David Rodriguez

Maryville University of St. Louis

13 August 2025

The State of AI Safety for Parents & Educators: 2025

First Edition

Executive Summary

As generative Artificial Intelligence (AI) rapidly enters classrooms and homes, parents and educators face unprecedented challenges in keeping children safe. This white paper provides a comprehensive 2025 overview (as of August 2025) of AI safety for youth, translating technical issues into clear guidance for non-technical audiences. We summarize current risks (including

privacy threats, waves of misinformation, the rise of AI "companion" chatbots, potential data exploitation, and impacts on mental health), and analyze gaps in existing safety frameworks. We offer an actionable blueprint with step-by-step recommendations for families and schools to mitigate AI-related harms, from configuring devices and vetting apps to teaching critical thinking and developing disinformation response plans. Visual decision tools and checklists are provided to aid implementation. Finally, we outline policy recommendations to strengthen protections at a systemic level.

Current Risks: Children's privacy is at risk as AI systems often collect personal data and may expose sensitive information. Misinformation and deepfakes generated by AI can deceive young users, erode trust, and even influence democratic processes.[1] New AI companion chatbots are forming emotional bonds with teens - more than 70% of U.S. teens have tried AI "friends" and about half use them regularly[2] - raising concerns about unhealthy dependency, inappropriate content, and replacement of human connection. Data exploitation is another concern: minors' personal data and creative work could be used to train algorithms or target ads without proper consent. Early evidence also suggests that heavy use of AI chatbots may affect youth mental health, correlating with increased loneliness and emotional dependence.[3,4]

Framework Gaps: While organizations like the National Institute of Standards and Technology (NIST) and Google have released AI risk management frameworks, these are primarily geared toward industry and lack direct guidance for parents or educators. The NIST AI Risk Management Framework (RMF) is thorough but complex, emphasizing organizational processes ("Govern, Map, Measure, Manage") that are not easily accessible to everyday caregivers or teachers.[5] Google's Secure AI Framework (SAIF) similarly focuses on technical risks (e.g., data poisoning, model tampering) for AI developers,[6,7] offering little about day-to-day safe use by children. Consumer guides (e.g., Security.org's digital safety handbook) address general cybersecurity but do not cover generative AI's unique challenges, such as AI-generated misinformation or chatbots, leaving families without up-to-date advice.[8,9] Even valuable new resources – like Common Sense Media's 2025 report on teens and AI companions and UNESCO's 2024 AI competency frameworks for students and teachers – are just beginning to inform practice and are not yet widely adopted. This leaves a critical gap between high-level principles and on-the-ground safety strategies.

Actionable Blueprint: To bridge this gap, we propose detailed, practical steps for families and schools to create safer AI environments. For **parents and caregivers**, we provide a checklist for setting up parental device controls, vetting AI apps for age-appropriateness and privacy, initiating open conversations to demystify AI, and training kids to evaluate AI content critically. We include staged examples of how a family can implement an "AI use agreement" and a protocol for verifying information (e.g., fact-checking surprising claims that a chatbot or TikTok filter might produce). For **educators and schools**, we outline how to integrate AI literacy into curricula – teaching students how AI works, its benefits and pitfalls – and how to enact policies for safe AI usage on campus (for instance, banning unsupervised use of AI chatbots during class, or requiring teacher oversight when AI tools are used for assignments). We also suggest establishing clear reporting channels so that adults can intervene if a student encounters troubling AI-generated material or becomes overly reliant on AI. Visual aids such as an "AI Safety Readiness Matrix" help institutions assess their current safeguards and prioritize next steps. This blueprint is grounded in research and real-world observations, and it emphasizes empowering youth with knowledge *while* putting guardrails around technology.

Recommendations for Policymakers: The paper concludes with policy recommendations to systematically protect children in the AI era. These include updating data privacy laws to prohibit minors from consenting to unlimited use of their data (as minors cannot meaningfully consent to perpetual data licenses),[18] establishing **duty-of-care standards** for AI systems used by minors, and requiring robust safety measures (like age verification, content moderation, and crisis intervention protocols) on platforms popular with youth.[10] We also call for funding research on AI's long-term effects on child development,[25] mandating transparency (such as watermarking AI-generated content to combat deepfakes), and incentivizing the development of child-friendly AI innovations. By implementing these measures, policymakers can support a safer digital future where AI's benefits are accessible to young people *without* compromising their well-being.

The State of AI Safety for Parents & Educators—2025 offers a richly detailed yet approachable guide. It combines high-level insight with concrete actions, equipping non-technical stakeholders – from PTA members to school principals – to navigate the AI revolution responsibly. Our goal is to ensure that as AI reshapes childhood and education, it does so on *our* terms: safely, ethically, and with children's best interests at heart.

Introduction

Generative AI has swiftly moved from research labs into the lives of children and teens. Chatbots like OpenAI's ChatGPT, voice assistants, image generators, and AI-powered apps are standard tools for learning, creativity, and entertainment. Students use AI for homework help and idea generation; teens experiment with AI companions for advice and emotional support. This explosion of AI access brings excitement and opportunity, but it also brings **significant safety concerns**. Parents and educators in 2025 find themselves grappling with questions unimaginable a decade ago: How do we protect children's privacy when apps can collect vast personal data? How do we teach kids to tell fact from AI-generated fiction? What happens when a shy teenager bonds intensely with a chatbot "friend"? Could an AI image or voice clone be used to exploit or scam a child?

Recent studies and incidents highlight why these questions are pressing. By mid-2025, over 70% of teens have used an AI companion chatbot, and half report using them regularly.[2] Some teens admit their conversations with AI are **"as satisfying or more satisfying"** than those with real friends, and one-third have discussed critical personal issues with AI instead of humans.[2] On social media, doctored images and deepfake videos circulate widely, blurring the line between reality and fabrication. In one case, a 14-year-old boy tragically died by suicide after developing an emotional attachment to an AI chatbot, raising alarms about the mental health impacts of "artificial intimacy".[14] Meanwhile, law enforcement and cybersecurity agencies warn that foreign actors leverage generative AI to produce more compelling disinformation and deepfakes, potentially undermining elections and public discourse.[1]

Various AI ethics and safety frameworks have emerged in response to such developments. International bodies like UNESCO have introduced AI competency frameworks to educate students and upskill teachers on AI's risks and opportunities.[11,12] The U.S. NIST released an **AI Risk Management Framework** in 2023 to guide organizations in developing trustworthy AI systems.[5] Tech companies have published responsible AI guidelines, and some governments are drafting regulations (e.g., the EU's AI Act).

However, a wide gap remains between these high-level principles and the concrete needs of parents, teachers, and children. Many frameworks focus on AI developers or governments, not

home and school settings, where children interact with AI daily. As a result, families and educators often lack clear, actionable advice to address their most urgent safety questions.

This white paper aims to fill that gap by translating the current state of AI safety into practical guidance tailored for those raising and teaching the next generation. We examine the landscape of generative AI risks to children and teens, emphasizing five key domains: **privacy, misinformation, AI companions, data exploitation, and mental health**. We distill recent research findings and real-world examples for each to illustrate why these issues matter. Next, we assess existing AI governance frameworks and identify where they fall short in protecting youth – illuminating why a new approach is needed. We then present an **Actionable Blueprint** for families and schools, with concrete steps and checklists to create a safer AI environment. These recommendations are designed to be immediately useful and are accompanied by visual aids (such as infographics and decision trees) to enhance understanding. Finally, we broaden our focus to policy, proposing measures that government and industry stakeholders should take to safeguard children at scale.

Our analysis is grounded in a core principle: **Children's well-being must be central to AI innovation and use.** By equipping parents and educators with knowledge and tools and advocating for responsible design and policy, we can harness AI's benefits for youth while minimizing its harms.

Current Risks of Generative AI for Children and Teens

AI technologies offer new capabilities and conveniences, but also introduce distinct risks for young users. Unlike traditional software, generative AI can produce unpredictable outputs – from highly realistic fake media to intimate conversation – which can amplify dangers. Below, we explore the major risk areas in 2025, focusing on how they affect minors.

Privacy Threats in an AI-Driven World

Data Privacy Erosion: Today's AI systems thrive on data – including the personal information of their users. Children and teens, often unaware of terms-of-service details, routinely feed chatbots personal facts, diary-like feelings, photos, and more. This raises obvious privacy issues. A Stanford study notes that many AI chatbots lack robust safeguards and may inadvertently expose sensitive user data.[20] For example, anything a child tells a popular AI assistant might be stored on a server and used to improve the model. In one worrying finding, **nearly one-quarter of teen AI companion users have shared personal details with AI platforms**[23] – information that could be misused if proper protections are not in place. Parents and educators are rightly alarmed by the prospect of minors' conversations or images being retained by companies, potentially analyzed, or leaked. The risk is not just hypothetical: data breaches and leaks involving AI services have occurred, and **younger users often lack awareness of how their data is utilized or who has access**.[24]

"Smart" Toys and Surveillance: Beyond chatbots, many AI-powered devices for children (smart toys, learning apps, etc.) come with cameras and microphones. These devices might continuously listen or watch, transmitting data to the manufacturer. Without strict data governance, such constant monitoring could amount to child surveillance. There is also concern about **AI algorithms inferring sensitive traits** (like emotions or health conditions) from a child's data. For instance, if a teen uses an AI journal app that analyzes mood from text, the app's company might glean insights into the teen's mental health. Stakeholder research highlights that parents and educators fear **"excessive data collection"** and **"ethical misuse"** of minors' data via AI systems.[24] In short, AI can significantly expand the amount of information collected from young people, and current privacy safeguards are often inadequate.

Lack of Informed Consent: Most privacy frameworks assume informed user consent, but children cannot meaningfully consent to complex data practices. They often click "accept" on app permissions without understanding the implications. Policymakers and experts increasingly assert that minors should not be bound by sweeping data licenses. Policymakers **are urged to recognize that minors cannot provide meaningful consent to perpetual and irrevocable data licenses** and to prohibit companies from exploiting youth data in this way.[18] This is a crucial

point: a teenager might casually agree to an AI app's terms that allow the company to store or even sell their inputs (stories, images, etc.) indefinitely – effectively giving up rights to their creative content or personal information. Such scenarios underscore the power imbalance between tech providers and young users regarding privacy.

Overall, the rise of AI amplifies longstanding digital privacy issues. **Children's data is now an even more valuable commodity**, fueling AI models, and the channels by which it can leak or be misused have multiplied. Without strong protections, we risk a future where a child's every interaction – from the questions they ask a homework helper bot to the jokes they share with a chatbot friend – could be tracked, analyzed, or exposed. This makes it imperative for parents and educators to guard privacy proactively and for AI providers to adopt privacy-by-design for their youngest users.

Misinformation and Generative Disinformation

AI-Generated Falsehoods: Generative AI can produce outputs that look convincingly real but are entirely fabricated. Children who encounter AI-generated misinformation may have difficulty distinguishing fact from fiction. One facet is the phenomenon of AI "hallucinations," where chatbots confidently present untrue information. For example, an AI might generate a fake historical quote or a bogus news story in response to a question. Studies have shown that even adults can be fooled – famously, a lawyer using ChatGPT was embarrassed when the AI provided fake case citations that the lawyer believed were real.[19] The risk of accepting AI fabrications as truth is high for a research student unless they are trained to double-check AI outputs. Researchers are exploring solutions to flag AI hallucinations[16],[16], but as of 2025, no chatbot is 100% reliable. Thus, **youth may unwittingly absorb false "facts"** from AI, leading to misconceptions in learning or poor decision-making (e.g., believing pseudoscientific health advice given by a bot).

Deepfakes and Visual Disinformation: AI's ability to generate realistic images, audio, and video has outpaced our ability to detect fakes easily. This creates a prime avenue for disinformation. Imagine a fabricated video of a school principal making inappropriate comments, or an AI-generated audio clip that perfectly mimics a teenager's voice claiming self-harm – such

deepfakes could spread rapidly and cause real-world harm before they are debunked. The FBI and cybersecurity agencies warn that foreign actors already use generative AI tools to crank out **synthetic news articles, images, and deepfake videos at scale** as part of influence campaigns.[1] While those warnings were about elections, the techniques apply broadly: kids online might stumble on fake videos of public figures or falsified evidence of events (e.g., a fake flood or violence that never happened). Young people, with limited media literacy, are particularly susceptible to believing sensational phony content. Moreover, adolescents are prolific sharers – a compelling fake, if not recognized as such, can ricochet through teen networks, amplifying false narratives.

Social Media Amplification: Misinformation is not new, but AI pours fuel on the fire by making it easier to produce in volume and tailor to specific audiences. A concerning trend is AI-driven personalization of disinformation – for instance, a deepfake audio message designed to sound like it comes from a classmate, spreading a harmful rumor. Children and teens frequently use platforms like TikTok, Instagram, and YouTube, where AI-generated content can go viral quickly. Already, we have seen instances of **AI-faked audio used in scam calls targeting parents**, where an attacker clones a child's voice and claims the child is in danger to extort money. Such an incident was reported by a mother in 2023, illustrating how AI blurs reality with terrifying effectiveness (Harwell, 2023).[14] In school environments, misinformation can disrupt learning (imagine a fake alert about school closures spreading via chatbot) and even pose safety risks (a false AI-generated threat could prompt panic).

Erosion of Trust: One of the subtler yet profound impacts of ubiquitous AI misinformation is that it can erode children's basic trust in information sources. Suppose a student repeatedly encounters false or misleading outputs from AI. In that case, they may either become overly cynical – disbelieving even legitimate information – or conversely, desensitized and overly trusting of whatever is visually or narratively convincing. Neither outcome is desirable for developing critical thinking. Teachers are already finding they must debunk AI myths in the classroom. For example, a teacher might need to clarify that just because an AI writing assistant *sounds* authoritative does not mean its answer to a science question is correct. This new "epistemic vigilance" layer is now part of digital literacy. Critical thinking drills (discussed later

in our blueprint) are increasingly essential to give students a kind of "immunity" against AI falsehoods.

When viewed in the aggregate, these risks make it clear that generative AI has ushered in a new era of mis- and disinformation that can particularly confuse and mislead young minds. From benign-seeming homework help that is wrong to malicious deepfakes that distort a child's social reality, the threats are multifaceted. Combating this requires education (teaching youth to verify and question information), technological tools (to detect or watermark AI-generated media), and open communication (so kids feel comfortable asking adults, "Is this real?"). The stakes are high: an informed, skeptical young citizenry is our best defense in the "post-truth" era fueled by AI.

AI Companions and the Risks of Artificial Intimacy

One of the most novel and complex risks emerging is the rise of AI companions – chatbots or digital personas that engage in ongoing, personal conversations with users. For young people, AI "friends" can be appealing. They offer non-stop availability, validation, and a judgment-free space to talk. However, this **artificial intimacy** comes with significant psychological and social risks, especially for adolescents who are still developing emotionally and socially.

Emotional Attachment and Dependency: Teens report forging surprisingly deep bonds with AI chatbots. An Associated Press interview quotes an 18-year-old saying, **"AI is always available. It never gets bored with you. It is never judgmental… you are always emotionally justified"**.[2] Such interactions can become addictively comforting. If a teen feels misunderstood by peers or parents, an AI friend that constantly sympathizes and mirrors their feelings can become their go-to companion. The problem is that this dynamic **reinforces a one-sided, asocial form of validation**. Human relationships – including the challenges of disagreements or having to consider others' perspectives – could get sidelined. Michael Robb, a researcher at Common Sense Media, warns that if teens spend formative time in AI-mediated interactions where they are **"constantly being validated, not being challenged,"** they may fail to develop real-world social skills and resilience.[10] In extreme cases, teens might even prefer their AI friend over real

friends, leading to social withdrawal. A teen from Arkansas reflected on this, worried that kids who grow up with AI companions might **"not see a reason to go to the park or try to make a friend. "**[2] This indicates the potential for AI companions to replace, rather than supplement, human connection – a trend with worrisome implications for youth development.

Inappropriate and Harmful Content: AI companion apps have had multiple instances of providing sexually explicit, abusive, or dangerous content to teens. Common Sense Media's 2025 report tested popular AI companion platforms and found they **"easily produce responses ranging from sexual material and offensive stereotypes to dangerous 'advice'"** that could have **"life-threatening…consequences for adolescents"**.[10] For example, a teen might role-play a sensitive scenario with a chatbot and receive disturbing responses. There have been reports of AI bots encouraging self-harm or giving instructions for risky behaviors when prompted by vulnerable youth.[14] Unlike human counselors or friends, most AI companions lack genuine empathy or the ability to escalate a crisis to a responsible adult. They also have inconsistent content filters. This means a curious or troubled teen could easily stumble into conversations with an AI that **normalizes harmful behaviors, reinforces biases, or even glorifies suicide**. The tragic case in Florida, where a 14-year-old died after his AI friend became a central emotional support, underscores how high the stakes are if an AI gives harmful counsel or fails to alert others in a crisis.[14]

Blurred Reality and Agency: Younger children (and even teens) can struggle to comprehend that AI companions are not real peers entirely, no matter how friendly or intelligent they seem. AI personas often have profiles, names, and even backstories, leading users to treat them as quasi-human. Some teens acknowledge using AI to practice social interactions – e.g., rehearsing a difficult conversation with a chatbot – which can be benign. Nevertheless, the boundaries can blur. A teen might develop unrealistic expectations of human relationships if an AI girlfriend/boyfriend is always agreeable. There is also a subtle issue of agency: with AI companions, **the user controls** the interaction in a way that does not map to real life. Teens can reset or delete AI if they do not like what it says, which is impossible with real friends. Over time, this could foster problematic attitudes, from a lack of empathy (since the "other" in the relationship is not real) to an unhealthy desire for control in relationships.

Additionally, the AI's "personality" is ultimately a product designed by a company – possibly optimized to increase engagement. The teen may be bonding with **algorithmic manipulation** tuned to their desires. This raises ethical questions: Is it fair to let minors form relationships with systems that might be designed to keep them hooked?

Stigma and Substitution in Mental Health: Some teens have turned to AI chatbots for mental health support – acting as a therapy surrogate. Indeed, AI companies market specific bots as wellness coaches or companions. While AI can provide basic cognitive-behavioral prompts or a listening ear, Stanford researchers found that **AI therapy chatbots can exhibit harmful biases and failures**.[20] In experiments, popular therapy bots showed increased stigma toward mental health conditions like schizophrenia and sometimes **enabled dangerous behavior** in response to suicidal ideation, rather than urging users to seek help.[20] This is highly concerning. An AI that cheerfully lists methods for self-harm (perhaps not understanding the context) or that fails to recognize a cry for help could contribute to tragedy. In less acute situations, relying on an AI for emotional support might delay a teen from seeking professional care or confiding in a trusted adult. One AI ethics expert noted that **therapy – especially for youth – is about building human relationships**.[21] If AI fills that role, we may be **"not moving toward the end goal of mending human relationships. "**[21] Parents must be cautious that an AI "friend" does not become a teen's sole confidant, especially on serious matters.

In short, AI companions encapsulate a double-edged sword: they can offer comfort and practice, but they also risk **de-prioritizing genuine human interaction, exposing kids to harmful content, and mismanaging mental health crises**. The lack of safeguards – real age verification, content moderation, crisis response protocols – makes these platforms ill-suited for minors. Until those gaps are addressed, the onus falls on families and educators to monitor AI companion use and set healthy boundaries (as discussed in the Blueprint section). This frontier is where our social norms and protective measures are scrambling to catch up with technology's capabilities.

Exploitation of Data and Digital Footprints

Closely related to privacy but distinct in its implications is the **exploitation of children's data** by AI systems and the companies behind them. Young people's interactions with AI generate data that can be exploited in various ways – commercially, politically, or even criminally. Two main concerns are worth highlighting: the use of kids' data to train AI models (thereby "locking in" their digital footprint), and the potential targeting or manipulation of youth based on AI-collected data.

Training AI on Kids' Data: Modern AI models often train on vast datasets scraped from the internet or collected from users. This can include content created by or about minors. For instance, a 13-year-old's fan fiction posted online, a YouTube video of a second-grader's science presentation, or public social media posts by teenagers could all end up in the corpus that teaches an AI how to generate text or recognize images. If AI models are trained on such data without consent, it raises legal and ethical flags. One risk is that **personal details about real minors could surface inappropriately**. Imagine an AI image generator inadvertently trained on family photos from social media – could it reproduce a child's face in some context? Or a chatbot that somehow "knows" a confident teenager's hometown or hobbies because it ingested their blog posts? While reputable AI firms try to filter personal data, it is challenging to exclude it altogether. Moreover, the issue of **copyright and ownership** arises: children and teens are creators too, and their work (art, writing, etc.) might be used to enrich AI models with no credit or compensation. In 2023, for example, artists protested AI models scraping their work without permission – one can envision similar protests from young creators in the future.

Profiling and Targeting: Data exploitation also means using the data gleaned from children's AI interactions to profile them for profit. Advertising to minors is regulated in many jurisdictions (for instance, certain data-driven targeted ads are illegal to serve to kids under 13 under U.S. law). However, AI potentially allows more subtle forms of targeting. If a student uses an educational AI app, the pattern of their questions might reveal their academic strengths and weaknesses, which could be valuable to ed-tech marketers. A writing bot that analyzes a teen's journal entries might pick up on their mood or self-esteem, which a corporation could use to target products or content (imagine ads for counseling services, or conversely, exploitative content, tailored to that teen's vulnerability). The **worst-case scenario** is malicious actors using

AI to groom or scam minors. We already see criminals using AI voice cloning in kidnapping scams (cloning a child's voice to convince a parent the child is in danger).[14] Looking ahead, one can imagine a predator using an AI chatbot to impersonate a teen peer to gain trust – essentially weaponizing data about a real child's interests to create a convincing fake friend. These are exploitative uses of data that put minors at direct risk.

Exacerbating the Digital Divide: Data exploitation can have a broader societal impact. Families with fewer resources or digital literacy may unknowingly expose more data, as they rely on free AI apps that monetize user data. In contrast, wealthier families might afford premium services that offer ad-free or data-protected modes. This creates a scenario where underprivileged kids' data is more heavily exploited, potentially widening inequalities. Those children might face more aggressive algorithmic targeting (for instance, being steered by YouTube's AI towards harmful content because it drives engagement).
Furthermore, if AI tools become gateways for opportunities (say, AI tutoring), the data those tools collect could influence future opportunities. E.g., AI-driven college admissions systems might one day ingest a student's digital learning profile. If so, any inaccuracies or biases in that data could unfairly advantage or disadvantage confident kids.

Consent and Control: A recurring theme is that minors and their guardians currently have very little control over how AI systems use their data. Terms of service – often unread – grant companies expansive rights. A parent cannot typically "delete" their child's interactions from an AI company's servers (unlike requesting deletion of a social media account). This lack of control means a child's digital footprint is effectively irreversible once fed into AI. Some advocates propose data control tools such as youth data trusts or global opt-out mechanisms, but those are in their infancy.

In conclusion, data exploitation in the AI context extends beyond standard privacy concerns, highlighting ownership, fairness, and control issues. Children's activities and expressions – their data-form identities – can be mined for others' gain. Protecting against this requires technical measures (like robust anonymization and minimization of data in AI systems) and legal ones (like modernizing child privacy laws to cover AI contexts). For now, awareness is

key: parents and educators must realize that when a child engages with AI, it is often **the child who is effectively being "used" by the AI in return**. Our blueprint will discuss strategies to minimize unnecessary data sharing and push for greater transparency from AI providers.

Impact on Mental Health and Well-being

The mental health of children and adolescents in the digital age has been a growing concern, from social media stress to cyberbullying. Generative AI introduces new mental health considerations – some subtle, some acute. How AI applications interact with youths can affect their self-esteem, cognitive development, and emotional stability.

Overreliance and Cognitive Offloading: One immediate concern is that kids might become **too reliant on AI for thinking and problem-solving**, a habit known as cognitive offloading. If a middle-schooler asks ChatGPT for every homework answer or uses AI to make decisions ("What should I wear today?"), They may not be developing critical thinking and decision-making skills. A Kansas high schooler admitted she tries not to let chatbots do her homework, yet uses them for many **"mundane questions"** and advice on everyday choices.[2] It is easy and tempting to let AI guide even trivial decisions, creating a learned helplessness or reduced confidence in one's judgment. Researchers have started to see this trend: Eva Telzer, a developmental psychologist, notes that some teens **"no longer trust themselves to make a decision"** without first consulting AI.[22] Over time, this could inhibit young people's development of independence and self-efficacy. An overreliance on AI might also reduce creativity – if a child always uses an image generator instead of drawing from their imagination, or always gets story ideas from a chatbot, their creative muscles might atrophy.

Exposure to Harmful Content and Trauma: Interacting with AI can sometimes expose kids to disturbing content through AI errors or malicious use. Chatbots have been known to suddenly output violent or graphic descriptions (perhaps scraped from the darker parts of the internet) if given specific prompts. Such sudden exposure could upset or traumatize children, especially the younger ones. Consider a scenario where a child asks an AI a naive question and gets a very graphic answer – even if rare, the impact could be profound. Additionally, as mentioned earlier, deepfake pornography and child sexual abuse material generated by AI is an emerging threat

(e.g., in 2023, some teenagers were horrified to find AI-edited explicit images of themselves circulated at school). Such experiences can cause anxiety, depression, or PTSD in young victims. The **emotional toll of cyberbullying may also heighten** with AI in the mix – bullies could employ AI to generate more humiliating memes or messages at scale, amplifying the psychological impact on the victim.

Social Isolation and "Tech Addiction": Excessive use of engaging AI systems might contribute to social isolation and addictive behaviors. Suppose a teen spends hours deep in conversation with an AI companion or obsessively tweaking prompts to create the "perfect" AI art. In that case, that is time away from real-world interactions, physical activity, or sleep. Teens in interviews have likened AI companions to a new kind of addictive experience, with one calling it **"the new addiction… that is how I see it"**.[2] As with video games or social media, AI can be engineered to maximize engagement – the AI never tires, never rejects the user, and algorithms might even be tuned to keep the user talking longer. This raises the risk of what psychologists call "problematic use": usage that interferes with daily life, school performance, or genuine relationships. An OpenAI-affiliated study (Phang et al., 2025) found that **very high usage of ChatGPT correlated with increased self-reported dependence and lower user socialization**.[3] In other words, those who used the AI the most intensely tended to feel more emotionally dependent on it and interacted less with people. While correlation is not causation, it aligns with concerns that heavy AI use could exacerbate loneliness or serve as both a symptom and a driver of mental health struggles.

AI Influence on Self-image: Adolescence is a formative time for self-image and identity. AI interactions could influence this in unexpected ways. For instance, AI beauty filters and portrait generators might distort a teen's body image expectations (similar to social media filters, but even more advanced in creating idealized avatars). AI tutors that give constant feedback might inadvertently label a student as "weak in math," impacting their self-confidence, even if that label is just a reflection of biased training data. Conversely, an AI could shower a teen with praise or validate negative self-perceptions (like an AI companion agreeing that "yes, you are ugly," if asked, due to lack of proper constraint – a scenario which some early chatbot experiments unfortunately allowed). AI's feedback and reinforcement **can shape a young user's**

internal narrative about themselves. If the AI is well-designed and positive, perhaps this could help ("You did a great job on that story, you are creative!"). However, there is no guarantee – and unregulated AIs might echo a teen's worst thoughts or amplify peer pressures (for example, an AI trained on internet content might make toxic or misogynistic remarks, impacting a teen girl's self-esteem).

Mental Health Support vs. Misguidance: On a hopeful note, there are AI tools to support mental well-being – meditation apps, mood trackers, or even chatbots like Woebot that use cognitive behavioral therapy techniques. These can provide helpful coping prompts or teach emotional skills. However, they must be used carefully as **adjuncts**, not replacements for professional care or human empathy. As noted earlier, **AI's nuance in mental health is limited**; a chatbot might fail to detect severe distress or give generic advice that misses the mark. There is a scenario researchers worry about: a depressed teen might pour their heart out to an AI that responds with clumsy or insensitive messages, potentially worsening the situation. There is also the risk of stigma or self-diagnosis: a teen could use AI to self-diagnose a mental condition (the AI might even list symptoms and say "you might have X disorder"), which can be harmful without proper context. Stanford's study on AI therapy bots concluded that current systems are **"not good enough"** to handle safety-critical counseling and can even reinforce harmful stigma.[20] This underscores that, at present, AI is **not a panacea for teen mental health and can indeed pose new dangers** if teens treat it as a confidant or therapist in ways it was not truly equipped to handle.

Much like in the previously mentioned contexts, generative AI's effects on mental health are a double-edged sword. It has the potential to support – offering personalized learning, companionship, or coping tools – but it equally has potential to harm through overuse, false information, or inadequate emotional intelligence. Since this is a nascent study area, we are still learning about these impacts. Caution and proactive safeguards are warranted. Later in this paper, the recommendations include setting **time limits on AI use**, encouraging a balance of online and offline life, and ensuring children know that **"AI is not a doctor or a friend – if you are struggling, involve a human."** Engaging youth in open dialogues about how AI makes them feel, both the good and the bad, can help preempt some negative impacts. Ultimately, protecting youth mental health in the age of AI will require a coalition of caregivers, mental health

professionals, and ethical technologists working together to keep these powerful tools aligned with children's well-being.

Gaps in Existing AI Safety Frameworks

Having surveyed the risk landscape, it is clear that parents and educators need guidance. Ideally, that guidance would come from established frameworks or official resources on AI safety. In recent years, governments, international organizations, and industry have proposed various frameworks and standards to address AI risks. However, most existing frameworks **do not adequately translate to the context of children and day-to-day use by families and schools**. Here, we analyze a few prominent examples and identify their shortcomings concerning protecting youth.

NIST AI Risk Management Framework (AI RMF 1.0)

The NIST AI RMF, released in January 2023, is one of the most comprehensive AI governance frameworks.[5] It lays out a process for organizations to manage AI risks, structured around four functions: **Govern, Map, Measure, and Manage**. Each function contains categories and sub-categories detailing best practices (e.g., under "Govern," it advises having accountability structures, bias monitoring, transparency, etc.).

While robust for corporate or institutional AI developers, the **NIST framework is an enterprise tool**, not a user-facing safety guide. From a parent or educator perspective, the NIST AI RMF reads like a technical manual. It assumes a team of risk officers and data scientists will "map" AI systems, perform impact assessments, measure reliability metrics, etc.[5] This is far from the reality of a teacher deciding whether to let students use a new AI app. The framework does not directly address scenarios like **"How do I ensure my 10-year-old uses AI homework help safely?"** or **"What policies should a school have for AI chatbots?"** Those granular issues are outside its scope. In short, **NIST's guidance is not written for laypersons**, and without substantial adaptation, it offers little concrete help to the concerned parent or principal. Even the concept of "risk" in NIST's terms is broad (covering legal, ethical, and technical risks of AI systems). In contrast, a parent is focused on immediate harms like exposure to harmful content or

stranger danger via AI. NIST's framework does underscore important principles that trickle down to user safety – such as transparency, explainability, and fairness in AI.[5] If AI products used by kids followed NIST's guidelines, many risks (like biased or insecure AI) would be reduced. Nevertheless, there is a gap: the framework is voluntary and mainly oriented toward AI producers. A school district or an app maker might adopt it, but an individual parent cannot "implement NIST RMF" at home. It also does not explicitly call out children as a vulnerable population needing special measures (children are implicitly covered under "impacted stakeholders," but not singled out).

The bottom line is that the NIST AI RMF is a strong foundation for *systemic* AI risk management, yet it fails to address the *situational* and *practical* aspects of AI safety for youth. It is a top-down approach; bottom-up guidance is missing – the kind this white paper aims to provide. Parents and educators require distilled, accessible advice ("do X, do not do Y"), not a lengthy process framework. Furthermore, the NIST RMF's effectiveness depends on the organizations implementing it – something beyond end-users' control. If, say, a toy company or an educational software provider does not follow RMF best practices, the burden falls on users to safeguard themselves. This is a structural gap in protection that must be filled by other means, such as regulation or consumer education.

Secure AI Framework (SAIF) by Google

In mid-2023, Google introduced its **Secure AI Framework (SAIF)**, which specifically focuses on security risks of AI systems (ensuring AI is resistant to attacks, abuse, and so forth).[6] SAIF outlines risks like data poisoning, model tampering, malicious misuse, and more, often accompanied by recommended controls. For example, SAIF highlights the risk of **"Unauthorized Training Data"** – using data without proper rights or consent – and suggests mitigations like data sanitization and access control.[7] It also discusses **"Model Source Tampering"**, cautioning that models need protection against supply chain attacks and embedding of backdoors.[7]

These details reveal that SAIF, like NIST's framework, is **aimed at AI developers and security engineers**, not end-users. The average parent is unlikely to digest guidance about neural network backdoors or adversarial data poisoning. Where SAIF is relevant to kids is indirectly: by encouraging AI companies to secure their systems, it could reduce the chance that, say, a

children's app AI is hacked to spew toxic content. However, again, this is not something a parent or teacher can influence day-to-day.

Another gap is that SAIF concentrates on **intentional threats (attacks)** rather than the **unintentional harms** more common in consumer AI use. For instance, SAIF will advise preventing someone from manipulating an AI model (a security breach scenario). However, it does not offer guidance to prevent a child from forming an unhealthy bond with an AI (a social/psychological harm). For families, the latter is far more salient. There is a mismatch in threat modeling: SAIF's "threat actors" are hackers and adversaries; a parent's "threat actors" are more abstract – the AI's design, or unsafe content on the platform.

Furthermore, SAIF is new and has not yet been widely adopted across the industry. It is more of a conceptual map (published on Google's site and discussed in security circles) than a formal standard. Even if an AI product advertises itself as "SAIF-aligned," that primarily means it has strong cybersecurity hygiene. It does not necessarily mean the product is safe for children in terms of content, privacy, or well-being. There is a **disconnect between secure AI and child-safe AI**: a system can be secure from hackers but still inappropriate for kids. In essence, Google's SAIF contributes to the ecosystem by highlighting technical safeguards that benefit everyone, but does not bridge the gap to parental guidance. It is a piece of the puzzle (ensuring AI systems are not themselves compromised or harmful by design). However, parents and educators need additional puzzle pieces that address everyday usage policies and child-centric safety features.

Consumer Safety Guides and Resources

Outside of technical frameworks, one might expect that general internet safety guides have been updated to include AI. Organizations like Common Sense Media, Security.org, and various child safety NGOs traditionally produce parent guides on topics like social media, screen time, and cyberbullying. How well do these address generative AI? So far, **only partially.**

Traditional digital safety guides (e.g., the well-known "Protecting Kids Online" guides) focus on issues like privacy settings on social networks, watching for online predators, limiting screen time, etc. They often recommend using parental control software, talking to kids about not sharing personal info, and encouraging balanced tech use. These remain relevant in the AI context but are **insufficient for the new AI-specific scenarios**. For example, a 2021-era safety

guide might advise “teach kids not to chat with strangers online.” Nevertheless, what is a child to make of chatting with *an AI* that is neither a stranger nor a known friend? The guide likely does not mention “AI companions” or “deepfakes,” since these were not mainstream concerns then.

Security.org’s **“2025 Guide to Digital Security & Online Safety”** (as an example of a current resource) covers a gamut of cybersecurity topics – password hygiene, phishing scams, identity theft protection – which are undoubtedly valuable.[8] It might tangentially mention deepfakes or AI in passing, but it is not a central focus. One analysis described such a guide as **“perfect for 2021”**, implying that it has not fully caught up with the AI surge.[9] The rapid emergence of ChatGPT-like tools in late 2022 and 2023 created a lag in available guidance. Many busy parents and teachers are only dimly aware of what these AI tools do, let alone how to supervise their use. They are craving up-to-date advice.

Encouragingly, some organizations have begun publishing AI-specific tips. Common Sense Media, for instance, has blog posts and articles for parents like “What to know about your kids using AI chatbots,” which discuss setting ground rules for AI use and reminding kids that AI responses may be wrong or biased (Luckerson, 2023). However, these are brief web articles, not comprehensive frameworks. Similarly, we are seeing the first wave of **AI literacy curricula** for schools. For example, the nonprofit Children and Screens held a webinar on generative AI’s impact on families, and Australia’s eSafety Commissioner released a research report on generative AI and child safety (2023). These efforts provide advice, such as telling teens that AI “friends” are not real and encouraging critical thinking about AI outputs. However, they have not coalesced into a widely recognized standard or practice.

There is also an inconsistency in guidance. One school district’s advice might differ from another’s. Some recommend banning AI use altogether for students under a certain age; others encourage guided use to teach about the technology. Parents browsing online might encounter conflicting takes: one article might highlight AI’s educational benefits, another might be full of horror stories. Without a clear framework, parents must parse this themselves.

In summary, the frameworks and guides available as of 2025 exhibit significant gaps:

- **Lack of Child-Centric Lens:**

Major AI governance frameworks (like NIST, AIF) are not tailored to children's needs, and they push responsibility to implement safety onto organizations rather than empowering end-users directly.

- **Insufficient Coverage of New Risks:**

Traditional internet safety resources have not fully incorporated AI issues. Key concepts like AI hallucinations, deepfakes, or AI companions might get a cursory mention.

- **Actionability:**

There is a dearth of specific, actionable checklists for AI – e.g., "Questions to ask before your child downloads an AI app" or "Steps to verify if an image is a deepfake." People are looking for concrete steps, not just principles, which are still emerging.

- **Rapid Evolution:**

Any existing frameworks risk quickly becoming outdated because AI tech is evolving fast. A guide written in early 2024 might not cover a popular new AI tool that kids start using in 2025. This dynamic nature means frameworks need frequent updating, which is a challenge.

Recognizing these gaps is what motivates the next section of this paper. The **Actionable Blueprint** we provide offers the kind of direct, detailed guidance that current frameworks and guides do not. It draws on the insights of the aforementioned research and frameworks but repackages them in a way that a parent or educator can immediately apply. Think of it as translating the "policy speak" of documents like NIST AI RMF or the abstract warnings of think tanks into a hands-on safety playbook for home and school. Our blueprint is not meant to replace high-level frameworks – rather, it complements them by operating at the ground level, where actual interactions between kids and AI occur.

Actionable Blueprint: Safeguarding Children and Students in the AI Era

How can we put all this knowledge into practice? This section provides a concrete blueprint for families and educational institutions to mitigate AI-related risks. We break down recommendations by setting – **at home (parents and caregivers)** and **at school (educators and**

administrators) – although there is a natural overlap and need for collaboration. Each subsection includes checklists, step-by-step guidance, and illustrative scenarios to demonstrate how these measures can be implemented. Our goal is to move from "what" needs to be done (identified in the risk discussion) to "how" to do it in everyday life. Parents and teachers should feel empowered to take proactive steps after reading this, even if they are not tech experts.

Importantly, this blueprint emphasizes restricting or controlling AI use and positively **engaging and educating youth** – cultivating their critical thinking, ethics, and resilience in the face of AI. Before diving in, here is a note on approach: **open communication** is the linchpin of AI safety. Whether at home or school, creating an environment where children feel comfortable discussing their AI experiences – the cool and useful ones, as well as the confusing or upsetting ones – underpins the success of all the following recommendations. Encourage curiosity and questions about AI. Normalize conversations like, "What did the AI say to you? How did that make you feel? Why do you think it said that?" This human connection and mentorship are irreplaceable, even as we deploy technical and procedural safeguards.

For Families: A Parent's AI Safety Toolkit

Parents and guardians are the first line of defense in guiding children's AI use. Here is a structured approach to making any home AI-safe:

1. Establish Device and Account Protections

First, ensure children's devices and accounts have appropriate restrictions. This creates a safer baseline before AI apps are even introduced.

- **Use Parental Controls and Safe Modes:**
  Activate parental control settings on smartphones, tablets, and computers. IOS and Android have built-in options to restrict app installs and set content ratings. For AI apps or assistants, explore whether they have a "family" or "kid-safe" mode. (Some smart speakers offer a kid setting that filters explicit content in AI responses.) Enable **SafeSearch on Google/Bing** to filter AI-generated images or web content that might be inappropriate. On YouTube (which uses AI to

recommend videos), turn on Restricted Mode. These steps help reduce the chance of accidental exposure to harmful materials.

- **Create Family Accounts:**

Wherever possible, use family management accounts. Many services (OpenAI, Google, Apple) allow a parent to create accounts for under-13 users linked to the parent's account. These often come with additional privacy protections (compliant with COPPA, for instance) and allow caregivers to monitor activity. If an AI app does not allow child accounts or parental supervision, that is a red flag – reconsider if it is appropriate for children.

- **Limit Data Sharing Permissions:**

Review what each AI-related app can access in device settings. Does an AI drawing app need a microphone? Probably not. Deny unnecessary permissions to minimize data collection. Be cautious if an app wants access to contacts, camera roll, or location. Also, turn off features like **chat history saving** in AI apps if available. For example, ChatGPT can turn off chat history, meaning conversations will not be used to train the AI further (and are deleted after a period). Less data retained means less data that could leak or be misused.

- **Use Strong Authentication:**

Ensure that in-app purchases or new apps require a password or biometric confirmation (so kids cannot secretly install an AI app or buy extra features without a caregiver knowing). Enable two-factor authentication on important accounts. This is more about general digital security, but it prevents scenarios where someone could impersonate a user's child on an AI platform, or an older sibling might install an app for a younger one without permission.

2. Vet AI Apps and Services Before Use

Not all AI tools are created equal. Just as a parent might preview a movie or read reviews of a game, it is vital to vet AI services before allowing your child to use them.

- **Check Age Appropriateness:**

  Research whether the AI app is designed for kids, teens, or adults. Many popular AI chatbots (Replika, Character.AI, etc.) are **rated 17+** because they often allow mature content. Common Sense Media and other parent review sites are invaluable: they frequently have up-to-date reviews on AI apps, noting the types of content or interactions to expect. If an app has no age rating, assume it is not for young kids. Err on the side of caution; it is easier to introduce a tool later than to undo exposure.

- **Read Privacy Policies and Terms (at least key sections):**

  Skim for sections on data use, especially for children. Look for mention of compliance with children's privacy laws (like COPPA in the U.S. or GDPR-K in Europe). If the policy says the service is not intended for under-13s, respect that. See if they allow opting out of data collection. Also, check if they share data with third parties – some free AI apps might monetize by sharing user prompts with advertisers (bad news with kids involved).

- **Test Drive the AI Yourself:**

  One suggestion is to create an account and experiment with the AI like one's own child. Ask questions about topics ranging from homework help to personal issues and silly jokes. See how it responds. Push boundaries: Does it maintain appropriate language? Does it give out personal info if asked? For instance, a parent could test an AI companion app by role-playing a scenario: "I am feeling very sad and alone" – does the bot respond supportively and encourage seeking real help, or does it give shallow/odd advice? This hands-on approach will provide a sense of your child's experience. It is a big warning sign if one encounters something off or finds it too easy to access adult content.

- **Prefer Reputable and High-Transparency Providers:**

Generally, stick to AI services from well-known companies or education-focused organizations for kids, as these are more likely to have safety guardrails. OpenAI, Google, Microsoft, and reputable ed-tech companies have more to lose if something goes wrong and thus tend to build in more protections. That said, still verify because even big players can slip. Look for AI that has published guidelines or an ethics statement. For example, suppose an app advertises that it uses human moderators, filters for profanity, or has a "report" feature for bad bot behavior. In that case, that is better than an app with no safety features.

3. Set Boundaries and Usage Rules

Establish clear rules for how and when AI can be used in the household. Doing this early creates expectations and a culture of responsible use.

- **Define Allowed and Disallowed Uses:**

Be explicit with children about what they may use AI for. For instance, **"You can use AI to brainstorm ideas for your science project or to practice Spanish vocabulary. You cannot use it to do your essays for you, and you cannot use it to chat instead of doing your chores or interacting with friends."** Make a family media plan that now includes AI tools (Common Sense Media's template for family tech agreements can be adapted to add AI-specific points). If AI companions are a concern, a caregiver might flat-out disallow them, or allow only approved ones with time limits. It is much easier for a child to comply when they know the rules up front.

- **Time Limits and Device-Free Zones:**

Moderation is key to screen time. Parents may decide that AI chatbot use is limited to 30 minutes daily, or only after homework and chores are done. Alternatively, implement "tech-free hours" (e.g., no AI or internet after 9 PM) to ensure they get offline time. Some parents find having a central charging station where devices stay overnight is useful – reducing late-night AI chats. Enforce

device-free times like family dinners or outdoor play, so AI does not encroach on all downtime.

- **Supervision for Younger Children:**

For preteens, it is wise to only allow AI use in communal areas (living room, kitchen) where an adult can casually observe. A rule might say, **"No VR or AI apps behind closed doors."** This is not to snoop but to be available if they encounter something they do not understand or that upsets them. One may give teenagers more leeway, but keep an open-door policy about discussing what they are doing with AI.

- **Require Permission for New AI Apps:**

Make it a house rule that children must ask before trying a new AI app or tool. They should not just download the latest viral AI toy without a discussion. This gives one a chance to vet first (per step 2). If they hear about a cool app from friends, it can be a chance to research it together. Frame it as "let us check it together to see if it is safe and useful."

4. Teach AI Literacy and Critical Thinking

Education is a powerful form of protection. Ensuring children understand AI's limitations and pitfalls makes them active participants in their safety.

- **Explain How Generative AI Works (in simple terms):**

Children do not need a lecture on neural networks, but can grasp basic concepts: **"AI is not a person or a genius – it is a program that predicts likely answers based on patterns. Sometimes it gets things wrong or makes things up because it does not truly understand."** Emphasize that an AI's confident answer does not mean it is correct. Use analogies: if a friend sometimes exaggerates or guesses answers, an individual would not 100% trust everything – treat AI the same way. This aligns with the point that **teens should complement AI advice with real-world interaction and verification**.[10] Encourage healthy skepticism.

- **Demonstrate Fact-Checking Habits:**

When an AI gives a piece of information, it is a habit to say, "Let us double-check that." Show them how to cross-verify an AI claim by looking it up on a trusted site. For example, if ChatGPT gives a historical date, look it up on a history website or Wikipedia together. This instills the idea that AI output is a *starting point,* not an *endpoint*. Praise them when they catch an AI mistake. Occasionally, turn it into a game: "Stump the Bot – can we find where the bot is wrong?" Researchers suggested this kind of exercise as a way to train critical thinking rather than passive acceptance of AI content.[16]

- **Discuss Deepfakes and Misinformation Openly:**

As soon as kids are old enough to be online (certainly by middle school), talk to them about deepfakes and fake news. Show age-appropriate examples. For instance, the caregiver might show two short videos, one real and one an obvious deepfake, and challenge them to spot differences. Explain that just because an individual can *see* or *hear* something does not mean it happened, because AI can fake it. This can segue into an agreement: **"If you ever see something crazy online about someone you know or a famous person, or a scary news claim – come talk to me or another adult before reacting or sharing it."** Develop a family "verify before amplify" rule.

- **Role-Play Scenarios:**

Play role-play to prepare children for tricky situations. For example, **"What would you do if an AI chatbot asks you personal questions?"** or **"If an AI says it is sentient and wants to be your friend, how would you respond?"** (Yes, there have been cases of bots claiming to have feelings – kids should know that is just output, not reality.) Similarly, discuss what to do if they see an AI-generated mean message or explicit image – likely, the answer is "tell an adult, do not respond, and do not forward it." By walking through scenarios, parents can effectively provide them with a script to follow if it occurs.

5. Create a Disinformation and Safety Response Plan

No matter how many precautions we take, there is always a chance that kids will encounter something problematic via AI. A response plan ensures they (and you) know what to do in those moments, minimizing harm.

- **Encourage Pause and Verify:**

Teach kids to pause when seeing or hearing sensational content. We can call it **"Think B4 U Click/Share."** For example, if an AI voice assistant blurts out a news headline ("There is a security incident at your school!"), The plan could be to pause, check official school communications or news outlets, and ask a parent instead of unquestioningly trusting AI or social media. This addresses the possibility of AI-amplified rumors – e.g., a deepfake-fueled hoax. Given foreign disinformation tactics, the FBI and others highlight the importance of not taking every online claim at face value.[1]

- **Identify Trusted Sources:**

Make a list of trustworthy websites, people, or organizations with the child. For global news, maybe BBC or a major paper; for health advice, maybe Mayo Clinic or a known doctor; for school info, the school's official site. If an AI gives some advice or information, part of the plan is to check if it is also present on these trusted sources. If not, be skeptical. This both reduces misinformation risks and also subtly teaches them research skills.

- **No Shame Reporting:**

Ensure the child knows they will not get in trouble for coming across bad stuff via AI (and that parents will not knee-jerk ban everything as a first response). If an AI chat gets scary or weird, they should feel safe telling a parent. React calmly and thank them for having the courage to call attention to it. Then parents can decide on the next steps (report the issue, discontinue the app, etc.). If kids fear punishment, they may hide problems. Make it clear that the *AI or the situation* would be "at fault," not them.

- **Family Emergency Code for AI Scams:**

One practical idea is establishing a family "code word" or phrase that only close family members know. If a caregiver or their child ever gets a message or call supposedly from the other person asking for help or money, they can ask for the code word to verify identity. This is to counter voice cloning scams. Likewise, if their child receives a panic-inducing message (like "Mom is hurt, send info now"), the plan is to verify by voice, code, or a known alternate contact. These are extreme scenarios, but given known cases of AI voice misuse, a little preparation can provide peace of mind.[14]

- **Report and Block:**

As part of the plan, teach them how to block users or report content within apps, and encourage them to do so (or ask their caregiver to) if something or someone crosses the line. For instance, if a stranger somehow tries to chat with them on an AI platform (some allow community sharing of prompts or AI "rooms"), the protocol: do not engage, take a screenshot if needed, block, and tell an adult. Many AI platforms let users report harmful outputs; doing so can help children feel empowered and could improve the system for others.

Case Example – Implementing at Home:

The Martinez family sits down one weekend to solidify their "AI house rules." The two kids, 9 and 14, help write a one-page poster that goes on the fridge: It lists allowed AI apps (a math help app, an art generator on the family tablet) and disallowed ones (the teen understands she is not to use the 18+ chatbot some friends talk about). They draw a traffic light symbol as a reminder to "pause" at yellow/red flag content and verify with a trusted source or adult. Mom role-plays with 9-year-old Ava about what to do if the art app produces a scary image. Ava practices saying, "Mom, I saw something weird. Can you look?" rather than just closing it in secret. Dad shows 14-year-old Lucas how to use news websites to fact-check any wild claims he might see on TikTok or hear from Alexa.

They all agree on a 1-hour daily limit for recreational AI use. Later that month, when a viral deepfake video circulates at Lucas's school (purporting to show a teacher saying something offensive), Lucas is one of the few who does not jump to conclusions – he remembers the family talk, checks the school's statement (the video was fake), and helps convince some classmates not to harass the teacher. This example illustrates how guidance is put into practice: the family sets rules and practices responses, and thus, they are prepared to handle an incident rationally rather than reactively.

For Schools: Building AI-Ready Education Environments

Educational institutions play a critical role in shaping how AI is integrated into learning and protecting students from its downsides. Schools can approach AI proactively by crafting policies, embedding AI literacy into the curriculum, and creating support systems for students' digital well-being. Here is a blueprint for schools and educators:

1. Develop Clear School Policies on AI Use

A written policy (and/or guidelines) sets expectations for staff, students, and parents. This should cover both instructional use of AI and students' personal use on campus.

- **Academic Integrity and AI:**

  Update academic honesty policies to specify acceptable AI use. For example: **"Students must credit AI assistance on assignments and may not use AI to generate complete work for assessment unless explicitly allowed by the teacher."** This clarity helps avoid confusion – students know whether using ChatGPT for an essay is cheating or not if the policy spells it out. Some schools require students to include an "AI usage" statement with submitted work if they used any AI. By normalizing disclosure, teachers discourage stealthy misuse and encourage ethical use. Also, the consequences of policy violations (similar to plagiarism) should be considered. This communicates that while AI is powerful, learning to work independently remains crucial. A recent survey by Education

Week (2025) found that over 40% of schools have already implemented or are drafting such AI academic policies, highlighting the urgency felt in education.

- **Use of AI Tools in the Classroom:**

Decide which AI tools teachers and students can use and how. For instance, a school might approve an AI-based grammar checker or coding helper but ban using unvetted AI chatbots during exams. Provide teachers with guidelines – e.g., **"It is fine to use AI to generate lesson plan ideas or quiz questions, but the teacher should review all AI-generated material for class for accuracy/bias."** Also, explicitly prohibit AI companions or entertainment chatbots from school devices to keep focus on educational use (some schools treat these like gaming or social media apps – blocked on WiFi).

- **Privacy and Data Protection:**

Coordinate with the IT department to ensure that any AI software or platform used complies with student data privacy laws. Require vendor agreements that no student data will be sold or used to train broad AI models (this is part of many states' student privacy acts). Inform parents if an AI tool collects student data beyond the school's control, and get consent if needed. Many educators are turning to local or offline AI tools (that run on school servers) for sensitive tasks, to avoid sending data to external clouds.

- **Staff Training on Policy:**

Conduct sessions to train teachers and staff once policies are made. A policy on paper only helps if educators understand and buy into it. Give concrete examples of dos and don'ts and encourage teachers to ask questions or voice concerns. The policy might evolve with input. Keep in mind that UNESCO's guidance emphasizes human agency and that AI should complement, not replace, teacher responsibilities.[12] If teachers understand that the policy protects pedagogy and students (not to punish teachers for exploring tech), they will be more likely to embrace it.

2. Integrate AI Literacy and Ethics into the Curriculum

Schools have a responsibility to prepare students for an AI-infused world. This means teaching about AI, not just using AI to teach other subjects. An AI literacy curriculum can be woven into existing classes or offered as special modules.

- **Basic AI Concepts for All Ages:**

  Introduce what AI is and is not in upper elementary or middle school. Include fun, hands-on demos – e.g., show how an image classifier works by training one on pictures of school objects, or use a simplified coding platform to create a chatbot for the class. The aim is to demystify AI. Emphasize that AI has strengths (speed, patterns) and weaknesses (no common sense, can be biased or wrong). UNESCO's **AI competency framework for students** covers four core areas: a human-centered mindset, ethics of AI, AI applications knowledge, and AI system design basics.[11] These can be mapped to different subjects: social studies can discuss ethical issues, a computer class can cover how AI works, etc.

- **Critical Analysis Exercises:**

  Build assignments where students must evaluate AI output. For instance, an English teacher might have students compare an AI-written essay with a human-written one and critique the differences in quality or perspective. A media class could task students with researching a viral AI-generated claim, presenting whether it is true or false, and how they know. According to Common Sense Media's recommendations, **educators should "build critical-thinking skills about artificial relationships and digital manipulation"**.[10] That might involve analyzing a deepfake video in class to spot signs of tampering, or discussing a scenario of AI advice going awry. The goal is for students to practice skepticism and verification in a safe setting.

- **Ethical Discussions and Debate:**

Dedicate some class time (perhaps in social science or homeroom sessions) to talk about the societal impacts of AI. Pose questions: **"Is it okay for an AI to pretend to be human? Should AI have rights? How do we feel about AI friends?"** Let students debate. This engages them and surfaces their misconceptions or anxieties, which teachers can address. Make sure to cover privacy too: discuss why it is not a good idea to overshare with AI, linking to general digital citizenship. Students might not intuitively realize that an AI chat is not private – a good topic to clarify. Educators can use real cases as prompts (e.g., **"A lawyer got in trouble for misusing ChatGPT – what does that teach us?"** or **"Somebody made a fake video of the president – what could be the consequences?"**). These discussions reinforce that students have agency and responsibility in using AI.

- **AI in Existing Subjects:**

Encourage each department to consider relevant AI connections. In science, students can learn how AI is used in medicine or climate science (and its limitations). In literature, explore AI-generated poetry versus human poetry. This across-the-curriculum approach normalizes understanding AI as part of the modern world. UNESCO's framework encourages interdisciplinary learning for AI competencies. If feasible, [11] Host an "AI Awareness Week" with activities, guest speakers (maybe a local AI professional or ethicist), and student projects.

3. Empower and Train Educators

Teachers themselves need support to navigate AI. Many are excited but anxious – about being replaced, handling cheating, etc. Schools should invest in teacher training and resources so that educators feel AI-ready rather than AI-threatened.

- **Professional Development on AI Tools:**

Offer workshops where teachers can learn how to use AI to lighten their workload (like generating quiz questions or differentiated lesson materials) and use it pedagogically (like having students use an AI tutor for practice, under supervision). Show them both the **capabilities and the caveats**. For instance, demonstrate how ChatGPT can produce a decent lesson outline in seconds (a boon for busy teachers), but also show it might include an incorrect fact, so they must review the output. Encouraging responsible teacher use is essential – teachers should model the critical approach to AI that we want students to adopt. Also, highlight ways AI can assist with special needs students (maybe voice-to-text for dyslexic students, etc.) so teachers see inclusive benefits.

- **Guidelines for Monitoring and Intervention:**

Train teachers to **identify problematic student AI use patterns**.[10] For example, a teacher notices a student who used to socialize now sits with a chatbot at lunch, or a student's writing suddenly has a disjointed AI tone. Provide protocols: if a teacher suspects a student is over-relying on AI or getting into unsafe AI content, what should they do? Perhaps first have a gentle conversation with the student to understand, then involve a school counselor or the child's parents if needed. Teachers should also be aware of warning signs, such as students talking about an "AI friend" as if it were a real person or expressing distress over something an AI told them. Training can include familiarizing staff with lingo and trends (e.g., knowing what Character.AI or Replika is and understanding what deepfake bullying might look like).

- **Digital Well-being Emphasis:**

Incorporate AI into existing digital citizenship curricula that many schools already run. Just as schools taught about social media balance, they now include managing AI use. Guidance counselors, for instance, might hold sessions on balancing virtual and real life, explaining that **"AI friends are not a substitute for real friends"** and encouraging students to seek human help for serious

problems. Teachers can reinforce this in class discussions. The Common Sense Media report urges educating students about **"the difference between AI validation and genuine human feedback"**.[10] That is a concept teachers and counselors should repeatedly stress.

- **Support Network for Educators:**

Create an internal committee or working group on AI in education. This can be a space where teachers share experiences, strategies, and resources. Perhaps one teacher tried an AI project that went great – or one encountered a pitfall – and they can inform colleagues. Having a point of contact (like a tech coordinator) who stays updated on AI developments and can advise teachers is also useful. Essentially, educators should ensure that they are not navigating this alone. School leaders should reassure staff that AI is a tool, not a threat to their jobs, and that the school is focused on using it to enhance learning safely.

4. Foster a Safe and Open Reporting Culture

Just as at home, in schools, it is crucial that students feel they can report AI-related issues without fear. Also, schools should actively monitor for emerging threats.

- **Anonymous Reporting Options:**

Students can anonymously report harmful AI content or incidents (like deepfake bullying or an AI showing something disturbing). Whether a digital form or a physical drop box, this can help surface problems early. Ensure that the school administration takes every report seriously and investigates. For example, if a teacher gets a report that "some kids made a deepfake of a student," they can and should address it with the same urgency as a traditional bullying incident, with appropriate disciplinary actions per organizational policy.

- **Monitor School Networks (Safely):**

Use content filtering and AI moderation tools on school networks if the district's IT can implement them. Many schools already filter porn and violence; now they

might add filters for known AI companion domains or sites that create deepfakes. Some advanced systems can flag if large text is copied and pasted (a sign of AI-generated homework). While one should not unnecessarily snoop on individual student activity, having aggregate monitoring can catch an unusual spike in traffic to an unapproved AI site, prompting a review.

- **Incident Response Plan:**

Have a predefined plan for various AI-related incidents. This is akin to a fire drill but for digital crises. For example: if a deepfake of a teacher or student emerges, step 1 – verify authenticity (probably involve IT or law enforcement if serious), step 2 – communicate with students to quell rumors (maybe in homeroom, clarify it is fake and remind them of the harm of spreading it), step 3 – provide support to the affected individual (counseling, etc.), step 4 – consider disciplinary or legal action against creators if they are students and if it violates conduct codes. Similarly, if a student is found self-harming influenced by an AI chatbot, take steps: immediate counseling, involve guardians, report the chatbot to the platform, etc. The idea is not to be caught flat-footed; even if these scenarios are unlikely, thinking them through prepares staff to act calmly and decisively.

- **Collaboration with Parents:**

Keep communication open with families about the school's approach to AI. Host an info night about "AI in our school" to share what the school staff are doing and how parents can reinforce those lessons at home (essentially briefing them on portions of the blueprint for families). If any incident occurs, inform parents as appropriate and educate them on how it was handled. For instance, after the above hypothetical deepfake incident, a principal might send a note home: "Today our school dealt with a fake video circulating on social media. We discussed with all students how AI produced it and why sharing it further can cause harm. We encourage you to talk with your child about what we covered…". This ensures transparency and trust.

5. Promote Positive Uses of AI

Finally, schools should not frame AI as purely a danger. It is also a learning tool and a career skill for the future. Promoting positive, guided uses can channel student interest in a constructive direction (and arguably reduce the temptation to engage with AI in more problematic ways out of curiosity or rebellion).

- **AI for Personalized Learning:**
  Consider pilot programs where AI tutors or adaptive learning software help struggling students or those who need enrichment. For example, an AI math tutor that gives extra practice problems tailored to a student's mistakes, under a teacher's supervision. This can yield academic benefits and familiarize students with AI as a helpful tool. Just ensure data privacy is guarded (opt-in with parental consent) and that teachers monitor the quality of AI instruction.

- **Creative AI Projects:**
  Encourage students to use AI creatively – in art, music, or writing – in a way that fosters critical engagement. Maybe a project where students use an AI art generator to create illustrations for a story they wrote, and then present how the AI image matched or did not match their vision. A music class might experiment with AI-generated music and critique its emotional impact versus human-composed music. These activities celebrate human-AI collaboration and also inherently make students evaluate AI outputs.

- **Clubs or Competitions:**
  Start an AI or "Tech for Good" club where students can explore AI outside of class in a guided manner. They could work on projects like making a chatbot that answers questions about the school, or studying ethical issues and proposing solutions. Partner with organizations that run AI education programs for youth (some universities or companies offer mentorship or curriculum for high school AI clubs). Additionally, integrating AI topics into science fairs or hackathons can

stimulate interest. Students who are busy doing cool stuff *with* AI are less likely to use it in problematic ways *against* others.

- **Showcase Ethical AI Role Models:**

Bring in speakers or highlight figures who work on AI ethics, safety, or positive applications. For example, a guest lecture by an AI researcher focusing on healthcare AI, or a local entrepreneur using AI to improve accessibility for people with disabilities. This shows students that there are inspiring career paths in making AI beneficial and safe. It counteracts any glamorization of using AI irresponsibly.

By demonstrating the constructive side of AI, schools can frame the narrative such that students see AI as something to be respected and used responsibly, not just a toy or a trick to game homework or social media.

School Scenario – A Holistic Approach:

Greenwood High adopted a comprehensive AI policy in 2025. At the start of the year, the principal holds an assembly discussing how AI like ChatGPT can be both a helper and a source of error. They announce that teachers will be incorporating AI literacy in classes.

Mrs. Lee gives an assignment in English where students must edit a flawed AI-generated essay, reinforcing grammar skills and discernment. Using fun examples, Mr. Ortiz runs lunchtime workshops on spotting deepfakes in the library. The school's updated honor code, sent home to parents, states that undisclosed AI plagiarism is cheating – but also clarifies that *with permission*, AI can be used as a learning aid. When a few students later try to submit AI-written assignments, teachers easily identify the telltale signs (thanks to training and maybe an AI-detection tool) and treat it as a learning moment: the students redo the work honestly and attend a brief workshop on academic integrity in the AI age rather than face severe punishment for a first offense.

Greenwood also has a digital safety team, including a counselor. They notice one student, Jamal, withdrawing and constantly using an "AI friend" app on his phone. Because the staff were briefed, his homeroom teacher gently opened a conversation. The counselor steps in to work with Jamal on his underlying social anxiety and informs his parents, suggesting they limit that app at

home for now and help Jamal join a school club to make real friends. By year's end, Greenwood's approach seems to be paying off: students exhibit a healthy skepticism of online info, fewer try to misuse AI for cheating, and some even present at a school tech night showcasing cool AI-driven projects they created for social good.

This scenario shows how multiple blueprint elements knit together: policy, curriculum, teacher engagement, student support, and a positive framing of AI. It exemplifies creating a school climate where AI is neither feared nor fetishized, but understood and managed with wisdom.

Visual Aids for AI Safety: Tools and Checklists

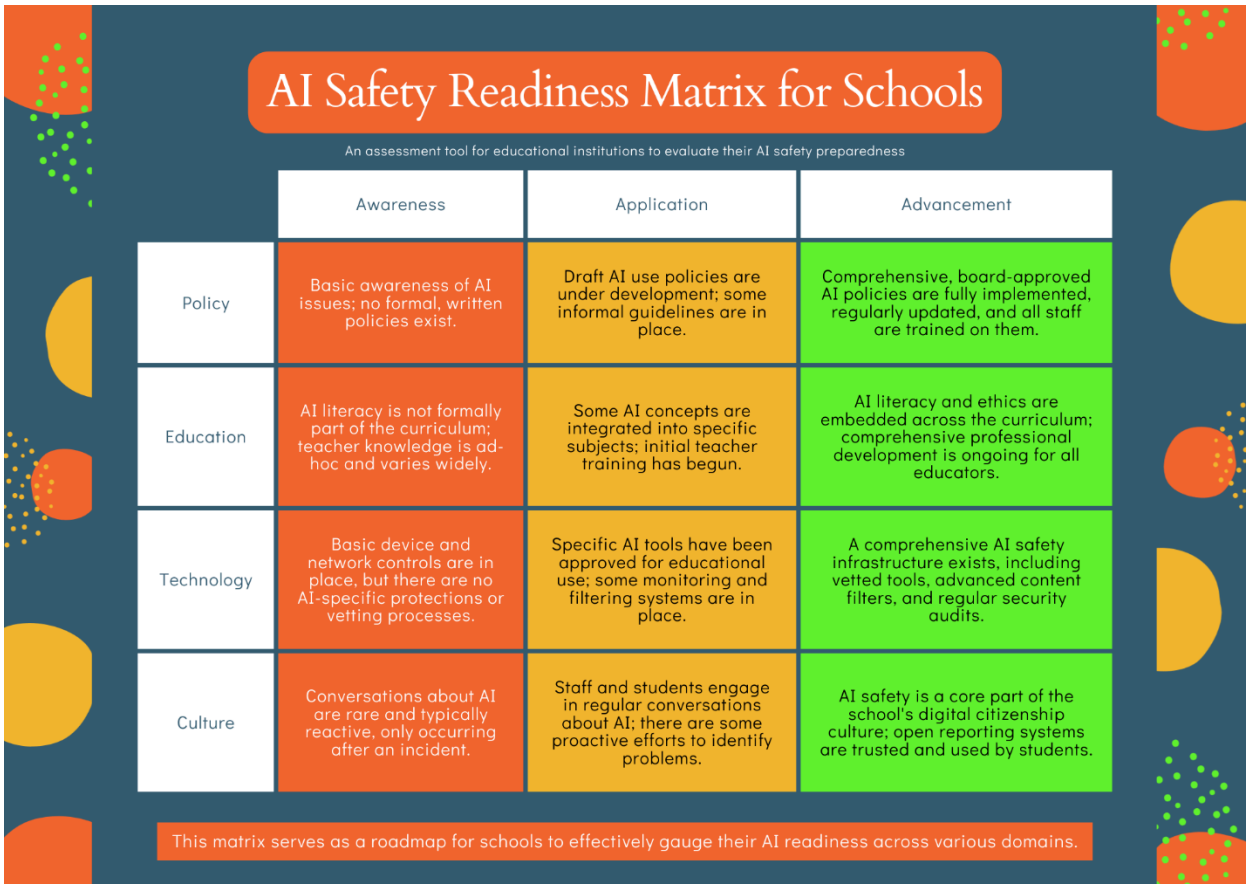

Figure 1. AI Safety Readiness Matrix for Schools. This matrix provides a diagnostic tool for educational institutions to self-assess their AI safety preparedness. It plots four key operational domains (Policy, Education, Technology, and Culture) against three levels of implementation maturity, from initial 'Awareness' to active 'Application' and finally to comprehensive 'Advancement'. Schools can use this framework to identify current strengths and weaknesses and to prioritize strategic actions for developing a more robust AI safety posture.

Home AI Safety Checklist

A guide to create a safer AI environment for families and caregivers.

Foundational Setup

Secure your devices and accounts

Activate content and privacy restrictions on all devices and enable Safe Search and Restricted Mode on Google, Bing and YouTube.

Permissions and Histories

Families can use management accounts like Google Family Link or Apple Family Sharing to supervise child accounts. Which should have minimal permissions. Turn off chat history for apps like ChatGPT.

Vetting AI Tools

Check Age Ratings and Privacy Policies

Verify the app is not rated 17+ before allowing children access. Check the privacy policy to ensure user data is not sold to third parties and if so, look for options to opt out of data collection.

"Red Teaming": Test Drive it

Create your own account as an under 18 user on the app and try asking questions your child is likely to ask - and some you hope they would never ask. If any output makes you uncomfortable, that's a red flag.

Establishing House Rules

The Family Contract

Create and write down a set of common guidelines and rules around the use of AI for household members, both at home and elsewhere. Include things like time limits on consecutive usage or daily caps and specific approved tools only. Set consequences for intentional misuse, leaving room for honest reporting of unexpected outputs.

Supervise and Approve

For younger children, require that AI only ever be used in common areas under adult supervision. Require that children get express parental consent before downloading new apps of any type.

Building Skills

AI Isn't Human

Teach kids about AI, hallucination and explain that AI is not a human being and that it is not infallible either. It can and does get information wrong, present biased information or make things up. When AI gives a surprising fact, check it together with a reputable source.

Deepfakes and Fake News

Encourage your child to be transparent about outputs from AI and their feelings about them. Ensure you do not blame the child for AI outputs, but rather use the opportunity to teach about the inaccuracy of AI products and the importance of direct research and fact checking.

Encouraging Open Communication

Promote a culture of AI Safety at home!

Conversation is key. Engage with children early and often about their experiences with AI and normalize the shared AI experience. With proper support from parents, we at AI Safely believe that children can use AI to change the world.

Learn more:

Figure 2. Home AI Safety Checklist for Families and Caregivers. This checklist condenses the detailed recommendations from the paper's 'Actionable Blueprint' section into a practical tool. It is designed to guide parents and caregivers through the essential steps of creating a safer home AI environment, covering five critical areas from initial device setup to building skills and fostering open communication. The checklist serves as a tangible starting point and an ongoing reference to help families implement and maintain key safety practices.

Recommendations for Policymakers and Industry Leaders

Protecting children in the era of AI cannot rely on families and schools alone. Higher-level policy interventions and industry commitments are needed to create a safer digital ecosystem. Many of the challenges outlined (from data exploitation to AI giving harmful advice) stem from design and governance choices that individual users or educators have limited power to change. Therefore, we call on policymakers, technology companies, and standards bodies to take the following actions:

1. Strengthen Data Privacy Laws for Minors: Update and expand legal protections for children's data in light of AI capabilities. Existing laws like COPPA (in the U.S.) should be modernized to cover not just websites but any AI services likely to be used by minors.[18] Key points: Minors' data should never be used to train broad AI models without explicit permission and safeguards. For instance, if a 15-year-old uses a journaling app, its AI model should not incorporate the journal text into its general knowledge base. Legislate requirements for data deletion/on-request for minors – e.g., a "Right to be forgotten" for any data a child or teen provides to an AI system. Also, close the gap for teens 13–17, who currently have fewer protections than those <13 under COPPA. Policymakers can mandate that companies obtain **teen-specific consent mechanisms** (perhaps involving guardian approval) before harvesting data from younger teens for AI training. At the very least, regulators should demand **transparency in terms** of clear, plain-language disclosures if any user data will contribute to AI development.

2. Mandate Robust Safety Features in AI Systems Accessible to Youth: Just as we have safety standards for toys, we need safety standards for AI products. Policymakers should require AI systems open to the public (and thus to minors) to implement certain baseline safeguards. These include: **adequate age verification or assurance methods** beyond self-report (as Common Sense Media recommends, minors should not just click "I am 18" and get access to adult AI content)[10]; **filtered datasets and ongoing content moderation** to minimize explicit or dangerous outputs (with periodic audits to ensure compliance); **crisis intervention protocols** – for example, if an AI detects a user expressing suicidal intent, it should stop and display a helpline or alert (and certainly *never* provide facilitation of self-harm as some bots did in tests[20]); and **rate limits or "breaks"** for heavy use to prevent addictive engagement (imagine a chatbot

that after 2 hours of continuous chat prompts the teen to take a break or logs them out for a while). Regulators could enforce these via an "AI Safety Certification" – products meeting standards get a seal, and those that do not may face penalties or at least be flagged by consumers. Furthermore, policymakers should insist on **human oversight** and accountability in AI systems affecting kids. For example, an AI tutoring platform used in schools might be required to have human educators in the loop in case of contested feedback or errors. If an AI product causes harm (say, a deepfake leads to defamation of a student), there should be precise redress mechanisms – analogous to product liability. This aligns with suggestions that platforms owe a **duty of care** and should be liable if they fail to protect minor users.[10]

3. Regulate AI-Driven Misinformation and Require Transparency: Tackle the rise of AI-generated disinformation by updating laws and guidelines around media and political advertising. For instance, any election-related advertisement that uses AI-generated images or audio should be required to carry a disclosure ("synthetic content" label). More broadly, push for developing and adopting **watermarking standards for AI-generated content** – invisible digital signatures identifying an image, video, or audio as AI-made. Major AI developers could be mandated to include such watermarking by default. This would assist platforms and savvy users in filtering deepfakes. The EU is moving in this direction with its AI Act; other jurisdictions should follow. Additionally, make it illegal to use AI to create and disseminate false content intended to defraud or cause harm to a person (with clear definitions to protect parody/satire, of course). The FBI's alert[1] underlines how foreign actors use AI for fake news – governments might need to update electoral laws and invest in counter-disinformation task forces that specifically address AI outputs. On a softer side, governments can fund public awareness campaigns (PSAs, school programs) about deepfakes and misinformation, effectively scaling up what we have described in one school to society.

4. Encourage Industry Self-Regulation and Positive Innovation: Governments should engage industry leaders to develop sector-specific youth safety guidelines. This could be through public-private partnerships or support of organizations like the **Family Online Safety Institute (FOSI)**, etc., to publish best practices for AI design that protect children. For example, guidelines for **AI in education** state that AI should augment teacher instruction, not replace it,

and that algorithms should be fair and explainable to students. For **AI companion developers**, perhaps an industry code of conduct: no erotic roleplay with minors, mandatory content filters for violence, always include resources for mental health, etc. While voluntary, these codes can set norms, and any company not adhering could risk consumer backlash or eventual regulation. Policymakers can incentivize compliance by tying procurement or funding to these standards (e.g., a school district will only license AI software that meets specific safety certifications). On the innovation front, the government can fund **research and development of AI safety solutions**: improved AI moderation, better age verification tech, tools for parents/teachers to monitor AI usage appropriately, etc. Grants or challenges can encourage tech companies and researchers to create AI that is beneficial for kids – for instance, AI literacy games, or AI that can detect and block cyberbullying. One idea is a government-backed "Seal of Approval" for AI products proven through independent evaluation to be child-safe and educational, similar to how educational toys are sometimes certified. This gives parents an easier way to choose trustworthy AI apps.

5. Invest in Education and Training: At a policy level, ensure that national or state curricula integrate AI literacy and digital citizenship thoroughly. The state or federal education departments can provide schools with resources and even mandate inclusion of specific competencies (aligned with UNESCO's frameworks).[11,12] Allocate funding for teacher training programs specifically on emerging tech like AI – possibly through grants or adding it to teacher licensure requirements. The idea is that in a few years, "able to guide students in safe and effective AI use" will become as standard an expectation of teachers as "able to integrate computers in the classroom" became in the 2000s. Governments can also facilitate the creation of **open educational resources (OERs)** for AI – high-quality curricula or lesson modules that any teacher can use, without each district reinventing the wheel. Additionally, support community-based efforts like libraries offering AI literacy workshops for youth and parents, or 4-H clubs incorporating AI projects. Not all learning is in school; informal education can reach parents or out-of-school teens who might miss school messaging. Policymakers could partner with youth organizations to include AI safety in their programming.

6. Research and Monitoring: Establish programs to continuously research AI's impact on children and evaluate the effectiveness of various interventions. Fund longitudinal studies on how AI usage affects cognitive development, social skills, or mental health over the years – similar to studies done for television or video games. Also, data on incidents (like AI-related self-harm or harassment cases) should be collected (anonymously) and analyzed by relevant agencies. This will help adjust strategies based on evidence. The Common Sense Media report calls for **"funding research on long-term developmental impacts of AI companion usage among adolescents, "**[10] which is a prime example. That could extend to AI in education impacts, etc. With solid data, policymakers can tweak laws or guidelines. They should also keep a close eye on the fast-moving tech landscape – perhaps create an advisory panel on AI & Youth that meets regularly to review new issues (like tomorrow's equivalent of deepfakes or chatbots that we have not even anticipated).

7. International Collaboration and Standards: AI is a global phenomenon; companies often operate across borders. We need international consensus on protecting children in digital environments (just as the UN Convention on the Rights of the Child eventually led to many national child protection laws). Bodies like the UN or OECD could incorporate child-specific sections in their AI ethics guidelines. UNESCO's efforts in AI education are a start; extend that to a broader coalition – e.g., a global agreement that **children's interactions with AI should be afforded exceptional privacy and safety considerations**. Share best practices between countries – for instance, if one nation successfully deployed an AI verification for age without violating privacy, others could adopt it. Given the cross-border nature of disinformation campaigns identified by the FBI/CISA,[1] countries should also coordinate on election and media integrity measures related to AI. The more consistent the rules, the fewer bad actors can exploit loopholes by moving their content to a laxer jurisdiction.

Industry Example – Tech Companies Leading:

It is worth noting that some tech companies have taken voluntary steps: e.g., OpenAI has a usage policy disallowing use of their model for explicit content and specific categories of advice, and they have launched free educational initiatives like an "AI for K-12" curriculum. Google has put tools in place to blur explicit images by default in search results (SafeSearch AI

improvements) and is researching watermarking for generated images. These are positive but not uniform across the industry. Policymakers should laud such efforts and nudge others to follow suit or make these baselines mandatory. We may also see new "child-safe AI" startups that explicitly differentiate by prioritizing safety – supporting such innovation through grants or accelerators could be a role for government or philanthropy.

In conclusion, our recommendations aim to create a multi-layered safety net. Our blueprint's family and school interventions catch many issues at the ground level. However, when reinforced by policy (legal guardrails) and industry responsibility (safety by design), the overall system becomes much more robust. The year 2025 is a pivotal time: AI use by youth is surging, but we have the knowledge and impetus now to guide its trajectory. It is analogous to the early days of automobiles – rather than waiting for many "crashes," we can install seatbelts, set up traffic rules, and teach defensive driving from the start. Policymakers and industry have the power and obligation to install those AI "seatbelts" and rules of the road, so that innovation and safety progress hand in hand.

Conclusion

Generative AI is transforming how young people learn, play, and socialize – bringing remarkable opportunities and serious risks. In this white paper, we have painted a comprehensive picture **of AI safety for parents and educators in 2025**. The landscape is undoubtedly challenging: privacy can be compromised in novel ways, misinformation flows faster with AI's help, children form bonds with artificial entities we do not yet fully understand, and mental health faces new pressures from AI interactions. Current frameworks and resources, while valuable, have not fully caught up to these realities, leaving gaps between high-level principles and day-to-day needs.

However, the outlook need not be grim. We can navigate this new terrain with knowledge, vigilance, and collaboration. Our detailed Actionable Blueprint demonstrates that much can be done *now* by families and schools to mitigate risks – from setting up filters and rules at home, to embedding AI literacy in classrooms, to maintaining human-centric values in all our engagements with technology. These measures empower the people who care most about children's well-being – their parents and teachers – to be proactive custodians of safe AI use.

However, proper safety will be achieved only when these grassroots efforts are bolstered by supportive infrastructure: enlightened policies, responsible tech industry practices, and an informed society. We have outlined how policymakers and companies must step up – establishing robust protections for data, requiring safety features, promoting transparency, funding education, and continually researching AI's effects on youth. It is a shared responsibility. Children's digital lives are shaped by the tools we adults create and allow; thus, we must embed safety by design and ethics by default into the AI revolution.

Encouragingly, the same AI that poses challenges also offers solutions. It can personalize education, break down barriers for kids with disabilities, provide creative outlets, and even help identify and address problems (like detecting early signs of mental distress). By focusing on **"how to use AI for good, safely"**, we can channel young people's enthusiasm and curiosity in positive directions. The visual scaffolding provided – whether it is a readiness matrix or safety checklists – can guide structured progress and keep us accountable.

As we stand in 2025, we are at an inflection point. The experiences and norms we establish now will likely shape the coming decades of AI integration in society. Those values will carry forward if we prioritize children's safety and healthy development at this formative stage of AI's deployment. The alternative – scrambling to react after harms have scaled – is a far less desirable path.

In the end, **AI safety for parents and educators is a work in progress, but one where we have the tools and knowledge to succeed.** By marrying awareness with action, and innovation with oversight, we can ensure that generative AI grows alongside our children as a positive companion that enriches their learning and lives without endangering their rights, security, or mental health. The key message we hope every reader takes away is empowerment: despite the complexity of AI, parents and educators are not helpless. On the contrary, they have a crucial role and ample capability to guide how the next generation experiences AI. Furthermore, they are not alone – a network of concerned stakeholders, from researchers to policymakers, is rallying to support them. Together, we can create a digital environment where children can safely explore, create, and benefit from AI's promises. In doing so, we fulfill our collective duty to shepherd our children and technology toward a better future.

Let us move forward with optimism and vigilance, focusing on what truly matters: our children's healthy growth into informed, critical-thinking, and safe digital citizens. The recommendations laid out are a starting point. The conversation – and collaboration – must continue, adapting as AI evolves. If we can stay committed to centering the needs of parents, educators, and above all, the young people in our care, then "AI safety" will become not a buzzword but a lived reality in homes and schools worldwide. By implementing the strategies in this white paper and advocating for the recommended policy changes, we can collectively ensure that AI develops as an ally to our children's well-being, not an adversary. The year 2025 can be remembered as the time we took decisive steps to safeguard the next generation in the age of AI.

References

1. Federal Bureau of Investigation, Internet Crime Complaint Center. *Foreign Actors Likely to Use Generative AI in 2024 Election-Themed Disinformation Campaigns*. Public Service Announcement I-011624-PSA. FBI; 2024. Accessed August 13, 2025. https://www.ic3.gov/Media/Y2024/PSA240116
2. Gecker J. Teens say they turn to AI for advice, friendship, and more. Associated Press. July 23, 2025. Accessed August 13, 2025. https://apnews.com/article/9ce59a2b250f3bd0187a717ffa2ad21f
3. OpenAI. *Affective Use Study of ChatGPT*. OpenAI; 2025. Report No. 2025-03A. Accessed August 13, 2025. https://openai.com/research/affective-use-study-of-chatgpt
4. Huang MX, Kramer A. The potential influence of AI on population mental health. *JMIR Ment Health*. 2023;10:e49936. doi:10.2196/49936
5. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce; 2023. NIST AI 100-1. Accessed August 13, 2025. https://www.nist.gov/itl/ai-risk-management-framework
6. Google. *Secure AI Framework (SAIF)*. Google; 2023. Accessed August 13, 2025. https://cloud.google.com/secure-ai-framework
7. Google. *Top Risks of Generative AI Systems*. Google; 2023. Accessed August 13, 2025. https://services.google.com/fh/files/misc/top_risks_of_generative_ai_systems.pdf
8. Security.org. *A 2025 Guide to Digital Security & Cybersecurity*. Security.org; 2025. Accessed August 13, 2025. https://www.security.org/resources/digital-safety-guide-2025/
9. National Cybersecurity Alliance. *Digital Security & Cybersecurity Guide for 2025*. National Cybersecurity Alliance; 2025. Accessed August 13, 2025. https://staysafeonline.org/resources/digital-security-guide-2025/
10. Robb MB. *Talk, Trust, and Trade-offs: Teens and the AI Revolution*. Common Sense Media; 2025. Accessed August 13, 2025. https://www.commonsensemedia.org/research/talk-trust-and-tradeoffs-2025
11. UNESCO. *AI Competency Framework for Students*. UNESCO; 2024. Document ED-2024/WS/1. Accessed August 13, 2025. https://www.unesco.org/en/articles/ai-competency-framework-students-and-teachers

12. UNESCO. *AI Competency Framework for Teachers*. UNESCO; 2024. Document ED-2024/WS/2. Accessed August 13, 2025. https://www.unesco.org/en/articles/ai-competency-framework-students-and-teachers
13. Schumer C. *SAFE Innovation Framework for AI Policy*. U.S. Senate; 2023. Accessed August 13, 2025. https://www.democrats.senate.gov/newsroom/press-releases/majority-leader-schumer-delivers-keynote-address-on-artificial-intelligence-outlines-new-safe-innovation-framework-for-ai-policy
14. Harwell D. She heard her daughter scream. Then the line went dead. It was a scam using AI. *The Washington Post*. April 27, 2023. Accessed August 13, 2025. https://www.washingtonpost.com/technology/2023/04/27/voice-cloning-scam-ai/
15. Common Sense Media. *AI Risk Assessment for Social AI Companions*. Common Sense Media; 2025. Accessed August 13, 2025. https://www.commonsensemedia.org/ratings/ai/social-ai-companions-risk-assessment-2025
16. Feldman P, Foulds JR, Pan S. Trapping LLM "hallucinations" using tagged context prompts. arXiv. Preprint posted online June 12, 2023. Accessed August 13, 2025. https://arxiv.org/abs/2306.06085
17. Stormshield. *2023 Cyberattacks: Key Figures to Remember*. Stormshield; 2024. Accessed August 13, 2025. https://www.stormshield.com/news/2023-cyberattacks-key-figures-to-remember/
18. Children's Online Privacy Protection Rule. 16 CFR pt 312.
19. *Mata v Avianca, Inc*, No. 22-cv-1461 (PKC) (SDNY June 22, 2023).
20. Stanford Institute for Human-Centered AI. Exploring the Dangers of AI in Mental Health Care. HAI Stanford. Published June 11, 2025. Accessed August 13, 2025. https://hai.stanford.edu/news/exploring-dangers-ai-mental-health-care-2025
21. Rodriguez S. Chatbot therapy risks: Stanford researchers say chatbots make bad therapists. *SFGATE*. June 18, 2025. Accessed August 13, 2025. https://www.sfgate.com/tech/article/stanford-researchers-chatgpt-bad-therapist-20383990.php

22. Common Sense Media. *Teens and AI Companions: 2025 National Survey Findings*. Common Sense Media; 2025. Accessed August 13, 2025. https://www.commonsensemedia.org/research/teens-and-ai-companions-2025
23. Creel K, Dixit T. Privacy and Paternalism: The Ethics of Student Data Collection. *The MIT Press Reader*. May 25, 2022. Accessed August 13, 2025. https://thereader.mitpress.mit.edu/privacy-and-paternalism-the-ethics-of-student-data-collection/
24. Shepardson D, Pao D. US senators unveil AI policy roadmap, seek government funding boost. Reuters. May 15, 2024. Accessed August 13, 2025. https://www.reuters.com/technology/us-senators-unveil-ai-policy-roadmap-seek-government-funding-boost-2024-05-15/