tailoring treatments – conclude with resignation over these obstacles. To transform AI into a daily tool for good, we must address and overcome these challenges.

I want to conclude on a different note, going back to basics. Standing in the front line, surrounded by all these techniques both for offending and for reaching out with preventative care, I often think that we must never forget that, at the center of all this, we are still humans, mostly the same despite millions of years on this planet.

AI mirrors us – it reflects our desires, our values, and our choices. It mimics human behavior, amplifying both our worst instincts and our best intentions.

Technology is not in control – we are. It’s our thoughts and actions that shape the future.

The question is not what AI will become, but rather what we, as a society, choose to make of it. The choice is ours.

Let us choose wisely.

Thank you.

BRIDGING WORLDS TO PROTECT CHILDREN: REFLECTIONS ON AI AND CROSS-SECTORAL COLLABORATION

SUSANNE DRAKBORG

World Childhood Foundation

■ <https://www.doi.org/10.65398/JKIT6973>

Introduction: More Than a Technical Problem

At World Childhood Foundation, we have worked for twenty-six years to prevent child sexual abuse. Since 2002 our focus has increasingly turned toward the digital sphere, where abuse has proliferated at unprecedented speed and scale, and where technical opportunities to protect children are underexplored. Since launching the Stella Polaris initiative in 2021, we have been immersed in the complex task of connecting artificial intelligence (AI) expertise with the deeply human realities of child protection.

The scale of the problem we face demands urgent innovation. Yet technical innovation alone is not enough. True progress depends on something far more difficult: genuine cross-sectoral collaboration. AI developers, law enforcement officers, researchers, and child protection professionals must not only come together – they must learn to understand one another. This requires humility, trust, and a sustained dialogue.

Cross-Sector Collaboration: A Non-Negotiable Foundation

There is a consensus that collaboration across sectors is necessary, but we often underestimate how difficult such collaboration truly is. Within World Childhood Foundation, we regard it not as an optional ideal, but as a non-negotiable precondition for progress.

The challenges are many. Professionals from different fields do not always share terminology, assumptions, or timelines. Sometimes confusion is immediate and obvious; other times, it is masked by agreement on shared terms that in fact conceal deep misunderstandings. Without people who can bridge these divides – who are fluent in both the languages of AI and of child rights – progress stalls.

One area where this is especially evident is data. There can be no AI without data, yet in the context of child sexual abuse material (CSAM), questions of data legality, quality, access, and bias are fraught.

The development of effective AI tools – whether for detection, classification, or early prevention – requires training on real and representative data. The legal and ethical boundaries around such data are complex and vary by jurisdiction, often limiting who can access it and under what conditions. Even when access is permitted, issues of consent, storage, and sharing must be navigated with extreme care. Working with child sexual abuse material – even in the service of protecting others – can take a significant emotional toll on law enforcement officers, forensic analysts, and technical experts, who may experience secondary trauma through repeated exposure. For survivors whose abuse has been recorded and distributed, the knowledge that these images may continue to circulate – even in restricted settings such as law enforcement investigations or AI model training – can be deeply distressing. This underscores the importance of involving survivors’ perspectives in discussions about data use and privacy safeguards.

What is clear, however, is that AI is necessary to combat child sexual abuse. At present, the few tools in use are largely reactive – employed only after a child has already been abused. While these retrospective tools are important in identifying victims, removing harmful material, and supporting law enforcement investigations, they do not prevent abuse from occurring in the first place. There remains an unconscionable lack of AI applications aimed at proactive, preventative intervention. This must change.

The current AI landscape is shaped by the dominance of a few large technology companies. These actors possess the technical capacity, data access, and infrastructure needed to develop advanced AI tools – but their efforts are often driven by internal priorities, commercial considerations, and reputational concerns rather than by the needs of children or the child protection sector.

Despite the high stakes, the development of AI solutions to combat child sexual abuse remains fragmented and uncoordinated. Many tools are built in isolation by smaller actors, with limited transparency or interoperability. In some cases, different actors work in parallel on similar problems without meaningful collaboration, leading to duplication of effort and missed opportunities for synergy.

Moreover, the very nature of child sexual abuse material as illegal and highly sensitive material limits the ability to develop effective AI solutions. Restrictions on handling such content – while essential for safeguarding victims and upholding legal norms – also create major barriers to innovation.

As new AI tools to combat child sexual abuse are being developed, it is essential to ensure they are designed with global applicability in mind. If future AI solutions are built without considering the resource constraints, legal frameworks, and linguistic diversity of low-income countries, they risk becoming tools of exclusion rather than protection. Creating effective AI for child protection must include a commitment to accessibility, adaptability, and fairness across geographies, so that children everywhere – not just in high-resource settings – benefit from technological advancements.

Finally, there is the concern about bias in AI tools used to detect abuse. Many of these systems have been trained primarily on datasets that reflect limited demographic diversity – often featuring images of Caucasian children and adults. This raises serious questions about how well these tools perform in detecting abuse involving individuals of different ethnicities or cultural backgrounds. If left unaddressed, such bias could result in deeply unequal protection for children across the globe.

Scaling Prevention: AI and the Perpetration Prevention Frontier

AI has a critical role to play not only in investigative work but also in perpetration prevention. It is increasingly clear that societies cannot arrest their way out of this problem. Moreover, the true ambition should not be to respond to abuse after the fact, but to prevent it from happening in the first place. This includes reaching individuals who are actively moving along pathways toward offending – such as those seeking out darknet forums specializing in child sexual abuse material, searching for guides on how to abuse children, or engaging in grooming-like behaviors online. These are not hypothetical risks; they are identifiable patterns of high-risk behavior that can be acted on – ethically and within legal boundaries – before harm is done.

AI can support this kind of prevention by helping law enforcement, clinical experts, and civil society actors detect such patterns more efficiently, and respond with timely, targeted interventions. In some cases, that may mean flagging users for investigation; in others, it may involve redirecting individuals to anonymous help platforms or therapeutic resources.

In the area of mental health and behavioral intervention, AI offers the potential to support both front-end and back-end efforts: from triage systems and anonymous self-help platforms to therapist-guided digital interventions. AI can help scale access, personalize care, and optimize limited resources – all while supporting those who may otherwise remain invisible in current systems.

One example of such innovation is the Prevent It initiative, supported by World Childhood Foundation. This online, therapist-guided program uses cognitive behavioral therapy to reach individuals at risk of consuming child sexual abuse material. It has been rigorously evaluated through randomized controlled trials, showing measurable reductions in child sexual abuse material viewing time among participants. As highlighted in a recent systematic review,¹ Prevent It represents one of the most promising approaches in this space, and its ongoing adaptation into multiple languages further expands its potential reach.

However, this field remains vastly under-researched and underfunded. The integration of AI in perpetration prevention requires not only technical innovation, but also robust safeguards, international legal clarity, and ethical frameworks grounded in child rights and trauma-informed care. It is a frontier that demands attention.

Elodie: Modernizing the Fight Against Child Sexual Abuse Through AI

A compelling example of cross-sectoral collaboration is the project called Elodie. This AI tool is being developed to assist law enforcement by generating textual summaries of child sexual abuse material, thereby reducing the emotional toll on investigators while also improving the efficiency, consistency and quality of documentation.

The project has evolved through a series of collaborative hackathons, partially with support from World Childhood Foundation. The initiative is bringing together experts from the private sector, law enforcement, and civil society. In a recent phase, the team succeeded in refining models to generate text summaries of child sexual abuse material – eventually enabling investigators to decrease the time spent looking at the material. In one powerful illustration of its capabilities, the AI tool was able to identify that an image depicted an infant boy, an adult male, and male genitalia. However, it could not describe the specific act taking place – because it has not had enough training on the full reality of such material. The extreme sensitivity, legality, and emotional gravity surrounding this imagery have meant that very few datasets are accessible to properly train AI models. The model's silence revealed not only its technical limitations, but also unin-

¹ Seto, M.C., Roche, K., Rodrigues, N.C., Curry, S., & Letourneau, E. (2024). Evaluating Child Sexual Abuse Perpetration Prevention Efforts: A Systematic Review. *Journal of child sexual abuse*, 33(7), 847–868. <https://doi.org/10.1080/10538712.2024.2356194>

tentionally carried profound symbolic weight: the tools we build struggle to confront what we, as humans, often cannot bear to fully acknowledge.

Because even we, as humans, rarely speak openly about what truly happens in these images. We use terms like “child sexual abuse” to shield ourselves and others from the unbearable specifics. And now, we have created machines that, too, cannot find the words. What appears poetic at first is, in fact, a serious problem: the inability of current AI systems to describe child sexual abuse material in sufficient detail has real-world consequences. Investigators are left without the technological support they desperately need to process vast amounts of evidence, slowing down the pursuit of justice and leaving children at continued risk. Our reluctance to face the full horror of these crimes has shaped not only our language but our technology – and in doing so, has compromised our ability to act.

If we are serious about protecting children, we must be equally serious about equipping those on the frontlines with better tools. Artificial intelligence has the potential to dramatically improve the speed, accuracy, and emotional sustainability of investigations into child sexual abuse. But realizing that potential will require investment, collaboration, and the moral courage to engage directly with the hardest realities. Funding for the development of AI tools trained specifically – and ethically – on real-world material is a necessity.

The Stella Polaris Initiative: Creating an Ecosystem for Innovation

If we are to build tools that match the scale and complexity of the problem, we must also build the conditions that make such tools possible. Stella Polaris was launched by World Childhood Foundation as a hub to accelerate the use of AI in combating child sexual abuse, particularly in Sweden but with international reach. Its core mission is to foster cross-sectoral innovation, to bridge gaps between technology and child rights, and to translate global momentum into local and scalable action.

To achieve this, Stella Polaris has focused on cultivating shared understanding and practical collaboration. Through curated workshops, match-making efforts, thematic summits, and sustained dialogue, we have brought together actors from across sectors to explore both the risks and the potential of AI in this field. At the heart of this work is a commitment to move beyond conversation – to initiate and support real-world pilots and test cases that surface ethical dilemmas, uncover technical challenges, and explore new solutions.

Much of this work has involved identifying and addressing critical knowledge gaps – including those related to perpetration prevention, the development of scalable AI tools, and the mapping of available datasets for responsible AI training.

Through these efforts, it has become increasingly evident that no single discipline or sector can navigate these challenges alone. Progress depends on building a shared fluency – where those working on the frontlines of child protection and those designing the systems that could support them can understand one another.

A Call for Commitment and Courage

The talent, ideas, and technical solutions needed to protect children already exist. What is needed now is long-term commitment – financial, political, and institutional. Those working in this field bring expertise, resilience, and a strong commitment to protecting children. To support their efforts, there must be sustained investment and the development of reliable infrastructure. Without these, promising initiatives like Elodie will remain prototypes rather than achieving the scale required for meaningful impact.

There is no shortage of technical innovation. What remains is the collective will to sustain, scale, and responsibly deploy these efforts. The use of AI in child protection has already shown remarkable promise, while also exposing critical gaps. It is a field that demands bravery as much as brilliance, humility alongside innovation. In an era where malicious actors increasingly leverage AI for harm, we must harness AI itself as part of the solution. While concerns about the broader implications of AI are valid and must be addressed, they cannot prevent action in response to urgent and immediate challenges. Thoughtful and responsible deployment of AI is necessary. The fight against child sexual abuse is not only technological or legal; it is fundamentally a human responsibility. And in that recognition lies a shared conviction: every child deserves a life free from harm.

AI, AFRICA, AND ITS YOUTH: OPPORTUNITIES AND CHALLENGES

DAYA REDDY

Professor Emeritus, University of Cape Town, South Africa
Member, Pontifical Academy of Sciences

■ <https://www.doi.org/10.65398/PNHO6105>

1. The digital revolution

The digital revolution has been one of rapid technological change, particularly during the first quarter of this century. Its impact has been immediate and transformational, and has been experienced across the broadest social and economic landscapes.

The most recent developments in artificial intelligence (AI), particularly in generative AI, in a short period represent perhaps the most dramatic phase to date of the digital revolution. AI has had a huge impact. In the educational sphere, to take a specific example, it has yielded unprecedented educational opportunities, including personalised learning and support, improvements in the effectiveness of course content, and teacher support.

At the same time the advent of AI has been accompanied by a multitude of ethical, social, and epistemic challenges.

Responses to these developments have been uneven, depending to a great extent on educational and technical preparedness, as well as on access to relevant resources. Africa is a case in point. On the one hand the continent has experienced rapid growth in internet access, and has seen pioneering digital developments such as mobile money platforms, which have achieved high levels of acceptance and use, replacing conventional payment methods. Conditions for full participation in the digital transformation are however not equitable, nor are they inclusive.

In this piece we examine Africa's response to AI; the conditions that would be conducive to the continent's embracing the opportunities that AI presents as an engine for social and economic development; and the circumstances under which the youth of Africa, equipped with digital and more particularly AI expertise, would be at the vanguard of 'leapfrogging' current technologies and approaches in accelerating development.

2. Africa today

Africa is the world's youngest continent, in the sense that 60% of its population are under the age of 25. It is projected that, by 2035, there will be

more young Africans entering the workforce each year than the rest of the world combined (Mpemba and Munyati, 2023).

This situation is viewed by many with optimism, the view being that Africa's "youth tsunami" is well placed to take centre stage in driving progress on the continent. Indeed, the African Union's Agenda 2063 presents a blueprint for a development agenda that is youth-led.

Nevertheless, several significant challenges confront the continent.

Comprehensive access to high quality education is a *sine qua non* for youth to play the role expected of it. There are however significant barriers to progress in the education sector. In 2020, 29% of school-age children in Sub-Saharan Africa were not being educated, albeit a decrease from 44% in 2000 (UNESCO, 2022). Enrolment numbers across the different regions and countries show much unevenness with regard to numbers attending school, and the quality of education provided. So, not only is the challenge one of comprehensively accessible education at all levels, starting with early childhood, but furthermore, one of ensuring a strong foundation of basic skills in education programmes, and of building into programmes those elements and skills that are necessary in the digital, high-tech world.

Infrastructural challenges abound. For example, around 600 million people in sub-Saharan Africa are living without electricity, accounting for more than 80% of the global population without access (WHO, 2024).

Nevertheless, improvements in telecommunication infrastructure and rapid adoption of mobile devices have led to a boost in internet access in Africa: the continent had around 570 million internet users in 2022,¹ a number that more than doubled compared to 2015. Of these, some 100 million users are in Nigeria. The continent has however yet to achieve its full digital potential: despite the rising number of users, the internet penetration rate stood at around 43 percent in 2021, below a global average of 66 percent (Statista, 2024).

Despite a number of encouraging signs of progress, digital inequality across the continent remains significant. Only 5% of children and young people aged 25 or younger in West and Central Africa, and just 13% in Eastern and Southern Africa, have internet access at home (UNICEF, 2020).

3. Africa must not be left behind

The question is: what to do to ensure that Africa is not left behind, is not in the position of having to play catch-up, is a partner or participant in the

¹ <https://www.statista.com/forecasts/1146636/internet-users-in-africa>

development of AI rather than a consumer and provider of cheap labour (for data labelling, for instance)? The importance of a constructive response to this question is a substantial pay-off: of an Africa in which there is wide and deep understanding of and experience in the use of AI, which is used to address in unprecedented ways and with unprecedented effectiveness the major challenges facing the continent, such as poverty, food security, inequality, and the consequences of anthropogenic climate change.

The goal then is one of bridging the digital divide so that Africa, in partnerships within the continent and with others around the world, becomes a full participant, contributor, and effective user of AI in addressing major challenges, and with youth playing a pivotal role.

There are a number of areas that require attention and action, and which are prerequisites to the successful harnessing of AI towards the social good.

Significantly greater investment in education will be a key component of a solid base. Investment would be directed to building capacity in addressing issues around the curriculum, from early childhood to high school to university, ensuring a proper place for solid basic skill sets together with the development of digital literacy; greater numbers of well qualified teachers, with provision for comprehensive in-service professional development programmes; and, given the relatively high numbers of children not attending school, interventions aimed at increasing school enrolments, and at ensuring that children complete their schooling.

Infrastructural targets range from the basic – for example working towards increasing the number of homes with access to electricity – to those essential to the digital world, such as inexpensive access to the internet and computer hardware. These and other needs are resource-heavy, so that public-private partnerships would be central to realising such goals.

With the assumption that steady progress can be made towards meeting these various prerequisites, AI-related goals must begin with the education sector. For AI to be a meaningful tool both students and educators must develop AI literacy, the ability to engage critically with AI, understand its limitations, and use it responsibly. Educational programmes should go beyond basic AI usage and emphasise critical engagement, and equipping students and educators with the ability to question and understand AI's biases, limitations, and ethical implications.

Resources exist that would guide both teachers and students in working towards these goals; for example, the competency frameworks for teacher and students released by UNESCO. The student framework aims to as-

sist educators in integrating artificial intelligence learning objectives into official school curricula, in this way preparing students to be responsible and creative citizens in the era of AI (UNESCO, 2024a); while the teacher framework guides teachers on AI use and misuse in education, and sets out the knowledge, skills, and values that would need to be mastered by teachers (UNESCO 2024b).

Providing the foundation for young people to learn, and access to personalized learning tools, are going to be of foundational importance.

Returning to infrastructure, the energy-related demands of AI pose significant environmental sustainability challenges. The anticipated rapid growth of AI will significantly increase energy demand, particularly for data centres and cloud computing. It will therefore be vital to give attention to sustainable energy provision in tandem with addressing the range of infrastructural needs of AI.

Cultural factors, and factors related to Africa's own social and developmental goals, should play a significant role in approaching questions pertaining, for example, to the manner in which AI literacy programmes are developed: there is a need to take cognisance of local educational contexts. Likewise, the reality is that by far the bulk of data is generated from sources in the Global North. It is essential that African entities develop and implement the means to address data and algorithmic bias through initiatives that include the local generation of datasets.

A further, significant, manifestation of inequality as it pertains to Africa and AI concerns the sidelining of African languages, in many of which current systems are not optimized to generate text, thus constituting an additional barrier to access. Organisations such as Masakhane (<https://www.masakhane.io>) are directing considerable energy and expertise towards building language capabilities in African indigenous languages and dialects.

A one-size-fits all approach to AI adoption is not feasible, given inequalities in relation to access to smart devices, internet access and affordable data rates. Thus, AI may exacerbate discrimination against marginalised groups. However, AI could also be a mechanism that can contribute to addressing such inequalities.

Adams (2024) has written persuasively on growing and ubiquitous inequality as the single biggest threat to the transformative promise of AI.

The use of AI poses potential risks. With regard to children for example, these include exposure to inappropriate content and false information, social media addictions, and lack of control (Research ICT Africa, 2024).

Furthermore, potential bias against marginalised groups that causes systems to treat those in different socioeconomic or ethnic groups differently.

It is therefore important to establish clear frameworks for responsible AI use, safeguarding the data privacy of young people and ensuring that AI tools do not perpetuate biases or inequalities. There exist important UN mechanisms such as the Convention on the Rights of the Child, and the Committee on the Rights of the Child, which monitors implementation of Convention by its States parties. The Convention affirms the right of every child to be respected, protected and fulfilled in digital environments.

Current AI policies and legal frameworks in AI do not provide sufficient mechanisms to protect children's data privacy, nor do they provide transparency and accountability for children online. There is often little oversight or regulation to ensure that the systems used are safe, ethical, and beneficial for the children involved (Research ICT Africa, 2024).

The need for and urgency of stronger regulatory frameworks to safeguard children online has seen action on the African continent, for example with the African Union's Child Online Safety and Empowerment Policy. The policy assesses the opportunities and risks related to digital access for children and the factors influencing child online safety, sets out the key cross-cutting issues to uphold children's rights in the digital environment and identifies ten related policy goals.

4. Bridging the digital divide – existing efforts

While the picture presented may be one of an Africa in which the adoption of AI is at best at an embryonic stage, it is worth noting many existing efforts aimed at the development of AI skills, as well as entrepreneurial AI-related activities.

Over 2,400 companies in Africa specialise in AI, with 41% being start-ups. And approximately \$2.02 billion has been invested in promoting AI activities on the continent (Afrilabs, 2024). Some examples of start-up activity include the employment of drone imagery and AI to enable early pest and disease detection (Dronelife, 2025); and the use of AI with text messaging to deliver personalised, accessible education to primary school students across Africa (ACTS, 2024).

Efforts to equip African youth with relevant digital skills include the EDISON Alliance, which aims to improve one billion lives through affordable and accessible digital solutions to challenges in health, finance, and education by 2025.

Other laudable AI-related initiatives include the Deep Learning Indaba, an annual meeting of the African machine learning and AI community. The Indaba is an educational charity whose mission is to strengthen AI in Africa, and to ensure African ownership of AI advances on the continent. The 2024 Indaba had over 600 participants from across Africa and beyond.

The AU Youth Envoy, in partnership with Google and regional governments, has led a major digital skills campaign aimed at equipping 100 000 young people with digital skills by 2024.

As a final example of infrastructure investment, Cassava Technologies, a global technology leader of African heritage, is committing more than \$700 million in a partnership with Nvidia, which controls more than 90% of the graphical processing unit (GPU) market globally, to build Africa's first "AI factory", with many thousands of GPUs forming the engines of data centres to be installed across Africa.²

5. A return to the matter of youth and AI³

We return to the question: to what extent are Africa's youth in a position to play a leapfrogging role, bypassing outdated systems and processes, embracing AI and exploiting it to lead the charge towards achieving developmental goals? How might Africa's demographic potential be unlocked? The youth dividend is a tantalizing and appealing notion: one of the continent's young people playing a vanguard role in harnessing the potential of AI so that it becomes a catalyst to driving innovation across agriculture, health-care, education, and other key sectors.

It is clear that such a payoff is not feasible without a number of pre- and co-requisites being addressed. The various prerequisites have been identified in an earlier section: these would ensure that the youth are equipped, educationally and otherwise, to seize the opportunities presented by digital and more particularly AI advances; that significant investment in education ensures a critical mass of well qualified young people; and that the essential infrastructure and related technical support, are in place.

Enabling actions would be central to generating momentum and achieving these goals. These actions would include policies and regulatory frameworks that actively promote youth-led innovation, with accompanying investments and budgetary provision. The success of such interventions will

² See <https://techcentral.co.za/cassava-african-ai-factory-720-million/261991/>

³ This section draws on (Ramaite, 2014) and (Lopes, 2012).

depend critically on the extent to which youth are full participants in developing the vision and formulating policy. Indeed, the success of an AI-centred strategy that is youth-led will depend in turn on the policy and other provisions that will allow the youth to exercise their agency, the better to seize the opportunities before them.

Such interventions will be highly capital-intensive, so that public-private partnerships as well as equitable partnerships with state and other entities beyond Africa would be central to achieving these goals, with strategies that include targeted investments and economic incentives. Ownership should be well defined, with local generation, collection and ownership of data, being key.

Rather than each country seeking to go it alone, a strategy based on shared resources in a distributed network of data centres, computer nodes, and training and research facilities, together with shared data, will render more viable access to cutting-edge infrastructure and training, and at the same time promote a beneficial culture of cooperation and mutual support, in the spirit of open scholarship and open science. The African Open Science Platform is an excellent example of such an African networked and cooperative structure. The Platform aims to position African scientists at the cutting edge of data intensive science by stimulating interactivity and creating opportunity, inter alia by building critical mass through shared capacities.⁴

By acting strategically, with purpose, and with the huge asset of Africa's youth and its potential in mind, African policymakers by acting decisively can create the conditions for the continent's youth to lead the way towards a more prosperous continent.

Acknowledgement

The author has benefited from communication with Carlos Lopes on this topic, and acknowledges the shared insights with thanks. The author also acknowledges with thanks comments by Jonathan Shock, which have led to the inclusion of further relevant and important material in this work.

⁴ See <https://council.science/news/the-national-research-foundation-of-south-africa-to-host-the-african-open-science-platform/>

References

- ACTS, 2024. African Centre for Technology Studies, Application of AI in Education, Nairobi, Kenya. Available at <https://ai4d.acts-net.org/wp-content/uploads/application-of-an-AI-education-in-Africa.pdf>
- Adams, R. 2024. *The new empire of AI: the future of global inequality*. Wiley.
- Afrilabs (2024). Mapping AI start-ups in Africa. Available at <https://www.afrilabs.com/ecosystem-insights>
- Dronelife (2025). <https://dronelife.com/2025/03/18/ai-drones-and-the-future-of-farming-a-game-changer-for-plant-disease-detection-and-food-security/>
- Lopes, C. (2012). Statement at the International Conference on Youth and Democratization in Africa, UN Economic Commission for Africa, 2012: available at <https://repository.uneca.org/handle/10855/46808>
- Mpemba, C. and Munyati, C. (2023). World Economic Forum, How Africa's youth will drive global growth. Available at <https://www.weforum.org/stories/2023/08/africa-youth-global-growth-digital-economy/>
- Ramaite, A.F. (2024). Can AI unlock Africa's youth demographic opportunities? ISS African Futures, available at <https://futures.issafrica.org/blog/2024/Can-AI-unlock-Africas-youth-demographic-opportunities->
- Research ICT Africa, (2024). Z Schroeder, D Nyakundi, J Fuller and L Dwolatzky, Children of the Digital Age: Exploring the Potential impacts of AI on Africa's Children, Cape Town. Available at <https://researchictafrica.net/research/exploring-the-potential-impacts-of-ai-on-africas-children/>
- Statista, 2024. Internet usage in Africa – statistics & facts. Available at <https://www.statista.com/topics/9813/internet-usage-in-africa/#topicOverview>
- UNESCO, 2022. Global Education Monitoring Report: Policy Paper, 2022; available at <https://www.unesco.org/gem-report/en/2022-out-school>
- UNESCO 2024a. AI student competency framework 2024: Available at <https://www.unesco.org/en/articles/ai-competency-framework-students>
- UNESCO 2024b. AI teacher competency framework 2024: Available at <https://www.unesco.org/en/articles/ai-competency-framework-teachers>
- UNICEF, 2020. How many children and young people have internet access at home? Available at <https://www.unicef.org/media/88381/file/how-many-children-and-young-people-have-internet-access-at-home-2020.pdf>
- World Health Organization, 2024. Tracking SDG 7: The energy progress report. Available at <https://trackingsdg7.esmap.org/downloads>

PANCHATANTRA 2.0: A CULTURALLY GROUNDED APPROACH FOR CHILDREN'S DIGITAL RIGHTS AND ONLINE SAFETY

SAGAR VISHNOI, PARISHRUT JASSAL & ALISHA BUTALA

Future Shift Labs

■ <https://www.doi.org/10.65398/CCFQ1845>

Abstract

Enhancing children's safety within the evolving digital landscape is paramount for fostering future-ready education. This paper directly addresses this imperative by synthesizing an AI-enhanced digital literacy framework aimed at cultivating critical engagement and ethical understanding among children, thereby contributing to their safety and preparedness for the future. The primary approach detailed herein is a comprehensive framework, which emphasises data rights and safeguarding in the digital age. This is complemented by a culturally grounded implementation strategy, an integral part of our approach, which notably leverages Indian ancestral stories, narratives from the Panchatantra, to make complex concepts relatable and instill ethical values. Key objectives of the framework include developing a nuanced understanding of data, data trails, promoting awareness of data collection's personal implications, knowledge of pertinent legal data rights (such as India's DPDP Act, 2023, GDPR, and UNCRC), and empowering children with ownership and agency over their personal data. The proposed framework is analyzed in relation to existing frameworks, including those by the Digital Education Council and UNESCO. Furthermore, the paper explores the dual role of AI, both as a potent educational tool for digital literacy and as a technology whose ethical deployment in educational settings requires stringent safeguards, particularly concerning data protection and bias mitigation, crucial for children's safety. Concluding with policy recommendations, strategies for curriculum development based on this framework, and future research directions, this work aims to contribute significantly to enhancing children's digital safety and equipping them as responsible, future-ready citizens.

1. Introduction

The pervasive integration of digital technologies and Artificial Intelligence (AI) into daily life has fundamentally reshaped how children learn,

interact, and perceive the world, bringing forth unprecedented opportunities alongside significant challenges to their safety and well-being online (UNICEF, 2021). Ensuring children are future-ready necessitates a robust approach focused on enhancing their safety in these digital environments, particularly concerning data privacy, algorithmic influence, and the cultivation of critical engagement skills. Children are increasingly navigating complex online environments, often without a comprehensive understanding of their digital footprint or the implications of their data being collected and utilized, which directly impacts their safety (Stoilova et al., 2021). To address this critical need for enhanced safety and future-readiness, this paper proposes a new AI-enhanced digital literacy framework. This approach, centered on the proposed framework, moves beyond mere technical proficiency to foster deep, critical, and ethically informed engagement with digital technologies among children. A core component of this approach is a culturally grounded implementation strategy, which leverages ancestral stories, such as those from the Panchatantra, to make the tenets of digital literacy and safety more accessible and impactful.

This report details the objectives of this proposed framework, with a particular emphasis on data rights and safeguarding as foundational elements for children's online safety. It further explores the culturally grounded implementation strategy, focusing on the Indian context. The report also examines existing digital literacy frameworks, such as those developed by the Digital Education Council and UNESCO, to identify synergies and distinctions with our proposed framework. Finally, it discusses the role of AI both as an educational tool for teaching these concepts and as a technology whose own deployment in educational settings requires careful ethical consideration and safeguarding to protect children. The ultimate aim of this framework is to equip children not only with the skills to use technology but also with the wisdom to navigate its complexities with agency, critical awareness, and a strong sense of digital safety.

2. Core Objectives of the AI-Enhanced Digital Literacy Framework

The proposed AI-enhanced digital literacy framework is designed around key objectives that transcend technical skills, aiming to cultivate a profound and critical understanding of digital technologies among children, thereby enhancing their online safety and agency. The framework prioritizes the following.

2.1. Understanding Data Trails

A primary objective is to enable students to comprehend how their personal data is captured, used, and stored across various digital platforms, as understanding this is crucial for their safety. This involves making the invisible visible, helping children understand that every online action, from website visits to social media posts and app usage, contributes to a “data trail”. This concept can be effectively taught by comparing digital trails to footprints left in the sand, a tangible analogy for younger learners. A practical exercise to achieve this is the “daily digital traceback”, where students reflect on their app usage and identify the types of data (e.g., location, images, behavioral patterns) these platforms collect. The “Data Diary” exercise, for example, requires students to log their digital activities over 24 hours, noting instances where sensors or electronic surveillance record their actions or information about them. This exercise helps students realize the frequency and extent of data collection in their daily lives. It is crucial to differentiate between general observation and actual data recording; for instance, “a camera recorded me entering the store” is valid data for this exercise, whereas “someone saw me going into the store” is not.

2.2. Data Awareness and Implications

Building on the understanding of data trails, the framework encourages children to explore the personal implications of data collection for their choices, autonomy, privacy, and safety. This involves fostering an awareness of how collected data affects their exposure to algorithmic influence. AI systems, integral to many digital platforms, require vast amounts of data to learn and improve, incentivising mass data collection, including sensitive information about children. This data can be shared or sold and may follow children throughout their lives. AI functions like surveillance, profiling, and automated decision-making are increasingly common in children’s lives, with applications that can range from “low-stakes” areas like targeted ads to “high-stakes” areas such as university admissions or child protective services, which can have significant safety implications. The complexity of AI can make it difficult to understand or contest these algorithmic decisions, potentially leading to unfair or discriminatory outcomes. The framework aims to unpack the concept of “autonomy” in this algorithmic age and how it is affected when data is constantly being tracked and monetized, potentially undermining a child’s ability to make independent choices.

2.3. *Knowledge of Legal Rights*

A critical component of the framework is introducing age-appropriate content about national and international legal frameworks governing data protection, which are essential for their safety and rights. This includes familiarizing students with their rights around consent, access to their data, and mechanisms for redress. In the Indian context, this involves understanding the Digital Personal Data Protection Act, 2023 (DPDP Act). The DPDP Act defines a child as an individual under 18 and mandates verifiable parental consent for processing their data. It also prohibits tracking, behavioral monitoring, or targeted advertising aimed at children, and emphasizes that data processing must not be detrimental to a child's well-being or safety (Ministry of Law and Justice, 2023). Data Fiduciaries have obligations to ensure data is erased when no longer needed and upon withdrawal of consent. Parents or guardians, acting as data principals for children, have the right to access details about processing activities, and request correction, updating, and erasure of personal data. The Act also establishes the Data Protection Board of India for grievance redressal (Ministry of Law and Justice, 2023). Internationally, frameworks like the General Data Protection Regulation (GDPR) in Europe and the UN Convention on the Rights of the Child (UNCRC) provide broader context. GDPR-K, the sections of GDPR pertaining to children, affords special protections due to children's lesser awareness of risks and establishes rules around the age of consent and the need for verifiable parental consent. UNCRC Article 16 specifically upholds a child's right to privacy, covering their personal space, conversations, and data, emphasizing the need to balance protection with respect for growing independence in the digital landscape for their overall safety. General Comment No. 25 of the UNCRC further elaborates on states' obligations to protect child rights in the digital environment (United Nations Committee on the Rights of the Child, 2021). Teaching these rights involves using age-appropriate language and methods, such as simple explanations for younger children and more in-depth discussions for older ones. Child-centric consent mechanisms, like using comics and videos, are particularly relevant in India, considering varying literacy rates.

2.4. *Ownership and Agency over Data*

The framework aims to empower children to identify and assert ownership of their data, which is key to their digital safety. This involves teaching them to recognise manipulative digital practices and make informed

choices about sharing or withholding personal information. Fostering digital awareness and agency is crucial, and this can be achieved by co-creating privacy and tech-use policies with students, families, and educators. Children need to understand that control over their digital footprint equates to control over their future opportunities, reputation, and safety. Practical exercises, like playing an open-source intelligence (OSINT) game, can vividly illustrate how scattered digital crumbs form a complete personal portrait, making the concept of data ownership tangible. The goal is to shift children from mindless sharing to deliberate and thoughtful online engagement, respecting both their own and others' privacy. This includes teaching them to manage privacy settings on various platforms and to think critically before posting. Ultimately, this objective seeks to cultivate a sense of empowerment, enabling children to navigate the digital world not as passive consumers but as active, informed agents who can protect their data, assert their rights, and ensure their safety.

3. Glance at Pivotal Digital Literacy Frameworks

To refine the proposed AI-enhanced digital literacy framework, particularly its focus on data rights, safeguarding, and enhancing children's safety, an analysis of existing prominent frameworks is essential. The Digital Education Council's AI Literacy Framework and UNESCO's AI Competency Framework for Students offer valuable perspectives.

3.1. Digital Education Council (DEC) AI Literacy Framework

The Digital Education Council (DEC) AI Literacy Framework adopts a human-centred approach, emphasising human skills like critical thinking, creativity, and emotional intelligence alongside AI competencies. It is designed to provide higher education institutions with structured guidance for developing AI literacy. The framework outlines five key dimensions of AI literacy: Understanding AI and Data, Critical Thinking and Judgment, Ethical and Responsible AI Use, Human-centricity, Emotional Intelligence, and Creativity, and Domain Expertise (Digital Education Council, 2023). Its "Ethical and Responsible AI Use" dimension directly addresses aspects of data rights and safeguarding by emphasising privacy principles and risk recognition. This aligns with the proposed framework's objectives concerning data awareness, legal rights, agency, and overall digital safety. However, the DEC framework, being targeted at higher education, may require adaptation for younger learners. The proposed framework's explicit

focus on exercises like “daily digital traceback” and understanding specific data trails from a child’s perspective offers a more granular approach for younger age groups compared to the broader principles outlined by DEC.

3.2. UNESCO’s AI Competency Framework for Students

UNESCO’s AI Competency Framework for Students aims to empower students to be skilled and responsible users of AI, as well as active contributors to inclusive and sustainable AI systems, fostering their safety and positive contribution. It emphasises critical judgment of AI solutions, awareness of citizenship responsibilities in the AI era, foundational AI knowledge, and inclusive, sustainable AI design, grounded in human rights and fundamental freedoms (UNESCO, 2022). By highlighting inclusivity, human agency, non-discrimination, and cultural sensitivity, the UNESCO framework aims to equip students with skills and values for responsible citizenship in an AI-driven world (UNESCO, 2022). This resonates strongly with the proposed framework’s emphasis on data awareness, legal rights, agency over data, and ensuring children’s safety. While the UNESCO framework mentions “Ethics of AI” as a core dimension, the proposed framework provides more specific objectives like “Understanding Data Trails” and “Ownership and Agency over Data”, which offer concrete areas for curriculum development concerning data rights and digital safety.

4. Panchatantra 2.0: Culturally Grounded Digital Literacy Through Storytelling

To ensure children’s understanding of digital rights and privacy is both engaging and deeply internalised, our framework integrates Panchatantra 2.0, a reimagined storytelling approach that draws on India’s rich narrative heritage to make abstract digital concepts relatable. Building on the centuries-old Panchatantra tradition, which used fables to teach moral reasoning (*nīti*) and practical life skills (Olivelle, 1997), this strategy translates modern digital challenges into familiar, culturally resonant story formats. This narrative-based pedagogy functions on multiple levels. It simplifies complex ideas, fosters emotional and ethical engagement, and enables active learning. Educational research confirms that by embedding concepts in character-driven stories, learners can more easily grasp abstract principles and improve knowledge retention (Haven, 2007). This culturally grounded method not only improves knowledge retention but also roots digital safety in a context that feels familiar and meaningful, bridging the cognitive gap

between a child's offline moral universe and their emerging online life to cultivate informed, ethical, and resilient digital citizens.

5. Role of AI in Digital Literacy and Safeguarding

AI serves a dual role in this context: it is both a powerful tool for teaching digital literacy and a complex technology that itself requires stringent ethical oversight. On the educational side, AI can personalise learning pathways, gamify complex exercises, and provide real-time feedback. However, the deployment of AI in classrooms necessitates robust safeguards around algorithmic bias, student surveillance, and data protection to ensure children's safety is never compromised (Schiff, 2021). This duality is central to the framework, which advocates for using AI to teach about AI, while simultaneously demanding ethical transparency in its application.

6. Policy Recommendations and Curriculum Development

- **Policy Recommendations:** Mandate age-appropriate digital safety and data literacy education as a core component of national school curricula. Require digital platforms, services, and educational technologies that process children's data to incorporate "Privacy by Design" principles, ensuring that data protection is the default setting, not an afterthought (Cavoukian, 2009). Establish clear regulatory oversight for the ethical use of AI systems in educational settings, with specific guidelines on data governance, transparency, and accountability.
- **Curriculum Strategies:** Develop tiered learning modules tailored to different age groups. Integrate practical, hands-on exercises like the "Data Diary" and collaborative OSINT games. Use culturally resonant storytelling, such as the proposed Panchatantra 2.0, to explain abstract legal rights and ethical principles. Provide comprehensive training for teachers and engagement resources for parents to create a supportive ecosystem for children's digital safety.

7. Conclusion and Future Directions

The proposed AI-enhanced digital literacy framework offers a comprehensive, ethically grounded, and culturally resonant approach to enhancing children's digital safety. This paper has argued that traditional digital literacy, which often prioritizes technical skill, is no longer sufficient to prepare children for the complex realities of a data-driven world. In response, our framework establishes a new paradigm centered on the indispensable pillars of understanding personal data trails, developing a critical awareness of al-

gorithmic implications, acquiring knowledge of essential legal rights under frameworks like India's DPDP Act and GDPR, and fostering a true sense of data ownership and agency.

A cornerstone of our approach is 'Panchatantra 2.0', an innovative pedagogical strategy that demonstrates how the rich heritage of storytelling can be reimagined to make abstract digital concepts accessible and ethically resonant. This culturally grounded method bridges the gap between timeless moral reasoning and contemporary digital challenges, offering a more granular and child-centric approach than existing high-level frameworks. The implications of this work extend far beyond the classroom; it is a blueprint for cultivating a new generation of empowered digital citizens who can navigate misinformation, protect their autonomy, and assert their rights. This represents a crucial shift from a strategy of mere protectionism to one of proactive empowerment, fostering critical consciousness over passive consumption of technology.

This research serves as a call to action for educators, policymakers, technology developers, and families to collaborate in building a world where children are equipped not just to use technology, but to command it with awareness and responsibility.

References

- Cavoukian, A. (2011). *Privacy by Design: The 7 Foundational Principles*. Information and Privacy Commissioner of Ontario. https://www.datatilsynet.no/globalassets/global/bilder/rettigheter-og-plikter/innebygd-personvern/7foundationalprinciples_annacavoukian2.pdf
- Digital Education Council. (2023). *AI Literacy Framework*. Retrieved from <https://www.digitaleducationcouncil.com/ai-literacy-framework>
- Stoilova, M., Livingstone, S. & Khazbak, R. (2021). *Investigating the risks and opportunities for children in a digital world: A rapid review of the evidence*. UNICEF Office of Research – Innocenti. Retrieved from <https://www.unicef.org/innocenti/media/5621/file/UNICEF-Investigating-Risks-Opportunities-Children-Digital-World-2021.pdf>
- Haven, K. (2007). *Story proof: The science behind the startling power of story*. Libraries Unlimited. DOI:10.5040/9798216019312
- Ministry of Law and Justice, Government of India. (2023). *The Digital Personal Data Protection Act, 2023 (No. 22 of 2023)*. The Gazette of India.
- Olivelle, P. (1997). *The Panchatantra: The book of India's folk wisdom*. Oxford University Press. <https://global.oup.com/academic/product/pacatantra-9780199555758?cc=in&lang=en>
- Schiff, D. (2021). Out of the laboratory and into the classroom: The future of artificial intelligence in education. *AI & SOCIETY*, 36(1), 331–348. <https://doi.org/10.1007/s00146-020-01033-8>
- UNESCO. (2022). *K-12 AI curricula: A mapping of government-endorsed AI curricula*. UNESCO. <https://doi.org/10.54675/ELYF6010>
- UNICEF. (2021). *The state of the world's children 2021: On my mind – Promoting, protecting and*

caring for children's mental health. UNICEF. <https://www.unicef.org/reports/state-worlds-children-2021>

United Nations Committee on the Rights of the Child. (2021). *General comment No. 25* (2021) on children's rights in relation to the digital environment. CRC/C/GC/25. <https://www.unicef.org/bulgaria/en/media/10596/file>

 SESSION 4.

**ETHICAL AND RESPONSIBLE AI:
ROLES AND RESPONSIBILITIES
OF STAKEHOLDERS AND REGULATION**

RETHINKING AI FOR CHILDREN: FROM TECHNOLOGICAL POTENTIAL TO HUMAN RESPONSIBILITY

VIRGINIA DIGNUM

Professor of Responsible Artificial Intelligence at Umeå University (Dept of Computing Science) and Director of the AI Policy Lab. Member of United Nations High Level Advisory Body in AI

■ <https://www.doi.org/10.65398/ANWN8379>

Abstract

As artificial intelligence systems increasingly shape the environments in which children grow, learn, and interact, the central question is no longer what AI *can* do, but whether and when it *should* be used at all. Drawing on global literature, policy frameworks, and the author's work with UNICEF, this paper explores the ethical, social, and developmental implications of AI in children's lives. It argues for a rights-based, inclusive approach to AI development, grounded in children's best interests. By first surveying the current state of AI applications in childhood and their attendant risks, we build toward a framework for responsible design and governance. We conclude with actionable policy and design recommendations that ask a fundamental question: *Is AI needed here?*

1. Introduction

Artificial Intelligence (AI) technologies are now deeply embedded in children's daily lives, through voice assistants, learning apps, health chatbots, and more. Yet, children are rarely part of AI policy debates or design processes as is evidenced in major AI guidelines such as those by the OECD,¹ the European Union,² UNESCO,³ or the, now revoked, U.S.⁴ that frequently reference general human rights but rarely mention children explicitly or consider their specific needs in AI governance.

These realities demand not only inclusion but urgent ethical scrutiny: what does it mean for children to grow up under the influence of AI systems they neither helped design nor can fully understand – systems that may not

¹ <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

² <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

³ <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

⁴ <https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/>

only exclude them, but also misrepresent, profile, or actively harm them through biased classifications, opaque decisions, or intrusive interventions? This paper seeks to move beyond the binary of AI as either innovation or intrusion and instead frame it as a matter of human responsibility. We begin by surveying existing research on the opportunities and risks AI poses to children across development, equity, education, and health. Building on this foundation, we examine the ethical and design challenges that AI introduces into childhood contexts and propose a values-led approach to AI governance. Ultimately, we argue that AI should be deployed only when it demonstrably upholds the rights and dignity of children.

2. Background

Artificial intelligence is reshaping childhood. From algorithmically curated content to AI-driven learning systems and health diagnostics, children are increasingly subject to digital systems they cannot fully comprehend or question [9, 12, 24]. Overexposure to AI systems may lead children to over-trust machines, blur distinctions between human and artificial agents, and adopt biased or commercialized worldviews embedded in training data [1]. This concern is heightened by the fact that children's unique developmental pathways make them especially vulnerable to such influences. Children are not miniature adults; their developmental trajectories are shaped by distinct cognitive, emotional, and social processes that unfold over time and are highly sensitive to environmental inputs. Moreover, children are still forming critical faculties such as judgment, skepticism, and self-regulation. This makes them particularly vulnerable to the affordances and limitations of AI systems, which often are presented as authoritative, responsive, or even relational entities. Recent literature emphasizes that AI is not a passive tool that merely supports childhood development, it is an active agent that shapes how children think, learn, interact, and understand the world [21, 27]. For example, AI-powered recommendation systems and conversational agents can influence what children see, how they reason, and even who they trust. As such, one pressing concern is that children may develop over-trust in machines, especially when AI systems present themselves as friendly, reliable, or omniscient. This can blur the distinction between human and artificial agents, undermining children's understanding of social relationships, accountability, and the difference between opinion and algorithmic inference. Additionally, when AI systems are trained on biased or commercialized data, they may subtly reinforce harmful stereotypes, materialistic

values, or culturally narrow perspectives, shaping children's worldviews in ways that are neither transparent nor easily reversible.

More worryingly, this technological transformation and its risks are not equally distributed. As UNICEF and ITU document,⁵ nearly two-thirds of the world's school-age children lack home internet access. AI-enhanced educational tools, mental health support systems, and child welfare analytics are mostly deployed in digitally privileged settings, leaving behind vast populations without access or influence. This means that the benefits of AI, whether in education, healthcare, or social support, are accessible only to a minority, and the design and data driving these tools rarely reflect the needs or realities of the global majority. This digital divide introduces a compounding inequity: not only are marginalized children excluded from potentially helpful technologies, but the systems themselves may embed assumptions that render these children invisible, misrepresented, or misjudged. And even worse, in some cases, these systems do not simply ignore children, they may actively harm them by misclassifying behaviors, reinforcing bias, or triggering inappropriate interventions, as seen in predictive child welfare algorithms or automated content moderation systems [8]. Researchers have argued that if equity and representation are not proactively embedded, AI could amplify rather than reduce social disparities [3, 16]. Training data and benchmarks remain predominantly Western, urban, and English-speaking, leading to a systemic lack of cultural and contextual relevance in how AI classifies, supports, or evaluates children globally.

In education, adaptive learning systems and intelligent tutoring technologies offer customization and scalability. Yet these systems raise questions about pedagogical oversight, transparency, and teacher displacement [20]. Studies suggest that while some AI systems can enhance learning outcomes, their long-term impact on motivation, social-emotional learning, and classroom dynamics is still under-researched. Similarly, in healthcare, AI shows promise in pediatric diagnostics, developmental screening, and mental health triage [14, 25]. But lack of clinical validation for diverse populations and concerns over data misuse remain critical barriers. Taken together, the literature paints a picture of both promise and peril. If AI is to genuinely enhance children's lives, its development must be grounded in empirical evidence, child development theory, and a commitment to justice. This re-

⁵ <https://www.unicef.org/reports/two-thirds-world-school-age-children-have-no-internet-access-home-2020>

quires not just technical safeguards but a participatory, child-centered paradigm for digital innovation.

3. Child-Centered Approaches to AI Governance

Ensuring that AI respects children's rights and developmental needs requires more than well-intentioned guidelines, it demands critical reflection on whose values are being encoded into these systems. As the push for child-centered AI governance grows, we must confront a foundational issue: the value assumptions embedded in AI models often reflect the priorities of a narrow segment of the global population. As illustrated in the 2022's Mozilla Internet Health Report,⁶ datasets used to train AI systems overwhelmingly originate from a handful of countries – most notably the United States. This concentration distorts the digital landscape: entire regions, particularly in the Global South, are effectively invisible in the data that shapes AI. As a result, the values, norms, and assumptions embedded in these systems reflect the perspectives of a narrow, Western-centric segment of the world. This “WEIRD” bias (Western, Educated, Industrialized, Rich, Democratic) [10] risks reinforcing algorithmic colonialism and marginalizing the lived experiences of most of the world's children. If AI is to support diverse childhoods equitably, value pluralism must be a foundational design and governance principle, not an afterthought.

In particular, the development of AI systems that interact with or affect children demands frameworks that are not only technically robust but also ethically grounded in child rights and developmental theory. A growing body of research emphasizes that children must be treated not merely as passive recipients of AI systems but as active stakeholders in their design and governance [11, 17]. Across domains, a common thread is the recognition of *children's agency and digital rights*. Research and international policy increasingly underscore that children have a right to participate in decisions affecting their digital lives and to access technologies that enhance rather than compromise their well-being [23].

The design of AI-powered social agents, such as the increasing use of educational or therapeutic assistants or robots, illustrates the need for *socially embedded frameworks*. These must account for children's cognitive and emotional development, emphasizing empathy, transparency, and safe interaction models [26, 6, 4]. Ensuring that these systems respect children's au-

⁶ <https://www.mozillafoundation.org/en/research/library/internet-health-report-2022/>

tonomy while supporting healthy socialization requires careful attention to design features such as embodiment, interactivity, and context awareness.

Commitment to *participatory design* has emerged as a best practice in child-computer interaction research. Involving children directly in co-design processes has been shown to improve relevance, inclusivity, and transparency in AI systems [19, 11]. Rather than tokenistic consultation, this approach seeks to center children's perspectives, enabling them to shape technologies in ways that align with their evolving capacities and lived experiences.

Equally promising mechanism is the use of *regulatory sandboxes*, that is, controlled, real-world environments in which AI technologies can be tested and evaluated for their impacts on children's rights, safety, and well-being prior to widespread deployment. These sandboxes enable iterative development with stakeholder feedback, allowing AI developers, educators, caregivers, and children to collaboratively assess appropriateness and potential harm [18, 5].

Together, these contributions underscore a shift toward AI governance models that are child-centric by design – integrating empirical insights, participatory ethics, and rights-based safeguards. Rather than simply adapting adult frameworks to child contexts, these approaches demand innovation in how we think about risk, responsibility, and agency in the digital age.

3.1. Participatory Design in Practice

Participatory design, where children are not merely consulted but actively involved as co-designers, has emerged as a cornerstone of ethical AI development for children. Moving beyond adult-centric assumptions, this approach recognizes children's evolving capacities, contextual knowledge, and lived experience as critical inputs to technology design [11, 19]. Research in child-computer interaction has demonstrated that participatory methods such as cooperative inquiry, storytelling, and iterative prototyping enhance both the usability and relevance of AI systems while fostering digital agency and critical reflection among children [6]. For example, projects like UNICEF's Innovation Fund⁷ pilots include child-led design workshops to shape technologies such as digital identity and safety tools, demonstrating that when supported appropriately, children can offer insightful perspectives on fairness, transparency, and ethical concerns in AI. However,

⁷ <https://www.unicef.org/innovation/FundCallDataAI>

true participation demands more than occasional consultation. It requires developmental adaptation, sustained engagement, and capacity-building for both children and adult facilitators to ensure that children's contributions are taken seriously and visibly influence outcomes [17].

Institutionalizing participatory design involves integrating it into AI development pipelines, including ethical review processes, funding criteria, and design team structures. As children are increasingly affected by AI systems, ensuring they co-create these systems is not just good practice – it is a matter of justice and rights. Moreover, inclusivity must be a guiding principle. Children from marginalized or underrepresented groups are often excluded due to linguistic, cultural, or infrastructural barriers. A rights-based participatory approach must therefore embed accessibility, equity, and power-sharing into the design process from the outset.

4. Toward Child-Centered AI: Design and Policy Recommendations

Building on the ethical reflections and developmental concerns raised throughout this paper, we identify a set of interlocking priorities to guide the responsible deployment of AI in children's lives. These recommendations are grounded in the recognition that AI in childhood presents a double ethical challenge: on one hand, the obligation to protect children from surveillance, misclassification, and harm; on the other, the responsibility to design systems that genuinely support and reflect their diverse needs, rights, and lived experiences. At the heart of this approach lies a foundational ethical checkpoint: not simply whether AI can be used, but whether it *should* be used at all. This question must precede technical considerations and frame the entire development process. It demands that we shift from a mindset of solutionism to one of reflective necessity. Before reaching for the AI “hammer,” we must ask whether the issue is truly a “nail”. Not all problems in children's lives require algorithmic solutions, and the presence of powerful technology should never substitute for thoughtful judgment about its appropriateness, impact, and proportionality.

Concerns about cultural bias and representational exclusion are not merely academic but must translate into concrete design and policy responses. If AI systems are to serve children equitably, they must be built on inclusive processes that proactively account for global diversity rather than retrofitting fairness as an afterthought. This means embedding value pluralism as a design principle from the outset, ensuring that training data, development teams, and governance frameworks reflect the multiplicity of childhoods across different cultural, social, and developmental contexts. In-

clusive AI for children is not just about avoiding harm; it is about enabling relevance, agency, and justice in the systems we create. Policy must therefore mandate not only transparency and oversight, but also meaningful participation from communities and children themselves in shaping technologies that affect their lives.

4.1. *Mitigating Dual Risks: Observation and Estimation*

Children's interaction with AI systems introduces a distinctive and deeply consequential set of dual risks: the risk of being constantly *observed*, and the risk of being *estimated*. These are not simply technical design flaws – they represent fundamental social, developmental, and ethical challenges that must be addressed through both governance and system design.

The first of these, *observation*, refers to the pervasive surveillance embedded in many AI-driven applications and platforms. These systems routinely track children's behaviors, preferences, locations, and interactions, often under the justification of personalization, safety, or enhanced service delivery. However, such ongoing data collection can compromise children's privacy and undermine their autonomy. More insidiously, it can normalize surveillance from an early age, habituating children to being constantly watched and disempowering them in their relationship to technology. This risks creating a digital environment in which children learn to self-censor, internalize external scrutiny, and accept asymmetrical power dynamics as a default condition of participation [15, 22].

The second, *estimation*, emerges from the use of predictive systems, where AI systems are used to infer a child's behavior, needs, or future outcomes based on their data footprint. These estimations, sometimes called “data doubles”, can shape how children are perceived and treated by educational, welfare, or health systems. When based on incomplete, biased, or context-insensitive data, such profiling can misrepresent a child's identity or potential, leading to inappropriate categorization, limited opportunities, or unwarranted interventions. Estimation risks are especially pronounced for children from underrepresented or marginalized populations. When children's identities, behaviors, or contexts are not adequately reflected in training data, AI systems cannot “see” them directly. Instead, such children are more likely to be estimated, subject to algorithmic assumptions drawn from other groups. This increases the likelihood of misclassification and exclusion, particularly in systems that aim to allocate support or assess risk [2]. In effect, the less visible a child is in the data, the more they become a

projection of someone else's profile, shaped by generalized norms that may not apply to their lived reality.

These risks intersect with the broader themes of this paper: the developmental vulnerability of children, the cultural bias in AI training data, and the limited transparency in automated systems. Importantly, they reveal that AI does not merely record or reflect children's realities, it actively constructs and constrains them through mechanisms that are often opaque and unaccountable. Crucially, these risks are not isolated but interconnected: surveillance feeds estimation, and estimation justifies further surveillance. Together, they reinforce a feedback loop in which children's realities are interpreted and constructed through opaque and potentially harmful algorithmic logics.

Protecting children from the harms of AI requires more than technical fixes; it calls for a reassertion of human judgment, dignity, and responsibility at every level of the digital ecosystem.

To mitigate the risks above, AI systems involving children must be subject to stringent ethical and developmental standards, as demanded by e.g. UNICEF. Data collection should be strictly minimized, limited to what is necessary and appropriate for the child's context. Predictive analytics in sensitive domains such as education, child protection, or mental health must be used with extreme caution, only when their fairness, relevance, and accuracy can be empirically validated. Systems must be designed with mechanisms for explainability and contestability, allowing decisions to be questioned, reviewed, and reversed. Above all, human oversight must be preserved at key decision points, ensuring that algorithmic judgments do not override contextual understanding or professional responsibility. Key safeguards should include:

- *Minimize data collection* by default, ensuring that only data strictly necessary for a given function is used, and that children's digital traces are not repurposed beyond their original context.
- *Enable explainability and contestability* in systems that affect life outcomes, especially in education and child protection, where misinference can have lasting effects.
- *Embed human oversight* and intervention points, particularly where AI outputs may lead to decisions about care, assessment, or risk.
- *Restrict use of profiling and predictive analytics* in sensitive domains unless their validity, fairness, and developmental appropriateness can be demonstrated for the child populations they target.

Ultimately, reducing the risks of observation and estimation requires acknowledging children not merely as data subjects but as rights-holding individuals whose evolving capacities demand protection, respect, and voice in digital systems. The responsibility lies with designers, regulators, and institutions to ensure AI serves children – not monitors or molds them into predefined algorithmic categories.

4.2. Centering “Question Zero”: Is AI Needed?

Too often, AI is introduced based on what technology can do, rather than what is genuinely needed. In the context of childhood, this mindset risks prioritizing efficiency, scale, or innovation over human connection, developmental appropriateness, and ethical responsibility. Before any AI system is developed or deployed for children, we must ask the foundational question: *Is AI needed here?*, also known as “Question Zero”.

This question goes beyond technical feasibility, it demands a shift from solutionism to reflective necessity [13]. It challenges designers, policymakers, and educators to consider whether a human-led, relational approach might better serve children’s needs. In many domains, from education to caregiving to mental health, peer-supported and community-driven models often outperform algorithmic solutions in promoting trust, resilience, and well-being.

“Question Zero” is not a barrier to innovation, it is a safeguard for dignity [7]. It reframes AI not as a default response but as a well-thought resort, used only when it clearly advances children’s rights, and children’s agency and development without introducing disproportionate risk. As emphasized throughout this paper, the ethical deployment of AI in childhood must begin not with the tools we have, but with the values we uphold. In this light, AI should not replace human judgment, but complement it, carefully, transparently, and only when it aligns with the best interests of the child. This perspective reflects a deeper ethical stance: that the presence of technology should never obscure the primacy of the human experience.

4.3. Design and Policy Recommendations

As argued in this paper, building AI systems that serve children justly and safely requires more than abstract ethics or voluntary codes, it demands structural commitments across both governance and design. This section outlines concrete policy and design actions that together form a roadmap for child-centered AI. These recommendations respond to the dual imper-

ative identified throughout this paper: to protect children from harm and to ensure that AI technologies genuinely enhance their rights, well-being, and opportunities. Crucially, these proposals operationalize the principle that ethical reflection must precede technical deployment – beginning with “Question Zero”, and extending through every stage of AI development, regulation, and use.

Policy Recommendations

Effective governance of AI in childhood contexts begins with legally binding safeguards that reflect the unique vulnerabilities and rights of children. Policies must be proactive, not reactive, shaping how AI systems are funded, deployed, and evaluated before harm occurs. This means integrating child-specific criteria into regulatory frameworks, data protection laws, and national AI strategies. It also means investing in the infrastructure and equity mechanisms necessary to ensure that all children – not only those in digitally privileged settings – can benefit from technology in ways that are safe, just, and developmentally appropriate. Concrete measures include:

- Mandate child rights impact assessments in all AI-related policies and funding programs.
- Expand data protection legislation with child-specific provisions, including developmental sensitivity and meaningful consent.
- Invest in equitable digital infrastructure to close access gaps, particularly in underserved regions.
- Ensure transparency and human oversight for all high-stakes AI applications affecting children.
- Align national AI strategies with international frameworks (e.g., UNICEF, OECD, UNESCO).

Design Principles

Design is not a neutral process, it encodes values, assumptions, and priorities into the systems that increasingly shape children’s lives. Responsible AI design for children must begin from a recognition of their evolving capacities, diversity, and need for protection and participation. This section outlines principles for embedding children’s rights into the very architecture of AI systems, from inclusive datasets to explainable interfaces, and from participatory co-design to developmentally appropriate safeguards. These are not optional add-ons but essential features for AI that is trustwor-

thy, empowering, and aligned with the lived realities of children globally.

- Implement “Children’s Rights by Design” as a foundational approach in AI development.
- Institutionalize participatory design practices involving children, educators, and caregivers from diverse contexts.
- Prioritize diversity in training data to reflect global childhoods and avoid systemic exclusion.
- Embed explainability, safety defaults, and opt-out mechanisms in all AI systems used by or for children.
- Provide AI literacy education for children and communities to foster informed agency and resilience.
- Establish formal mechanisms for ongoing child consultation and feedback in the AI lifecycle, beyond initial design, through advisory panels, school-based labs, or digital participation platforms.

5. Operationalizing Child-Centered AI

Translating child-centered AI principles into real-world practice requires structured, institutional, and iterative approaches. While the previous sections established the ethical foundations, the following framework outlines how these principles can be systematically implemented, monitored, and continuously improved across diverse contexts. Towards this, we outline practical steps for governments, developers, educators, and communities to embed children’s rights into every stage of AI development and deployment. As a preliminary proposal, it is intended as a starting point for further discussion, adaptation, and validation across different policy, cultural, and institutional settings. We welcome critique, collaboration, and refinement to ensure its relevance and effectiveness in practice.

Implementation pathways: Governments and international bodies should integrate child-specific safeguards into existing AI governance mechanisms. This includes:

- Requiring Child Rights Impact Assessments (CRIA) as a condition for funding or regulatory approval.
- Embedding “Children’s Rights by Design” in national AI strategies, modeled after existing frameworks like UNICEF’s policy guidance.
- Mandating AI audits for education, health, and welfare systems that affect children.

Phased adoption model: Recognizing that capacities vary, institutions can adopt a four-phase model for implementing child-centered AI:

- Phase 1: Awareness and capacity building: Train stakeholders on children’s rights in AI contexts.
- Phase 2: Ethical and legal alignment: Integrate rights-based criteria into procurement, standards, and governance.
- Phase 3: Participatory Development: Institutionalize child co-design and stakeholder consultation throughout AI development.
- Phase 4: Auditing and Iteration: Establish continuous feedback, monitoring, and course correction mechanisms.

Institutional embedding: To prevent child participation and oversight from becoming tokenistic, dedicated mechanisms are essential:

- Youth advisory councils within regulatory and policy bodies.
- Ethics and co-design panels embedded in EdTech and pediatric AI initiatives.
- Public registers of child-facing AI systems subject to oversight and transparency.

Monitoring and evaluation: Operational success requires clear metrics and feedback loops. Suggested indicators include:

- Percentage of AI projects with documented CRIA.
- Representation of marginalized children in design processes.
- Rate of contestation or reversal of AI-driven decisions involving children.
- Measured impact on child well-being, trust, and understanding.

Global coordination and benchmarking: To ensure progress is shared and accountable, international bodies such as UNICEF, UNESCO, and OECD should:

- Facilitate peer learning across countries.
- Maintain open benchmarking databases of child-related AI initiatives.
- Support under-resourced regions in capacity building and implementation.

6. Conclusion

The integration of AI into children's lives is not merely a technological shift, it is a societal turning point that demands deep ethical, developmental, and policy reflection. As this paper has argued, the central challenge is not how to make AI more powerful or pervasive, but how to ensure it is justified, rights-respecting, and child-centered in both design and deployment.

Protecting children from the potential harms of AI cannot be addressed through technical fixes alone. It requires a broader reorientation toward human judgment, dignity, and responsibility – from the earliest design stages to final implementation. Crucially, it demands that we ask “Question Zero”: not whether AI can be used, but whether it should be used in the first place.

Responsible AI for children is not a constraint on innovation, it is its ethical foundation. By embedding children's rights, voices, and diverse developmental needs into the core of AI governance and design, we can shape technologies that truly serve the best interests of every child – across cultures, contexts, and generations.

Upholding children's rights includes not only shielding them from harm but also amplifying their voices as co-creators of their digital futures.

References

- [1] Valentina Andries and Judy Robertson. “alexa doesn't have that many feelings”: Children's understanding of ai through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence*, 4:100100, 2023.
- [2] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2022.
- [3] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, 2021.
- [4] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robots that teach. *Science Robotics*, 1(1), 2016.
- [5] Vicky Charisi and Virginia Dignum. Operationalizing ai regulatory sandboxes for children's rights and well-being. In *Responsible AI in Practice*. Taylor & Francis, 2023.
- [6] Vicky Charisi, Laura Malinverni, Marie-Monique Schaper, and Elisa Rubegni. Creating opportunities for children's critical reflections on ai, robotics and other intelligent technologies. In *Proceedings of the 2020 ACM interaction design and children conference: extended abstracts*, pages 89–95, 2020.
- [7] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer, Cham, 2019.
- [8] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, 2018.
- [9] Dagmar Mercedes Heeg and Lucy Avraamidou. Young children's understanding of ai. *Education and Information Technologies*, pages 1–24, 2024.
- [10] Joseph Henrich. *The WEIRDest People in the World: How the West Became*

- Psychologically Peculiar and Particularly Prosperous*. Farrar, Straus and Giroux, New York, 2020.
- [11] Ole Sejer Iversen, Rachel Charlotte Smith, and Christian Dindler. Child as protagonist: Expanding the role of children in participatory design. In *Interaction Design and Children 2017*, pages 27–37. Association for Computing Machinery, 2017.
- [12] Philip Hui Li and John Chi-Kin Lee. Ai, brain, and child: navigating the intersection of artificial intelligence, neuroscience, and child development, 2025.
- [13] Simon Lindgren and Virginia Dignum. Beyond ai solutionism: toward a multidisciplinary approach to artificial intelligence in society. In *Handbook of critical studies of artificial intelligence*, pages 163–172. Edward Elgar Publishing, 2023.
- [14] Hannah Lonsdale, Ali Jalali, Luis Ahumada, and Clyde Matava. Machine learning and artificial intelligence in pediatric research: current state, future prospects, and examples in perioperative and critical care. *The Journal of Pediatrics*, 221:S3–S10, 2020.
- [15] Deborah Lupton and Ben Williamson. The datafied child: The dataveillance of children and implications for their rights. *New Media & Society*, 19(5):780–794, 2017.
- [16] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, 2018.
- [17] Rahmonova Sanoat Shuhrat Qizi, Toshpulatova Shakhnoza Shukhratovna, and Musayeva Amina Karamatovna. Implementation of education and protection of children’s rights in the age of technology. *SPAST Reports*, 1(7), 2024.
- [18] Inioluwa Deborah Raji, Andrew Smart, Rebecca White, et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.
- [19] Janet C. Read and Tilde Bekker. The nature of child computer interaction. In *Proceedings of the International Conference on Interaction Design and Children*, pages 1–8, 2011.
- [20] Neil Selwyn. The future of AI and education: Some cautionary notes. *European Journal of Education*, 57(4):542–547, 2022.
- [21] Gazal Shekhawat and Sonia Livingstone. Ai and children’s rights: a guide to the transnational guidance. *Parenting for a Digital Future*, 2023.
- [22] Emmeline Taylor. The rise of the surveillance school. In *Routledge Handbook of Surveillance Studies*, pages 225–231. Routledge, 2012.
- [23] The Alan Turing Institute. Mapping study on the rights of the child and artificial intelligence: Legal frameworks that address AI in the context of children’s rights. Technical report, November 2024. Approved by the Steering Committee for the Rights of the Child (CDENF) during its 9th plenary meeting, Strasbourg, 28–30 May 2024.
- [24] Simone van der Hof and Eva Lievens. The child’s right to informational self-determination in the digital age: Challenges and perspectives. *Human Rights and International Legal Discourse*, 12:26–44, 2018.
- [25] Gerrit van Schalkwyk. Artificial intelligence in pediatric behavioral health. *Child and Adolescent Psychiatry and Mental Health*, 17(1):38, 2023.
- [26] Joachim von Braun, Margaret S Archer, Gregory M Reichberg, and Marcelo Sánchez Sorondo. Ai, robotics, and

- humanity: opportunities, risks, and implications for ethics and policy. *Robotics, AI, and Humanity: Science, Ethics, and Policy*, pages 1–13, 2021.
- [27] Ying Xu, Yenda Prado, Rachel L Severson, Silvia Lovato, and Justine Cassell. Growing up with artificial intelligence: Implications for child development. In *Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence*, pages 611–617. Springer Nature Switzerland Cham, 2024.

PROTECT, PROVIDE, PARTICIPATE: CHILDREN FIRST IN THE ARTIFICIAL INTELLIGENCE ERA

JOSIAH KIARIE KIMANI

AI for Science MSc Student, African Institute for Mathematical Sciences (AIMS), South Africa, josiah@aims.ac.za

■ <https://www.doi.org/10.65398/FHVB7362>

Abstract

Artificial Intelligence (AI) is rapidly transforming different sectors and landscapes of our lives across the globe, offering both unprecedented opportunities and profound risks especially for children (Elliott, 2019). This paper, drawing on my experiences as a youth and AI for Science MSc Student, explores the insights and ethical imperatives of protecting, providing for, and enabling the participation of children in the AI era. I examine the duality of AI – its capacity to enhance safety, education, and creativity/innovation, as well as its potential to harm and exacerbate risks such as privacy violation, exploitation, and manipulation. I propose actionable recommendations for a children-first approach by emphasizing the need for robust safeguards, equitable access, and youth involvement in AI governance and policy-making. This perspective aligns with the outcomes of the 2025 Vatican conference *Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children*, held at Casina Pio IV in the Vatican City in March 2025, advocating for collaborative, ethical, and inclusive AI development that promotes and prioritizes the dignity and well-being of every child.

1. Introduction

It is important to be cognisant of how AI is reshaping and influencing childhood by providing opportunities for growth as well as risks that particularly, if not addressed are disastrous. (Zhou, 2025) Childhood in the current AI era is different in ways unimaginable a decade ago. AI's impact on children is multifaceted as the effect is age-group specific spanning early childhood development to adolescence. From personalized learning platforms to AI-powered safety tools, children are increasingly interacting with intelligent systems that influence their education, socialization, and well-being. However, these technological advances also introduce complex risks-ranging from privacy breaches to psychological harm that disproportionately affect young users.

This paper articulates a youth perspective on how we can protect, provide for, and empower children in the AI era, in addition to the conference outcomes and current research and development.

2. Protect: Designing AI to Safeguard Children

AI must be designed with proactive, robust safeguards to ensure the safety and dignity of children. As a young African researcher, I observe that children across Kenya, the continent, and the world increasingly interact with AI-driven systems – particularly through social media, educational platforms, and digital games. These systems, while offering benefits, expose children to threats such as cyberbullying, online grooming, data exploitation, and manipulative content.

Recommendation algorithms, for instance, often prioritize engagement and commercial metrics over user well-being, inadvertently exposing young users to harmful material or addictive behaviours (Montag & Diefenbach, 2023). Social media platforms must be held accountable for the unintended consequences of AI-driven personalization that prioritize profits over protection.

Protective strategies must include:

- **Transparent algorithms:** Developers should publish the objectives and consequences of recommendation engines, especially those interacting with minors.
- **Safety-by-design principles:** Platforms must integrate child safety features as core functionalities, not as optional add-ons.
- **Real-time harm detection:** AI can be leveraged to detect predatory behaviours or cyberbullying patterns and alert guardians or moderators promptly. Griffeye Analyze DI Pro and Project Arachnid are examples of how AI can be used in law enforcement to distinguish real vs AI-generated CSAM (UNICRI, 2024) as well as detect new CSAM variants (PAS, 2025).
- **Privacy by Design:** Children’s data must be protected through stringent privacy measures, minimizing data collection and ensuring informed consent. Regulatory frameworks should mandate data minimization and clear parental controls.
- **Strict assessments for educational AI tools:** EU AI Act (2024) requires watermarking AI-generated content to combat deepfakes, promote transparency and accountability.

- **Regular monitoring and evaluation of AI systems:** Ethical frameworks that guide AI development should mandate regular monitoring and evaluation of systems for potential harm.

Globally, and particularly across the Global South where regulatory frameworks may be weaker, these protections are vital to prevent disproportionate harm to vulnerable populations (Pontifical Academy of Sciences, 2025). As AI evolves, continuous adaptation of protective frameworks remains crucial to balance innovation with child safety.

3. Provide: Equitable Access and Inclusive Innovation

Beyond protection, AI must provide opportunities for all children to thrive. In regions like Africa, AI holds the potential to democratize access to quality education, healthcare, and social services. Adaptive learning systems can personalize education for children in remote or under-resourced communities, and assistive technologies can empower children with disabilities (Innocenti, 2023).

At AIMS, we see firsthand how AI can transform education for African students, enabling access to world-class resources and personalized learning experiences. However, the promise of AI can only be realized if intentional efforts are made to bridge the digital divide. Without deliberate investment, AI risks exacerbating inequalities between children with abundant resources and those without. Reliable internet access, affordable devices, and culturally relevant content remain scarce for millions of African children.

Proposals for ensuring equitable provision include:

- **Development of low-bandwidth AI educational tools:** Tailored for rural areas with limited connectivity.
- **Public-private partnerships:** Governments and technology companies must collaborate to subsidize digital infrastructure in underserved regions.
- **Open-access resources:** Sharing of educational AI models under open licenses to promote inclusivity.

If designed and deployed equitably, AI can be a force for good, expanding children's access to knowledge, health services, and creative expression regardless of geography or socio-economic status.

4. Participate: Empowering Children and Youth in Shaping AI

Children and young people are not passive users of technology – they are active agents who experience, adapt, and shape digital realities. However, they are often excluded from decision-making spaces where technologies that impact their lives are designed and regulated.

Youth participation must move beyond tokenism. As digital natives, young people bring invaluable insights into how AI systems affect mental health, social relationships, and learning. Involving youth in AI ethics committees, governance bodies, and product design processes ensures more relevant, ethical, and user-centred outcomes (5Rights Foundation, 2024).

Key strategies to foster meaningful youth participation include:

- **Institutionalizing youth participation in advisory boards** within AI research institutes, technology companies, and policy frameworks.
- **Integrating digital and AI literacy programs** in education curricula that empower children to critically engage with AI. UNICEF’s AI Literacy Toolkit teaches 12-17-year-olds to identify deepfakes through interactive simulations (UNICEF, 2020).
- **Funding youth-led innovation projects** that explore ethical, inclusive AI development. A notable example is the (World, 2025) AI for Good Scholarship 2025, offered by The Brandtech Group in partnership with One Young World, which selects and funds five exceptional young innovators each year who are using AI to address social equity, environmental sustainability, and the responsible use of technology. Similar efforts, such as the Commonwealth AI Innovation Fund proposal and (UNESCO, 2023) UNESCO’s Global Youth Grant Scheme, also provide grants, mentorship, and capacity-building for youth-led projects, emphasizing the importance of empowering young people to create AI solutions that are ethical, inclusive, and aligned with sustainable development goals.

My own academic journey at AIMS – a pan-African network nurturing STEM excellence – has demonstrated the transformative power of empowering young people. Extending such opportunities globally can ensure that children’s voices are not an afterthought, but a driving force in shaping ethical AI.

5. Conclusion: A Common Commitment for a Child-First Digital Future

The rapid evolution of AI demands that we act decisively to centre children's rights and well-being at the heart of technological innovation. Protecting children requires proactive safeguards; providing for them demands equitable and inclusive deployment; enabling their participation ensures that AI serves rather than dominates humanity's future.

Echoing the Vatican conference's final call to action, I emphasize that protecting, providing, and promoting participation are not optional luxuries – they are urgent ethical imperatives. As a young scientist, I commit to contributing to an AI future that upholds human dignity, fosters opportunity, and empowers every child to flourish in the digital age.

As we move forward, let us remember that the true measure of technological progress is not in what it makes possible, but in how it serves the most vulnerable among us—especially children.

AI must serve children – not the other way around.

References

- (UNICEF), U.N. (2020). *Virtual consultation workshop report: AI and children in East Asia and the Pacific*. United Nations Children's Fund (UNICEF), Office of Global Insight and Policy.
- (UNICRI), U.N. (2024). *AI for Safer Children*. <https://unicri.org/topics/AI-for-Safer-Children>. UNICRI.
- 5Rights Foundation. (2024). *The Digital World of Children: Policy Recommendations for AI Development*. <https://5rightsfoundation.com/>
- Elliott, A. (2019). *The culture of AI: Everyday life and the digital revolution*. Routledge. London: Routledge.
- Innocenti, U. (2023). *AI and Children: Opportunities, Risks and Policy Recommendations*. Florence, Italy: UNICEF Office of Research-Innocenti. Retrieved from <https://www.unicef-irc.org/>
- Malgieri, G. (2024). Vulnerability in the EU AI Act: building an interpretation. Available at SSRN 5058591.
- PAS. (2025). AI enhances detection of child abuse cases. *Neuroscience News*. <https://neurosciencenews.com/ai-child-abuse-detection-28741/>
- UNESCO. (2023). *Financing youth innovation for green solutions*. <https://www.unesco.org/en/articles/financing-youth-innovation-green-solutions>
- World., T.B. (2025). *AI for Good Scholarship 2025: Empowering young innovators to address global challenges*. <https://opportunitiesforyouth.org/2025/04/28/ai-for-good-scholarship-2025-empowering-young-innovators-to-address-global-challenges/>
- Zhou, X. (2025). Innovation and Reconstruction of Early Childhood Education Models Driven by Artificial Intelligence Technology. *Journal of Artificial Intelligence Practice*, ISSN 2371-8412 Vol. 8 Num. 1.

ETHICAL RESOURCES FOR HELPING ARTIFICIAL INTELLIGENCE BENEFIT THE COMMON GOOD

BRIAN PATRICK GREEN, PHD

Director of Technology Ethics - Markkula Center for Applied Ethics - Santa Clara University, USA

■ <https://www.doi.org/10.65398/EZWL9307>

In her speech to the Academy, Marianna Rozo-Paz, mentioned the importance of hope, and of not giving in to hopelessness. This caught my attention, because the day before that she had asked me “What gives you hope?” After pondering a moment, I replied: “What gives me hope is that all of humanity’s problems can be solved.” And all of humanity’s problems can be solved – we have the technology, the skill, the power, and more. But there’s just one problem: we have to want to solve them.

Therefore, our problem is not one of capacity, but of motivation. Motivation requires believing that one is in a hopeful situation, not a hopeless one. We have to believe that we are in a situation that can actually be solved in a good way, because if we are hopeless we become demotivated and in-activated, and then we have already lost – lost before we have even tried.

Indeed, Pope Francis chose hope as the theme of the 2025 Jubilee Year. In 2015 the Jubilee theme was mercy, but now he diagnosed that hope is what humanity needs most.

What does hope have to do with ethics? I assert that ethics gives hope.

Ethics provides hope in several ways, for example, by building social trust, by building social community and fellow-feeling, and also in a particular way that I try to promote: by taking seemingly intractable ethical problems and making them tractable through the “tools” of applied ethical thinking.

Here is a thought experiment. Imagine a different world, one where everyone just does the right thing all the time, with no prompting. Free individuals would simply always choose to do the right thing, with no external regulation of behavior necessary. With a few other changes, this world might seem heaven-like.

However, we do not live in such a world. In our world, people make mistakes and even intentionally do wrong, and so we do need external forms of regulation to govern behavior. Brought into the very specific context we are examining here; this means that the technology industry cannot be perfect on its own. In fact, many leaders in the industry are not really

trying. Self-governance – just choosing to do the right thing – has failed. And if self-governance fails, then higher levels of regulation should step in.

However, following the principles of subsidiarity, self-regulation is still something of an ideal that ought to be pursued, because the people closest to a problem are often the ones best placed to solve it. Some or even much of the regulation can come from within the tech industry itself, from among those who are willing, whether individual employees, leaders, or institutionalized practices. Government can even mandate self-regulation, as some professional societies do, like medical doctors, lawyers, and some engineers.

But if self-regulation of behavior – or any regulation of behavior – is ever to have any hope of succeeding, people need to have the right tools at their disposal for considering and solving their own ethical dilemmas, and for providing guidance to others who may ask them for help. My work at the Markkula Center for Applied Ethics at Santa Clara University is to give people the practical, useful, resources that they need in order to make better decisions in whatever context they find themselves in – but particularly in the technology industry, and even more particularly with AI.

For example, the Markkula Center Framework for Ethical Decision Making is a process for taking any ethical issue and making it tractable, and hopefully resolvable in a way that not only makes sense to many people, but also is actually good.¹ It has five steps:

1. Recognize the ethical problem – what values are at risk?
2. Get the facts about the case – like in a courtroom, the facts matter
3. Evaluate options through six ethical lenses: human rights, justice, utilitarianism, common good, virtue ethics, and the ethics of care
4. Mentally test the option you think is best – can it respond to challenges?
5. Implement your decision in the best way that you can.

The Framework is an easy tool for almost any ethical situation. But if you are working in technology, there are some other more specific tools that can help too.

¹ Markkula Center for Applied Ethics, “A Framework for Ethical Decision Making,” *Markkula Center website*, November 8, 2021, <https://www.scu.edu/ethics/ethics-resources/a-framework-for-ethical-decision-making/>

For those actually developing technologies, the Ethics in Technology Practice resources provide best practices, case studies, and a seven-part toolkit for thinking about ethics when designing technology. The seven tools are ethical risk sweeping, ethical post-mortems and pre-mortems, expanding the ethical circle, case-based analysis, remembering the benefits of creative work, thinking about the terrible people, and closing the ethical loop.² This toolkit was integrated at some Alphabet companies in 2018.³

Next, diving deeper into the corporate context, we have three World Economic Forum case studies on ethical practices at Microsoft,⁴ IBM,⁵ and Salesforce.⁶ These cases, from 2021 and 2022, are one snapshot in time at these companies, describing how they did corporate tech ethics at that time. Indeed, in the years since writing these cases much has changed in the tech industry, and not all of it for the better, as competition, employee layoffs and disempowerment, and changing political winds have altered the corporate landscape.

Thanks to these cases being developed, as well as many other resources, in 2023 the Markkula Center published *Ethics in the Age of Disruptive Technologies: An Operation Roadmap*.⁷ Known more briefly as *The ITEC Handbook*, this is a resource for companies wanting to integrate ethical thinking about

² Shannon Vallor, Brian Green, and Irina Raicu, “Ethics in Technology Practice,” *Markkula Center website*, June 22, 2018, <https://www.scu.edu/ethics-in-technology-practice/>

³ Kent Walker, “Google AI Principles updates, six months in,” *Google Blog*, December 18, 2018, <https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/>

⁴ World Economic Forum and Markkula Center for Applied Ethics at Santa Clara University. “Responsible Use of Technology: The Microsoft Case Study,” *The World Economic Forum website*, February 2021, <https://www.weforum.org/whitepapers/responsible-use-of-technology-the-microsoft-case-study>

⁵ World Economic Forum and Markkula Center for Applied Ethics at Santa Clara University. “Responsible Use of Technology: The IBM Case Study,” *The World Economic Forum website*, September 2021, https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_The_IBM_Case_Study_2021.pdf

⁶ World Economic Forum and Markkula Center for Applied Ethics at Santa Clara University. “Responsible Use of Technology: The Salesforce Case Study,” *The World Economic Forum website*, September 2022, https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_Salesforce_Case_Study_2022.pdf

⁷ Jose Roger Flahaux, Brian Patrick Green, and Ann Gregg Skeet, *Ethics in the Age of Disruptive Technologies: An Operational Roadmap (The ITEC Handbook)*, Markkula Center for Applied Ethics, 2023, www.scu.edu/institute-for-technology-ethics-and-culture/itec-book-pdf

technology from top to bottom in their organization. Like the Framework for Ethical Decision Making, *The ITEC Handbook* has five steps, which we characterize as going on a road trip:

First, leadership must choose to go on this journey: to integrate ethics into their corporation. Without this step, without leadership on board, ethics efforts are doomed to failure.

Second, you need to figure out where you are: do a corporate ethical culture assessment. This assessment gives you a measure of what employees think about ethics and culture in your organization, as well as a baseline for beginning work.

Third, you must decide where you want to go: establish your goals/destination and what ethical success looks like. This might involve principles, trainings, applications, product goals, etc.

Fourth, determine your route and start driving: implement and institutionalize changes in technology design and human resources. Anyone can say ethical words, but actually operationalizing those good intentions is an entire additional layer of very hard work.

Fifth, remain vigilant that you are making progress: get measurable metrics for improvement and move towards them.

The ITEC Handbook was written with the blessing of the Dicastery for Culture and Education, but it is argued completely secularly, as technology companies are.

As a last point, I would return to the Vatican's work on AI ethics and mention again what Cardinal Parolin mentioned previously. *Antiqua et Nova* is a major next step for the Vatican's dialogue on AI.⁸ Built upon years of reflection by Pope Francis and others, the document lays out some basics for thinking about right and wrong when it comes to AI, as well as the deeper anthropological and theological issues raised by AI. And if you read *Antiqua et Nova* and want even more, then I would suggest *Encountering Artificial Intelligence: Ethical and Anthropological Investigations*, written by the AI Research Group of the Centre for Digital Culture if the Dicastery for Education of the Holy See.⁹

⁸ Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, "Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence," *The Holy See website*, January 28, 2025, https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_dcf_doc_20250128_antiqua-et-nova_en.html

⁹ Matthew J. Gaudet, Noreen Herzfeld, Paul Scherz, Jordan J. Wales, eds., and the AI Research Group of the Centre for Digital Culture of the Dicastery for Culture and Edu-

These books and tools might seem like slim grounds for hope, but these tools for making ethics tractable have themselves gotten traction with those developing new technologies. Some of these tools have been scaled up at companies totaling 160,000 employees, and others have been read by US congressmen; so I think there might be at least some grounds for hope. And I hope that you might find the tools useful – and even hopeful – too.

Acknowledgements

The author would like to thank the Pontifical Academy of Sciences and its wonderful leadership and staff, its supporters and patrons, and all of the attendees for their support and interest in ethics.

References

- Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, “Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence,” *The Holy See website*, January 28, 2025, https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_dcf_doc_20250128_antiqua-et-nova_en.html
- Flahaux, Jose Roger, Brian Patrick Green, and Ann Gregg Skeet, *Ethics in the Age of Disruptive Technologies: An Operational Roadmap (The ITEC Handbook)*, Markkula Center for Applied Ethics, 2023, www.scu.edu/institute-for-technology-ethics-and-culture/itec-book-pdf
- Gaudet, Matthew J., Noreen Herzfeld, Paul Scherz, Jordan J. Wales, eds., and the AI Research Group of the Centre for Digital Culture of the Dicastery for Culture and Education, *Encountering Artificial Intelligence: Ethical and Anthropological Investigations*, Theological Investigations of Artificial Intelligence Book Series, Vol. 1, Pickwick / Wipf & Stock, Eugene, Oregon, 14 December 2023, <https://jmt.scholasticahq.com/article/91230-encountering-artificial-intelligence-ethical-and-anthropological-investigations>
- Markkula Center for Applied Ethics, “A Framework for Ethical Decision Making,” *Markkula Center website*, November 8, 2021, <https://www.scu.edu/ethics/ethics-resources/a-framework-for-ethical-decision-making/>
- Vallor, Shannon, Brian Green, and Irina Raicu, “Ethics in Technology Practice,” *Markkula Center website*, June 22, 2018, <https://www.scu.edu/ethics-in-technology-practice/>
- Walker, Kent, “Google AI Principles updates, six months in,” *Google Blog*, December 18, 2018, <https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/>
- World Economic Forum and Markkula Center

- for Applied Ethics at Santa Clara University. "Responsible Use of Technology: The Microsoft Case Study," *The World Economic Forum website*, February 2021, <https://www.weforum.org/whitepapers/responsible-use-of-technology-the-microsoft-case-study>
- World Economic Forum and Markkula Center for Applied Ethics at Santa Clara University. "Responsible Use of Technology: The IBM Case Study," *The World Economic Forum website*, September 2021, https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_The_IBM_Case_Study_2021.pdf
- World Economic Forum and Markkula Center for Applied Ethics at Santa Clara University. "Responsible Use of Technology: The Salesforce Case Study," *The World Economic Forum website*, September 2022, https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_Salesforce_Case_Study_2022.pdf

RISKS AND OPPORTUNITIES OF AI FOR CHILDREN: A COMMON COMMITMENT FOR SAFEGUARDING CHILDREN

**Statement by the Pontifical Academy of Sciences,
Casina Pio IV, Vatican City, March 24, 2025**

Summary

The Pontifical Academy of Sciences, in collaboration with the World Childhood Foundation, chaired by Queen Silvia of Sweden, and the Institute of Anthropology of the Pontifical Gregorian University, convened a high-level workshop to address the opportunities and risks AI presents for children. The conference, held at Casina Pio IV in the Vatican on March 21-22, 2025, included political leaders, AI expert scientists, youth representatives, United Nations and international organizations, civil society, corporate leaders, survivors of abuse, and church officials. While AI has the potential to enhance safety, education, and well-being, it also introduces serious risks.

Key Risks AI Poses to Children:

- AI-Driven problematic behavior & Mental Health Issues: Algorithms maximize engagement, leading to excessive screen time, anxiety, and sleep disorders.
- Privacy Violations & Data Exploitation: AI systems collect children's personal data without stringent safeguards.
- Manipulation & Behavioral Targeting: AI-powered advertising influences children's consumption habits, often without them realizing it.
- Social & Emotional Development Concerns: Over-reliance on AI chatbots and AI-powered toys may limit real-life social interactions.
- Sexual Exploitation & Grooming: AI has been misused to generate child exploitation materials and facilitate online grooming.
- Educational Risks: Excessive digital exposure negatively impacts attentional, linguistic, and cognitive processes, as well as academic performance.

AI's Positive Potential for Children:

- Educational Benefits: AI-driven learning tools can personalize education, expanding access to quality education in particular in underserved regions.
- Protecting from Harmful Content: AI can filter out explicit material and monitor online interactions for safety.
- Early (Mental) Health Detection: AI-powered tools and chatbots can help identify health issues, incl. signs of anxiety or depression in children.
- Enhanced Safety & Surveillance: AI can be used to detect risks in public spaces, schools, and online environments.
- Disaster or Disease Prevention: Rapid response coordination targeted on vulnerable children.
- Promoting Creativity and Innovation: Used responsibly, AI can enhance children's creativity by providing new tools for artistic expression, storytelling, and problem-solving.

A Call to Action for Ethical AI Development:

AI must be developed responsibly, prioritizing children's well-being. Stakeholders – including governments, parliamentarians, tech companies, educators, civil society, and faith-based institutions – must work together to establish and enforce ethical regulations, AI governance frameworks, and child-centered policies and interventions. Key recommendations include:

1. Strengthening international governance and regulations to protect children from AI-driven harm.
2. Developing AI systems that prioritize children's safety, privacy, and dignity.
3. Promoting AI's benefits while ensuring transparency, accountability, and fairness.
4. Elevating youth voices in AI governance discussions, and in the design and development of AI systems and products.
5. Global collaboration to ensure that AI serves, rather than endangers, the intellectual potential, dignity and well-being of every child in the digital age.

Risks and Opportunities of AI for Children: A Common Commitment for Safeguarding Children

The Pontifical Academy of Sciences, in collaboration with the World Childhood Foundation, chaired by Queen Silvia of Sweden, and the Institute of Anthropology of the Pontifical Gregorian University, convened a high-level workshop to address the opportunities and risks AI presents for children. The conference, held at Casina Pio IV in the Vatican on March 21-22, 2025, included political leaders, parliamentarians, AI expert scientists, youth, survivors of abuse, UN and International Organizations, and leaders from civil society and the corporate sector.

Background and Objectives

Artificial Intelligence (AI) is reshaping the digital world at an unprecedented pace, profoundly influencing the lives of children. While AI holds promise in enhancing safety, education, and well-being, it also introduces significant risks. The rapid advancement of AI, accelerated by the widespread digital adoption during the COVID-19 pandemic, has exposed children to increasing online risks, including deep fakes, cyberbullying, and sexual exploitation. AI should serve as a tool to protect, provide, and promote participation, ensuring that children are safeguarded and empowered in the digital world. To address these urgent challenges and opportunities, the Pontifical Academy of Sciences, together with the World Childhood Foundation and Institute of Anthropology of the Pontifical Gregorian University, has convened a high-level workshop with representatives from academia, the technology industry, faith-based institutions, civil society, and youth. We recognize that AI must be developed and deployed with the utmost ethical responsibility to ensure that it protects – rather than endangers – children. AI systems are inseparable from digital platforms, including social media, which shape children’s social, cognitive, and emotional development. It is our moral duty to ensure that these technologies uphold the dignity, rights, cognitive and social-emotional development, and safety of every child.

Emerging Research Insights on AI Risks and Opportunities for Children

The workshop is based on new scientific insights from various relevant disciplines. Assessing the risks and opportunities of AI requires interdisciplinary cooperation and close interaction between sciences and other stakeholders.

AI Risks for Children

AI can be misused in ways that directly put children's safety, cognitive potential, and well-being at risk. It has been exploited to generate and disseminate child sexual abuse material, facilitate online grooming, and enable human trafficking. AI-driven surveillance, while often implemented for safety, can unintentionally infringe upon children's right to privacy, fostering a culture of excessive monitoring that undermines their autonomy and dignity. Additionally, biased AI algorithms may perpetuate existing social inequalities, disproportionately affecting vulnerable children. The increasing use of AI-driven algorithms by social media platforms has raised new concerns about online protection of children. AI algorithms designed to maximize user engagement often prioritize sensational, emotionally charged, or addictive content, which can negatively affect children's mental health and academic performance. The disproportionate impact of AI-driven practices in different groups of children must be recognized in risk patterns, particularly children from the Global South (which account for 75% of the world's children) and those with disabilities – in order to inform targeted and effective policy measures.

1. **Increased Screen Time and AI-Driven-disordered-technology-use:**

AI-powered algorithms in digital platforms, including social media, video games, and educational applications, are designed to maximize user engagement. Studies indicate that AI-driven content recommendations may lead to disordered technology use, reducing children's physical activity, sleep and academic performance. Disordered use of technology has been associated with more anxiety and depression, although discussions regarding causality between variables are still a matter of debate. Further, in the research field it is discussed if the data business model aiming at the prolongation of online time takes too much of children's time, which is needed for distinct age-dependent psycho-developmental tasks including sufficient play time and direct social interaction.

2. **AI Chatbots and Emotional Dependency:** The rise of AI-powered chatbots, such as generative AI companions, has introduced new concerns regarding children's emotional development. Over-reliance on AI-driven interactions may hinder children's ability to develop real-life social skills, resulting in lower empathy levels and difficulties with peer relationships.

- 3. Collection and Exploitation of Children's Data:** AI-driven platforms collect vast amounts of personal data, including behavioral patterns, location, and preferences. Many AI-powered apps designed for children lack stringent data protection measures, potentially exposing them to targeted advertising, identity theft, and unauthorized surveillance. Often, data are collected without proper consent.
- 4. AI-Powered Manipulation and Behavioral Targeting:** AI-driven marketing strategies are increasingly used to influence children's preferences and consumption habits which can lead to commercial exploitation. Studies show that children are particularly vulnerable to AI-powered advertising, as they may struggle to differentiate between organic content and AI-generated advertisements. AI-driven advertising increases children's likelihood of making impulsive purchasing decisions, raising ethical concerns about the manipulation of young minds. AI-generated content can mislead children, spreading fake news or inappropriate material. Deepfakes make it harder for children to distinguish real from fake, right from wrong.
- 5. AI-Powered Toys and Their Influence on Child Development:** The growing market for AI-powered toys, such as interactive robots and smart assistants, raises concerns about their impact on children's social and cognitive development. Excessive reliance on AI-based interactions may limit opportunities for real-world socialization, leading to weaker verbal communication skills and less interest in cooperative play compared to peers who interact more frequently with human caregivers and friends.
- 6. Potential Reduction in Human Interaction and Learning:** AI-driven social media algorithms personalize and filter content to maximize engagement, often limiting children's exposure to diverse perspectives. This may contribute to the development of "digital echo chambers," where children are only exposed to information that reinforces their existing beliefs. Additionally, excessive digital exposure is negatively linked to attention and academic performance in reading and writing. This could reduce critical thinking skills and hinder their ability to engage with diverse viewpoints in real-world social interactions.
- 7. Cyberbullying:** AI systems can potentially promote harmful content, including cyberbullying and exploitation, either directly or indirectly.

Since AI is designed to prioritize engagement, harmful interactions such as hate speech or harassment may be amplified. This poses a potential risk to children's safety online, leading to emotional stress and, in some cases, long-term psychological trauma. Moreover, AI-driven platforms frequently promote content that fuels negative comparisons, fostering a distorted sense of self-worth. This is particularly concerning for adolescents, who are still developing their self-identity and emotional resilience. The algorithms can reinforce unrealistic beauty standards, materialism, or toxic behaviors, leading to heightened feelings of inadequacy or anxiety. This problematic behavior with social media also can interfere with sleep patterns, physical activity, and face-to-face social interactions, all of which are vital for a child's healthy development.

8. **Sexual Abuse and Exploitation:** AI has also been exploited by predators to facilitate online sexual abuse and exploitation of children. Through AI-powered chatbots, deepfakes, and virtual worlds, offenders can prey on vulnerable young people. Moreover, AI can be used to generate explicit content, such as deepfake pornography, which can involve the non-consensual use of minors' images. The proliferation of such content increases the risk of exploitation, and the ability of AI to create hyper-realistic images and videos without the need for direct physical contact makes it difficult for authorities to trace and stop offenders. These developments complicate efforts to protect children from online sexual abuse.
9. **Broader AI Risks for Youth:** in addition to child specific risks, broad risks for children shall be pointed out too. These include catastrophic risks from advanced AI systems that could fundamentally affect children's futures, including global security risks or alignment failures in increasingly autonomous systems. AI will also transform economies, potentially creating severe disruption to job markets that today's children will enter.

AI Opportunities for Children

AI also presents opportunities to protect and empower children. Advanced AI systems can detect and prevent online risks in real time, identifying harmful behavior and removing illegal content. In regions with limited access to traditional education and child protection services, particularly in the Global South, AI offers innovative solutions to expand access to learning, health care, and social support.

1. **Protection from Harmful Content:** AI can filter out harmful content and monitor online interactions to protect children from exposure to inappropriate material. For instance, AI algorithms can detect and block explicit content, cyberbullying, or hate speech before children encounter it. Platforms like YouTube have already implemented AI-based tools to remove harmful content, and similar systems can be adapted for other social media platforms. These tools help protect vulnerable children from being exposed to explicit material or violent content that may negatively impact their mental health. In high-income countries with access to advanced technology, these AI systems can be integrated into mainstream platforms to help parents and guardians monitor their children's online activity, giving them greater control over what their children are exposed to. In low-income countries, these tools can be particularly valuable in situations where children may not have parental supervision or access to safe online spaces.
2. **Early Detection of Health Issues:** AI systems can identify early signs of health issues in children and youth. By analyzing patterns in children's social media posts, online behavior, or interactions with virtual assistants, AI can help detect changes in mood or behavior that might signal depression, anxiety, or other mental health concerns. This field is also known as following a digital phenotyping or mobile sensing approach. Such health surveillance systems can then alert parents, teachers, or counselors to intervene early and provide necessary support. In low-income countries, where mental health resources are often scarce, AI could provide a cost-effective way to identify and address mental health issues in children before they become severe. AI-powered apps or chatbots could offer children emotional support, guidance, and coping mechanisms, serving as an initial resource while directing them to professional care if needed. Future AI might provide continuous personalized health monitoring and interventions tailored to each child's unique developmental trajectory.
3. **Enhanced Safety Measures and Surveillance:** In both high and low-income countries, AI can be used to improve child safety in public spaces, schools, and homes. AI-driven surveillance systems can monitor public spaces to identify potential risks or dangerous situations involving children. In schools, AI-based systems can help identify bullying,

violence, or unsafe environments, allowing for prompt intervention by school authorities. These systems use facial recognition, behavioral analysis, and predictive modeling to detect and respond to risks. Additionally, in regions with higher levels of poverty or conflict, AI can assist in tracking missing children or identifying children at risk of exploitation. By using AI-powered tools such as facial recognition, geolocation tracking, and behavioral patterns, authorities can locate and rescue children who are at risk of being trafficked or involved in child labor.

4. **Educational Benefits and Access:** AI has the potential to transform education for children in both high and low-income countries. In low-income areas, where access to quality education may be limited, AI-powered tools can provide personalized learning experiences, enabling children to learn at their own pace and receive individualized support. AI-driven educational apps can offer interactive lessons, quizzes, and feedback, making learning more engaging and accessible. AI-powered tools assist children with disabilities through speech recognition, translation, and assistive technologies. Moreover, AI can help bridge the digital divide by providing remote learning opportunities to children in underserved areas. In high-income countries, AI can enhance existing educational platforms, creating more adaptive and inclusive learning environments. By tailoring lessons to a child's learning style and progress, AI can help ensure that all children, regardless of income, receive the support they need to succeed. Advanced AI tutors could eventually provide education comparable to having the world's best educators available to every child, potentially transforming educational outcomes globally. AI could transform inclusion for children with disabilities through neural interfaces, advanced prosthetics, or cognitive assistance.
5. **Creativity and Innovation:** AI can empower children to not only express their creativity but also think critically about real-world challenges and develop innovative solutions. AI-powered tools can allow children to create interactive stories, generate music, or design digital art, expanding their ability to experiment with different forms of expression. Beyond artistic creativity, AI can help children become problem-solvers by enabling them to develop tools that test their knowledge, build interactive learning experiences, or even design AI-driven applications to address issues they care about, such as mental health, environmental sustainability, or social inclusion. For example, children can use AI to analyze environmental da-

ta, propose sustainable solutions for their communities, or develop chatbots that provide mental health support to their peers. By engaging with AI as a tool for both creativity and impact, children can cultivate critical thinking, innovation, and a sense of agency in shaping a better future. However, it is essential to guide them in using AI ethically and responsibly, ensuring that technology enhances their imagination rather than limiting independent, critical and democratic thought.

The emerging research highlights both the benefits and risks of AI's growing presence in children's lives. These impacts require a *Child Safety First* ethical approach.

A Call to Action: Ethical AI for Child Protection and Support

Any call to action in the field of AI needs to consider the fast evolving and changing nature of AI. This calls for attention to AI's transformative benefits and risks for children's futures. It is crucial that researchers, regulators, educators, and parents work together to ensure its use does benefit children and minimizes harm, ultimately creating a more secure and supportive digital world for younger generations. The development and application of AI must be guided by ethical principles that prioritize children's best interests. This requires robust regulatory frameworks, ongoing interdisciplinary dialogue, and responsible AI governance. Workshop participants propose fostering a shared commitment to ethical AI development and deployment, grounded in the following key pillars:

1. Shared Evidence-Based Understanding of the Diverse Impacts of AI on Children

1. Developing a comprehensive, evidence-based understanding of AI's role in children's safety, dignity, and development.
2. Exploring AI's influence across education, governance, child online protection, and industry.
3. Examining AI's impact on youth well-being, including cognitive, emotional, and social development.
4. Identifying both the risks and opportunities of AI for children, ensuring that technology serves as a force for good.

2. Initiatives for Research Partnership and Leadership in AI that Serves Children

5. Building bridges between corporate and public goods related AI research, overcoming constrained access to massive proprietary datasets.
6. Data trusts or cooperatives, with academic-led data commons where individuals or institutions donate or license their data under ethical governance.
7. Academic leadership in creating open, independent benchmarks that even corporate models must be evaluated against, i.e. transparency, fairness, and robustness. This can also include establishing sandboxes for testing and assessing not only the data, models and algorithms but also the regulations, standards and responsibilities.

3. Joint Commitment: Actionable Steps for Child Well-Being

8. While regulations have made significant strides in addressing privacy concerns and limiting harmful content, the challenge remains to protect children from the profound risks posed by AI-driven algorithms designed to engage them. Policymakers worldwide must continue to develop and enforce comprehensive measures that protect children's data, limit their exposure to potentially harmful content, and regulate AI-driven social media platforms when they may foster addictive behavior, and embed children's rights and ethics by design.
9. Given the global nature of AI's risks to Child Welfare, it is essential to enhance international learning about regulations and controls, as well as sharing implementation and enforcement strategies.
10. Outlining concrete measures that stakeholders – governments, tech companies, educators, religious communities, and civil society – can take to enhance child protection, leveraging AI to combat child sexual exploitation, including the detection and removal of abusive material.
11. Promoting AI-driven innovations that expand educational opportunities and improve access to reliable information for children and youth, while ensuring that reliance on such innovations does not disrupt or diminish the development of critical cognitive processes.
12. Fostering multi-stakeholder coalitions and collaborative efforts that promote concrete actions to protect children in the digital age by bringing together governments, science, tech companies, educators, civil so-

ciety, child advocacy groups, and youth, among other stakeholders, through sustained joint projects and initiatives that advance responsible AI solutions for children. These could establish mechanisms for regular monitoring, evaluation, and accountability in AI deployment, and guide standards and policies.

4. Defined Responsibilities: The Role of Stakeholders

13. Recognizing the unique responsibilities of scientists, the tech industry, policymakers, parliamentarians, civil society, religious communities (including the Catholic Church), international organizations, and young people themselves as well as parents and families in shaping ethical AI design and use.
14. Addressing the urgent need for rigorous, evidence-based research to inform AI policies that safeguard children's rights, and strengthen related criminal justice systems. Encouraging research specifically on making increasingly autonomous systems safe and aligned with human values.
15. Exploring AI's evolving role in all relevant systems connected to children for preventing child abuse, advancing education, and supporting digital health initiatives for young people.

6. Ethical and Responsible AI: The Path Forward

16. Advocating for AI policies and regulations that prioritize children's protection in the digital environment, developing sound AI standards and reviewing of related existing legislation.
17. Elevating the voices of young people in discussions on AI governance with digital ethics, and the design of AI systems.
18. Examining the opportunities AI presents for children with disabilities, enabling greater inclusion and accessibility.
19. Promoting AI-driven governance models that serve the best interests of children, ensuring fairness, transparency, and accountability in digital systems. These should adapt to Anticipatory Governance, i.e. governance mechanisms specifically designed to adapt to rapidly evolving capabilities.

As AI continues to shape the digital landscape, we must remain steadfast in our commitment to ensuring that technology serves humanity – especially

its most vulnerable members. The Pontifical Academy of Sciences calls upon all stakeholders to work collaboratively, guided by moral responsibility and ethical principles, to harness the power of AI for the good of all children. Only through shared commitment and decisive action can we ensure that AI safeguards the dignity, safety, and well-being of every child in the digital age.

LIST OF PARTICIPANTS

Prof. Joachim von Braun

President, Pontifical Academy of Sciences, Bonn University, Germany

His Eminence Cardinal Peter K.A. Turkson

Chancellor, Pontifical Academy of Sciences, Vatican City

His Eminence Cardinal Pietro Parolin

Secretary of State, Vatican City

Her Majesty Queen Silvia of Sweden

Her Royal Highness Princess Madeleine of Sweden

Msgr. Dario E. Viganò

Vice Chancellor, Pontifical Academy of Sciences, Vatican City

Ms. Sulyna Abdullah

Chief, Strategic Planning and Membership and Special Advisor to the Secretary-General International Telecommunication Union (ITU)

Ms. Lynn-Everdene Akisarl

Sub-Saharan Africa, Health Workforce, Emergencies, Mental Health

Ms. Yalda Aoukar

Co-founder and Managing Partner, Bracket Foundation

Prof. Dr. Cecile Aptel

Deputy Director, UNICEF Innocenti Global Office of Research and Foresight

Dr. Preetika Banerjee, PhD

Student in Health Economics and Outcomes Research, University of Washington; Research Fellow at the DTH-Lab

Ms. Leanda Barrington-Leach

Executive Director, 5Rights Foundation

Ms. Jacqueline F. Beauchere

Global Head of Platform Safety, Snap Inc.

Mr. Kenneth Bengtsson

Chair World Childhood Foundation

Hon. Brando Benifei MEP

Rapporteur of the EU AI Act, President of the EU Parliament-US Delegation

Mr. Charles Mingo Bennström

Communications Manager, World Childhood Foundation

Prof. Ernesto Caffo

President, Chair Professor of Child and Adolescent Psychiatry,
Telefono Azzurro

Ms. Kerstin Claus

Independent Commissioner for Child Sexual Abuse Issues
(UBSKM), Federal Government of Germany

Mr. Juan Carlos Cruz

Philanthropist/Survivor Activist, Member of the
Pontifical Council for the Protection of Minors

Ms. Michelle DeLaune

President & CEO NCMEC, USA

Ms. Brigitte De Lay

Director, Prevent Child Sexual Abuse Programme

Prof. Virginia Dignum

Professor of Responsible Artificial Intelligence at Umeå University
(Dept of Computing Science) and Director of the AI Policy Lab;
Member of United Nations High Level Advisory Body in AI

Ms. Susanne Drakborg

Senior Program Manager, World Childhood Foundation

Mr. Iain Drennan

Executive Director, WeProtect global alliance

Ms. Sohayla Eldeeb

DTH Regional Youth Champion, Digital Transformation for Health Lab

Sr. Prof. Gill Goulding

Regis College, University of Toronto, Canada

Dr. Brian Patrick Green

Director of Technology Ethics, Markkula Center for Applied Ethics, Santa Clara University, USA

Ms. Paula Guillet de Monthoux

Secretary General, World Childhood Foundation

Prof. Arturo Harker Roa

Director of Imagina, Universidad de los Andes

Dr. Pedro Hartung

CEO, Alana Foundation, Brazil

Ms. Julia Hiller

Office of the Independent Commissioner for Child Sexual Abuse Issues (UBSKM), Federal Government of Germany

Ms. Britta Holmberg

Deputy Secretary General, World Childhood Foundation

Mr. Martin Hullin

Director, Digitalisation and the Common Good, Bertelsmann Stiftung, France

Ms. Ida Jallow

Senior Liaison Officer, ITU, Switzerland

Dr. Salman Khan

Regional Youth Champion for Central and Southern Asia, Digital Transformation for Health Lab (DTH-Lab)

Mr. Josiah Kiarie Kimani

AI for Science MSc student, African Institute for Mathematical Sciences (AIMS), South Africa

Prof. Jürgen A. Knoblich

Academician, Pontifical Academy of Sciences, Scientific Director, Institute of Molecular Biotechnology, Vienna, Austria

Ms. Saanika Kodial

Brave Movement, Mumbai, India

Ms. Imane Lakbachi

Digital Transformations for Health Lab (DTH-Lab), Focal Point at the WHO Youth Council and Director of Network Engagement at IYAFP Casablanca, Morocco

Mr. Sean Litton

President and Chief Executive Officer, Tech Coalition

Mr. Guillaume Landry

Executive Director, ECPAT International

Mrs. Jane Leek

Child Protection Programme Lead, Porticus/Philanthropy, Netherlands

Hon. Neema Lugangira, MP Tanzania

Chair of the African Parliamentary Network on Internet Governance, Chair of the Parliamentary Network on World Bank and IMF, UN IGF MAG Member

Mr. Nicolas Makharashvili

Director Safe Futures Hub / SVRI

Dr. Javier Medina Vásquez

Deputy Executive Secretary of the UN Economic Commission for Latin America and the Caribbean, Chile

Prof. Dr. Christian Montag

Centre for Cognitive and Brain Sciences, Institute of Collaborative Innovation, University of Macau, Macao, China

Mr. Similoluwa Okunowo

Google DeepMind Scholar, African Institute for Mathematical Sciences (AIMS), South Africa

Ms. Erika Olsson

Project Coordinator, World Childhood Foundation

Prof. Andrew Przybylski

Professor of Human Behaviour and Technology, Oxford Internet Institute

Dr. Emilio Puccio

Secretary General, European Parliament Intergroup on Children's Rights, Board member EU Child-Participation Platform and the EU Better Initiative for Kids (BIK+)

Ms. Alexandra Rácz

Postdoctoral Student, Angelicum

Dr. Christoffer Rahm, MD, PhD, Docent

Centre for Psychiatry Research, Department of Clinical Neuroscience
Karolinska Institutet & Stockholm Health Care Services, Region Stockholm

Ms. Vicky Isabelle Rateau

Programme Officer with the Prevent Child Sexual
Abuse Programme, Oak Foundation

Prof. Sergio Rodríguez López-Ros

Pro Rector of Universitat Abat Oliba CEU de Barcelona;
Special Envoy of the Sovereign Order of Malta, Spain

Prof. Mariana Rozo-Paz

Policy, Research and Project Management Lead at the Datasphere
Initiative, #Youth4OurDataFuture Project, Colombia

Prof. Christian Schmahl

Medical Director, Central Institute of Mental Health, Mannheim, Germany

Baroness Joanna Shields, OBE

Founder WeProtect Global Alliance; CEO and Founder Precognition

Prof. Wolf Singer

Academician, Pontifical Academy of Sciences; Ernst
Struengmann Institute for Neuroscience in Cooperation
with Max Planck Society, Frankfurt, Germany

Ms. Mama Fatima Singhateh

United Nations Special Rapporteur on the Sale,
Sexual Exploitation and Sexual Abuse of Children

Mr. Vilhelm Skoglund

Founder, Astralis Foundation, Sweden

Ms. Melissa Stroebel

Vice President, Research & Insights Thorn, USA

Ms. Laura Taddei

Director of Operations and Strategy, Precognition, UK

Dr. Beck Tench

Center for Digital Thriving,
Harvard Graduate School of Education, USA

Ms. Charlotte Thibault

Regional Youth Champion DTH-Lab

Dr. Yuriy Tykhovlis

Postdoctoral researcher, Pontifical University of Saint Thomas

Mr. Sagar Vishnoi

Future Shift Labs, India

Dr. Emily Weinstein

Co-Director – Principal Researcher Center for Digital Thriving, Harvard Graduate School of Education, USA

Mr. Sebastian Wijas

Pontifical University of Saint Thomas Aquinas, Rome, Italy

Prof. Maryanne Wolf

Academician, Pontifical Academy of Sciences; Director, UCLA Center for Dyslexia, Diverse Learners, and Social Justice

Ms. Elaine Zhang

The Datasphere Initiative #Youth4OurDataFuture Project, Duke University, Durham, North Carolina, USA

Prof. Hans Zollner, SJ

Director of the Institute of Anthropology, Pontifical Gregorian University, Rome