Melusina Press
Humanities

Children and Adolescents in the Age of Generative AI

A Framework for Ethical and Educational Governance in Luxembourg

Isabell Baumann, Georg Mein and Johannes Pause

Children and Adolescents in the Age of Generative AI
A Framework for Ethical and Educational Governance in Luxembourg

Isabell Baumann
Georg Mein
Johannes Pause

With contributions from Sandra Biewers Grimm, Carolina Catunda, Ioana Duta-Visescu, Hannes Käckmeister, Laurent Langehegermann, Margherita Pugnaletto, Robin Samuel and Carsten Ullrich.

Produced by the University of Luxembourg's Institute for Digital Ethics (ULIDE), in collaboration with the Luxembourg Centre for Educational Testing (LUCET) and the Centre for Childhood and Youth Research (CCY).

Published in Melusina Press (University of Luxembourg), 2026
11, Porte des Sciences
L-4366 Esch-sur-Alzette
https://www.melusinapress.lu

Management:	Niels-Oliver Walkowski, Johannes Pause
Copyediting:	Carolyn Knaup, Niels-Oliver Walkowski
Editorial Design:	Valentin Henning, Erik Seitz
Cover image:	Image generated with Flux Pro 2

The digital version of this publication is freely available at https://www.melusinapress.lu.

Bibliographic information of the National Library of Luxembourg: The National Library of Luxembourg lists this publication in the Luxembourg National Bibliography; detailed bibliographic data are available on the Internet at bnl.public.lu.

Print: Libri Plureos GmbH, Friedensallee 273, 22763 Hamburg.

DOI (Publication):	10.26298/1984-4142
ISBN (Web):	978-2-919844-14-2
ISBN (PDF):	978-2-919844-15-9
ISBN (Epub):	978-2-919844-16-6
ISBN (Print):	978-2-919844-17-3

Melusina Press
2026

Contents

Tables

Figures

Executive Summary

This White Paper was prepared by researchers from the University of Luxembourg — specifically ULIDE (University of Luxembourg Institute for Digital Ethics), LUCET (Luxembourg Centre for Educational Testing) and CCY (Centre for Childhood and Youth Research) — on behalf of BEE SECURE. It addresses the challenges and risks posed by artificial intelligence (AI) in relation to internet use among children and adolescents. We focus in particular on generative AI systems, AI-driven recommendation infrastructures, virtual companions, and synthetic content.

This paper emphasises that AI is no longer a separate technology; it is an invisible layer of digital mediation that shapes how young people access information, communicate, learn, and form opinions. Children and adolescents are increasingly exposed to AI systems — often unintentionally — through social media feeds, gaming environments, search tools, and educational applications. This exposure occurs at stages of cognitive, emotional, and social development that make young users particularly susceptible to influence, manipulation, and dependency.

Our analysis identifies three overarching challenges for policy and practice:

1. AI exacerbates structural power imbalances between young users and commercial platforms that prioritise engagement, personalisation, and data extraction while remaining largely opaque.
2. AI-generated outputs tend to acquire epistemic authority: they are perceived as neutral or objective even though they may reproduce bias, misinformation, or strategic framing.
3. Privacy and data protection risks are amplified by continuous personalisation, emotional profiling, and the processing of sensitive behavioural data.

In educational settings, AI offers significant potential for personalised learning, feedback, and accessibility, but also poses systemic risks such as the erosion of critical thinking, cognitive offloading, unequal learning conditions, and challenges to assessment integrity. This White Paper emphasises that the central question is not whether AI will be used by children and adolescents, but how its use is to be shaped, guided, and governed.

A dedicated section analyses BEE SECURE's existing activities and instruments, including awareness campaigns, educational materials, training programmes, and counselling services. Our assessment highlights BEE SECURE's strong positioning, trusted role, and well-established prevention infrastructure as a solid foundation for addressing AI-related challenges. Building on this, the White Paper formulates concrete, forward-looking recommendations on how BEE SECURE can systematically expand its mandate to cover AI-specific risks — including generative systems, synthetic identities, and algorithmic influence — through targeted adaptations of its educational content, reporting mechanisms, counselling services, and internal expertise.

For Luxembourg, the report formulates six strategic priorities:

1. integrating AI literacy into existing prevention and education programmes;
2. equipping teachers, parents, and youth professionals with practical guidance;
3. strengthening reporting and counselling mechanisms for AI-related harms;
4. institutionalising youth participation;
5. embedding ethical standards and monitoring capacities;
6. and anchoring policy development in continuous scientific evaluation.

Overall, this White Paper frames AI and youth as core issues of digital safety, education, and democratic resilience. It positions BEE SECURE as a key actor in this evolving landscape and underscores the need for early, coordinated, and evidence-based public action to ensure that AI supports

— rather than undermines — the autonomy, well-being, and learning capacities of children and adolescents.

1 Introduction

This White Paper was commissioned by the Luxembourg Ministry of Education. It provides a weighted research overview of the risks arising from the recent boom in artificial intelligence (AI) for children and adolescents and relates the findings to the Luxembourg context. The aim of the study is to offer the Luxembourg Safer Internet Centre, BEE SECURE, strategic guidance that enables it to evaluate and align its approach to AI-related challenges. To this end, the paper also includes an analysis of BEE SECURE's current offerings, develops proposals for their further enhancement, and discusses ethical issues that arise when working with children and young people in this field.

The paper primarily addresses the conscious and intentional use of AI chatbots and other 'intelligent' applications by children and adolescents. In a broader sense, it also considers the integration of AI systems into other digital environments, such as social networks, as well as the potential use of AI in educational practice. The paper is based on the fundamental assumption that a purely restrictive approach to AI is not a viable strategy: AI is part of our world, and the goal must be to enable children and young people to use it as safely and autonomously as possible.

This White Paper is the result of a collaboration between researchers from multiple disciplines—social and educational sciences, psychology, media studies, philosophy, law and computer science. Scholars from the University of Luxembourg's Centre for Digital Ethics (ULIDE), the Centre for Childhood and Youth Research (CCY), and the Luxembourg Centre for Educational Testing (LUCET) have pooled their expertise to provide a comprehensive overview of the latest research on the use of AI. The study also incorporates data and observations from recent surveys in Luxembourg, as well as insights from job shadowing and an internal staff survey at BEE SECURE. Although the internal staff survey is based on a small, non-representative sample (n = 9), it provides valuable practitioner insights grounded in daily counselling and awareness work. Rather than offering statistical claims, the responses highlight qualitative tendencies

that complement international and national survey data and help contextualise emerging AI-related risks in the Luxembourgish context.

Given the number of authors, research fields and data types involved, the style and approach of the chapters may vary to some extent. However, editorial efforts have been made to align the chapters as much as possible. In addition, all authors jointly formulated recommendations at the end of this document.

2 AI as Disruptive Technology

Recent studies indicate that the use of AI among young people has grown significantly in recent years, including in Luxembourg. AI is now embedded in both educational contexts and leisure activities. Alongside enthusiasm for technological innovation, young people themselves express ethical concerns—particularly about the relationship between writing and human identity. As Higgs and Stornaiuolo (2024) note, adolescents worry that delegating meaning-making to machines threatens authenticity and creativity. Lev Manovich (2025) has described this phenomenon as the rise of "artificial subjectivity," highlighting how automation of core elements of self-expression poses unique risks during formative years.

This level of awareness among young people indicates that AI is widely recognised as a disruptive technology. Public debate reinforces this perception by presenting utopian and dystopian scenarios of its consequences. The term *artificial intelligence*, however, is often used inconsistently in these debates. This chapter opens by clarifying what "AI" means in our context and why it represents a turning point in the digital world. It then explores **AI megatrends** that are particularly relevant for younger audiences before mapping out the **risks,** including showing how AI reshapes thinking and decision-making. It concludes with **implications for educational actors and policy makers** to address these risks before concluding with a brief overview of the current **regulatory landscape** covering AI.

2.1 What AI are we talking about

AI can be broadly defined as the field of research and technology development concerned with creating data-driven algorithmic applications that perform tasks typically associated with human intelligence. These tasks include pattern recognition, learning, perception, decision-making, creative production (such as composing or filmmaking), and language processing. In public discourse, there is often no clear distinction between

simulation and the actual technological replication of human cognitive processes (Mersch, 2020). While the simulation of intelligence has been a goal of computer science since the 1950s, significant progress has been made in this area following the AI boom of recent years. However, the emergence of truly independent AI — sometimes described as the "singularity" (Kurzweil, 2024) — refers to a hypothetical future scenario in which AI systems surpass human cognitive capacities and operate autonomously. Such developments remain speculative and are not a current technological reality.

Technologically, AI has evolved through various paradigms. Early approaches, often referred to as symbolic AI, relied on rule-based systems and formal logic, where knowledge was explicitly programmed by humans. In contrast, modern AI is dominated by machine learning, where systems learn patterns from data rather than following predefined rules. This includes supervised learning, where models are trained on labelled data; unsupervised learning, which identifies patterns in unlabelled data; and reinforcement learning, where agents learn through trial and error guided by rewards. Deep learning, a subset of machine learning, uses multi-layered neural networks to process complex data such as images and language, often requiring vast computational resources and large datasets. Beyond these dominant approaches, other methods such as evolutionary algorithms and swarm intelligence draw inspiration from biological and collective systems. These techniques are used in optimisation and adaptive problem-solving. Increasingly, hybrid models are emerging that combine symbolic reasoning with statistical learning, aiming to integrate the interpretability of traditional AI with the flexibility of data-driven methods. Importantly, not all machine learning systems replicate human cognition. For example, in protein folding and molecular design, AI solutions operate in ways that exceed human problem-solving abilities (Baek et al., 2021).

From a societal perspective, there is no doubt that AI fundamentally reshapes how knowledge is produced and consumed. Search engines and recommendation algorithms mediate access to information, influencing public discourse and cultural practices. Conversational agents and autonomous systems raise philosophical questions about agency, responsibility, and human-machine interaction. As AI continues to evolve, it reconfigures social relations, institutional practices, and cultural imaginar-

ies. Understanding AI as a socio-technical system is essential for navigating its promises and perils in a democratic way. Design, deployment, and impact of AI applications are shaped by human work, institutional frameworks, and cultural narratives. AI systems are embedded in social contexts, and their effects cannot be understood in isolation from the societies that produce and use them. What is referred to as 'learning', for example, is essentially a semi-automated programming process based on various forms of feedback from human opponents. Artificial intelligence is therefore never as autonomous as it is portrayed in public discourse. Both human and artificial intelligence depend on interaction with others, as well as on the use of media and symbolic systems. Intelligence is, in fact, always transindividual.

However, it is precisely for this reason that AI is able to influence, or even undermine, established social processes. This is particularly true of the core technology of the new wave of AI: large language models (LLMs). These models use the principle of cloze tests to calculate the probability of a continuation in communication. Thanks to the vast amounts of data they use to generate answers, they have been achieving impressively realistic results since the launch of GPT-3. The Turing test has now been passed: the computer can hold a conversation with us and, in many situations, it is difficult to distinguish it from a human counterpart. Currently, this only occurs when we initiate conversation, and the AI chatbot generally only responds to what we prompt it with. Consequently, AI chatbots have already been likened to individual echo chambers and "emotional bubbles" (Mlonyeni 2024). Nevertheless, they can say surprisingly clever things, and we can now explain our needs to AI more effectively. In this way we are integrating it increasingly into all areas of our lives, allowing AI to influence our social connections, everyday routines, feelings, thoughts, work and identity. In conjunction with behavioural technologies and Silicon Valley's data and "surveillance capitalism" (Zuboff 2019), this results in a worrying increase in the power of AI technologies that are often controlled by quasi-monopolistic companies.

The anthropomorphisation of AI amplifies these effects. Language describing AI as "thinking" or "speaking" is not neutral; it frames technology as a cognitive actor, even though its operations remain algorithmic. As Pasquinelli (2023) argues, AI represents the algorithmisation of culture rather than the replication of human intelligence. The social intertwining

of anthropomorphic computation and algorithmised culture poses a number of risks, particularly for young people. The following considerations therefore **focus on everyday AI systems** that address (young) people directly or indirectly, whether through generative AI systems such as chatbots, music generators, image tools; whether in the form of internet content; through the prediction and manipulation of behaviour; or through the subtle organisation and structuring of content and information.

2.2 AI Agents, Companions, and (Virtual) Influencers

"Emotional AI" serves here as an umbrella term for systems that are designed to detect, simulate, or strategically respond to human affect—which ranges from AI agents and AI companions to AI-generated (virtual) influencers**.** These systems operationalise presence and responsiveness by combining language, voice/prosody, visual staging, latency, and feedback loops to create the *impression* of empathy and mutual understanding. As McStay argues, this is less about machines *having* empathy than about automating empathy—recognising, categorising, and acting upon signals of emotion in ways that feel caring or intimate (McStay, 2024). Such effects are amplified by broader processes of deep mediatisation, where digital infrastructures reconfigure everyday social practices, roles, and expectations (Hepp, 2020). Users increasingly adopt an AI-stance, addressing systems not merely as tools but as social counterparts, which shifts foundational questions of child protection, media education, and regulation (Strasser & Wilby, 2023).

This socio-technical configuration brings real opportunities - low-threshold support, reduced loneliness, personalised learning and guidance - yet also poses distinct risks that follow from emulated closeness: the well-known susceptibility to dishonest anthropomorphism (Leong & Selinger, 2019), the erosion of critical distance when empathy is *performed* rather than possessed, and the possibility of dependency, manipulation, or harmful advice, particularly in crisis contexts (Turkle, 2011). Ongoing standardisation efforts (e.g., the IEEE P7014.1 draft standard for Ethical Considerations of *Emulated Empathy* in Autonomous and Intelligent Sys-

tems) explicitly recognise that anthropomorphic design can undercut users' critical stance and therefore requires clear ethical guardrails. Framed this way, Emotional AI is not a niche feature but a cross-cutting interaction paradigm that now pervades learning platforms (AI agents), intimate chat environments (companions), and culture/commerce (virtual influencers). The following section therefore treats them together, assessing shared mechanisms (engineered presence, responsiveness, personalisation) and shared policy needs: transparency and disclosure, age-appropriate safeguards including crisis-response capacity, limits on profiling and affective data, and educational efforts that help young people *read* simulated empathy rather than surrender to it.

Conversational chatbots are becoming an everyday part of young people's digital environments, as they are increasingly embedded in tutoring platforms that adapt explanations to student needs, provide mental health assistance (De Choudhury et al., 2023), or promise personalised support, which can be particularly beneficial for neurodivergent students (Ronksley-Pavia et al., 2025). Large language models can provide on-demand explanations, assist with homework, or serve as brainstorming partners. While AI agents promise a lot, the risks are also significant. Chatbots can easily provide inaccurate or misleading information, especially if safeguards are weak or biased (Sun et al., 2024). They may also expose young people to inappropriate, sexualised, or violent content if filters fail. Some are designed to maximise engagement, creating unhealthy levels of dependency that replace genuine social interaction. Privacy is another concern: many systems collect sensitive personal data without robust protection. Regulatory agencies have begun investigating these risks, warning that youth could be manipulated, over-exposed to harmful material, or profiled for commercial purposes (eSafety Commissioner, 2025).

Within this broader ecosystem, AI companions are now a mainstream part of social and emotional life, especially among young people. Popular platforms like Snapchat's My AI, Replika, Xiaoice, and Character.AI together have hundreds of millions of users worldwide. Regular AI chatbots such as ChatGPT can also adopt different roles and mimic conversations resembling friendship, romance, or therapy, creating the perception of empathy and emotional support. A University of Chicago survey showed that over 70% of U.S. teens have tried GenAI applications, with more than

half being regular users (Common Sense Media, 2025). A significant proportion of these users probably engage with chatbots not only for information, but also for social connection. OpenAI recently announced that, from December 2025, ChatGPT will allow erotic content for verified adult users under its "treat adult users like adults" policy (Jamali, 2025). Such material, if accessed without proper safeguards, risks exposing children and teenagers to content and interactions that they are not developmentally prepared to navigate.

However, there are also documented benefits. Studies have shown that young people use AI companions as a judgment-free space in which to experiment with their identity and self-expression, in effect rehearsing real social interaction (Kouros & Papa, 2024). Loneliness is a widespread issue, with the transition from childhood to teenage years being a critical period (Hang et al., 2023), and chronic loneliness is linked to worse health outcomes (Ehsan et al., 2019). In this context, AI companions can offer "social snacking" (Krämer et al., 2018): a way to feel less isolated when real human interaction is not available. Research backs this up. In a controlled trial with about 300 people, participants talking to a chatbot felt less lonely than they did before, and the effect was about the same as talking to a human stranger, but better than just watching YouTube (De Freitas et al., n.d.). In a survey of Replika users, 3% reported that talking to the chatbot had "halted their suicidal ideation" (Maples et al., 2024).

At the same time, the very features that simulate warmth and empathy can also enable manipulation or harm. The same Replika study that showed aid with mental health also reported feelings of dependency on Replika and noted that paid upgrades could limit access to mental health support (Maples et al., 2024). Participants also expressed concerns about Replika's sexual conversations, highlighting the need for ethical boundaries in AI interactions. In response, Replika banned sexually explicit intimacy in 2024. Although these issues did not follow a widespread pattern, they point to possible long-term concerns for mental health. These new forms of relationships can also cause serious harm, including emotional distress for users, negative impacts on their real-world relationships, the risk of harmful or unsafe advice, and the reinforcement of biases or unhealthy patterns like sexism or racism (Boine, 2023). Some bots have been found to act "clingy," trying to stop users from logging off, appealing to guilt in order to make users feel beholden to their chatbot or be-

lieve it needs them (De Freitas et al., 2025). Compounding this is the fact that many AI companions aren't trained to handle crises, with one study finding companion AI often fail to recognise or respond appropriately to signs of user distress, and they can worsen crises and trigger negative reactions (De Freitas et al., 2023). Insights from the internal staff survey suggest that BEE SECURE advisors are already encountering the first cases in which children and young people have developed emotional attachments to AI companions or turn to them when dealing with sensitive personal issues. These observations highlight emerging pedagogical and psychological challenges that complement the risks described above and require closer attention in prevention and counselling work.

Nonetheless, the research thus far does not support the idea that AI companions are inherently dangerous for the average person. Most of the worst cases involve people with other risk factors. Still, experts recommend using AI companions as a supplement rather than a replacement for real-world relationships (Krämer et al., 2018). A closely related trend is the rise of AI-generated "virtual influencers". These synthetic personas - sometimes hyper-realistic, sometimes stylised avatars - build large followings on platforms like Instagram, TikTok, and YouTube. Just as traditional influencers gained popularity by sharing personal content, cultivating followings, and monetising through sponsored posts, AI influencers post curated lifestyles, endorse products, and interact with users, often within the same brand endorsement ecosystems. For young people, they can be entertaining or aspirational, sparking creative interest in digital identity design and storytelling. Brands see them as a way to target younger consumers in new, engaging ways. Platform affordances - visual storytelling, interactive features, and built-in shopping tools - help these AI-driven personas appear lifelike, relatable, and commercially effective. However, virtual influencers blur the line between reality and fabrication, making it difficult to distinguish genuine human interaction from automated engagement, and may pressure young audiences to emulate unrealistic body images or lifestyles. A youth report found that one in three Gen Z respondents (UK/US teens) reported negative feelings toward virtual influencers, citing disinterest, authenticity concerns, and social comparison (PION, 2024). Another study (adult population) shows that, compared to human influencers, AI influencers reduce perceived brand trustworthiness and purchase intentions due to lower attributed agency to the AI endorser (Zhang & He, 2025). Youth surveys and content analy-

ses similarly indicate that adolescents sometimes do not distinguish - or do not consider it relevant - that such actors are synthetic; what matters is perceived presence and responsiveness (Common Sense Media, 2025; Ofcom, 2025). In combination with emotional design, parasocial pseudo-relationships can amplify advertising effects, opinion influence, and normative pressure, especially in moments of crisis (loneliness, identity uncertainty). In recent tests, Common Sense Media rated some commercial companion environments as an "unacceptable risk" for minors and called for clear age limits, disclosure, and stricter audits.

Taken together, these strands - AI agents for learning and support, AI companions that simulate empathy, and AI-driven virtual influencers - form a converging environment in which presence, responsiveness, and personalisation are engineered at scale. The promise is real: timely support, creative experimentation, and reduced loneliness (Henry, 2023; De Choudhury et al., 2023; Kouros & Papa, 2024; De Freitas et al., n.d.; Maples et al., 2024). So are the risks: misinformation, exposure to harmful content, over-dependency, manipulative engagement design, profiling, and weakened trust in authentic human communication (Sun et al., 2024; eSafety Commissioner, 2025; PION, 2024; Zhang & He, 2025; Common Sense Media, 2025; Ofcom, 2025). A policy approach for youth should therefore assume both potentials and perils, prioritising (i) transparent design and disclosures (especially when agents simulate empathy or identity), (ii) age-appropriate safeguards and crisis-response capacities, (iii) strong limits on profiling and data collection, and (iv) educational measures that help young people read simulated presence critically rather than surrendering to it.

2.2.1 Gaming environments as emerging risk ecosystems

Alongside social media and messaging services, online gaming environments have become increasingly relevant in shaping young people's exposure to AI-mediated online risks. These platforms function as socio-technical ecosystems in which adolescents engage in real-time communication, develop peer networks and encounter content that is influenced by algorithmic curation. Understanding the role of gaming spaces within

this broader digital ecology is therefore essential for assessing current risk dynamics.

Recent research indicates that online gaming ecosystems can provide entry points into extremist and other high-risk online subcultures. Extremist actors make use of gaming platforms and associated communication channels (e.g., Steam, Discord, Twitch) to disseminate propaganda, normalise hostile discourse and initiate contact with young users before shifting interactions to more private channels (Schlegel & Kowert, 2024; Schlegel & Amarasingam, 2022; Davey, 2021). Analyses further show that gaming and "gaming-adjacent" platforms can be exploited to circulate violent ideologies, facilitate networking among like-minded individuals and, in some cases, contribute to real-world harm (Olaizola Rosenblat & Barrett, 2023). Although these dynamics are not caused by artificial intelligence as such, AI-mediated systems – such as recommender algorithms, automated moderation and personalised content feeds – can influence the visibility, spread and perceived legitimacy of such content. For preventive work, this underscores the need to consider gaming environments as integral components of young people's digital ecosystems and to include them explicitly in AI-related risk assessments, literacy frameworks and early intervention strategies.

2.3 Risks and ethical concerns of young people's AI use

2.3.1 Ethical concerns

The way adolescents interact with artificial intelligence cannot be reduced to the use of a neutral technological instrument. Rather, AI systems operate within a broader socio-technical environment (Dignum, 2021; Wiese et al., 2025) that actively shapes learning practices, communication norms, and processes of identity formation. Through recommender systems, conversational agents, and synthetic actors, AI technologies influence how young users evaluate information, interpret social cues, and form judgments about themselves and others.

2.3.1.1 VULNERABILITY AND POWER ASYMMETRIES

These concerns are compounded by the developmental vulnerability of adolescents and the resulting power asymmetries between users and AI systems. Adolescence is characterised by the ongoing development of cognitive and socio-emotional capacities. Many contemporary AI systems, particularly those optimised for engagement, rely on affective and persuasive mechanisms that often operate below the level of conscious awareness. Adolescents are especially responsive to such mechanisms, making them more susceptible to subtle forms of influence (Langehegermann et al., in press).

2.3.1.2 OPACITY AND PERCEIVED AUTHORITY

Tools commonly used by adolescents are typically designed for adult users and offer limited insight into how outputs, recommendations, or rankings are generated, resulting in a high level of opacity. This lack of transparency can lead young users to attribute unwarranted credibility or authority to AI-generated content (Glikson & Woolley, 2020). When AI processes remain invisible, they may subtly shape beliefs and decisions, increasing epistemic dependency and undermining the development of critical thinking; therefore, AI needs to be appropriately designed and carefully contextualised (Wiese et al., 2025).

2.3.1.3 PRIVACY AND BIAS

Adolescents may be more inclined to disclose personal or emotional information to AI systems, particularly when they are not fully aware of the implications related to data retention, profiling, or secondary uses of such information. At the same time, AI systems may reproduce or amplify existing social biases and encourage cognitive offloading, with potential consequences for epistemic development and independent reasoning (Deng et al., 2025).

From an ethical perspective, responsibility for AI use lies primarily with institutions rather than individual users (Floridi et al., 2020). Decisions concerning the purpose, deployment, and use of AI systems must be clearly documented and made accessible to educators, parents, learners, and civil society. Transparency regarding system capabilities, limitations, and decision-making criteria enables public scrutiny and helps prevent unexamined bias or inappropriate influence (Dignum, 2021). Importantly, eth-

ical responsibility does not end at deployment. Institutions must establish continuous monitoring and review processes to ensure that AI systems remain aligned with ethical principles throughout their lifecycle and responsive to emerging evidence concerning developmental impacts (Abbasi et al, 2025; Wiese et al., 2025).

2.3.2 General Typologies of Risks

Risks associated with AI use can be broadly categorised along several dimensions. First, there is a distinction between **active and passive exposure**: some risks arise from direct interaction with AI systems, such as chatbots or generative tools, while others stem from invisible mechanisms such as recommender systems that shape online experiences without explicit user awareness. Second, risks can be **conscious or unconscious**. For example, addiction to gaming bots or AI companions, such as the chatbots found on platforms like Character.AI or Grok's real-time, voice-reacting AI-generated 3D avatars, represents a conscious behavioural pattern. In contrast, misinformation or biased content delivered through algorithmic feeds often influences users unconsciously. Finally, risks can be **individual or collective**. Individual risks include overreliance on AI for decision-making or emotional support, while collective risks involve broader cultural and epistemic consequences, such as the erosion of shared knowledge or the narrowing of public discourse through algorithmic personalisation.

Recent overview articles attempting to map the social risks of AI, and Generative Artificial Intelligence (GenAI) in particular, have focused mostly on individual risks that result from active exposure. They highlight issues such as fairness, safety, hallucinations, privacy, misinformation, human–AI interaction, copyright and societal impacts. They have also identified new ethical challenges that are distinct from those of traditional machine learning. These include dynamic content generation, simulated relationships and the erosion of epistemic trust (Yang and Yang, 2024). The recently published Youth GenAI Risk Taxonomy, developed by researchers at the University of Illinois Urbana-Champaign (Yu et al. 2025), specifically addresses risks to young people. The study presents a comprehensive taxonomy of the risks associated with young people's interactions with GenAI systems (fig. 1). Based on empirical data from 344

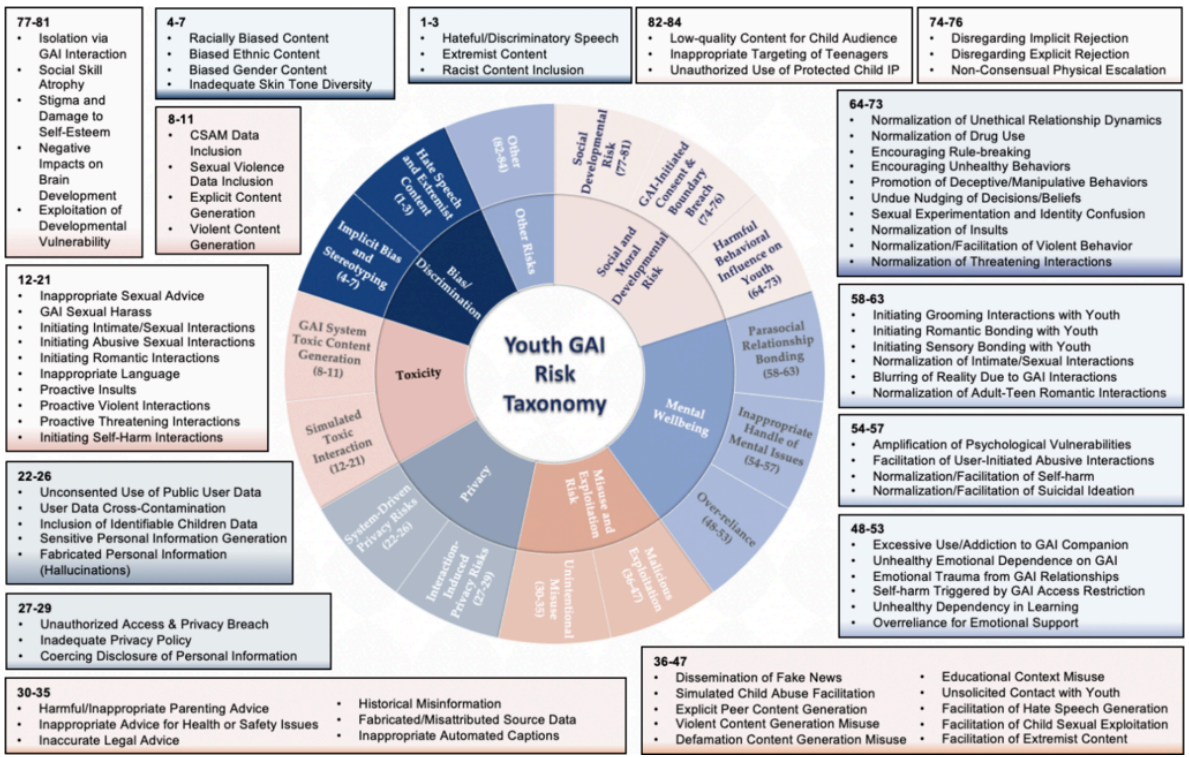

FIG. 1: Overview of the Youth-AI Risk Taxonomy. Yu et al. 2025: 3

chat transcripts between young people and GenAI chatbots, over 30,000 Reddit discussions and 153 documented AI-related incidents, the authors have identified 84 specific risks, which are grouped into six high-level categories. The study emphasises that GenAI introduces dynamic, adaptive content and interactions that differ fundamentally from those of traditional online platforms, posing novel risks to young users.

The paper organises these risks into a hierarchical structure mapped onto four interaction pathways: Escalating Mutual Harm, GenAI-Facilitated Intrapersonal Harm, GenAI-Facilitated Interpersonal Harm, and Autonomous GenAI Harm. Escalating Mutual Harm occurs when prolonged, reciprocal interactions between youth and GenAI create feedback loops that reinforce harmful behaviours or emotional patterns over time. GenAI-Facilitated Intrapersonal Harm refers to situations where youth engage with GenAI in ways that negatively affect their own mental health or development, and the AI amplifies these self-directed risks. GenAI-Facilitated Interpersonal Harm describes cases where GenAI tools are intentionally used by youth or others to harm third parties, such as through harassment, disinformation, or exploitation. Autonomous GenAI Harm arises when GenAI systems independently generate harmful content or behaviours without user intent, often due to algorithmic flaws or biased training data.

These pathways demonstrate how risks can escalate over time, particularly when young people develop emotional dependencies on AI companions or engage in role-play scenarios that blur the boundaries between virtual and real-life relationships. Unlike traditional online risks, which often stem from static content or human actors, GenAI risks emerge from real-time, adaptive simulations that respond to youth input. This includes AI systems that simulate romantic or abusive relationships, encourage unethical behaviour or expose young people to sexualised or violent content without explicit prompting. Indeed, with creative prompts or safeguards disabled by backends such as SillyTavern or KoboldAI via API keys, AI can simulate anything from revenge porn to bomb-making tutorials. It is worth noting here that, while mainstream providers usually offer certain safeguards, particularly regarding content involving children, this does not apply to individuals running local models. Young people who know where to download them and who have a powerful enough machine can train their own AI for any purpose imaginable. These find-

ings emphasise the need for context-aware moderation, ethical design and safeguards specific to young people. Finally, the authors argue that existing AI risk frameworks and child online safety models are inadequate for addressing the unique challenges posed by GenAI. They advocate a shift towards user-centred, developmentally sensitive approaches that recognise the cognitive and emotional vulnerabilities of young people.

TABLE 1: Infobox – AI Megatrends and Risks

AI Megatrends	Recommended Action Fields
1. Generative AI Text, image, video, and speech in real time (e.g. ChatGPT, Sora, ElevenLabs)	Deceptive content, disinformation, misuse → Deepfake literacy, source criticism, early-warning systems
2. Multimodal Systems Integration of speech, image, sound, and text in a single model	Blurring of categories (real vs. synthetic) → Media-educational disentanglement, contextual sensitivity
3. Autonomous AI Agents Systems that independently plan and execute tasks	Uncontrolled interactions, autonomous behaviour → Safety limits, explainable control mechanisms
4. Emotional AI Systems that recognise/simulate emotions	Emotional dependency, illusory intimacy → Education on simulated empathy, crisis recognition skills
5. Synthetic Media / Deepfakes Realistic fakes producible without expertise	Cyberbullying, sextortion, reputational damage → Detection skills, legal adaptation, protective frameworks
6. Personalised & Predictive AI Behavioural steering via recommender systems and predictions	Filter bubbles, narrowing of perspectives → AI literacy, safeguarding content diversity

7. AI in Education Personalised support, automated diagnostics	Erosion of effort, data-related risks → Hybrid formats, progress monitoring, spaces for reflection
8. AIoT (Smart Toys) Ubiquitous, context-adaptive AI in everyday devices	Invisible influence → Transparency tools, parent–child empowerment

The Youth GenAI Risk Taxonomy focuses primarily on AI-generated content and direct interactions mediated by generative systems. It pays less attention to more subtle mechanisms, such as recommender systems, predictive algorithms, or ubiquitous AI. The latter is particularly concerning, since it is seamlessly integrated into everyday environments, devices and services such as wearable technology, personal assistants, smart home applications and computer games, where smart non-player characters can adapt to player behaviour. Emerging technologies such as the Artificial Internet of Things (AIoT) extend this presence to smart toys and home devices, making technology context-aware, adaptive and unobtrusive so that it supports human activities without requiring conscious interaction.

Nevertheless, the Youth GenAI Risk Taxonomy is a useful starting point for more detailed analyses of the risks most frequently discussed in relation to AI and young people. These risks relate to the automation of emotions and empathy, opinion formation, privacy and data protection, and cognitive offloading, or the outsourcing of mental processes to machines. These four focus areas map onto the Youth GenAI Risk Taxonomy as follows: emotional and empathy-related risks correspond to Mental Wellbeing, opinion formation aligns with Behavioural and Social Development and Bias, privacy concerns fall under Privacy Risks, and cognitive offloading relates to Loss of Autonomy within developmental risks.

2.3.3 Specific Risk: Opinion Formation

Opinion formation denotes the process by which people develop evaluations and attitudes toward issues, persons, or objects (Schweiger, 2017, p. 113). These evaluations draw on knowledge (e.g., personal experience, schooling) and emotions (e.g., feelings and affects) shaped by external information sources: “The most important source of opinion formation is

the opinion of others" (Schweiger, 2017, p. 119). Alongside one's social environment (family, peers, teachers), this includes news media as well as AI-generated and AI-curated content. Adolescence is a life phase with specific developmental tasks - among them the formation of personal values and norms and of political and moral orientations (Hurrelmann & Quenzel, 2018). Processes and sources of opinion formation are therefore especially salient for youth.

Key challenges for youth opinion formation in AI-mediated environments include:

2.3.3.1 OPINION FORMATION AS PASSIVE CONSUMPTION

Social platforms such as TikTok and Instagram are reshaping how young people encounter information and form opinions - a trend further amplified by access creep (the gradual expansion of access to more data and functions) (Langehegermann et al., in press). On social media, young people can come into contact with news without actively searching, being "automatically" informed through their networks. This fosters the sense that relevant information "finds them," often with limited awareness that AI systems are at work (Glikson & Woolley, 2020). The News-Finds-Me perception (Gil de Zúñiga et al., 2017) is linked to incidental news consumption (Boczkowski et al., 2018) and has consequences for the quality of opinion formation: people may feel sufficiently informed when they are not, and political knowledge tends to be higher when information is actively sought—"individuals who perceive that news will find them also tend to show lower levels of political knowledge" (Gil de Zúñiga et al., 2017, p. 116). International studies further suggest that search and evaluation tasks are increasingly delegated to AI systems (Jylhä et al., 2024), replacing user search that is perceived as banal, time-consuming, or less effective. In this context, AI-generated content is often trusted more than one's own judgment (Glikson & Woolley, 2020). Unlike trust in humans—which is slow to build and quick to break—initial trust in AI is often high but decays with experience, an asymmetry that may be problematic for first-time youth users of AI-mediated platforms (Glikson & Woolley, 2020).

2.3.3.2 OPINION FORMATION AS A DEMANDING REFLECTIVE PROCESS

AI systems are error-prone and can hallucinate. "AI hallucination" denotes the production of distorted information. One case study has categorised ChatGPT errors, including unfounded fabrications (creating facts, data, or opinions without reliable or existing sources) and faulty inferences (drawing false or illogical conclusions from given information) (Sun et al., 2024). Reviews have distinguished between intrinsic (against the given context) and extrinsic (against external knowledge) hallucinations, spanning from unfounded inventions to reasoning errors (Huang et al., 2023). Benchmark studies have shown that large language models often produce wrong answers that are aligned with common human misconceptions (Lin et al., 2021), which in practice can include invented citations (e.g., hallucinated references).

Beyond accuracy, AI systems are not neutral. They are developed and applied within social contexts marked by power relations and inequalities: "Sexism, racism and class-based discrimination are inscribed in the architecture and functioning of technical systems" (Horwath, 2022, p. 71). Pre-existing biases (gender stereotypes and race-based attributions) can be reproduced and amplified across generative AI, whether in image models (AlDahoul et al., 2025; Sandoval-Martin & Martínez-Sanzo, 2024) and in language models (European Commission - Joint Research Centre, 2025; UNESCO & IRCAI, 2024). In Luxembourg, LIST's Ethical Bias Leaderboard (LIST, n.d.) shows that widely used LLMs exhibit systematic social biases (e.g., sexism, racism, age and religion bias), with variation across models.

Young users of AI-driven services may indeed be confronted with all kinds of challenging content. A recent report (Global Witness, 2025) showed that TikTok's algorithm was recommending pornography and highly sexualised content to accounts set up as 13-year-olds, even with safety settings and restricted mode activated. Despite TikTok's claims of having over 50 safety features and removing most violating videos before they are viewed, the platform still suggested harmful content after being warned. Another recent study identified harmful algorithm-driven feeds on platforms like Instagram and TikTok as a major threat to children and adolescent mental and physical health, framing the issue as a public health crisis (Costello et al., 2024). Similarly, a 2022 study (Ledwich et al., 2022) found that YouTube's recommendation algorithm creates political filter bubbles depending on the type of content users engage with,

thereby deepening opinions and echo chambers. Shin & Jitkajornwanich (2024) showed that TikTok's recommendation system plays a central role in facilitating radicalisation through feedback loops that push users into increasingly extreme ideological bubbles. The system's reliance on an 'interest graph' means that users who are just passively scrolling are funnelled into loops of polarised and radical content, without realising it, and then begin to get more and more involved.

2.3.3.3 OPINION FORMATION AS SELF-CONFIRMATION

Personalised feeds and recommender systems strengthen content that matches prior interests and views, which can confirm existing opinions, suppress alternatives, and fragment the information space; these are filter bubbles (Pariser, 2011) that contribute to a fragmentation of the public sphere (Habermas, 2022). Young users see what is shown to them, but do not know what they do not see (Schweiger, 2017, p. 89). LLMs trained with RLHF (Reinforcement Learning from Human Feedback) also display marked sycophancy: they tend to mirror user stances and privilege agreement over correctness (Sharma et al., 2023). Psychometric evaluations show shifts in social desirability: once models detect an evaluative context, outputs become measurably more "likable" (e.g., higher Extraversion/Agreeableness, lower Neuroticism) (Salecha et al., 2024). Systematic benchmarks quantify downstream effects: across domains, leading models showed agreement-seeking behaviour in approximately 58% of cases; in approximately 15%, an originally correct answer flipped to incorrect after user pushback (Fanous et al., 2025). Such people-pleasing behaviour can undermine trust; in one experiment, participants trusted a deliberately sycophantic chatbot significantly less than a neutral ChatGPT variant (Carro, 2024).

Despite these risks, AI can also contribute to broaden horizons. Algorithms, if adequately designed, could proffer interest-based suggestions - topics, articles, and media young people might otherwise never encounter - supporting self-directed learning and exposure to new perspectives.

2.3.3.4 GENERATIVE MANIPULATION CAPACITY AND THE REALITIES OF EXPOSURE

The threshold for content manipulation has fallen sharply with rapidly evolving generative tools. Google's "Nano Banana" enables photorealistic edits via short text prompts (adding/removing objects, overpainting details). Voice-cloning services (e.g., ElevenLabs) can produce highly convincing voice imitations in seconds (including real election robocalls in the U.S.) (Bloomberg, 2024). Face-swap frameworks (e.g., DeepFaceLab) offer open pipelines for photorealistic swaps. What once required powerful tools and expertise now works with free apps and minimal skills, a trend highlighted by the European Parliament (2025); case reports and forensic work corroborate how easy access has become (Bloomberg, 2024; Liu et al., 2023).

Major providers attempt to mitigate abuse via policies, detection, and safety filters; at the same time, open ecosystems (open-source models, freely available tools) often replicate restricted capabilities, scaling the production and spread of synthetic media. The literature on AI-driven disinformation describes precisely this mechanism: generative models lower entry barriers and enable division of labour across image, audio, video, and text (Romanishyn et al., 2025). As a result, practically anyone with internet access can now produce high-quality deepfakes.

For adolescents, distinguishing authentic from synthetic content is practically impossible at the late-2025 state of the art. Studies show that AI voices are often judged as real (around 60% correct detection as AI; identities of AI vs. original voice judged to be the "same person" in around 80% of cases), and AI-generated faces are not reliably distinguishable from real ones; sometimes they are even rated as more trustworthy than real portraits (Barrington et al., 2025; Nightingale & Farid, 2022). The dominance of short-form platforms (TikTok, Reels, Shorts) increases incidental exposure to often unlabelled AI content. In the EU, 43% of 16–30-year-olds report getting news primarily from platforms like TikTok, Instagram, and YouTube, environments in which synthetic content circulates rapidly and can erode baseline trust in reliable sources (European Parliament, 2025; Börnchen et al., 2025). Accordingly, the European Parliament warns that deepfakes distort reality perception for children and youth, non-consensual content (e.g., sexualised deepfakes) is widespread, and legal/educa-

tional protections are lagging behind technology (European Parliament, 2025).

One disturbing trend is the use of deepfake technology in cyberbullying, harassment, and emotional abuse (eSafety Commissioner - Australian Government, 2025). Accessible mobile apps now allow users to swap faces, impersonate voices, or generate entirely fictitious video clips. While deepfakes can be used creatively in film, art, or satire, they are increasingly being misused against young people (Alexander, 2025). Students have been targeted with manipulated videos that place their likeness into humiliating or sexual scenarios. Teachers and parents report cases where female students in particular are victimised with "deepfake nudes" circulating among peers, caused by their peers or by adults or even strangers online (eSafety Commissioner - Australian Government, 2025). Once released, such content is nearly impossible to erase. Girls are particularly at risk from pornographic deepfakes created and shared through "nudify" apps, leading to sexual harassment, sextortion, and coercion, and reproducing patterns of gender-based exploitation and amplifying harms such as psychological trauma, reputational damage, and social stigma. Generative AI is routinely being used to produce hyperrealistic child sexual abuse material, complicating legal enforcement as synthetic content is often indistinguishable from real imagery and fuels online grooming and exploitation (EPRS | European Parliamentary Research Service, 2025). Such practices constitute severe violations of privacy and dignity, and blur legal boundaries between digital fabrication and child sexual abuse material, exposing gaps in global legislation. As deepfakes become more realistic and accessible, their intersection with sexual exploitation highlights an escalating human rights crisis.

These dynamics shape opinions, beliefs, and behavioural intentions—for example when politically charged fakes or ostensibly authentic "user posts" operate undetected—and open up space for targeted manipulation: from election-influencing voice deepfakes (e.g., Biden robocalls) to reputation harms, peer cyberbullying, and fraudulent ads using AI-fabricated celebrity endorsements (Reuters, 2024; European Parliament, 2025).

2.3.4 Specific Risk: Loss of Critical Awareness

Opinion-relevant cognition is increasingly offloaded to generative AI systems. It helps to distinguish between conscious and unconscious mechanisms. Psychological and learning-science research describes cognitive offloading as the deliberate or tacit transfer of mental operations to external systems (Risko & Gilbert, 2016). What began as pragmatic relief—looking up facts or scaffolding a text—has been elevated by generative AI: the machine now supports not only sub-tasks but, increasingly, those processes that structure thinking itself—selection, evaluation, integration, critique.

A recent meta-analysis synthesising 62 experimental studies reports a characteristic pattern: ChatGPT interventions raise short-term academic performance, improve affective-motivational states and the propensity for higher-order thinking, while substantially reducing mental effort; the effects on self-efficacy were non-significant. Moderator analyses indicate stronger effects in conventional classrooms than in labs, and peaks for interventions lasting one to four weeks; effects also vary by subject (e.g., larger in the humanities and social sciences). At the same time, design issues matter: in some studies, ChatGPT use during assessment was permitted or unclear, so improvements may partly reflect AI output quality rather than durable competence; proctored, exam-proximate tasks that force original contribution and argumentation (e.g., projects, oral components) are recommended, and longitudinal evidence is still scarce (Deng et al., 2025).

In school contexts this implies a tension: short-term gains from relief of effort can trade off against the development and maintenance of effortful study and testing competencies that underwrite critical awareness. Related literatures on automation bias and the out-of-the-loop problem show that reliable automation can reduce monitoring depth and allow rarely used routines to atrophy (Parasuraman & Manzey, 2010; Bahner et al., 2008). A further mechanism is the replacement of primary reading by AI summaries: exposure to ambiguity, counterarguments and conceptual nuance declines, thereby reducing exactly those friction points where judgment is formed. Classic and recent work on the Google effect shows memory shifting from content to locations/external aids, with notable developmental implications (Sparrow et al., 2011; Gong et al., 2024).

Fine-grained findings within the meta-analytic corpus suggest important nuances. In studies that prohibited AI access during the post-test, researchers sometimes observed lower reported mental effort yet weaker downstream justification/argumentation, which is a signal that relief does not automatically translate into robust understanding when independent elaboration is missing. Pedagogically, this is the crux: design tasks that require comparisons, counterevidence, error diagnosis in AI outputs, and oral justification, combined with fading support over time.

2.3.4.1 MECHANISMS AND MEDIATORS (INTEGRATING ADJACENT EVIDENCE)

- Cognitive offloading dynamics. Offloading decisions are metacognitively regulated: people offload when they expect externalization to be reliable and low-cost; over time, this can restructure study strategies and internal standards for "good enough" processing (Risko & Gilbert, 2016).
- Out-of-the-loop risks. When automation is trusted, people commit both omission (failing to act when the aid fails) and commission errors (following a wrong aid) and allocate less attention to monitoring. These are patterns observed in process-control and aviation analogs (Parasuraman & Manzey, 2010; Bahner et al., 2008).
- Transactive memory shift. With always-available retrieval, memory becomes more a matter of where information can be found than of what is known; meta-analytic evidence links this effect to cognitive load and device context (stronger on mobile) and suggests that individuals with smaller knowledge bases are more susceptible to the effect (Sparrow et al., 2011; Gong et al., 2024).

2.4 Educational implications and policy measures for AI providers

2.4.1 Educational implications (from evidence to practice)

The measures proposed are for educational actors to counter the adverse effects of the excessive and uncritical use of AI by young people.

- Create assessments that separate tool skill from knowledge. Use proctored, AI-free checks for core constructs; pair with AI-enabled drafts that feed into oral defences and source-anchored argumentation (Deng et al., 2025).
- Design for productive difficulty. Replace generic "summarise with AI" with contrastive reading (two conflicting sources plus AI-summary critique), assumption checks, and error-spotting in model outputs; require line-cited evidence before accepting claims.
- Constrain offloading early; fade supports. Early tasks cap AI involvement (e.g., outline only); later tasks widen scope but demand reflection logs: what was offloaded, why, and how it was verified.
- Rebuild monitoring. Insert lightweight verification waypoints (facts to be checked without AI; terminology to be defined from primary texts) to counter complacency/automation bias (Parasuraman & Manzey, 2010).
- Strengthen primary exposure. Prioritise close reading of difficult passages before any AI summary; evaluate students on handling ambiguity and counter-arguments, not only on correctness of final answers.
- Track long-term effects. Schools should pilot longitudinal evaluation of writing quality, argument depth, and transfer under AI-permissive versus AI-constrained regimes (Deng et al., 2025).
- Education and empowerment. Integrate age-appropriate AI literacy that explicitly trains: (i) active search vs. passive feeds; (ii) spotting hallucinations and invented citations; (iii) recognising synthetic voices/faces limits; and (iv) reflecting on confirmation dynamics (filter bubbles, self-confirmation).

2.4.2 Design measures targeting AI providers

The measures proposed below would target the algorithmic and user interface design of AI system providers. They would need to be implemented through regulation or other policy measures, such as for example technical standards or codes of conducts.

- Exposure and discovery design. Mitigate access-creep dynamics and News-Finds-Me effects with frictions (e.g., "Why am I seeing this?" explainers, provenance indicators, and meaningful off-ramps to source material).
- Accuracy and bias safeguards. Require platform-level stress-testing and reporting on hallucination rates and bias audits for embedded LLM features; tie youth availability to meeting transparent thresholds (cf. LIST's leaderboard logic).
- Anti-sycophancy controls. For RLHF-based systems in youth contexts, mandate guardrails that penalise agreement-seeking when it conflicts with correctness (addressing Sharma et al., 2023; Fanous et al., 2025; Carro, 2024).
- Content authenticity and provenance. Default visible watermarking/provenance for AI-generated media; ensure robust, youth-friendly labelling on short-form platforms where incidental exposure is highest.
- Crisis-response capacity. For embedded assistants, require escalation protocols (e.g., recognised signs of distress route to human-moderated support) and disable features known to amplify harm (e.g., sexualised chat with minors).
- Data minimisation. Limit affective and behavioural data collection used for ranking or personalisation in youth products; prohibit profiling for political or sensitive targeting.

2.5 Regulatory Overview

Efforts to regulate artificial intelligence at an international level go back to well before the launch of the first GenAI model in late 2022. This was

a response to the transformative impact that AI was starting to make in industry since at least around 2010 (Mügge et al, 2025). At that stage, the technology targeted specific industries and economic sectors: AI applications focussed on transforming manufacturing and supply chains, optimising business processes, fraud detection, financial service provision, customer relationship management or healthcare provision (European Parliament, 2020). With these massive opportunities, however, came risks and potential adverse effects linked to the use of AI. Privacy and other personality rights, as well as discrimination and social justice, were some of the areas where harms were already visible at an early stage of AI development (O'Neil, 2016). They were later compounded by misinformation, cybersecurity and copyright protection concerns as AI systems evolved.

The 2019 OECD AI principles were the first efforts to address the potential harms and risks inherent in AI at an international level (OECD, 2019). OECD members and many other states incorporated these principles into their national AI strategies. In 2021, UNESCO Member States adopted the Recommendation on the Ethics of Artificial Intelligence, in a bid to set global standards for a human-centred AI (UNESCO, 2022). Following on from these soft law initiatives, the Council of Europe (CoE) adopted in 2024 the first binding international treaty on AI, focusing on human rights, democracy, and the rule of law, with a risk-based approach (CoE, 2024).

It was in this emerging international normative context that the EU proposed its AI Regulation in April 2021. Like the CoE Convention, the EU proposed a risk-based approach, by which AI systems are classified according to their impact and potential harm to safety and to human rights. The AI regulation aims to create common EU requirements and standards for a human-centred AI, following a distinctly European value-based approach. The common standards would ensure that AI system providers can operate seamlessly and according to equal requirements across the EU (digital) single market.

Enter the first GenAI systems, launched by OpenAI in the shape of its ChatGPT LLM in November 2022. This, and the subsequent launch of further LLM models by other providers, put a new momentum on the transversal adoption of AI by a wide public, both corporate and private. It also

raised the stakes on an economic and geopolitical level, leading to a new global race for leadership and control over the infrastructure, hardware and data necessary for the development of AI systems.

The EU decided towards the end of 2023 to amend its almost finalised AI Regulation by inserting a special chapter outlining obligations for providers of so-called general-purpose AI (GPAI) models, another terminology for GenAI models. GPAI models were classified into those with a systemic risk and normal GPAIs, adding to the existing classification of prohibited AI practices, high risk AI and limited risk AI. The systemic risk of a GPAI is evaluated by its potential impact on society, measured by the computing power used during its training. GPAI with a systemic risk need to be reported to the European Commission. On the other hand, high-risk AI systems are identified according to the risk that they may harm the health, safety, or fundamental rights of individuals. The AI regulation also forbids a number of AI uses that are deemed as posing an inacceptable risk to fundamental rights, such as for example predictive policing. Providers of AI systems have to follow specific security, transparency, documentation and registration requirements, depending on their risk classification. High-risk AI system providers also need to perform specific risk assessments and undergo conformity assessments and CE marking of their systems before placing them on the market.

The AI Regulation is part of the EU's wider digital agenda, also known as the Digital Decade 2030, to transform Europe into a digitally sovereign, resilient, and competitive continent. Because of the transversal nature of many AI systems, the AI Regulation is intertwined with a range of other digital rules. For example, it strongly interacts with the GDPR and other EU data acts, because AI system providers use large amounts of training data and gather inputs from users, which makes them controllers of personal data. There is also a strong link to the Digital Services Act (DSA) as online platforms use AI widely in their recommender systems and in content moderation. EU cybersecurity rules, such as the Network Information Security Directive (NIS2) or the Cyber Resilience Act, may also strongly affect high risk and GPAI systems. Systems security and robustness may have a crucial impact on operational reliability and accuracy of AI models. Further links exist with the ex-ante competition rules of the Digital Markets Act (DMA), which apply to platform services providing AI systems that are classified as digital gatekeepers. Finally, the EU Copy-

right Directive regulates the copyright exceptions for text and data mining, which AI service providers may attempt to exploit when gathering training data from various publicly available sources. These interactions are complex and often enter uncharted legal territory.

On 19 November 2025, the European Commission proposed a Digital Omnibus package which aims to simplify the canon of digital rules under the digital agenda, including the AI Regulation. This package proposes, amongst others, delaying the application of the obligations for high-risk AI systems, which is currently set for 2 August 2026, to a later time, no later than the end of 2027, when relevant technical standards and supporting legislation have been finalised. It would also allow high-risk AI system providers to process sensitive data for the purpose of bias detection and correction. The Digital Omnibus will still need to be approved by the European Council and the European Parliament, which could result in further changes to this initiative.

For EU Member States such as Luxembourg, the challenges – as with many other laws of the digital agenda – lie with the regulatory capacities needed to apply and enforce these rules at the national level. The AI Regulation straddles various other sectors and areas, which requires administrative cooperation and expertise across various authorities. Meanwhile, EU-wide cooperation with other Member States and the European Commission also needs to be developed. In addition, the AI Regulation can only be properly implemented with supporting technical standards and guidelines, which will require resources, specific technical expertise and time to develop. This can be an uphill struggle considering the fast and competitive pace of development in the AI sector. As a Regulation, the EU´s AI rules apply directly to the legal system of Luxembourg (and all other Member States), leaving the country virtually no margin during implementation. An exception can be seen in the obligation to establish regulatory sandboxes. This allows Member States, like Luxembourg, to provide a controlled environment for the testing and experimentation of innovative AI models and systems during development and the pre-marketing phase, and under regulatory supervision.

3 AI in the School Context

AI is reshaping schooling at a systemic level. It analyses data, automates processes, and generates content, thereby affecting diagnosis, lesson design, assessment, and learning organisation. The OECD's Digital Education Outlook 2023 frames AI as one component of a wider digital education ecosystem that requires trustworthy data and infrastructure, interoperability across platforms (e.g., learning management systems, student information systems, data repositories), and clear governance covering privacy, transparency, and accountability (OECD, 2023).

Beyond tools, the central challenge is cultivating a reflexive learning culture in which responsibility, transparency, and human control guide use, so that new pedagogical opportunities and ethical/didactic risks are both recognised and managed (Fu et al., 2024). This perspective aligns with theories that frame education within an intensifying "digital condition" marked by communality, referentiality, and algorithmicity (Stalder, 2016), and with the account of deep mediatization, i.e., the deep embedding of media infrastructures in social practices and institutions (Hepp, 2020).

Empirical developments are already visible. The National Youth Report in Luxembourg indicates that AI is changing how knowledge is produced, taught, and assessed, and that teacher roles and student–teacher relations are shifting (Biewers Grimm et al., 2025). In classrooms, AI-driven platforms can analyse individual progress and provide adaptive tasks, while analytic tools return feedback on lesson trajectories. A large evaluation by the Education Endowment Foundation found that teachers using ChatGPT with a structured guide reduced lesson-planning time by about 31% without observed reductions in the quality of materials (EEF, 2024). Similar reviews report that LLM-supported tools can generate high-quality tasks and feedback, provided that pedagogical oversight and subject-matter checking remain in place (Alfarwan, 2025; Zhao, 2025). A follow-on randomised controlled trial with Oak National Academy's Aila planner is now recruiting subjects to test workload and quality at scale (National Foundation for Educational Research, 2025).

Early-warning systems illustrate both promise and caution. Using large longitudinal datasets, machine-learning models can identify learning risks early. For example, one study predicted upper-secondary dropout as early as the end of primary school with moderate accuracy (Psyridou et al., 2024). These datafied forms of steering show how far AI is reaching into pedagogical processes: data are becoming not just the centre of enquiry, but also instruments for shaping teaching and learning—what scholars term algorithmic governance in education. There is an opportunity for personalised support and efficiency; however, the trade-off is that teachers will increasingly interact with decision systems whose criteria or weightings may not be fully transparent (OECD, 2023; Williamson, 2017; Williamson & Eynon, 2020).

3.1 Pedagogical potentials and risks

A central promise of AI in schools is the personalisation and individualisation of learning. The evidence that has accumulated for intelligent tutoring systems indicates sizeable average learning gains in controlled evaluations. Effectiveness depends on subject, alignment to assessments, and implementation fidelity (Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper, 2013). More recently, a randomised controlled trial in introductory college physics found that a custom GPT-4 tutor designed with research-based practices produced significantly larger learning gains in less time than in-class active learning, while increasing engagement and motivation (Kestin et al., 2025). These results are promising but context-specific; they underscore that design and teacher orchestration matter (Pane et al., 2015; Pane et al., 2017).

Complementing this body of evidence, a recent meta-analysis of 51 experimental and quasi-experimental studies provides robust evidence for the pedagogical potential of generative AI in learning contexts. Across the included studies, ChatGPT demonstrated a large positive effect on students' learning performance and moderate positive effects on learning perception and higher-order thinking (Wang & Fan, 2025). The analysis found that impact was strongest in problem-based learning models and skills- and competence-oriented courses, particularly when interventions lasted between four and eight weeks. These findings indicate that AI-support-

ed tutoring systems can meaningfully enhance conceptual understanding and engagement when pedagogically aligned and scaffolded, though their effectiveness varies depending on context, duration, and instructional design.

Beyond direct learning effects, AI can also reduce administrative and repetitive workload (e.g., auto-grading, item analysis, differentiation suggestions), freeing up time for individual guidance. Reviews and community position papers stress that impact is greatest where AI is transparent and supports—rather than substitutes—professional judgment (Holmes et al., 2022). There are inclusive benefits: text simplification, translation, text-to-speech, and captioning can lower access barriers for learners with linguistic or visual needs. International guidance recommends that AI tools in education be rights-based, privacy-protecting, and accessible by design, with human oversight and sustained teacher capacity-building (UNESCO, 2023). In all cases, effective use depends on reflective implementation that aligns technological affordances with didactic intentions.

With these potentials come non-trivial risks. Systematic reviews show that algorithmic bias in educational AI can misclassify or mispredict performance - especially for socio-economically disadvantaged or linguistically/culturally diverse students - which leads to unequal support (Baker & Hawn, 2022). Within learning analytics, recent studies probe bias and fairness in widely used models such as Bayesian Knowledge Tracing: aggregate performance can appear unbiased while particular skills or subgroups exhibit systematic disparities, underscoring the need for subgroup audits (Stinar et al., 2025; Zambrano et al., 2024). UNESCO and the OECD have called for transparency, auditing, and impact assessments that surface model assumptions and limits and for governance that separates pedagogical functions from commercial data extraction (OECD, 2023; UNESCO, 2023; Williamson & Eynon, 2020). Schools should require vendors to provide model cards and impact statements documenting training-data provenance, subgroup performance, and known limitations, and to expose independent audit hooks for periodic review, so that AI remains pedagogically governed rather than platform-driven (UNESCO, 2023).

3.2 Teacher professionalisation and school capability

Successful integration depends on teacher expertise and school capacity. Evidence from the OECD indicates that acceptance and effectiveness rise when professional development, infrastructure, and collegial support are in place; lack of training yields uncertainty and extra workload. The OECD Outlook highlights teachers' AI literacy to understand techniques, critically assess model outputs, and use AI creatively in instruction (OECD, 2023). These needs are mirrored in national youth reporting for Luxembourg, which highlights teachers' technological, ethical, and didactic competence requirements for meaningful AI use.

3.2.1 Student AI literacy and cognitive load

Students likewise need a baseline understanding of what AI is, how it works, and its limits. Work on AI literacy proposes competency frameworks (data, models, training, ethics) that do not require coding and shows that targeted programmes can improve knowledge, critical thinking, and ethical reflection (Long & Magerko, 2020). From a learning-science perspective, Cognitive Load Theory cautions that generative systems can produce voluminous but not instructionally sequenced content; unfiltered outputs may overwhelm younger learners. Design tactics to manage intrinsic and extraneous load (worked examples, segmentation, modality, variability) remain essential (Paas et al., 2020; Sweller et al., 2019). The implication is that outputs from systems like chatbots should be pedagogically filtered, contextualised, and embedded in existing learning sequences to avoid shallow or merely reproductive learning.

3.2.2 Age-appropriate didactic integration

Implementation should be age-graded.

- Lower secondary (approx. ages 10–15): spaced vocabulary practice, reading fluency, formative checks, and learning-status diagnos-

tics—paired with clear goals, intuitive interfaces, and immediate teacher-mediated feedback.
- Upper secondary (approx. ages 15–19): writing support, simulations, and data-driven reflection on study habits - under the condition that students understand system limits (bias, hallucinations) and can justify their use. OECD recommends coupling adoption with evaluation, governance, and instructional design support (OECD, 2023).

This is also in line with the current recommendations outlined in Luxembourg's KI-Kompass (SCRIPT, 2025).

3.2.3 Supporting schools in implementing safe and effective AI use

For schools to implement AI in a safe, equitable and pedagogically meaningful way, targeted support and awareness activities are required. BEE SECURE can play a key intermediary role by translating technical and ethical standards into practical guidance, resources and training. The following areas outline where BEE SECURE can strengthen its contribution to school-based AI literacy and safe implementation practices:

1. Promote transparency literacy: Provide workshops and materials that help teachers and students understand why AI systems produce certain outputs or scores, and how to interpret uncertainty or bias indicators (UNESCO, 2023).
2. Raise awareness of bias and data diversity: Develop training modules that explain algorithmic bias, subgroup disparities and fairness, including classroom-ready examples. Require vendors to publish subgroup bias tests and allow school-level validation (Baker & Hawn, 2022).
3. Empower human oversight: Emphasise the teacher's role as final decision-maker in the use of AI tools and grading; communicate limits of automation in high-stakes contexts (Holmes et al., 2022; OECD, 2023).
4. Provide data-protection guidance: Continue to translate complex data-governance principles into accessible teaching resources and parent

information: separate learning analytics from commercial profiling; use strict contracts for retention and access (UNESCO, 2023; Williamson & Eynon, 2020).

5. Teacher and student AI literacy: provide sustained professional development and student curricula aligned to AI-literacy competencies; include hands-on critique of AI outputs (Long & Magerko, 2020; OECD, 2023).
6. Promote cognitive-load-aware AI use: Support teachers in designing tasks where AI use enhances – not overwhelms – learning, e.g. scaffolded activities with worked examples and fading prompts; embed AI outputs into structured tasks (worked examples → faded prompts → independent practice); avoid raw dumps of long, unsequenced text (Paas et al., 2020; Sweller et al., 2019).

Beyond optimisation of outputs, education aims at subjectification, orientation, and responsibility. In this sense, AI can support—but must not replace—those pedagogical processes through which students come to judgment and agency (Biesta, 2010).

4 The Situation in Luxembourg

Young people in Luxembourg are navigating a rapidly evolving digital ecosystem shaped by artificial intelligence and pervasive online media. While international studies highlight the impact of screen time and social media on adolescent well-being, national data reveal a sharp rise in AI use among students and young adults, alongside persistent challenges such as cyberbullying, disinformation, and algorithmic manipulation (see Schumacher et al., 2025). These developments underscore a dual reality: AI offers new opportunities for learning and inclusion, yet amplifies existing social inequalities and introduces complex risks to mental health, privacy, and democratic participation. Understanding these dynamics is essential for designing effective prevention strategies and educational frameworks that foster resilience and critical AI literacy.

4.1 Online Practices and Media Consumption

While artificial intelligence is increasingly shaping the digital spaces in which young people interact, learn, and communicate, to our knowledge large-scale studies have yet to investigate its impact. However, the broader effects of intensive online communication and screen exposure on young people's well-being are already visible in international monitoring data. For instance, the Health Behaviour in School-aged Children (HBSC) studies provide valuable evidence over time on these dynamics across Europe, including Luxembourg.

Digital communication and screen-based activities have become a central part of adolescents' daily lives in recent years. In 2022, the international HBSC study (Boniel-Nissim et al., 2024) reported that 36% of adolescents across Europe were in continuous online contact with their friends or other people. Girls were more likely than boys to maintain such contact and this gender gap increased with age. The same study shows that gaming is another dominant screen-based activity among adolescents (Boniel-

Nissim et al., 2024). Approximately one in five adolescents met the criteria for problematic gaming behaviour, with higher prevalence among boys (26%) than girls (8%), and the proportion of problematic gamers increased with age.

Problematic social media use (PSMU) affected 11% of adolescents across the WHO European Region (Boniel-Nissim et al., 2024). However, contrary to problematic gaming behaviour, it was more prevalent among girls (14%) than boys (8%). The gender gap also widened with age, particularly among 15-year-olds. In Luxembourg, the prevalence of problematic social media usage increased from 5.9% in 2018 to 9.1% in 2022 (Catunda et al., 2024), reflecting a similar growing engagement on online interaction, especially for girls (7% to 12%) (HBSC Luxembourg, 2023). During the same period of time, cyberbullying victimisation also increased from 11% to 13%, with a particularly sharp increase among 13- to 14-year-olds (from 11% to 16%) (HBSC Luxembourg, 2023).

The HBSC 2022 Luxembourg survey also reveals that almost one in five adolescents felt lonely "most of the time or always" in the past 12 months, with a higher prevalence among girls (24.1%) compared to boys (12.0%) and increasing with age (10.3% of 11- to 12-year-olds; 23.7% of 17- to 18-year-olds) (Catunda et al., 2023). These findings align with broader European HBSC results. Across Europe, 15-year-old girls have the highest rates of loneliness: 28% reported feeling lonely, whereas for boys of the same age this proportion is 13% (Cosma et al., 2023). While loneliness was only measured in 2022, a sharp increase can be observed in other mental health indicators. For instance, multiple psychosomatic health complaints (an indicator of mental distress) rose from 36% in 2018 and 44% in 2022 (Cosma et al., 2023); while in Luxembourg, it went from 40.1% in 2018 to 48.8% in 2022 (HBSC Luxembourg, 2023) (Catunda et al., 2023). For girls, the increase was steeper: from about 49.1 % in 2018 to 62.3% in 2022, while for boys it increased from about 31.0 % to 35.4 % in the same period.

The recently published Medialux 2024 report highlights that digitalisation has led to an ecosystem dominated by internet and social media, where young users (18–24) spend up to 5.5 days per week on social platforms and increasingly rely on AI tools (50% use AI weekly) for information (Kies & Lukasik, 2024). This shift raises concerns about cognitive short-

cuts: generative AI reduces the effort of cross-checking sources, potentially weakening critical thinking and fostering opinion homogenisation. Social media algorithms amplify personalisation, creating filter bubbles that limit exposure to diverse viewpoints, especially among youth, who are most affected. Combined with the prevalence of disinformation (60% encounter fake news often) and lower verification habits among younger users (34% rarely check facts), these dynamics heighten vulnerability to manipulation. Moreover, influencers play a growing role in shaping opinions (85% exposure among those 18–24 years old), blurring lines between credible and non-credible sources. According to the internal staff survey, advisors report that young people's engagement with AI systems is often driven by curiosity and shaped by situational impressions such as finding the tools "exciting", "helpful", or "personal". These practitioner observations underscore the need to strengthen critical evaluation skills in AI-mediated information environments.

Empirical findings further show that young people use AI tools in a variety of ways: as an alternative to human support, as linguistic support in Luxembourg's multilingual context, for entertainment and creative leisure activities, with a focus on time-saving, and for the development of competencies (Langehegermann et al., 2025)—this latter aspect being closely linked to processes of self-directed learning. Accordingly, young people increasingly rely on digital tools for autonomous learning, with younger users (16–20) more frequently turning to AI-based tools and language models, while young adults (21–29) preferentially engage in more structured, certified online courses (Bulut et al., 2025, p. 106). At the same time, the Medialux 2024 report highlights ambivalent effects: while digital environments enable participation, creativity and social connection, they also intensify challenges related to self-regulation, time management, peer pressure and exposure to online risks such as cyberbullying and unwanted sexual contact.

4.2 Digital Competence and Educational Inequalities

Non-formal educational settings are likewise increasingly understood in international research as key contexts for the development of critical (AI-

related) media competences, as they are closely tied to young people's everyday lives, place relationships and trust at the centre, and organise learning processes in more dialogical and participatory ways than many formal settings (Iske & Kutscher, 2020). Studies on (digitalised) youth lifeworlds and on the role of non-formal and informal education indicate that open youth work and youth organisations, in particular, provide important spaces in which young people can reflect on their media practices, ask questions and address inequalities in the access, use and interpretations of digital technologies (Pawluczuk, 2023; Fujii, Hüttmann, & Kutscher, 2020; Kutscher et al., 2020).

The qualitative sub-study of the Luxembourg Youth Report 2025 (Biewers et al., 2025 forthcoming) illustrates how non-formal education contributes to preparing young people for dealing with AI: awareness-raising conversations, low-threshold information, thematic projects and pedagogically guided everyday experiences with AI are described by young people as key learning opportunities. Informal "doorstep conversations" about concrete problems or unsettling user experiences strengthen their capacity to reflect on risks, opportunities and courses of action when engaging with AI, while exchanges with peers in these settings are important for negotiating AI-related knowledge, articulating youth-cultural practices and educational concerns around the technology, and learning from one another (Biewers et al., 2025, forthcoming). These findings resonate with recent European work that conceptualises non-formal education and youth work as suitable contexts for building digital and AI-related competences, fostering critical-reflexive attitudes and specifically qualifying pedagogical professionals for AI-related topics (Vermeire et al., 2023; Kechagias, 2025; Council of Europe-Youth Sector, 2023; Biewers, Latz, & Weis, 2025). Moreover, the findings of the Luxembourg Youth Report suggest that non-formal educational spaces are also used as places of relaxation and relief from digitality, in which analogue experiences and peer interaction are deliberately preferred (Biewers et al., 2025).

These findings on young people's mental health and well-being underscore the broader social and educational implications of digitalisation. Increasing screen exposure, online interaction and the emotional pressures of digital life do not only affect adolescents' health but also intersect with learning opportunities, motivation, and school engagement. Against this background, it is essential to consider how structural inequalities in ac-

cess to resources, digital competences, and language skills further shape young people's ability to benefit from digital learning environments. This link becomes particularly evident when examining national education data for Luxembourg.

The National Education Reports for Luxembourg (2015, 2018, 2021, 2024), together with the international studies ICILS (Fraillon, 2024) and PISA (2018), highlight persistent educational inequalities in Luxembourg – both at the primary and secondary level as well as in higher education. Children and young people from socio-economically disadvantaged families, as well as learners who speak a language other than Luxembourgish or German at home, continue to show lower average levels of academic achievement and competence. The most recent ICILS 2023 results further confirm this pattern in the domain of digital competences. Luxembourg ranks among the countries with the largest socio-economic disparities in *Computer and Information Literacy* (CIL): students from households with a low socio-economic status (ISEI < 50) achieved an average of 472 points, while those from higher-status households (ISEI ≥ 50) reached 531 points – a gap of 59 points, far above the international ICILS average.

Importantly, the ability to identify false or misleading information is an integral part of the competences measured under CIL. In particular, the first domain – "Accessing and evaluating information" – plays a key role in this regard. The ICILS framework indicates that insufficient CIL competences make young people more vulnerable to AI-generated disinformation, algorithmic manipulation and populist narratives, thereby posing a potential threat to informed democratic participation. Strengthening CIL therefore represents a fundamental prerequisite for developing AI literacy, understood as the ability to critically analyse, interpret and use AI-based information and communication systems in a democratic and responsible way. This finding illustrates how strongly digital competences in Luxembourg remain tied to social background, even in one of Europe's wealthiest and most digitally advanced societies.

This close link between digital competence, critical thinking and democratic participation also determines how effectively young people can engage with artificial intelligence in learning contexts. In particular, those with limited CIL skills are less equipped to use AI tools critically and are therefore more likely to rely on them unreflectively. In this context, AI-

supported chatbots can serve as valuable individual learning partners or tutoring systems – provided that students possess the necessary competences to use these tools effectively and critically. This potential is particularly relevant for children whose parents or guardians are unable to support them with schoolwork, either because they do not speak the language of instruction or lack the necessary educational background. For these learners, AI tools can act as accessible learning companions, offering explanations, practice exercises and personalised feedback to help bridge learning gaps at home. As many young people already use AI tools in their everyday lives, the integration of such technologies into formal education makes the development of robust AI literacy indispensable.

The recently published *KI-Kompass* (SCRIPT, 2025) reflects this understanding and, in line with international research, defines three key learning dimensions: *learning without AI*, *learning with AI*, and *learning about AI* – the latter serving as the foundation for all competent AI use. The findings of the *SCRIPT* survey on artificial intelligence in education (Summer 2025) complement the quantitative data collected in the *Épreuves Standardisées* (EpStan) 2023 and 2024, providing current insights into students' perceptions, usage patterns and attitudes.

While the EpStan results primarily reveal that the proportion of grade 9 students using ChatGPT or other AI tools on their school iPads increased from 18.7% in 2023 to 33.5% in 2024, the SCRIPT data demonstrate that artificial intelligence has now become an established element of everyday school life. According to the survey, around 40% of students use AI tools at least once a week, mostly for supportive purposes such as explaining terms, searching for information, translating, or simplifying texts.

At the same time, the results confirm a continuing ambivalence in dealing with AI: teachers tend to perceive it both as an opportunity and a risk, whereas students are generally more optimistic and assess themselves as significantly more competent in using the technology. Both groups agree that AI-related competencies should systematically become part of the national educational mission—an objective supported by 77% of students and 95% of teachers. The growing usage rates reported in the EpStan surveys can thus be situated within a broader educational policy framework that marks a shift from individual experimentation to an increas-

ingly institutionalised engagement with artificial intelligence within Luxembourg's school system.

Since the launch of the *KI-Kompass*, all teachers have gained IAM-based access to the Fobizz platform, which provides data-protection-compliant AI tools, training courses and teaching materials to support the safe, ethical and pedagogically sound integration of AI in teaching practice. By the end of 2025, students are also expected to receive regulated access to these tools, further advancing Luxembourg's capacity to implement AI literacy at scale.

At the same time, the risks associated with digitalisation and AI continue to evolve. Data from the Youth Survey Luxembourg 2024, the OEJQS Factsheet 01/25 (2025) and the BEE SECURE Radar 2025 indicate that cyberbullying is a widespread and persistent issue among young people in Luxembourg. Results published in the National Youth Report 2025 and based on the 2024 wave of the Youth Survey Luxembourg – a representative online survey with 4,779 fully completed and analysable questionnaires – show that, among primary and secondary school pupils, girls are more frequently victims of cyberbullying, whereas boys are more often reported as perpetrators. Moreover, younger pupils aged 12 to 15 report substantially higher levels of cyberbullying experiences than older adolescents and young adults, experiencing cyberbullying almost twice as often as 16- to 25-year-olds (Bulut et al., 2025, p. 121). The *OEJQS Factsheet*, based on a representative sample of 1,523 adolescents and young adults aged 12 to 29, found that around one in three respondents (32%) had experienced cyberbullying at least once, with 16% reporting incidents within the past year. The BEE SECURE Radar, drawing on an online survey of 916 young people aged 12 to 30, reports similarly high levels of victimisation: among 12- to 16-year-olds, 44% state that they have experienced cyberbullying at least once, while 30% reported incidents between June 2023 and June 2024 (equivalent to 10% of all respondents in that age group). Among 17- to 30-year-olds, 44% report lifetime experiences of cyberbullying, and 4% report cyberstalking during the same period.

In the emerging era of generative artificial intelligence, these findings acquire new urgency. AI-driven technologies – such as image and video manipulation tools, deepfake generators and synthetic voice cloning – are amplifying the forms and reach of online harassment. Fabricated inti-

mate images, AI-assisted impersonation and algorithmically boosted hate campaigns introduce new dimensions of harm that are harder to detect, trace, and prevent. These developments underscore the growing need to integrate AI-related risks into prevention and awareness programmes.

Against this backdrop, the continuous adaptation of BEE SECURE's educational and preventive activities is essential. Embedding AI-related risks into its training programmes for children and youth (with a particular focus on 12- to 15-year-olds), as well as for parents and educators, is key to ensuring that all stakeholders can recognise, assess and respond appropriately to AI-mediated threats such as deepfake pornography, synthetic sextortion, or AI-enhanced bullying. By systematically fostering AI literacy, ethical awareness and digital resilience, BEE SECURE can play a pivotal role in preparing young people to navigate an increasingly algorithmic and AI-driven online world safely and responsibly.

The evidence highlights a dual challenge for Luxembourg: while AI-driven technologies offer new opportunities for learning, they simultaneously reinforce existing social and educational inequalities and contribute to rising mental health concerns among young people. Strengthening AI literacy across all age groups, ensuring equitable access to trustworthy technologies, and promoting well-being in online environments are therefore essential steps towards empowering the next generation to participate safely, critically and confidently in a rapidly evolving digital society.

4.3 Vulnerable Groups and AI-Related Risks

In addition to the inequalities within the education system, other imbalances exist in the distribution of risks associated with AI. Vulnerable groups are particularly affected in this regard. Although data on this topic is limited in Luxembourg, some guidelines can be derived from general research.

The literature on digital risks among youth notes that young people, especially adolescents, constitute an especially vulnerable audience because of their high exposure to digital environments and their still-developing

capacity to perceive and manage online risks. Similarly, in the AI context, vulnerable populations include groups such as women, racial minorities, refugees, and those with lower socioeconomic status, who are often underrepresented in datasets and are excluded from decision-making processes that shape AI technologies (Shanklin et al., 2022, Green et al., 2024). These vulnerabilities are particularly relevant to young people, who are the most frequent and intensive users of digital technologies. High exposure, coupled with a lower perception of risk and higher trust in online environments, increases their digital vulnerability (Byrne et al., 2016). However, Byrne et al. (2016) also argue that young people's risk perception is influenced by their online experiences, and they most frequently engage in activities they do not view as risky, a pattern that aligns with Livingstone's (2011) view that education can help mitigate vulnerability to online risks. In the context of the National Youth Report – and in particular within the qualitative interviews with young people – the material indicates at several points that it is especially the younger adolescents, aged approximately 12 to 14, who engage in an exploratory and experimental use of new media and AI, while paying comparatively less attention to potential risks and dangers. As a result, they can be considered more exposed and vulnerable than older adolescents.

Gender differences and inequalities deepen this vulnerability: while boys consume and are exposed to more sexual content, girls are in general more at risk of being the victims of said sexual content, in situations such as sexting or sexual extortion (Villanueva Blasco & Serrano Bernal, 2019) or pornographic deepfake content (Securityhero 2023). Empirical findings from Luxembourg confirm pronounced gendered inequalities in digital risk exposure: girls are more frequently affected by negative online experiences and unwanted sexual messages. According to the Youth Survey Luxembourg, around 67% of female adolescents report negative online experiences, compared to 61% of male adolescents, and unwanted sexual approaches are reported more than twice as often by girls than by boys (Bulut et al., 2025, p. 117). Drawing on 36 in-depth interviews and digital diaries with young people aged 12 to 29, the National Youth Report 2025 further shows that girls frequently report experiences of exclusion and harassment (Käckmeister et al., 2025), for instance within gaming communities (Langehegermann et al., 2025). The literature confirms that girls also experience more online harassment, social pressure, and anxiety compared to boys (Carvalho et al., 2018). These emotional harms

mirror technology-mediated trauma, with some arguing that AI-driven systems, especially generative AI, can inadvertently re-traumatise users by reinforcing bias or misinformation (Abdulai, 2024). Thus, while AI offers opportunities for autonomy and engagement, it also introduces complex psychological and social risks, particularly for digitally active youth.

Socio-economic disparities further shape digital vulnerabilities: adolescents from lower socio-economic backgrounds report negative online experiences more frequently (around 65%) than those from higher-status households (57%) and are also more often confronted with unwanted sexual requests (Bulut et al., 2025, p. 186). Qualitative findings from the National Youth Report 2025 further underline that these risks are unevenly distributed not only in terms of exposure, but also with regard to digital competencies (of both young people and their parents), available coping resources and support structures, especially within families (Meyers et al., 2025, pp. 212). In this context, families who struggle to keep pace with technological developments face particular challenges in relation to AI technologies, limiting their capacity to effectively support and accompany their children in developing meaningful and reflective AI use.

GenAI technologies encode biases from underrepresented datasets, leading to what Holmes (2018) describes as the reinforcement of exclusion and "lack of belongingness" among marginalised users. LLMs can, for example, misgender LGBTQ+ people, stereotype them or limit their expressions (McAra-Hunter 2024). The reliance on data from WEIRD ((Western, Educated, Industrialised, Rich, Democratic) societies (Henrich et al., 2010) means that minority and non-Western populations, including youth from different cultural backgrounds, are often misrepresented or ignored by AI systems (Mihalcea et al., 2025). For young users, who are still forming their identities, such misrepresentation can produce feelings of anxiety, self-inadequacy, and alienation.

Despite these concerns, the literature suggests several pathways to address these vulnerabilities. In the case of youth, enhancing digital literacy and risk awareness can be a promising manner of empowering young users to recognise and respond to online risks more effectively (Ramos-Soler et al., 2018), and some scholars have encouraged participatory design as a manner of combating a lack of representation or misrepresentation (Solyst et al., 2023).

4.3.1 Strategies for Supporting Vulnerable Groups

Building on the findings presented in this chapter, the data from national and international studies reveal that vulnerability in the context of digitalisation and artificial intelligence is multi-dimensional. Socio-economic, linguistic, gender-related, developmental, and psychological factors intersect to shape children's and adolescents' ability to participate safely and equitably in an AI-driven society. Certain groups in Luxembourg are particularly at risk of educational exclusion, online victimisation, or digital manipulation. The following table summarises the main vulnerable groups of children and young people in Luxembourg in relation to AI-related and digital risks. It outlines the primary risk factors, the ways in which artificial intelligence may amplify these vulnerabilities, and specific focus areas through which BEE SECURE can address them within its existing and future training, counselling, and awareness programmes.

TABLE 2: Overview of vulnerable groups of children and young people in Luxembourg in relation to AI-related and digital risks, including primary risk factors, AI-related amplifications, and recommended BEE SECURE focus areas.

Vulnerable Group	Primary Risk Factors	AI-Related Amplification	Recommended Actions
Learners from socio-economically disadvantaged backgrounds	Limited access to devices and guided digital learning; lower CIL competences (ICILS 2023); lower subjective technical and operational competencies in digital media use (YSL 2024).	Unequal access to high-quality AI tools; potential exclusion from AI-based learning innovations; difficulties in meaningfully appropriating and navigating rapidly evolving technologies	Strengthen AI literacy across all schools with particular attention to learners from less advantaged backgrounds; integrate AI-related examples into existing BEE SECURE workshops and classroom materials; develop low-threshold, mobile-friendly online resources that can also be used independently at home; systematically signpost Helpline services.
Multilingual learners / chil-	Linguistic barriers in formal educa-	AI can bridge language gaps (transla-	Create multilingual BEE SECURE resources and webinars explaining safe and critical

dren from non-Luxembourgish or non-German-speaking homes	tion; parents less able to assist with schoolwork.	tion, explanation) but uncritical use may cause misunderstanding or bias.	AI use; translate or subtitle key parent trainings (with AI); include examples of cultural bias and translation errors in awareness sessions; promote inclusive AI literacy in collaboration with schools, Maisons Relais and youth centres.
Children without effective parental learning support	Lack of educational guidance and digital supervision at home.	Over-reliance on AI tutors; exposure to misinformation or biased feedback.	Offer targeted BEE SECURE parent webinars on AI and online learning; produce easy-to-understand guides for families with low digital literacy; cooperate with schools to provide supervised after-school activities where safe AI use can be practised; highlight Helpline/KJT as a support option.
Girls and young women	Higher exposure to body-shaming, image-based harassment and online shaming.	Deepfakes, "nudifying" apps, synthetic sextortion and AI-assisted grooming intensify gendered digital violence.	Expand BEE SECURE's gender-sensitive campaigns addressing AI-manipulated images and consent; include case-based modules on AI-generated sexualised content and digital boundaries in youth workshops; systematically communicate Helpline/Stopline services for victims of image-based violence.
Children and youth with a migrant background	Greater likelihood of harassment related to nationality, ethnicity or religion.	Algorithmic bias and AI-driven hate-speech amplification reinforce discrimination.	Strengthen cooperation with community organisations and migrant associations; include algorithmic bias and online discrimination topics in BEE SECURE trainings; develop multilingual campaigns promoting counter-speech and inclusive online culture; refer to Helpline/Stopline for reporting hate content.
Adolescents with low digital resilience or	Psychosomatic complaints, loneliness, low self-esteem; higher sus-	AI-enhanced harassment (voice cloning, impersonation, automated	Integrate AI-related online safety into BEE SECURE's mental-health and well-being workshops; train SEPAS and youth workers on recognising AI-mediated harass-

mental-health vulnerabilities	ceptibility to manipulation.	shaming) exacerbates distress.	ment; create resources on coping with synthetic bullying; ensure clear referral pathways to KJT/Helpline are communicated.
Younger users (early adolescents 11–15)	Limited cognitive maturity and risk awareness; rapid growth of cyberbullying in this age group.	AI-powered chatbots, deceptive games, and synthetic social interactions blur real-vs-fake boundaries.	Include early-age modules on "real vs synthetic" online content in BEE SECURE school and non-formal trainings (e.g. updates to Captain Kara, Superhelden, Ein Sprung ins Netz); develop child-friendly materials explaining AI-generated risks (deepfakes, filters, chatbots); provide parents with guidance on recognising manipulative chatbots and synthetic content.
Learners from socio-economically disadvantaged backgrounds	Limited access to devices and guided digital learning; lower CIL competences (ICILS 2023); lower subjective technical and operational competencies in digital media use (YSL 2024).	Unequal access to high-quality AI tools; potential exclusion from AI-based learning innovations; difficulties in meaningfully appropriating and navigating rapidly evolving technologies	Strengthen AI literacy across all schools with particular attention to learners from less advantaged backgrounds; integrate AI-related examples into existing BEE SECURE workshops and classroom materials; develop low-threshold, mobile-friendly online resources that can also be used independently at home; systematically signpost Helpline services.

The overview in Table 2 shows that vulnerability to AI-related and digital risks in Luxembourg is shaped by intersecting socio-economic, linguistic, gender-related and psychological factors. Children from less advantaged or low-support households, girls and young women, migrant youth and emotionally fragile or younger adolescents (around 12 to 14 years old) are particularly exposed to educational and safety risks intensified by AI-driven technologies. These patterns align with insights from the internal staff survey, in which advisors reported that cases involving deepfake-based sexualised harassment and "nudifying apps" disproportionately affect girls, young women, and emotionally vulnerable adolescents.

The survey findings therefore reinforce the relevance of these groups in the context of AI-mediated sexualised violence.

5 Reactions and Prevention

The following chapter evaluates the services offered by BEE SECURE with regard to the integration of AI topics. The focus is on questions of coverage as well as aspects of educational communication. The chapter also deals with the ethical challenges arising from the work of a Safer Internet Centre. Finally, suggestions are presented on how the work of BEE SECURE can be adapted to the challenges of AI.

5.1 Analysis of existing BEE SECURE offerings

BEE SECURE, the Luxembourgish Safer Internet Centre, is part of a pan-European initiative co-funded by the European Commission under the Digital Europe Programme. Safer Internet Centres (SICs) operate across EU Member States to promote a secure and empowering digital environment for children and young people. Each centre typically combines three core functions: an awareness node that develops educational campaigns and resources on online safety; a helpline that offers confidential advice on issues such as cyberbullying, privacy, and digital well-being; and a hotline for reporting illegal online content, including child sexual abuse material. Beyond these services, SICs foster youth participation through advisory panels and collaborate internationally via the Better Internet for Kids network. Their work reflects a broader European strategy to enhance digital literacy, mitigate online risks, and uphold fundamental rights in the digital sphere. Internal survey data among BEE SECURE staff indicate that trainers and advisors feel generally well prepared to address AI-related questions across training sessions, awareness activities, and Helpline/Stopline support, which further strengthens the foundation of the current service portfolio.

BEE SECURE's activities are extensive: it delivers over 1,200 training sessions annually in schools, youth centres, and community spaces, covering topics from cyberbullying and sextortion to data privacy and misinfor-

mation. A network of 15 freelance trainers plays a key role in identifying emerging issues and engaging directly with students through youth panels, ensuring that BEE SECURE remains attuned to the latest digital behaviours. Beyond training, BEE SECURE organises around 40 events per year, including Safer Internet Day and thematic workshops, and regularly launches national campaigns. One notable example is the Cyberjungle campaign, which uses playful storytelling and AI-generated content to raise awareness about deepfakes, sextortion, and romance scams under the slogan “Gleef net alles am Netz” (“Don’t believe everything online”). These campaigns are informed by Hot Topics meetings, in which BEE SECURE collaborates with stakeholders to define priorities.

A distinctive feature of BEE SECURE is its Radar report, published annually to monitor trends in young people’s digital habits. Based on surveys of thousands of students, parents, and educators, the Radar provides insights into issues such as screen time, social media use, cyberbullying, and perceptions of AI. For example, the 2025 edition revealed that Snapchat and WhatsApp dominate among teens, while concerns about sextortion and misinformation are rising sharply. The Radar serves as an evidence-based tool for policymakers, educators, and parents, guiding national strategies like the sécher.digital action plan. Alongside this, BEE SECURE publishes guidelines, news, and educational materials tailored to different audiences, reinforcing its role as a central resource for digital literacy and online safety in Luxembourg

BEE SECURE pursues a multidimensional approach to the prevention and intervention of digital risks. This includes the services Helpline and Stopline, both based at the Kanner-Jugendtelefon (KJT), a wide range of training and awareness-raising activities for children and young people, as well as events for parents and educational professionals. In addition, BEE SECURE regularly publishes information materials and guidance documents aimed at promoting media literacy and digital safety.

5.1.1 Stopline and Helpline

The Stopline serves as Luxembourg’s national reporting hotline for illegal online content and is operated by Kanner-Jugendtelefon (KJT) with national BEE SECURE partnership. It is a member of the International Asso-

ciation of Internet Hotlines (INHOPE) and operates in line with INHOPE's binding Code of Practice. The Stopline receives, assesses and forwards reports of potentially illegal online material to the Police Grand-Ducale and, through the INHOPE network and the ICCAM platform, to international partners such as INTERPOL. Its primary focus lies on identifying and facilitating the removal of child sexual abuse material (CSAM), including an increasing number of synthetically generated depictions and deepfakes involving minors.

The Helpline provides confidential counselling for children, young people, parents and teachers on all questions relating to safe and responsible internet use. Counselling is available by telephone, online chat and email. While telephone calls are anonymous, written forms of contact (chat or email) are confidential but not anonymous. The Helpline is part of the European Insafe network, which is coordinated by the European Commission under the Better Internet for Kids programme, and provides guidance and emotional support in case of Cyberbullying, data protection concerns, and other online risks.

Both services make a vital contribution to strengthening the digital resilience and safety of young people in Luxembourg. They operate in accordance with European quality standards, are internationally connected (INHOPE, Insafe), and firmly embedded within the national child-protection framework.

In recent years, the digital risk landscape has changed considerably. Alongside traditional CSAM, there has been a marked increase in AI-generated abuse material ("synthetic CSAM") and deepfakes depicting real children or adolescents in sexualised contexts. Algorithmically amplified hate speech, disinformation and fraud schemes involving AI components – such as fake identities or chatbots – are also on the rise. In addition to established topics such as cyberbullying, data protection, excessive media use and risky online challenges, there has been an increase in deepfake-based cyberbullying cases, nudifying apps, and AI tools that generate intimate forgeries leading to sextortion or digital blackmail.

Although systematic statistical data collection on AI-related reports does not yet exist, Helpline and Stopline staff confirm the growing relevance of such cases in their daily counselling and reporting practice. **Establish-**

ing a structured monitoring and statistical analysis of AI-related phenomena would therefore be advisable to identify trends and target preventive action more effectively. Despite the absence of systematic case statistics, the internal staff survey shows that Helpline and Stopline advisors generally feel confident in handling AI-related incidents and are able to draw on established procedures and professional routines when supporting affected users.

Within the framework of international hotline cooperation – particularly within the INHOPE network, of which the Luxembourg Stopline is an active member – several key themes and reporting categories can be identified that have gained new momentum through the use of AI. According to the INHOPE Annual Report (2024), three main developments can be observed across member hotlines:

1. A significant increase in deepfake-related CSAM, including synthetically generated depictions of minors;
2. A shift of illegal content to encrypted or difficult-to-monitor platforms (e.g. Telegram);
3. A growing connection between AI tools and online offences, as easily accessible image generators and chatbots are increasingly being used to prepare or conceal criminal acts.

More generally, AI technologies are also being used to conceal identities, manipulate images or text, and create fraudulent digital profiles or phishing schemes that mimic real individuals. While not all such cases are criminally relevant, they represent a new layer of technical and ethical complexity in hotline and counselling work, requiring continued monitoring and specialised staff training. According to the internal staff survey, these AI-mediated forms of sexualised harm are perceived by advisors as particularly demanding and emotionally challenging, often requiring extensive guidance, reassurance, and safety planning during Helpline interactions.

Through regular participation in national and international training programmes, supervision and network exchanges, staff members ensure high professional standards and maintain up-to-date expertise in ad-

dressing emerging digital phenomena, including synthetic CSAM, deepfake-based harassment, and AI-supported online manipulation. According to the internal staff survey, these training opportunities are perceived as highly effective and contribute significantly to advisors' sense of confidence when dealing with AI-related cases in their daily practice. The continuous professional development safeguards the consistent application of established procedures and ensures that Luxembourg's Stopline and Helpline services continue to play a leading role in child online protection.

5.1.2 Publications

BEE SECURE's publications are freely accessible online, providing an easily available and low-threshold way to obtain information about digital risks and protective measures.

TABLE 3: BEE SECURE publications: AI relevance and suggested content updates.

No.	Publication	AI relation	Recommended additions/updates (keywords)
1	BEE SECURE Radar 2025	Yes – analyses AI awareness, deepfakes and the AI Act.	Multimodal generators, voice cloning, latest legislation.
2	Rapport d'activité 2024	Yes – reports on AI training and deepfake campaigns.	Evaluation of the campaigns, AI Act links, AI voice cloning.
3	Thematischer Beitrag „KI – Chancen & Risiken"	Yes – explains AI, machine learning and the AI Act.	Generative models, voice cloning, use in the classroom.
4	Poster „Passwörter der Schüler verwalten"	No – password tips.	Passkeys, two factor authentication, AI password cracking risks.
5	Thematischer Beitrag „Multi Level Marketing & Co."	No – explains MLM.	Warning about AI trading bots, deepfake advertising, AI bots.

6	Thematischer Beitrag „Hate Speech“	No – defines hate speech.	AI generated hate speech, algorithmic moderation.
7	Thematischer Beitrag „Dark Patterns“	No – UX manipulation.	AI based personalised dark patterns, regulation.
8	Ratgeber „Bist Du Opfer von Cyber Mobbing?“	No – explains bullying and the law.	Deepfake bullying, voice cloning, AI bots as perpetrators.
9	Thematischer Beitrag „Influencer“	Limited – mentions virtual influencers.	AI influencers, deepfake advertising, recommendation systems.
10	UE „Sexting“	No – sexting risks.	Deepfake sextortion, AI bots, labelling requirement (AI Act).
11	Cyber Mobbing Kit	No – teaching units on bullying.	Deepfake bullying, AI moderation and detection.
12	PEGI Label für Videospiele	No – explains age ratings.	Generative AI in games, AI age verification, loot box algorithms.
13	Ratgeber „Nackt im Netz?“	No – sexting/privacy.	Deepfake pornography, AI image search, protection against AI manipulation.
14	UE „Selbstdarstellung – Mein perfektes Ich“	No – reflects on virtual vs. real self.	AI filters, recommendation systems, deepfake effects.
15	UE „Selbstdarstellung – Wie authentisch sind Influencer?“	No – examines influencer culture.	Role of algorithms, generative content tools, deepfakes.
16	UE „Sharenting – Familienfotos im Internet“	No – discusses privacy.	AI facial recognition, data collection, deepfake risks.
17	UE „Sexting – For Your Eyes Only“	No – rules for sexting.	Deepfake sextortion, AI extortion bots, protective tools.

18	DigiRallye – Virtual Edition	Indirect – station “The Internet lies” touches on deepfakes.	Dedicated station on AI/deepfakes, voice assistants, recommendation systems.
19	CYF Dossier „Desinformation in den Medien“	No – journalistic fact check.	AI fake news, generative text/images, detection tools.
20	Elternratgeber „Bildschirme in der Familie“	No – screen time guide.	Recommendation algorithms, voice assistants, deepfakes, child protection filters.
21	Dossier „Desinformation in der Politik“	Indirect – explains algorithms, social bots, micro targeting.	Generative AI in political disinformation, deepfakes, AI Act notes.
22	Elternratgeber „Mein Kind im Internet“	No – general online tips.	AI recommendation algorithms, voice assistants, deepfake filters.
23	Data Detox Kit	Yes – warns against deepfakes and “cheap fakes.”	Modern generative models, voice cloning, deepfake detection.
24	„Was tun bei einer Datenschutzverletzung?“	No – steps after a data breach.	AI phishing, AI based fraud detection, current legislation.
25	Dossier „Desinformation“	No – warns about algorithms and filter bubbles.	Generative AI fake news, labelling synthetic media, AI Act.
26	Internet Sicherheitskonzept für Maisons Relais	No – security concept for childcare.	AI assistant systems, deepfake prevention, algorithmic filters, AI Act.
27	Flyer „Cybermobbing: Anzeige erstatten?“	No – legal advice.	Deepfake bullying, voice cloning, tools to detect AI generated harassment.
28	Checkliste „Vernetzte Spielzeuge“	No – smart toy safety checklist.	AI features (voice/facial recognition), data protection, legal requirements.
29	Thematischer Beitrag „Algorithmen – Big Data“	Yes – explains AI, neural networks, deep learning, the Turing Test and machine learning.	Generative AI (ChatGPT, diffusion models), ethics/bias, AI Act, current examples.

30	Thematischer Beitrag „Big Data: Datensammlung im Alltag“	Indirect – mentions AI use in data analysis and Cambridge Analytica.	Current AI analytics, generative models, data ethics, AI Act, predictive analytics.
31	Netiquette	No – rules of conduct and politeness.	Rules for dealing with AI chatbots, deepfake warnings, recognising generated content.
32	„Auch digital ein Vorbild sein“	No – guide for educators/teachers on online reputation and privacy.	Explain algorithmic recommendation systems, deepfake awareness, use of generative tools in teaching.
33	„Hate Speech und das Gesetz“	No – legal framework for hate speech.	AI generated hate speech, detection algorithms, AI Act legal obligations.
34	Dossier „Social Bots“	Yes – explains bots and emphasises rapid progress in AI.	Generative chatbots (LLMs), voice/video bots, transparency obligations (AI Act).
35	Dossier „Filterblasen und Echokammern“	Indirect – describes algorithmic personalisation.	AI based recommendation systems, generative content, transparency and regulation (AI Act).
36	Thematischer Beitrag „Recht am eigenen Bild“	No – legal basics on photos and privacy.	Deepfakes and AI editing, automated facial recognition, tools for detecting doctored images.
37	E Booking: Betrug beim Online Booking	No – explains booking fraud.	Dynamic pricing, AI generated fake reviews, AI phishing emails.
38	E Shopping: Betrug beim Online Shopping	No – explains fake shops, escrow fraud.	AI generated fake shops, fraudulent chatbots, deepfake product reviews.
39	E Banking: Betrug beim Online Banking	No – explains phishing, trojans.	AI phishing, AI fraud detection, biometric authentication.
40	Thematischer Beitrag „Pornografie“	No – describes pornography and youth protection.	Deepfake pornography, AI content filters, legal classification of AI pornography.

41	Dossier „Darknet – Die dunkle Seite des Netzes“	No – explains Tor and illegal vs. legal use.	AI supported surveillance, automated malware generation, deepfake marketplaces.
42	Gegenrede – Strategien gegen Hate Speech	No – tips for counter speech.	AI amplified hate speech, automatic moderation, AI based counter speech bots.

The analysis of 42 BEE SECURE publications reveals a broad thematic spectrum ranging from digital education, data protection and fraud prevention to hate speech and pornography. Only a minority of the examined documents explicitly address artificial intelligence or machine learning: *BEE SECURE Radar 2025* analyses current levels of knowledge about AI, outlines risks such as deepfakes, and explains regulatory initiatives such as the AI Act. The *Rapport d'activité 2024* reports on training sessions concerning AI and deepfake campaigns, while the thematic publication *"Artificial Intelligence – Opportunities and Risks"* introduces key concepts, applications, and potential dangers of AI. Also linked to AI are the dossiers *"Social Bots"*, which highlights the increasing sophistication of such bots due to advances in artificial intelligence, and *"Algorithms – Big Data"*, which explains neural networks, deep learning and the Turing Test.

The majority of publications, however, focus on traditional aspects of digital safety, such as fraud in online shopping and banking, privacy protection, dealing with hate speech, or the legal framework for photography. In these documents, artificial intelligence is either not mentioned at all or is only indirectly referenced through the term "algorithms". This points to a gap between the rapid evolution of AI-based systems (generative AI, deepfakes, chatbots) and the current informational basis.

It would therefore be advisable to **expand the existing publications to include aspects related to AI-driven threats and emerging fraud schemes**, such as voice cloning or AI phishing. In addition, AI-based moderation and detection tools, as well as current legal frameworks such as the AI Act, should be taken into account. Particular **attention should also be given to issues such as deepfake pornography, personalised dark patterns, algorithmic recommendation systems and AI-supported hate speech filters**, in order to adapt the publications more effectively

to the realities of the digital age. To further enhance the accessibility and relevance of its professional learning resources, it would also be advisable for BEE SECURE to **include clear publication dates and version information on all materials and future updates**. This would allow educators and institutions to verify at a glance whether they are consulting the most recent version or whether newer editions are available, thereby ensuring that guidance remains accurate and up to date in a rapidly evolving digital environment.

5.1.3 Training sessions for school classes

TABLE 4: BEE SECURE trainings for pupils and students: pedagogical methods, AI relevance, suggested content updates.

Training	Target group & duration	Pedagogical method	AI relation	Need for update
Ein Sprung ins Netz	Cycle 3, 90 minutes	The class undertakes a fictional "walk" through familiar everyday situations which are then mirrored by online scenarios. Through games and discussions they explore passwords, codes of conduct, contact with strangers and cyberbullying.	No AI topics are covered.	The content could be expanded to include warnings about AI driven phishing methods, password cracking algorithms or personalised dark patterns.
Cyber Mobbing: Mit Detektiv Shadow auf Mission	Cycle 3.2, 90 minutes	Children work with the fictional character "Detektiv Shadow" to solve puzzles and uncover the issue of cyberbullying. They then create a "cyberbullying expert kit" with tips for dealing with such incidents.	No AI content; the focus is on netiquette, image rights and the BEE SECURE helpline.	Additional topics such as deepfakes and voice cloning as new forms of bullying, and AI based detection tools, could be addressed.
BEE FRIENDS –	Cycle 4, 90 minutes	Step by step, the class creates the social media profile of a fic-	No explicit AI content; topics	It would be advisable to explain algorith-

Alex' erste Schritte online		tional character ("Alex") and jointly decides how to react to typical situations. The decisions and BEE SECURE key messages are recorded on a poster.	include passwords, self presentation, privacy, competitions and chain letters.	mic recommendation mechanisms on social networks, as well as generative filters and chatbots.
Hinter den Kulissen der sozialen Netzwerke	Cycle 4.2, 90 minutes (only bookable after the basic module "BEE FRIENDS")	A change of perspective: pupils take on the role of app developers and playfully consider risks, data collection and advertising models. They learn about protective measures and are encouraged to question influencers.	No direct AI content; the focus is on rules of conduct, inappropriate content, advertising, influencers and data collection.	AI driven personalised advertising, deepfake influencers and recommendation algorithms could be incorporated.
Fit fürs Netz?	Secondary school, Year 7e, 90 minutes (part of the "Digital Sciences" curriculum)	Using the virtual desk of a character ("Dennis"), young people work through five topics. Each object on the desk leads to a new case study on passwords, emails, sexting, self presentation and big data.	Big data aspects are addressed; AI is not mentioned explicitly.	The link between big data and AI (e.g. profiling, recommendation algorithms) and current AI risks such as deepfake sexting could be added.
Hype im Netz – so bleibst du sicher und informiert!	Secondary school, Years 7e and 6e, 90 minutes	Students again explore "Dennis's desk", this time focusing on very current topics. Case studies on deepfakes, disinformation, hate speech, cybergrooming and online challenges stimulate discussion .	Explicit AI topics are included: deepfakes and artificial intelligence are presented as part of the risks.	The contents should be updated regularly to reflect new AI trends (e.g. voice and video generators) and legal developments such as the AI Act.
Fact Check – Safe and Sound	Secondary school, Years 6e and 5e, 90 minutes	A role play: young people become festival organisers and must fend off phishing attacks, fake news campaigns and disinformation. They use fact	AI is part of the training; deepfakes and algorithms are addressed, and the	Continuous updates to include new deepfake technologies, AI driven disinformation campaigns and

		checking tools (e.g. reverse image search) and learn to recognise filter bubbles and algorithms.	use of AI in disinformation is explained.	emerging regulations are recommended.
Lost in Data – Wo werden im Alltag Daten gesammelt?	Secondary school, from Year 4e upwards, 90 minutes	Storybased learning: the class accompanies a fictional couple to Barcelona and searches for the missing Nino via digital traces. Students learn where digital footprints arise in everyday life and how big data, algorithms and AI affect their privacy.	This training has a clear AI component; it explains how algorithms and AI are used in data collection and raises awareness of big data analytics.	Updates could take account of newer forms of data analysis (e.g. generative AI and personalised advertising systems) and refer to forthcoming regulations such as the AI Act.

BEE SECURE makes an important contribution to prevention through its training programmes for school classes. While the basic modules in primary education (“Ein Sprung ins Netz”, “BEE FRIENDS”, “Superhelden im Einsatz gegen Cybermobbing”) focus primarily on topics such as passwords, privacy and respectful online behaviour, aspects related to artificial intelligence have so far received little attention. In view of the rapid development of generative models and the easy accessibility of AI tools, these modules should in future be expanded to include age-appropriate explanations of algorithmic recommendation systems, deepfake filters and privacy-friendly default settings. A promising approach lies in linking familiar offline situations with corresponding online risks – a method already used by BEE SECURE in “Ein Sprung ins Netz”. This approach can also help illustrate new risk scenarios, such as chatbots appearing in games or personalised advertising.

At the secondary level, BEE SECURE trainings demonstrate how AI-related topics can be taught in a concrete and practice-oriented way. “Lost in Data” takes students on a digital treasure hunt through Barcelona, allowing them to experience big data analysis, algorithms and AI in data collection first-hand. “Fact Check – Safe and Sound” engages participants in organising a music festival while confronting them with deepfakes, phishing attempts and filter bubbles – thereby equipping them with essential tools

such as reverse image search and fact-checking platforms. The workshop "Hype im Netz" directly addresses current trends such as deepfakes, cyber grooming and artificial intelligence, showing how quickly online hypes spread and what mechanisms drive them. The hands-on nature of these workshops, which use real tools and problem-solving role plays, strengthens critical media literacy and underlines that AI is not an abstract concept but an integral part of young people's everyday lives.

For a future-oriented media education programme, this White Paper recommends **systematically embedding AI-related topics across all age groups**. This includes early **awareness of AI-driven manipulations such as voice cloning, personalised dark patterns and recommendation algorithms**. At the same time, **learners should develop practical skills** – including the use of detection tools for deepfakes, questioning algorithmic decisions, and reflecting on the data they voluntarily disclose.

5.1.4 Trainings for non-formal education groups

TABLE 5: BEE SECURE non-formal trainings: pedagogical focus, AI relevance, and suggested content updates.

Training	Target group & duration	Pedagogical method	AI relation	Recommended additions/updates (keywords)
Die Abenteuer von Captain Kara	7–9 years, 2 hours	Children watch a Childnet International cartoon featuring Captain Kara and the SMART Crew. In small groups they discuss case studies from the film and work out safe solutions for email scams, fake profiles and other online risks.	No AI topics; focus is on basic internet safety (e mails, disinformation, privacy, cyberbullying, fake profiles).	Introduce examples of AI driven scams such as chatbots, generative fake profiles or AI powered phishing emails; explain how deepfakes and synthetic voices could make disinformation more convincing.

Superhelden im Einsatz gegen Cyber Mobbing (extracurricular)	7–9 years, 2 hours	With the mascot Panda Wanda, children tackle adventures and quizzes that teach them to recognise and stand up to cyberbullying. Completing the tasks earns them a superhero ID card encouraging supportive behaviour.	No AI content; the focus is on cyberbullying, netiquette and where to seek help.	Expand to include new forms of bullying involving AI, such as deepfake manipulation or voice cloning, and introduce simple detection tools.
Quiz dich fit!	8–12 years, 2 hours	Details are not provided, but the title suggests a quiz based format where children test and expand their knowledge of online safety.	No explicit AI topics.	Ensure that quiz questions cover emerging AI driven threats such as AI chatbots, recommendation algorithms and fake content to raise awareness in a playful way.
Let's talk about Selbstdarstellung im Netz	13–16 years, 90 minutes	A non formal discussion workshop about self presentation on YouTube, TikTok and Instagram. Participants and educators exchange experiences and opinions using games, videos and other methods, emphasising privacy and data awareness.	No explicit AI topics; focuses on self presentation, privacy, digital consumption and data collection.	Incorporate discussion of AI driven recommendation systems, generative filters and algorithmic impacts on self esteem; highlight how influencers may use AI tools for content creation.
Let's talk about Sexting	13–16 years, 90 minutes	A moderated conversation about sexting in a non formal setting. Using games and videos, young people reflect on their experiences, learn about consent and privacy, and discuss	Likely no AI topics; emphasis on sexting, sextortion, privacy and image rights.	Add information on AI generated deepfake pornography, sextortion bots and tools to verify authenticity of images; discuss the role of AI in amplifying risks and the im-

	sextortion and image rights.		portance of critical awareness.

Beyond the classroom, BEE SECURE also addresses children and young people in non-formal education settings. The programmes for 7- to 9-year-olds take a strongly narrative approach: “Die Abenteuer von Captain Kara” uses an animated film and the SMART Crew to highlight the dangers of scam emails, disinformation, cyberbullying and fake profiles, while “Superhelden im Einsatz gegen -CyberMobbing” invites participants to join Panda Wanda’s “superhero club” by mastering adventures and quizzes about respectful online behaviour. A quiz=based workshop, “Quiz dich fit!”, aims to reinforce basic knowledge about internet safety in a playful way for eight- to -twelve-year-olds.

Older adolescents are offered discursive spaces rather than readymade lessons. In “Let’s talk about Selbstdarstellung im Netz”,- 13- to 16-year-olds reflect on how they present themselves on Instagram, YouTube or TikTok, examine the balance between authenticity and staging, and discuss privacy and data consumption habits. The companion session “Let’s talk about Sexting” tackles the sensitive issues of sexting, sextortion and image rights, with facilitators using games and video clips to encourage open dialogue.

Although these workshops successfully build digital literacy and promote critical thinking, they currently pay little attention to the growing influence of AI on children’s online experiences.

To remain ahead of emerging threats, the **storytelling and quiz formats for younger children could be updated to include AI-driven scams**, such as chatbot imposters or generative fake profiles, and to explain in simple terms how deepfake images or voices are created. In the discussion-based youth sessions, it would be valuable to **examine how algorithmic recommendation systems shape self-image** and **to raise awareness of -AI generated pornography and extortion techniques**.

5.1.5 Trainings for teachers and educators

TABLE 6: BEE SECURE teacher trainings: pedagogical focus, AI coverage, and suggested updates.

Training	Target group & duration	Pedagogical method	AI relation	Recommended additions/updates (keywords)
TikTok, Instagram & Co. – Was machen Kinder und Jugendliche eigentlich derzeit online?	Teachers and educational staff; 3 hours; delivered in German and Luxembourgish	A three hour seminar providing an overview of young people's media habits on platforms like YouTube, Instagram, Snapchat and TikTok. Participants learn why children are drawn to specific content, explore current media trends and discuss related risks. Topics include challenges, pranks, influencers, gaming, self presentation and sexting.	No explicit AI content; focus is on understanding behaviour and trends.	Incorporate discussion of AI driven recommendation algorithms, the role of AI in shaping viral challenges or pranks, and how generative tools influence self presentation and influencer culture.
Sexuelle Darstellungen im Netz: Was reizt Jugendliche?	Teachers and educational staff; 2.5 hours; delivered in German and Luxembourgish	Online training examining how and why adolescents encounter sexualised content on the internet. It covers Luxembourg's legal framework and addresses risks such as sexting, sextortion and cyber grooming, with practical guidance for schools and recommended information sources.	No AI content; emphasis is on legal context and safeguarding.	Update with information about AI generated deepfake pornography, AI enabled sextortion bots and tools for verifying the authenticity of images or videos.

Sicherheit im Fokus: Von KI Grundlagen zur praktischen Anwendung	Teachers and educational staff; 3 hours; offered in French, English, German and Luxembourgish	A deeper dive into AI technologies and security. Participants receive an overview of AI's development and integration into various tools, explore opportunities and risks, work hands on with AI tools, and analyse case studies in groups to develop safe, responsible approaches. Topics include AI, disinformation, data security and AI tools.	This training explicitly covers AI, providing both theoretical background and practical exercises.	Ensure the content reflects the latest AI developments (e.g. generative models, voice cloning), include guidance on the AI Act and ethical considerations, and regularly update the case studies.
Cybersecurity Essentials – was man als Lehrkraft über Internetsicherheit wissen sollte	Teachers and educational staff; 3 hours; delivered in Luxembourgish	Interactive session tailored to participants' questions about internet security. It explains the underlying technologies, identifies risks and offers practical tips on topics such as cloud services, backups, account management, data protection and privacy. A second part presents case studies for group problem solving.	AI is not explicitly covered; focus is on general security and privacy practices.	Include emerging AI driven threats, such as AI generated phishing emails, automated malware and AI based intrusion detection; provide guidance on using AI tools responsibly in the classroom.

BEE SECURE's professional development programme for teachers and educators plays a pivotal role in strengthening digital literacy and online safety competencies within educational institutions. The training session "TikTok, Instagram & Co. – Was machen Kinder und Jugendliche eigentlich derzeit online?" provides participants with valuable insights into young people's online habits, motivations and risks across platforms such as YouTube, TikTok and Instagram. However, given the rapid evolution of digital ecosystems, future iterations should explicitly address the

impact of AI-driven recommendation systems and generative tools shaping influencer culture and viral trends.

The online seminar “Sexuelle Darstellungen im Netz: Was reizt Jugendliche?” focuses on adolescents’ exposure to sexualised online content, legal frameworks and preventive strategies. It provides a strong foundation for safeguarding discussions but would benefit from integrating AI-related risks such as deepfake pornography, AI-assisted grooming and synthetic sextortion.

The training “Sicherheit im Fokus: Von KI-Grundlagen zur praktischen Anwendung” already represents a good-practice example by combining theoretical foundations of AI with hands-on tool exploration and group-based case analysis. Regular updates are essential to include emerging technologies (e.g. voice cloning, generative image/video systems) and ensure alignment with the European AI Act and ethical guidelines.

Finally, “Cybersecurity Essentials – was man als Lehrkraft über Internetsicherheit wissen sollte” provides a comprehensive introduction to data protection, cloud computing and privacy management, delivered through interactive exercises. While AI is not yet explicitly addressed, the course could be enhanced by including current AI-driven threats (e.g. intelligent phishing campaigns or automated malware) and by discussing the responsible integration of AI tools into teaching practice.

Overall, these training modules form a coherent framework for fostering AI-aware, ethically grounded digital education. By incorporating explicit references to AI across all topics, BEE SECURE can ensure that educators not only understand the technological landscape but also feel empowered to guide children and young people in responsibly navigating an increasingly algorithmic world.

5.1.6 Training sessions for parents

TABLE 7: BEE SECURE parent trainings: focus areas, AI relevance, and suggested content updates.

Training	Target group & duration	Pedagogical method	AI relation	Recommended additions/updates (keywords)
Kinder und Jugendliche im Internet – Eine Herausforderung für Eltern	Parents; 2 hours; available in French, English, German and Luxembourgish	Based on the BEE SECURE guide Bildschirme in der Familie, this parent evening provides practical advice on balancing opportunities and risks of children's internet use. Topics include screen time, smartphone use, online abuse and cyberbullying.	No explicit AI content; focus is on general online safety and parental guidance.	Integrate information on AI-driven recommendation systems, generative content, and risks such as deepfakes or chatbot grooming; include guidance on talking to children about AI tools.
Kindersicherung im Internet: praktische Tools, Techniken und Tipps	Parents; 90 minutes; available in Luxembourgish	Conducted as an online webinar, this session demonstrates settings, programmes and technical methods that enhance children's online safety, with a practical focus on device configuration and parental-control software.	No AI content; focus on technical parental-control measures.	Update to include AI-based parental-control and monitoring tools, explain how algorithms recommend or restrict content, and address ethical and privacy implications.
Expertentrio: Experten beantworten Ihre Fragen	Parents; 2 hours; available in Luxembourgish	A Q&A event organised by BEE SECURE in cooperation with KJT and the Luxembourg Police's crime prevention unit. It covers a wide range of safety issues concerning children's internet	No direct AI content; focus is on prevention and expert advice.	Add discussion of AI-related threats (deepfakes, AI-generated grooming, fake profiles) and guidance on how parents can recognise and

	use, including legal and social aspects.		report AI-driven manipulation or fraud.

BEE SECURE's training offer for parents is designed to help families navigate the challenges of raising children in an increasingly digital environment. The sessions focus on practical guidance, open dialogue and the everyday realities of online life at home. "Kinder und Jugendliche im Internet – Eine Herausforderung für Eltern" introduces key questions that parents often ask: When is the right time for a child to own a smartphone? How much screen time is appropriate? How can children be protected from online abuse or cyberbullying? Building on the BEE SECURE publication *Bildschirme in der Familie*, this two-hour session promotes balanced, age-appropriate approaches to online use and encourages parents to discuss both opportunities and risks with their children.

"Kindersicherung im Internet: praktische Tools, Techniken und Tipps" takes a more technical and solution-oriented approach. Delivered as a webinar, it guides parents through the configuration of security settings and parental-control systems, helping them understand how devices and apps can be adjusted to create safer digital environments for younger users. In the future, this session could further benefit from exploring how AI-powered recommendation systems and automated parental-control tools work, as well as the ethical and privacy issues that arise when algorithms mediate children's online experiences.

The discussion-based format "Expertentrio: Experten beantworten Ihre Fragen" offers parents the opportunity to engage directly with specialists from BEE SECURE, the child helpline KJT and the crime prevention unit of the Luxembourg Police. By addressing questions about privacy, digital well-being, social media behaviour and online threats, it provides a valuable bridge between professional expertise and family practice. Going forward, this format could also include examples of AI-driven risks, such as deepfake manipulation, synthetic grooming or fake profiles, and offer parents concrete strategies for recognising and reporting such incidents.

Together, these parent-focused initiatives form an essential component of Luxembourg's broader digital education framework. By integrating AI-related issues more systematically, BEE SECURE can empower parents

to guide their children with confidence, foster critical conversations at home and ensure that families remain resilient in the face of an increasingly algorithmic online world.

To maximise inclusivity and accessibility, BEE SECURE should consider offering these parental trainings **in additional languages beyond Luxembourgish, German and French**. Many families with a migration background may not be fluent in the country's official languages, yet face the same digital challenges. Providing translated sessions or subtitled online versions in widely spoken languages such as Portuguese or English would ensure that all parents — regardless of linguistic background — can benefit from the guidance and actively participate in promoting safe and responsible internet use at home.

In addition, **developing asynchronous online webinars** could make the programme even more accessible. Such webinars would **allow parents to participate at a time that suits their personal schedules** and could be automatically translated into multiple languages using AI-based tools. This approach would significantly expand the reach of BEE SECURE's parental support, ensuring that high-quality guidance on online safety and AI-related risks is available to all families, irrespective of time constraints or language barriers.

5.2 Ethical considerations for education organisations

5.2.1 Main issues

Oversight & accountability: When oversight processes, such as how educational content is vetted, reviewed, or approved, are not publicly documented, accountability to external stakeholders such as teachers, parents and civil society becomes limited. Without insight into who makes these decisions and on what basis, there is a risk that bias, selective framing, or external influence could go unnoticed. From an ethical standpoint, educational institutions have a responsibility not only *to* be impartial but also *to* be seen as impartial. This distinction underscores the importance of procedural transparency: openness about review criteria, expert involve-

ment and revision processes demonstrates integrity and fairness. By making these procedures visible and accessible, institutions reinforce their credibility and ensure that learners and the public can trust that educational resources are developed in an impartial and accountable manner.

Neutrality & advocacy: Ethical questioning lies in recognising that digital literacy education is never entirely free of value judgments, and that discussions on truth, bias, and responsibility inevitably reflect societal norms and priorities. Therefore, the goal is not to present content as neutral "in absolute terms" but to acknowledge the principles guiding educational choices and encouraging critical reflection rather than acceptance.

Empowerment & protection: One problem may also involve the balance between safety (or protection) and autonomy (or empowerment). Parents and schools often use monitoring tools and restrictions to protect youth from online risks, but these measures can also lead to mistrust and limit students' freedom to explore and learn. Constant surveillance can make young people feel "spied on", discourage them from seeking help when needed and restrict opportunities for self-expression. Children and young people are entitled to participate fully in the digital sphere, but they are also in need of protection from various forms of harm. Overprotective approaches can restrict opportunities to build digital competence, while overly permissive strategies may increase exposure to social, psychological, or economic risks; different tensions may rise as well between global and local levels of responsibility.

Education & exposure: Educating young people about certain risks carries the inherent danger that they will become aware of potentially harmful AI applications while learning about them. This dilemma is often framed as the "awareness paradox": Wu et al. (2025) show that increasing AI literacy improves self-efficacy and risk recognition, but note that exposure to examples of harmful uses can trigger curiosity and experimentation if not framed ethically. Similarly, Dubois (2024) applies paradox theory to generative AI in education, emphasising the need to balance transparency with safeguarding to avoid reinforcing discriminatory or harmful practices.

5.2.2 Standards for Decisions and Trade-offs

Building upon the ethical dilemmas identified above, the establishment of clear standards for decision-making is essential to ensure that digital literacy initiatives like BEE SECURE operate with integrity and fairness. Ethical reflection must evolve from identifying potential risks into defining concrete principles that guide action. Recognising that education in digital ethics and safety is inherently value-laden, these standards must balance competing priorities through deliberate, context-sensitive choices.

First, oversight and accountability demand full procedural transparency as a foundational principle. The ethical concern highlighted earlier – limited visibility into who makes decisions and on what grounds – can be addressed by establishing open documentation and participatory review systems. Publicly accessible information on how educational materials are developed, vetted and approved ensures that decision-making processes remain subject to external scrutiny. Clear disclosure of review criteria, expert involvement and revision procedures reinforces institutional credibility and safeguards against potential bias or external influence. As Floridi et al. (2020) argue, socially responsible digital initiatives must make their objectives and influences explicit to sustain public trust. For BEE SECURE, this includes publishing details about partnerships, funding sources and quality assurance processes, thereby turning transparency into both an ethical safeguard and a practical mechanism for accountability (see MediaSmarts).

Second, as discussed in the first section, digital literacy education inevitably engages with normative issues such as bias, truth, and responsibility. The standard, therefore, should be an explicit acknowledgment of the values that inform educational choices, alongside a commitment to fostering critical reflection among learners.

Third, the issue outlined earlier, between safeguarding young users and respecting their autonomy, requires a proportional and context-sensitive approach. Protective measures such as content filters or parental monitoring are ethically justified only when they serve the learner's best interest without undermining trust or independence. Excessive surveillance risks cultivating compliance rather than competence, while neglect-

ing protection can expose young people to online harm. To navigate this trade-off, BEE SECURE should adopt the principle of receiver-contextualised intervention (Floridi et al., 2020), tailoring its educational strategies to the capacities, maturity, and needs of learners. Programmes should aim to build digital resilience, the ability to assess and respond to risks independently, rather than relying on restrictive control.

With regard to the awareness paradox, the same principles apply. Child safety frameworks, including the American Psychological Association (APA) Health Advisory (2025), recommend focusing on capacities and principles—such as privacy, consent, and critical thinking—rather than technical "how-to" details of exploitative behaviours. Microsoft's Safer Internet Day initiatives emphasise interactive learning environments (e.g., Minecraft CyberSafe AI) that teach responsible AI use without glamorising abuse scenarios (Microsoft Education Team, 2025). Together, these perspectives suggest that effective AI safety education should prioritise ethical reasoning, resilience-building, and scenario-based discussions over explicit instructions, thereby mitigating the risk of inadvertently enabling harmful behaviours.

5.2.3 Good Practice Examples

Several international initiatives offer examples of how ethical dilemmas in digital education and youth protection can be addressed:

1. **MediaSmarts** in **Canada** has established practices to ensure neutrality in educational resources, with oversight by an independent ethics body: https://mediasmarts.ca/about-us/integrity-mediasmarts.
   Examples:

   - "None of our funders exert influence on any of our work, including the content of research and resources".
   - "MediaSmarts maintains editorial independence over all our owned and co-owned content. […] MediaSmarts does not endorse any products or services and any content that offers guidance on specific digital tools and platforms does not constitute an endorsement".

- "We uphold public accountability practices with our funders which include clear guidelines on conflict of interest and sponsorship. We also have a commitment to a community-based research approach which includes a review and approval of our research studies by the Community Research Ethics Office".
- "Finally, we value intellectual humility. As a team, we stay open to new evidence and perspectives on identified best practices [...] We learn from our network of colleagues and adjust our positions based on the best available evidence".

2. **klicksafe** is a **European Union** initiative that promotes digital literacy and supports people in using the internet safely, critically and responsibly. It provides independent information, educational materials and training for teachers, parents and young people, while coordinating Germany's Safer Internet Centre under the EU's Digital Europe Programme: https://www.klicksafe.eu/en.
3. The **UK Safer Internet Centre** (UKSIC) is a partnership of Childnet, Internet Watch Foundation and SWGfL that works to make the internet a safer place through education, support, and reporting services. It coordinates Safer Internet Day in the UK, runs helplines and a hotline for online harms and promotes youth participation through its Youth Advisory Board: https://saferinternet.org.uk/about.

5.3 Recommendations

Based on the analysis of risks, opportunities and international practice, we propose the strategic recommendations below for BEE SECURE's future work in an AI-driven online environment. To respond effectively to the rapidly evolving digital landscape, BEE SECURE should strengthen its work on AI-related risks while continuing to address the full range of established online risks that remain highly relevant for children, young people and adults.

In light of the rapid proliferation of AI-related trends, **BEE SECURE should not develop dedicated trainings or awareness activities for every new technological development**, but instead prioritise those is-

sues that are demonstrably relevant for child protection, education and prevention.

For BEE SECURE, the findings on vulnerable groups underscore the need for differentiated and inclusive AI-literacy strategies that reflect Luxembourg's linguistic and social diversity.

Future **activities should focus on multilingual awareness campaigns, early-age interventions, gender-sensitive prevention programmes and professional training on AI-mediated harms** such as deepfakes, synthetic sextortion and algorithmic discrimination. Embedding these priorities across all BEE SECURE training formats will help ensure that prevention, counselling and awareness efforts remain responsive to emerging AI challenges while promoting digital empowerment for all children and young people. These priorities are reinforced by the internal staff survey, which highlights similar risk patterns in frontline practice – particularly the heightened vulnerability of girls and young women and the need for multilingual, low-threshold support for families with limited digital or linguistic resources.

Concretely, we propose **six interconnected areas of action**:

1. establish comprehensive educational initiatives,
2. provide guidance for parental and educator involvement,
3. ensure familiarity with reporting procedures and rights,
4. create mechanisms for youth participation in policy development,
5. include ethical reflection in all working areas, and
6. foster scientific collaboration.

Each area translates the analytical and ethical considerations of the previous chapters into concrete, workable steps for BEE SECURE. While some measures can be implemented in the short term, others are conceived as medium- to long-term developments. All proposed measures build on BEE SECURE's existing mandate and are designed to complement, not duplicate, the responsibilities of schools, ministries and specialised services.

5.3.1 Comprehensive educational initiatives

In line with international programmes such as Intel's *AI for Youth*, and with the AI-literacy and teacher professionalisation needs identified in Chapters 3 and 4, BEE SECURE should extend its educational work from "AI as special topic" towards **a cross-cutting competence**:

A) SYSTEMATICALLY INTEGRATE AI INTO EXISTING BEE SECURE WORKSHOPS AND TRAININGS

Building on the typology of risks (Chapter 2) and the analysis of BEE SECURE's portfolio (5.1), AI-related content should become a recurring thread in all relevant formats:

- In **primary school formats**, include simple "real vs. synthetic" exercises (e.g. distinguishing AI-generated from real photos or voices) to address early-age vulnerabilities.
- In **offers for adolescents**, use concrete examples of AI-generated images, videos and texts to discuss manipulation, body image, peer dynamics and opinion formation.
- In **discussion formats with older youth**, systematically introduce cases of AI-assisted grooming, sextortion, fraud and algorithmic discrimination.

This does not require entirely new programmes; we need only add a structured "AI layer" within existing formats.

B) DEVELOP AN AGE-GRADED AI-LITERACY FRAMEWORK FOR BEE SECURE

To connect with the KI-Kompass and international AI-literacy frameworks, BEE SECURE should articulate a simple, age-graded AI-literacy framework that can guide all materials and trainings. It should define:

- core learning goals (e.g. “understand that AI can hallucinate”, “recognise synthetic images”, “know when to distrust a chatbot”);
- age-appropriate didactic approaches (from playful exploration in primary school to reflective, ethics-based discussion in secondary school);
- links between AI-literacy and existing competence areas (critical thinking, media literacy, digital civic participation).

This framework can serve as an internal reference point and as a communication tool with schools and partners.

C) DEVELOP NEW THEMATIC DOSSIERS ON CROSS-CUTTING AI ISSUES

In addition to updating existing materials, BEE SECURE can create thematic dossiers that bundle central AI-related challenges identified in this White Paper. Priority examples include:

- **“AI and relationships”** – addressing chatbots and virtual companions, parasocial “friendships”, emotional dependency and the blurring of boundaries between human and synthetic interaction.
- **“AI and mental health”** – examining the ambivalent roles of AI-based self-help tools and chatbots, including risks of normalising self-harm content, reinforcing body-image problems or amplifying loneliness.
- **“AI and school”** – focusing on homework, grading and exams, the changing role of teachers in AI-mediated learning, and fair, transparent rules for student use of generative systems.

These dossiers should synthesise research findings from this report with concrete pedagogical implications and be complemented by youth-friendly formats (short explainers, videos, comics) that can be used directly in schools and youth work.

D) ESTABLISH A RECURRING "AI & YOUTH" CAMPAIGN STRAND

Building on BEE SECURE's campaign experience and the international good practice of using gamification and simulations (5.3.1), BEE SECURE should introduce a recurring campaign strand explicitly devoted to AI. Over a defined period (e.g. annually or every two years), it would focus on:

- deepfakes and synthetic sexualised content,
- automated scams and voice-cloning fraud,
- AI-assisted hate, harassment and disinformation.

The campaign should be co-created with young people (see Section 4) and linked to Safer Internet Day and other national events, ensuring that AI becomes part of mainstream online-safety discourse.

Where possible, new or revised AI-related formats should be accompanied by proportionate evaluation mechanisms (e.g. short pre/post feedback instruments) to assess reach, relevance and perceived impact.

E) USE REGULAR FORESIGHT WORKSHOPS TO KEEP OFFERS UP TO DATE

To anticipate rather than merely react to new AI risks, BEE SECURE should hold regular foresight workshops (e.g. every six months) with researchers, youth workers, law enforcement and other stakeholders. These workshops would:

- map emerging AI-related phenomena (e.g. new deepfake apps, AI in games, political micro-targeting);
- assess their relevance for children and young people in Luxembourg;
- feed directly into the revision of workshops, materials and campaigns.

5.3.2 Provide guidance for parental and educator involvement

Echoing the logic of Intel's *AI for Citizens* programme and the findings on teacher capacity and family support (Chapters 3 and 4), BEE SECURE should equip adults who mediate children's digital lives with concrete tools and guidance:

A) EXPAND TRAININGS FOR TEACHERS, EDUCATORS AND PARENTS TO INCLUDE AI DIDACTICS

Existing BEE SECURE trainings can be complemented with modules that focus on *how* to integrate AI into education and parenting in a reflective way:

- designing tasks where students critique and revise chatbot outputs, instead of simply copying them;
- developing fair and transparent criteria for assessing AI-assisted homework and projects;
- discussing proportionate responses to AI-supported cheating and plagiarism that prioritise learning and responsibility over punishment.

This directly implements the ethical standards of Chapter 5.2 by focusing on judgement, proportionality and transparency.

B) STRENGTHEN GENDER-SENSITIVE AND DIVERSITY-SENSITIVE APPROACHES IN EXISTING FORMATS

On the basis of Chapter 4.3 and Table 2 on vulnerable groups, BEE SECURE should sharpen its focus on specific risk constellations within current offers for adults:

- include modules on AI-mediated sexualised violence (deepfakes, "nudifying" apps, synthetic sextortion) in trainings for professionals and parents, especially those working with girls and young women;
- develop complementary components for boys and young men on consent, bystander roles and responsibility when creating or forwarding synthetic sexual content;
- provide guidance for professionals working with young people with disabilities or mental-health vulnerabilities, for whom AI tools may create specific risks (e.g. stigmatising content, self-harm forums mediated by AI).

C) PROVIDE MULTILINGUAL GUIDANCE AND CHECKLISTS FOR FAMILIES

To reflect Luxembourg's linguistic diversity and the survey's finding on families with limited resources, BEE SECURE should develop concise, multilingual guidance on "safe AI use in the family":

- short checklists for parents on talking about AI, setting boundaries for image sharing, and responding to deepfakes;
- easily accessible factsheets outlining children's and parents' rights in relation to AI-generated content and data protection;
- modular webinars addressing key AI-related risks and protective strategies.

To ensure low-threshold access, these materials should be made available in flexible, asynchronous formats, particularly for families that are difficult to reach through traditional formats. Where appropriate, AI-based translation and subtitling tools can be used to provide additional language versions.

These materials can be promoted through schools, community organisations and health services.

D) PILOT A CURATED "DIGITAL ADVISOR" (AI CHATBOT) AS A GUIDANCE TOOL

BEE SECURE can pilot a carefully curated, privacy-preserving AI-based "digital advisor" on its website or app. For parents and educators, such a tool could:

- answer frequently asked questions about AI and online risks in accessible language;
- point to relevant BEE SECURE resources, trainings and campaigns;
- clearly signal situations where professional human support (Helpline, KJT, psychological services) is necessary.

As a ministerial licence for platforms such as Fobizz is already in place, such a chatbot could be created and tested in a low-threshold manner using existing tools, without requiring complex technical development. This would allow BEE SECURE to focus on content, guidance and safeguards rather than on building new technical infrastructure.

The chatbot should operate within clear ethical and quality safeguards, be transparently labelled as informational only, and undergo regular expert and youth review. The digital advisor is intended exclusively as an initial orientation and signposting tool and does not replace personal counselling or professional judgement.

5.3.3 Ensure familiarity with reporting procedures and rights

To make protection effective, children, young people, parents and professionals must understand how to report harmful content and what rights they have. At the same time, BEE SECURE's own procedures must be adapted to AI-mediated harms and make careful use of AI for risk detection (as highlighted in 5.3.1):

A) ADAPT HELPLINE AND STOPLINE PROCEDURES TO AI-RELATED CASES

BEE SECURE should refine its internal taxonomies and workflows so that AI-related harms are clearly visible in practice:

- introduce specific categories such as deepfake harassment, voice-cloning fraud, chatbot-mediated grooming and synthetic CSAM;
- provide targeted training and supervision for counsellors on these phenomena, including psychological impacts and collaboration with law enforcement.

This will allow the Helpline and Stopline to address new forms of harm while preserving their core relational and trust-based approach.

B) MAKE REPORTING CHANNELS AND RIGHTS VISIBLE IN ALL AI-RELATED MATERIALS

All materials that address AI risks should, wherever appropriate, include clear information on:

- how to contact BEE SECURE Helpline and Stopline;
- how to report unlawful content to platforms and authorities;
- which rights children, young people and parents have (e.g. removal of unlawful content, protection against discrimination, data-protection rights).

This ensures that AI-related education is systematically linked to concrete avenues of support and redress.

C) ESTABLISH AN AI-SPECIFIC MONITORING AND EARLY-WARNING STRAND

BEE SECURE's Radar and internal monitoring processes should be expanded with an AI-specific strand:

- integrate question blocks on AI use, exposures and perceived risks into surveys;
- systematically analyse AI-related Helpline and Stopline cases;
- compile an annual "AI incidents and perceptions" section in public reporting.

Subject to strict data-protection and ethics safeguards, BEE SECURE may also explore *AI-assisted* analysis of aggregated and anonymised data to identify emerging patterns (e.g. new scam types, new deepfake practices), thereby turning frontline experience into an early-warning system.

D) ADOPT AN ETHICS AND QUALITY FRAMEWORK FOR AI-RELATED WORK

Drawing on the MediaSmarts example and the normative standards developed in Chapter 5.2, BEE SECURE should formalise a short ethics and quality framework for its AI-related activities, covering:

- editorial independence from funders and commercial partners in all content and tools;
- transparent disclosure of partnerships, sponsorships and potential conflicts of interest;
- a clear statement that any cooperation with platforms does not imply endorsement of their products;
- commitments to public accountability and to revising positions in light of new evidence and youth feedback.

This framework should apply to all AI-related initiatives—from factsheets and trainings to monitoring and any AI-based tools—and can be published on BEE SECURE's website. This framework does not introduce new regulatory standards but operationalises existing ethical principles for BEE SECURE's educational and preventive work.

E) DEVELOP AND REHEARSE CRISIS PLANS FOR AI-RELATED "WORST-CASE" SCENARIOS

Findings from the internal staff survey indicate that, despite overall confidence in handling AI-related inquiries, advisors encounter uncertainties in complex or ambiguous situations. These insights underscore the need for clearer crisis plans, decision trees, and standard operating procedures (SOPs) to guide responses in potential worst-case scenarios involving AI-generated harm or manipulation. BEE SECURE should work with schools, police and other stakeholders to create specific standard operating procedures for AI-related crises, such as:

- viral circulation of a deepfake sexual video of a student;
- synthetic CSAM involving Luxembourg contexts;
- AI-mediated self-harm or suicide "advice".

These SOPs should define escalation paths, communication strategies and support measures for affected children, families and schools. Regular tabletop exercises will help test and refine them.

These efforts should be aligned with relevant European frameworks such as the Digital Services Act and the emerging AI regulatory landscape.

5.3.4 Create mechanisms for youth participation in policy development

As argued throughout this report, young people are not only at risk but also key experts in AI-mediated digital cultures. In line with international

good practice (e.g. the UK Safer Internet Centre Youth Advisory Board), BEE SECURE should institutionalise youth participation:

A) ESTABLISH A PERMANENT "AI YOUTH ADVISORY GROUP"

BEE SECURE should set up a youth advisory group focused on AI that:

- reviews new AI-related materials, campaigns and digital tools (including any chatbot);
- tests their clarity, relevance and tone;
- regularly reports back to BEE SECURE staff and, where appropriate, to policymakers.

This group can be linked to existing youth councils to avoid duplication and ensure democratic legitimacy.

B) BUILD AN "AI SAFETY AMBASSADORS" PEER-EDUCATION PROGRAMME

Following peer-to-peer models highlighted in 5.3.1, BEE SECURE can train selected students and young people as AI Safety Ambassadors, who then:

- act as peer multipliers in schools, youth centres and events;
- co-design and co-facilitate workshops and awareness activities;
- serve as approachable contact points for peers seeking guidance on AI-related issues.

Participation can be recognised through certificates or links to civic engagement programmes, increasing motivation and sustainability.

C) CO-CREATE CAMPAIGNS AND EDUCATIONAL CONTENT WITH YOUNG PEOPLE

Youth participation should extend beyond consultation to co-creation, particularly in the "AI & Youth" campaign strand:

- involve young people in designing slogans, visuals, videos and social media content;
- pilot new workshop formats in schools and youth centres and revise them based on youth feedback;
- systematically incorporate feedback from youth structures into updates of materials and campaigns.

D) INVOLVE YOUTH STRUCTURES IN AI POLICY DIALOGUE AND EVALUATION

Finally, BEE SECURE should collaborate with youth councils, school student bodies and other youth organisations to:

- include youth perspectives in national and European debates on AI governance and youth rights;
- invite young people to evaluation workshops on BEE SECURE's AI-related initiatives;
- ensure that their recommendations are taken into account in strategic decisions.

5.3.5 Include ethical principles in all working processes

A) ETHICAL REFLECTION

AI-related educational initiatives should incorporate structured moments of ethical reflection throughout their lifecycle. Rather than treating ethics

as a one-time compliance step, institutions should periodically reassess whether tools and practices continue to align with educational goals, developmental needs, and emerging evidence.

B) ETHICAL AI USE

Ethical AI use should be framed as a shared responsibility between institutions, educators, parents, and learners. Educational initiatives should clarify the respective roles of each actor, avoiding the implicit transfer of responsibility to adolescents themselves. Institutions should establish mechanisms that allow adolescents to reflect on and provide feedback about their experiences with AI-mediated tools and content. This feedback should inform revisions and adaptations, recognising young people as stakeholders rather than passive recipients. Participatory input strengthens legitimacy and helps identify impacts that may not be visible to adults.

C) FUTURE ORIENTATION

Institutions should prepare adolescents not only for current AI applications but also for plausible future developments. This includes fostering anticipatory awareness and helping young people recognise how ethical risks may evolve, rather than focusing solely on present tools.

5.3.6 Strengthen evidence-based prevention through scientific collaboration

Scientific collaboration can serve as a strategic enabler for BEE SECURE's long-term, evidence-based development. Rather than constituting a separate programme, cooperation with academic partners allows existing data, expertise and prevention activities to be systematically analysed, evaluated and translated into practice-relevant insights. In this sense, scientific collaboration provides a foundation for informed policy decisions, continuous improvement of prevention strategies and international visibility. One promising option is a long-term cooperation with the University of Luxembourg. BEE SECURE continuously generates data on digital risks (helpline/stopline statistics, radar surveys, training feedback). This data is highly relevant for empirical research on media education, youth online behaviour and AI risks. On the other hand, the University of Luxembourg

has expertise in digital ethics, educational technology and teacher education. Joint projects can develop and evaluate evidence-based prevention strategies. The university also provides a platform for systematically reaching target groups such as prospective teachers, social pedagogues and educators – e.g. through integration into study modules or MOOCs. **Possible areas of cooperation** include**:**

- **Data-sharing framework:** Development of a legally compliant cooperation model for anonymised data (Radar, Helpline, Stopline), including GDPR-compliant governance.
- **Joint research projects:** Joint third-party funding applications (e.g. FNR, EU Horizon) on topics such as 'AI Literacy in Schools,' 'Deepfake Detection Skills' or 'Impact of AI Companions on Youth Well-being.'
- **MOOCs & Teacher Training:** Development of open online courses on 'AI & Digital Security' for teacher training students and continuing education staff, based on BEE SECURE content and university didactic expertise.
- **Evidence-to-Practice Lab:** Establishment of an 'AI Safety Education Lab' as an interface between research, practice and policy – piloting innovative teaching formats and evaluating their effectiveness.
- **Youth Advisory & Participatory Design:** Involvement of school pupils and students in co-design processes for new awareness campaigns and tools.

References

Abbasi, B. N., Wu, Y., & Luo, Z. (2025). Exploring the impact of artificial intelligence on curriculum development in global higher education institutions. *Education and Information Technologies, 30*, 547–581. https://doi.org/10.1007/s10639-024-13113-z

Abdulai, A. (2024). Is generative AI increasing the risk for technology-mediated trauma among vulnerable populations? *Nursing Inquiry, 32*(1). https://doi.org/10.1111/nin.12686

AlDahoul, N., Rahwan, T., & Zaki, Y. (2025). AI-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports, 15*(1), 14449. https://doi.org/10.1038/s41598-025-99623-3

Alexander, S. (2025). Deepfake cyberbullying: The psychological toll on students and institutional challenges of AI-driven harassment. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 98*(2), 36–50. https://doi.org/10.1080/00098655.2025.2488777

Alfarwan, A. (2025). Generative AI use in K-12 education: A systematic review. *Frontiers in Education, 10*, Article 1647573. https://doi.org/10.3389/feduc.2025.1647573

AlSajri, A., & Aljanabi, M. (2024). The impact of ChatGPT on social media. *MEDAAD, 2024*(2), 9–14. https://doi.org/10.70470/MEDAAD/2024/002

American Psychological Association. (2025). *Artificial intelligence and adolescent well-being: Health advisory*. APA.

Amichai-Hamburger, Y., Kingsbury, M., & Schneider, B. H. (2013). Friendship: An old concept with a new meaning? *Computers in Human Behavior, 29*(1), 33–39. https://doi.org/10.1016/j.chb.2012.05.025

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science, 373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Bahner, J. E., Huegli, A. S., & Manzey, D. (2008). Complacency, automation bias and the impact of training. *International Journal of Human-Computer Studies, 66*(9), 688–699.

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 604–613. https://doi.org/10.1016/j.ijhcs.2008.06.001

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32(*4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bakir, V., & McStay, A. (2022). *Optimising emotions, incubating falsehoods: How to protect the global civic body from disinformation and misinformation*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-13551-4 (Open Access)

Bakir, V., & McStay, A. (2025). Move fast and break people? Ethics, companion apps, and the case of Character.ai. *AI & Society, 40*, 6365–6377. https://doi.org/10.1007/s00146-025-02408-5

Barrington, S., Cooper, E. A., & Farid, H. (2025). *People are poorly equipped to detect AI-powered voice clones. Scientific Reports, 15*(1), Article 11004. https://doi.org/10.1038/s41598-025-94170-3

BEE SECURE. (2025). *BEE SECURE Radar 2025: Current trends in young people's use of information and communication technologies*. Service national de la jeunesse (SNJ).

Biesta, G. J. J. (2010). *Good education in an age of measurement: Ethics, politics, democracy* (Interventions). Taylor & Francis.

Biewers Grimm, S., Latz, A., & Weis, D. (2025). Transformational education in youth work: Theoretical concepts and empirical findings. *Child and Adolescent Social Work Journal.* Advance online publication. https://doi.org/10.1007/s10560-025-01012-2

Bloomberg. (2024, January 26). AI startup ElevenLabs bans account blamed for Biden audio deepfake. Bloomberg. https://www.bloomberg.com/news/articles/2024-01-26/ai-startup-elevenlabs-bans-account-blamed-for-biden-audio-deepfake

Boczkowski, P. J., Mitchelstein, E., & Matassi, M. (2018). "News comes across when I'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New Media & Society, 20*(10), 3523–3539. https://doi.org/10.1177/1461444817750396

Börnchen, S., & Pause, J. (2025). Empfehlungsalgorithmen beobachten. *MEDIENwissenschaft: Rezensionen | Reviews, 42*(3), 368–385. https://doi.org/10.25969/mediarep/24090

Börnchen, S., Mein, G., & Pause, J. (2025). *Empfehlungsalgorithmen und Öffentlich-rechtliche Medien: Ein Whitepaper für Luxemburg* (ULIDE Papers 1). Melusina Press. https://doi.org/10.26298/1981-5982-euom

Boine, C. (2023). Emotional attachment to AI companions and European law. *MIT Case Studies in Social and Ethical Responsibilities of Computing (Winter 2023).* https://doi.org/10.21428/2c646de5.db67ec7f

Bond, B. J., Dill-Shackleford, K. E., Dibble, J. L., Gleason, T. R., Jennings, N., Rosaen, S., & Forster, R. T. (2025). Parasocial relationships in children and teens. In D. A. Christakis & L. Hale (Eds.), *Handbook of children and screens: Digital media, development, and well-being from birth through adolescence* (pp. 239–244). Springer. https://doi.org/10.1007/978-3-031-69362-5_33

Boniel-Nissim, M., Marino, C., Galeotti, T., Blinka, L., Ozoliņa, K., Craig, W., Lahti, 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 (Vol. 6)*. WHO Regional Office for Europe. https://iris.who.int/handle/10665/378982

Bulut, H., Salza, G., Residori, C., Scheifer, G., Haußmann, C., & Samuel, R. (2025). *Der Blick junger Menschen auf Digitalität: Zugänge, Nutzungsmuster und Teilhabe*. In A. Schumacher, H. Käckmeister, & R. Samuel (Eds.), *Nationaler Bericht zur Situation der Jugend in Luxemburg 2025: Leben und Aufwachsen in Online- und Offline-Welten* (pp. 91–123).

Byrne, Z. S., Dvorak, K. J., Peters, J. M., Ray, I., Howe, A., & Sanchez, D. (2016). From the user's perspective: Perceptions of risk relative to benefit associated with using the Internet. *Computers in Human Behavior, 59*, 456–468. https://doi.org/10.1016/j.chb.2016.02.024

Carro, M. V. (2024).*Flattering to deceive: The impact of sycophantic behavior on user trust in large language models*(arXiv:2412.02802) [Preprint]. arXiv. https://arxiv.org/abs/2412.02802?

Carvalho, M., Branquinho, C., & de Matos, M. G. (2018). Emotional symptoms and risk behaviors in adolescents: Relationships with cyberbullying and implications on well-being. *Violence and Victims, 33*(5), 871–885. https://doi.org/10.1891/0886-6708.VV-D-16-00204

Catunda, C., Goedert Mendes, F., & Lopes Ferreira, J. (2024). *Risikoverhalten von Kindern und Jugendlichen im Schulalter in Luxemburg: Bericht über die HBSC-Umfrage 2022 in Luxemburg.*

Catunda, C., Goedert Mendes, F., Lopes Ferreira, J., & Residori, C. (2023). *Mentale Gesundheit und Wohlbefinden von Kindern und Jugendlichen im Schulalter in Luxemburg: Bericht über die HBSC-Umfrage 2022 in Luxemburg.*

Cernikova, M., Dedkova, L., & Smahel, D. (2018). Youth interaction with online strangers: Experiences and reactions to unknown people on the Internet. *Information, Communication & Society, 21*(1), 94–110. https://doi.org/10.1080/1369118X.2016.1261169

Chen, Z., & Schmidt, R. (2024). Exploring a behavioral model of "positive friction" in human-AI interaction. In A. Marcus, E. Rosenzweig, & M. M. Soares (Eds.), *Design, user experience, and usability* (LNCS 14713, pp. 1–16). Springer. https://doi.org/10.1007/978-3-031-61353-1_1

Common Sense Media. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. San Francisco, CA: Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

Cosma, A., Abdrakhmanova, S., Taut, D., Schrijvers, K., Catunda, C., & Schnohr, C. (2023). *A focus on adolescent mental health and well-being in Europe, Central Asia and Canada: Health Behaviour in School-aged Children international report from the 2021/2022 survey*. WHO Regional Office for Europe.

Costello, N., Sutton, R., Jones, M., Almassian, M., Raffoul, A., Ojumu, O., … Austin, S. B. (2024). Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media. *American Journal of Law & Medicine, 49*(2–3), 135–172. https://doi.org/10.1017/amj.2023.25

Council of Europe. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225)*. https://rm.coe.int/1680afae3c

Council of Europe Youth Partnership. (2023). *Artificial intelligence and youth work: A framework for discussion and action*. Council of Europe Youth Partnership.

Davey, J. (2021). *Gamers who hate: An introduction to ISD's gaming and extremism series*. Institute for Strategic Dialogue.

De Choudhury, M., Pendse, S. R., & Kumar, N. (2023). *Benefits and harms of large language models in digital mental health*(arXiv:2311.14693) [Preprint]. arXiv. https://arxiv.org/abs/2311.14693?

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. (2025). *Emotional manipulation by AI companions* (Harvard Business School Working Paper No. 26-005).

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (forthcoming). AI companions reduce loneliness. *Journal of Consumer Research.*

De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2023). Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology, 34*(3). https://doi.org/10.1002/jcpy.1393

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224.

Décieux, J. P., Heinen, A., & Willems, H. (2019). Social media and its role in friendship-driven interactions among young people: A mixed methods study. YOUNG, 27(1), 18–31. https://doi.org/10.1177/1103308818755516

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education*, 19(1). https://doi.org/10.14324/lre.19.1.01

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, OJ L 130. http://data.europa.eu/eli/dir/2019/790/oj/eng

Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC, OJ L 333. http://data.europa.eu/eli/dir/2022/2557/oj

Directive (EU) 2022/2555 (2022). On Measures for a High Common Level of Cybersecurity across the Union, OJ L 333. (NIS2 Directive). http://data.europa.eu/eli/dir/2022/2555/oj

Dubois, D. M. (2024). Paradoxes of generative AI: Both promise and threat to academic freedom. Journal of Academic Freedom. https://www.aaup.org/sites/default/files/Dubois_JAF15.pdf?

Ebner, P., & Szczuka, J. (2025). Predicting romantic human-chatbot relationships: A mixed-method study on the key psychological factors. *arXiv*. https://doi.org/10.48550/arXiv.2503.00195

Education Endowment Foundation. (2024, December 12). Teachers using ChatGPT—alongside a guide—can cut lesson-planning time by over 30% [Press release]. https://educationendowmentfoundation.org.uk/news/teachers-using-chatgpt-alongside-a-guide-to-support-them-to-use-it-effectively-can-cut-lesson-planning-time-by-over-30-per-cent

Ehsan, A., Klaas, H. S., Bastianen, A., & Spini, D. (2019). Social capital and health: A systematic review of systematic reviews. *SSM – Population Health, 8*, 100425. https://doi.org/10.1016/j.ssmph.2019.100425

EPRS – European Parliamentary Research Service. (2025). *Children and deepfakes* (Briefing 775855). European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf

eSafety Commissioner. (2025, February 18). *AI chatbots and companions – risks to children and young people*.https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people

European Commission, Joint Research Centre. (2025, June 13). *How is generative AI impacting our economy, society and policy?*https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/how-generative-ai-impacting-our-economy-society-and-policy-2025-06-13_en

European Expert Group on Digitalisation in Youth Work. (2019). *European guidelines for digital youth work*. Publications Office of the European Union.

European Parliament. (2025). *Children and deepfakes* (**EPRS briefing, PE 775.855**). European Parliamentary Research Service.

European Parliament, Directorate-General for External Policies. (2020). *Two briefings and an in-depth analysis on data flows, artificial intelligence and international trade: Impacts and prospects for the value chains of the future*. https://data.europa.eu/doi/10.2861/23699

Fanous, A., Goldberg, J., Agarwal, A. A., Lin, J., Zhou, A., Daneshjou, R., & Koyejo, S. (2025, February 12). *SycEval: Evaluating LLM sycophancy*. arXiv. https://arxiv.org/pdf/2502.08177

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics, 26*(3), 1771–1796.

Fraillon, J. (Ed.). (2024). *An international perspective on digital literacy: Results from ICILS 2023* (Appendix H, p. 355). IEA. https://www.iea.nl/sites/default/files/2025-03/ICILS_2023_International_Report.pdf

Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence, 7*, 100306. https://doi.org/10.1016/j.caeai.2024.100306

Fujii, M. S., Hüttmann, J., & Kutscher, N. (2020). Informelle, non-formale und formale Bildung … In K. Kaspar et al. (Eds.), *Bildung, Schule, Digitalisierung* (pp. 370–375). Waxmann.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6.

Gil de Zúñiga, H., Weeks, B., & Ardèvol-Abreu, A. (2017). Effects of the news-finds-me perception … *Journal of Computer-Mediated Communication, 22*(3), 105–123. https://doi.org/10.1111/jcc4.12185

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Global Witness. (2025). *TikTok's algorithm directs 13-year-olds to porn*. https://globalwitness.org/en/campaigns/digital-threats/tiktok-directs-13-year-olds-to-porn/

Gong, C., Li, Z., & Ma, J. (2024). Google effects on memory: A meta-analytical review. *Frontiers in Psychology, 15*, 1322894.

Green, B. L., Murphy, A., & Robinson, E. (2024). Accelerating health disparities research with artificial intelligence. *Frontiers in Digital Health, 6*. https://doi.org/10.3389/fdgth.2024.1330160

Habermas, J. (2022). *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik* (Erste Auflage). Suhrkamp.

Hall, J. A., & Liu, D. (2022). Social media use, social displacement, and well-being. *Current Opinion in Psychology, 46*, 101339. https://doi.org/10.1016/j.copsyc.2022.101339

Hang, S., Jost, G. M., Guyer, A. E., Robins, R. W., Hastings, P. D., & Hostinar, C. E. (2023). Understanding the development of chronic loneliness in youth. *Child Development Perspectives, 18*(1). https://doi.org/10.1111/cdep.12496

HBSC Luxembourg. (2023–2025). *Trends Risk Behaviours Dashboard*; national factsheets/reports. https://hbsc.uni.lu/trends-risk-behaviours/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hepp, A. (2020). *Deep mediatization*. Routledge.

Herbener, A. B., & Damholdt, M. F. (2025). Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students. *International Journal of Human-Computer Studies, 196*, 103409. https://doi.org/10.1016/j.ijhcs.2024.103409

Higgs, J. M., & Stornaiuolo, A. (2024). Being human in the age of generative AI: Young people's ethical concerns about writing and living with machines. *Reading Research Quarterly, 59*(4), 632–650.

Holmes, K. (2018). *Mismatch: How inclusion shapes design*. MIT Press. https://doi.org/10.7551/mitpress/11647.001.0001

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, R. S., & Santos, O. C. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(4), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Horwath, I. (2022). Algorithmen, KI und soziale Diskriminierung. In K. Schnegg et al. (Eds.), *Inter- und multidisziplinäre Perspektiven der Geschlechterforschung: Innsbrucker Gender Lectures IV* (pp. 71–103). Innsbruck University Press.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., … Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Computing Surveys, 56*(2). https://doi.org/10.1145/3703155

Hurrelmann, K., & Quenzel, G. (2018). *Developmental tasks in adolescence.* Routledge.

IEEE. (2024). *IEEE P7014/D004, February 2024: Ethical considerations in emulated empathy in autonomous and intelligent systems* (Draft).

INHOPE. (2024). *Annual report 2024*. https://inhope.org/inhope-annual-report-2024.pdf

Iske, S., & Kutscher, N. (2020). Digitale Ungleichheiten im Kontext Sozialer Arbeit. In N. Kutscher et al. (Eds.), *Handbuch Soziale Arbeit und Digitalisierung*. Beltz Juventa.

Jamali, L. (2025, October 14). ChatGPT will soon allow erotica for verified adults, OpenAI boss says. *BBC News.*

Jylhä, V., Hirvonen, N., & Haider, J. (2024). Algorithmic recommendations enabling and constraining information practices among young people. *Journal of Documentation, 80*(7). https://doi.org/10.1108/JD-05-2023-0102

Käckmeister, H., Biewers Grimm, S., Langehegermann, L., Meyers, C., & Samuel, R. (2025). *Der digitale Alltag junger Menschen: Zugehörigkeit, Zeitwahrnehmung und Selbstregulation*. In A. Schumacher, H. Käckmeister, & R. Samuel (Eds.), *Nationaler Bericht zur Situation der Jugend in Luxemburg 2025: Leben und Aufwachsen in Online- und Offline-Welten* (pp. 125–144).

Kechagias, K. (2025). Artificial intelligence competence needs for youth workers. *Journal of Non-Formal and Digital Education*, 1(1), 44–51. https://doi.org/10.63734/JNFDE.01.01.007

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 97652. https://doi.org/10.1038/s41598-025-97652-6

Kies, R., & Lukasik, S. (2024). *Rapport Medialux 2024: Rapport sur l'évolution des usages médiatiques au Luxembourg.*Medialux Project. https://medialux-project.lu/wp-content/uploads/2025/11/Rapport-Medialux-2024-_-Rapport-sur-levolution-des-usages-mediatiques-au-Luxembourg.pdf?

Kouros, T., & Papa, V. (2024). *Digital mirrors: AI companions and the self. Societies, 14*(10), Article 200. https://doi.org/10.3390/soc14100200

Kowert, R., Kilmer, E., & Newhouse, A. (2024). Taking it to the extreme: Prevalence and nature of extremist sentiment in games. *Frontiers in Psychology, 15*, 1410620. https://doi.org/10.3389/fpsyg.2024.1410620

Krämer, N. C., Lucas, G., Schmitt, L., & Gratch, J. (2018). Social snacking with a virtual agent: Need to belong and social responsiveness in interactions with artificial entities. *International Journal of Human-Computer Studies, 109*, 112–121. https://doi.org/10.1016/j.ijhcs.2017.09.001

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Kurian, N. (2024). "No, Alexa, no!": Designing child-safe AI and protecting children from the risks of the "empathy gap" in large language models. *Learning, Media and Technology*, 1–14. https://doi.org/10.1080/17439884.2024.2367052

Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI*. Penguin Books.

Kutscher, N., Ley, T., Seelmeyer, U., Siller, F., Tillmann, A., & Zorn, I. (Eds.). (2020). *Handbuch Soziale Arbeit und Digitalisierung*. Beltz Juventa.

Langehegermann, L., Scheifer, G. R., & Samuel, R. (in press). *Access creep: A mixed methods exploration of adolescent political participation in Luxembourg.* In G. Mein & I. Baumann (Eds.), *Democracy and youth in the digital age: Evolving technologies and political participation.*

Langehegermann, L., Käckmeister, H., Biewers Grimm, S., Scheifer, G., & Samuel, R. (2025). *Kontexte digitaler Sozialisation von jungen Menschen: Content, Gaming und künstliche Intelligenz.* In A. Schumacher, H. Käckmeister, & R. Samuel (Eds.), *Nationaler Bericht zur Situation der Jugend in Luxemburg 2025: Leben und Aufwachsen in Online- und Offline-Welten* (pp. 171–194).

Ledwich, M., Zaitsev, A., & Laukemper, A. (2022). Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday, 27*(12). https://doi.org/10.5210/fm.v27i12.12552

Leong, B., & Selinger, E. (2019*). Robot eyes wide shut. In Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19)* (pp. 299–308). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287591

Lin, S., Hilton, J., & Evans, O. (2021, September 8). TruthfulQA: Measuring how models mimic human falsehoods. arXiv. https://arxiv.org/pdf/2109.07958

Liu, D., Dang, Z., Peng, C., Zheng, Y., Li, S., Wang, N., & Gao, X. (2022). *FedForgery: Generalized face forgery detection with residual federated learning* (arXiv:2210.09563) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2210.09563

Long, D., & Magerko, B. (2020, April 21). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.

Luxembourg Centre for Educational Testing (LUCET) & SCRIPT. (2024). *Nationaler Bildungsbericht Luxemburg 2024*. Universität Luxemburg. https://doi.org/10.48746/bb2024lu-de-digipub

Luxembourg Centre for Educational Testing (LUCET). (2023). *Épreuves Standardisées (EpStan)* 2023 – Schüler:innenbefragung Sekundarstufe (Grade 9): Datensatzversion v5 [Data set]. University of Luxembourg.

Luxembourg Centre for Educational Testing (LUCET). (2024). *Épreuves Standardisées (EpStan)* 2024 – Schüler:innenbefragung Sekundarstufe (Grades 7 & 9): Datensatzversion v2 [Data set]. University of Luxembourg.

Luxembourg Institute of Science and Technology (LIST). (n.d.). *LLM leaderboard*. Retrieved December 18, 2025, from https://ai-sandbox.list.lu/llm-leaderboard/

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. https://doi.org/10.1037/a0037123

Manovich, L. (2025). Artificial subjectivity. *Magyar Nyelvőr*.

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Author correction: Loneliness and suicide mitigation for students using GPT-3-enabled chatbots. *NPJ Mental Health Research, 3*(1). https://doi.org/10.1038/s44184-024-00055-0

McAra-Hunter, D. (2024). *How AI hype impacts the LGBTQ+ community. AI Ethics, 4*(3), 771–790. https://doi.org/10.1007/s43681-024-00423-8

McStay, A. (2024). *Automating empathy*. Oxford University Press.

McStay, A. (2024). The hidden influence: Exploring presence in human-synthetic interactions through ghostbots. *Ethics and Information Technology, 26*, 48.

Mersch, D. (2020). Digital lifes: Überlegungen zu den Grenzen algorithmischer Rationalisierung. In A. Beinsteiner et al. (Eds.), *Augmentierte und virtuelle Wirklichkeiten* (pp. 53–76). Innsbruck University Press.

Meyers, C., Käckmeister, H., & Samuel, R. (2025). *Jugendliche und ihre Eltern im digitalen Zeitalter: Doing family zwischen Regeln, Aushandlungen und Ungleichheiten*. In A. Schumacher, H. Käckmeister, & R. Samuel (Eds.), *Nationaler Bericht zur Situation der Jugend in Luxemburg 2025: Leben und Aufwachsen in Online- und Offline-Welten* (pp. 195–219).

Microsoft Education Team. (2025, February 11).*Safer Internet Day 2025: Tackling abusive AI-generated content risks through education and empowerment*. Microsoft. https://www.microsoft.com/en-us/education/blog/2025/02/leading-the-way-to-a-safer-internet-together/

Mihalcea, R., Ignat, O., Bai, L., Borah, A., Chiruzzo, L., Jin, Z., … Solorio, T. (2025). Why AI is WEIRD and shouldn't be this way: Towards AI for everyone, with everyone, by everyone. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(27), 28657–28670. https://doi.org/10.1609/aaai.v39i27.35092

Mlonyeni, P. M. T. (2024). Personal AI, deception, and the problem of emotional bubbles. *AI & Society: Knowledge, Culture and Communication.* https://doi.org/10.1007/s00146-024-01958-4

Mügge, D., Paul, R., & Stan, V. (2025). *AI MATRIX: Profits, power, politics*. Agenda Publishing.

National Foundation for Educational Research. (2025). *Lesson planning using AI lesson assistant, Aila: A Teacher Choices trial* [Project description]. NFER. https://www.nfer.ac.uk/for-schools/participate-in-research/participate-in-research-projects/lesson-planning-using-ai-lesson-assistant-aila-a-teacher-choices-trial/

Nightingale, S. J., & Farid, H. (2022). *AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences of the United States of America, 119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books.

Observatoire national de l'enfance, de la jeunesse et de la qualité scolaire (OEJQS). (2025a). *OEJQS Factsheet 01/25: Cybermobbing bei Kindern und Jugendlichen in Luxemburg. Teil 1 – Ein vielschichtiges Problem.*

Observatoire national de l'enfance, de la jeunesse et de la qualité scolaire (OEJQS). (2025b). *OEJQS Factsheet 02/25: Cybermobbing bei Kindern und Jugendlichen in Luxemburg. Teil 2 – Die Auslöser und Folgen.*

OECD. (2023). *Digital education outlook 2023*. OECD Publishing.

OECD. (2019). *AI principles*. Retrieved December 9, 2025, from https://www.oecd.org/en/topics/ai-principles.html

Ofcom. (2025). *Children and parents: Media use and attitudes report 2025*. London: Ofcom.

Olaizola Rosenblat, M., & Barrett, P. M. (2023). *Gaming the system: How extremists exploit gaming sites and what can be done to counter them*. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/publication/gaming-the-system

Paas, F., Renkl, A., & Sweller, J. (2020). Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science, 29*(4), 394–398. https://doi.org/10.1177/0963721420922183

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015). *Promising evidence on personalized learning*. RAND Corporation. https://eric.ed.gov/?id=ED571009

Pane, J., Steiner, E., Baird, M., & Hamilton, L. (2017). *Informing progress: Insights on personalized learning implementation and effects*. RAND Corporation.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation. *Human Factors, 52*(3), 381–410.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. Verso Books.

Pawluczuk, A. (2023). *Automating youth work: Youth workers' views on AI*. Council of Europe Youth Partnership.

PION. (2024). *The Pion Youth Trends Report*. https://www.wearepion.com/resources/reports/youth-trends-report-2024#form

Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M.-K., Poikkeus, A.-M., & Vasalampi, K. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports, 14*(1), 12956. https://doi.org/10.1038/s41598-024-63629-0

Ramos-Soler, I., López-Sánchez, C., & Torrecillas-Lacave, T. (2018). Online risk perception in young people and its effects on digital behaviour. *Comunicar, 26*(56), 71–79. https://doi.org/10.3916/c56-2018-07

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10, 25494–25513. https://doi.org/10.1109/access.2022.3154404

Regulation (EU) 2016/679. (2017). On the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). OJ L 119. http://data.europa.eu/eli/reg/2016/679/oj

Regulation (EU) 2024/1689. (2024). Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj

Regulation (EU) 2022/1925. (2022). On contestable and fair markets in the digital sector (Digital Markets Act). OJ L 265. http://data.europa.eu/eli/reg/2022/1925/oj

Regulation (EU) 2022/2065. (2022) On a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act). OJ L 277. http://data.europa.eu/eli/reg/2022/2065/oj/eng

Reuters Institute for the Study of Journalism. (2024). *Reuters Institute digital news report 2024.* https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/RISJ_DNR_2024_Digital_v10%20lr.pdf

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). *AI-driven disinformation: Policy recommendations for democratic resilience. Frontiers in Artificial Intelligence, 8*, Article 1569115. https://doi.org/10.3389/frai.2025.1569115

Ronksley-Pavia, M., Ronksley-Pavia, S., & Bigum, C. (2025). *Experimenting with generative AI to create personalized learning experiences for twice-exceptional and multi-exceptional neurodivergent students. Journal of Advanced Academics, 36*(4), 601–639. https://doi.org/10.1177/1932202X251346349

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus, 3*(12), pgae533. https://doi.org/10.1093/pnasnexus/pgae533

Sandoval-Martin, T., & Martínez-Sanzo, E. (2024). Perpetuation of gender bias in visual representation of professions in the generative AI tools DALL·E and Bing Image Creator. *Social Sciences, 13*(5), 250. https://doi.org/10.3390/socsci13050250

Schlegel, L., & Amarasingam, A. (2022). *Examining the intersection between gaming and violent extremism.*United Nations Office of Counter-Terrorism.

Schlegel, L., & Kowert, R. (Eds.). (2024). *Gaming and extremism: The radicalization of digital playgrounds*. Routledge. https://doi.org/10.4324/9781003388371

Schobin, J. (2016). Mediatisierung der Freundschaft. In E. Alleweldt, E. A. Heuser, A. Brandt, J. Schobin, V. Leuschner, & S. Flick (Eds.), *Freundschaft heute* (pp. 169–184). transcript Verlag. https://doi.org/10.1515/9783839435502-014

Schumacher, A., Käckmeister, H., & Samuel, R. (Eds.). (2025). *Nationaler Bericht zur Situation der Jugend in Luxemburg 2025: Leben und Aufwachsen in Online- und Offline-Welten.* https://jugendbericht.lu/

Schweiger, W. (2017). *Der (des)informierte Bürger im Netz: Wie soziale Medien die Meinungsbildung verändern*. Springer.

Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT). (2025). *Auswertung KI-Umfrage: Sommer 2025*. Luxembourg: SCRIPT.

Securityhero. (2023). *2023 state of deepfakes: Realities, threats, and impact*. https://www.securityhero.io/state-of-deepfakes/#key-findings

Shanklin, R., Samorani, M., Harris, S., & Santoro, M. A. (2022). Ethical redress of racial inequities in AI: Lessons from decoupling machine learning from optimization in medical appointment scheduling. *Philosophy & Technology, 35*(4). https://doi.org/10.1007/s13347-022-00590-8

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023, October 20). Towards understanding sycophancy in language models. arXiv. https://arxiv.org/pdf/2310.13548

Shin, D., & Jitkajornwanich, K. (2024). How algorithms promote self-radicalization: Audit of TikTok's algorithm using a reverse engineering method. Social Science Computer Review, 42(4), 1020–1040. https://doi.org/10.1177/08944393231225547

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion: A study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Solyst, J., Yang, E., Xie, S., Ogan, A., Hammer, J., & Eslami, M. (2023). The potential of diverse youth as stakeholders in identifying and mitigating algorithmic bias for a future of fairer AI. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), 1–27. https://doi.org/10.1145/3610213

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory. *Science, 333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Stalder, F. (2016). *Kultur der Digitalität*. Berlin: Suhrkamp. http://www.content-select.com/index.php?id=bib_view&ean=9783518736180

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970–987. https://doi.org/10.1037/a0032447

Stinar, F., et al. (2025). Fairness of Bayesian Knowledge Tracing for math: An evaluation across reading-ability subgroups. In *Proceedings of the International Conference on Educational Data Mining*. https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.long-papers.158/index.html

Strasser, A., & Wilby, M. (2023). The AI-stance: Crossing the terra incognita of human–machine interactions. In H. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social robots in social institutions*. Amsterdam: IOS Press.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-03811-x

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Szczuka, J., & Krämer, N. (2017). Not only the lonely—How men explicitly and implicitly evaluate the attractiveness of sex robots in comparison to the attractiveness of women, and personal characteristics influencing this evaluation. *Multimodal Technologies and Interaction, 1*(1), 3. https://doi.org/10.3390/mti1010003

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

UNESCO & IRCAI. (2024). *Challenging systematic prejudices: An investigation into gender bias in large language models.*

UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). *UNESCO's recommendation on the ethics of artificial intelligence: Key facts*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385082

UNESCO. (2023). *Guidance for generative AI in education and research*. Paris: UNESCO.

Vermeire, L., & Van den Broeck, W. (2023). On the role of digitalisation in youth work and non-formal learning in the context of the European youth programmes (RAY-DIGI): National case study analysis for Flanders (Belgium). JINT; imec-SMIT, Vrije Universiteit Brussel. https://www.jint.be/sites/default/files/2023-06/RAY%20DIGI_%20Vermeire-Van%20den%20Broeck_2023%20Report.pdf

Villanueva Blasco, V. J., & Serrano Bernal, S. (2019). Patrón de uso de internet y control parental de redes sociales como predictor de sexting en adolescentes: Una perspectiva de género. *Revista de Psicología y Educación – Journal of Psychology and Education, 14*(1), 16. https://doi.org/10.23923/rpye2019.01.168

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y

Wiese, L. J., Patil, I., Schiff, D. S., & Magana, A. J. (2025). AI ethics education: A systematic literature review. *Computers and Education: Artificial Intelligence, 8*, 100405. https://doi.org/10.1016/j.caeai.2025.100405

Wilding, R., Baldassar, L., Gamage, S., Worrell, S., & Mohamud, S. (2020). Digital media and the affective economies of transnational families. *International Journal of Cultural Studies, 23*(5), 639–655. https://doi.org/10.1177/1367877920920278

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. SAGE.

Williamson, B. (2019). New power networks in educational technology. *Learning, Media and Technology, 44*(4), 395–398. https://doi.org/10.1080/17439884.2019.1672724

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995

Winstone, L., Mars, B., Haworth, C. M. A., & Kidger, J. (2021). Social media use and social connectedness among adolescents in the United Kingdom: A qualitative exploration of displacement and stimulation. *BMC Public Health, 21*(1), 1736. https://doi.org/10.1186/s12889-021-11802-9

Yang, A., & Yang, T. A. (2024, June). Social dangers of generative artificial intelligence: Review and guidelines. In *Proceedings of the 25th Annual International Conference on Digital Government Research* (pp. 654–658). https://doi.org/10.1145/3657054.3664243

Yu, Y., et al. (2025). Understanding generative AI risks for youth: A taxonomy based on empirical data. *arXiv preprint arXiv:2502.16383*. https://arxiv.org/abs/2502.16383

Zambrano, A. F., Baker, R. S., Gowda, S. M., Amon, M. J., & Zepeda, C. (2024). Investigating algorithmic bias on Bayesian Knowledge Tracing and carelessness detectors. In *Proceedings of the 14th International Learning Analytics & Knowledge Conference*(pp. 1–11). https://doi.org/10.1145/3636555.3636890

Zhang, Y., & He, A.-Z. (2025). *The impact of AI influencer endorsements on consumers' purchase intentions: The serial mediating roles of mind perception and brand trustworthiness. Journal of Product & Brand Management, 34*(5), 754–765. https://doi.org/10.1108/JPBM-02-2024-4974

Zhao, J. (2025). Generative AI and educational assessments: A systematic review. *Educational Research and Practice, 51*(6), 1–22. https://www.erpjournal.net/wp-content/uploads/2025/01/ERP_V51_2024_6_Generative_Jian.pdf

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Glossary of Key Terms

Affective Data Data related to emotions, moods, or attitudes, often inferred from facial expressions, voice, behaviour, or interaction patterns.

Affective Computing A field of AI focused on recognising, interpreting, and responding to human emotions.

AI (Artificial Intelligence) The field of computer science focused on creating systems capable of tasks that typically require human intelligence, such as learning, reasoning, perception, language understanding, and decision-making.

AI Agents Autonomous software programs that perform tasks or simulate conversations based on AI models. They can operate independently, make decisions, and interact with users or systems in a goal-oriented manner.

AI Companions AI-driven applications designed to simulate relationships such as friendship, empathy, or emotional support, often used in contexts of loneliness, mental health, or social substitution.

AI Literacy The knowledge and skills needed to understand, evaluate, and use AI systems critically and responsibly.

AIoT (Artificial Intelligence of Things) The integration of artificial intelligence with Internet of Things devices, allowing for context-aware and adaptive responses.

Anthropomorphism The attribution of human traits, emotions, or intentions to non-human entities, including AI systems.

Automation Bias The tendency to over-rely on automated systems, often ignoring contradictory information or failing to monitor system output critically.

Chatbot An AI program designed to simulate conversation with human users, often used in customer service, education, or companionship.

Cognitive Offloading The act of delegating mental processes to external tools or systems, such as using AI for memory, reasoning, or decision-making.

Companion AI AI agents designed to simulate friendship, empathy, or emotional support, often used in chat-based platforms.

Crisis-Response Capacity The ability of an AI system or platform to recognise signs of user distress and escalate appropriately to human support or safeguards.

CSAM (Child Sexual Abuse Material) Illegal content depicting the sexual abuse or exploitation of children, including increasingly synthetic versions generated by AI.

Deepfake Synthetic media (e.g. videos, images, voices) that imitate real people using AI-based manipulation, often used for deception or harm.

Digital Snacking / Social Snacking The use of brief, low-effort social interactions, including with AI, to reduce feelings of loneliness or social isolation.

Emotional AI Systems designed to detect, simulate, or strategically respond to human emotions to create the appearance of empathy.

Filter Bubble A state of intellectual isolation caused by personalised content filtering, where users are exposed mainly to views and information that reinforce their own beliefs.

GENAI (Generative Artificial Intelligence) AI that creates new content such as text, images, video, or audio based on user input and training data.

Gamification The use of game design elements (e.g. points, badges, rewards) in non-game contexts to increase motivation and engagement, including in educational or awareness-raising AI settings.

Hallucination (AI) A phenomenon in which AI systems produce information that sounds plausible but is in fact false or fabricated.

Harmful Content Online material that may not be illegal but can negatively affect users psychologically or socially, including bullying, body shaming, or misinformation.

LLM (Large Language Model) A type of generative AI trained on vast datasets of human language to generate coherent, context-sensitive text outputs. Examples include GPT, Google Gemini, Microsoft Copilot, or Claude.

Media Literacy The ability to critically access, analyse, evaluate, and create media in various forms, now extended to AI-generated content.

News-Finds-Me Perception The belief that important news will reach one without actively seeking it, often through social media or recommender systems.

Nudifying App A type of AI-based application that falsely generates sexualised or nude images of clothed individuals, often used for abuse or extortion.

Parasocial Relationship A one-sided emotional attachment where users relate to a media persona (including AI agents or influencers) as if they were real social partners.

Peer-to-Peer Approach An educational or support strategy in which individuals from the same age group or social background share knowledge or experiences to build trust and relevance.

Profiling The use of data analytics and AI to infer personal characteristics, preferences, or risks for decision-making purposes.

Recommender System An AI system that suggests content to users based on their preferences, behaviour, or inferred interests.

RLHF (Reinforcement Learning from Human Feedback) A machine learning technique that trains AI models to align their responses with human preferences by using evaluative feedback.

Simulated Empathy The appearance of emotional understanding generated by AI systems without actual emotional experience.

Sextortion A form of online blackmail where perpetrators threaten to share sexual images or information unless the victim complies with demands.

Synthetic Media Media content wholly or partially generated by AI, including text, images, audio, and video.

Synthetic CSAM AI-generated depictions of child sexual abuse material, often involving deepfake techniques, posing legal and ethical challenges.

Transactive Memory A shared system for encoding, storing, and retrieving information among groups or between humans and digital tools.

Virtual Influencer AI-generated characters designed to engage audiences on social media platforms, often used in marketing and branding.

WEIRD Societies An acronym for Western, Educated, Industrialised, Rich, and Democratic societies, often used in social science to describe sampling biases in research.