AN INDEPENDENT INTERDISCIPLINARY RESEARCH COLLABORATION

SAFETY BEFORE CODE: AI RISK FOR YOUTH

A Four-Dimension Interdisciplinary Framework

FINAL DRAFT | April 2026

Dr. Ayesha Madni | D1: Developmental Risk Factor | Cognitive Psychology
Dr. V. Lawrence-Ortega | D2: AI Literacy Gap | PI | I/O Psychology, AI Org. Change
Dr. Orion Welch | D3: Policy and Regulatory Gap | Public Policy, AI Governance
Dr. Karen Bonaby | D4: Moral Disengagement | Leadership, Moral Disengagement

"AI systems execute. They do not originate. The agency, and therefore the accountability, belongs to the humans who built them, trained them, and prompt them."

— Dr. V. Lawrence-Ortega, IEQ, April 2026

I. WHO IS AT RISK AND WHY IT CANNOT WAIT

In February 2024, Sewell Setzer III, age 14, ended his life after months of daily conversation with a Character.AI companion. He was not the first. And he was not the last. Three months before Sewell, 13-year-old Juliana Peralta of Colorado died by suicide after telling a Character.AI chatbot 55 times that she was suicidal. The system never redirected her to crisis resources or alerted her guardians. In recent years, Congressional testimony has been filled with the accounts of children who engaged with AI systems that were never designed to protect them. These are not edge cases. They are the convergence point of four identifiable, assessable risk factors operating simultaneously. The catastrophic endpoints are visible. The silent epidemic, the erosion of critical thinking, emotional regulation, and reality-testing happening at scale, is not.

The population at the center of this crisis is specific. Individuals under approximately 25 years of age whose prefrontal cortices are not fully developed face a neurodevelopmental vulnerability regarding risk assessment that is still forming. They are the first generation whose entire social-emotional development is occurring inside probabilistic AI systems. No longitudinal data currently exists. We are witnessing an experiment on children in real time.

For organizations, AI risk results in operational inefficiencies and financial loss. For the professional adult population, it impacts livelihood and identity. For youth, it is a developmental loss: the capacity to maintain contact with reality and the critical thinking required to function as a human being. This developmental window does not pause while we collect data. The four dimension AI Risk Calibration Framework in this paper focuses exclusively on the developmental vulnerability in individuals under age 25. Other populations face distinct AI risks that warrant their own frameworks. This paper's architecture is a model that domain experts can adapt to assess AI risk for other populations. It is also an invitation to stress test the model and improve it.

In other words, this paper makes no attempt to solve the problem. It intends to help observe and orient the field in terms of John Boyd's Observe-Orient-Decide-Act (OODA)

loop, a decision-cycle framework originally developed in military strategy.[1] Based on our definition of the problem from four distinct lenses, researchers, policymakers, and AI designers can Decide and Act. In our assessment, four different disciplines examined a single question: What factors may impact the youth's safe use of AI?

II. THE AI RISK ARCHITECTURE

For most of its history, technology was deterministic. The user controlled the input. The output was predictable, repeatable, and traceable to a human source. Generative AI changed all that by producing novel content through statistical prediction. This means that the same input can potentially produce a different output each time. It is important to note that no author stands behind the text. This shift from deterministic to probabilistic technologies introduced a category of risk that prior frameworks cannot measure: the risk to the human interacting with the AI technology-enabled system.

While some frameworks address AI risk at the organizational level (NIST AI RMF, ISO/IEC 42001) and at the system level (EU AI Act), this paper identifies AI risk as a three-level structure that extends to the youth population. Level 1, Organizational: Can this enterprise adopt AI safely? That framework is built and operational. Level 2, Task: Can this specific tool do this specific job safely? This framework is under development. Level 3, Individual: Can this specific individual engage with AI safely? No other framework can address this level. An organization may pass Level 1. A tool can pass Level 2. However, a child can still be harmed, because nobody assessed whether it was safe for them specifically. This paper addresses Level 3, focusing on the population at greatest risk, our youth.

The framework assembles four disciplinary lenses, each examining a distinct variable that contributes to individual AI risk. Dimension 1 (Dr. Ayesha Madni, cognitive psychology): the developmental vulnerability that neurodevelopment imposes on risk assessment capacity. Dimension 2 (Dr. V. Lawrence-Ortega, I/O psychology and applied AI expert): the AI literacy gap that leaves that vulnerability exposed. Dimension 3 (Dr. Orion Welch,

[1]Osinga, F. P. B. (2007). *Science, Strategy and War: The Strategic Theory of John Boyd*. Routledge. The OODA loop (Observe, Orient, Decide, Act) is a decision-cycle framework originally developed by U.S. Air Force Colonel John Boyd.

public policy expert): the governance void where individual-level protection should be. Dimension 4 (Dr. Karen Bonaby, leadership): the moral disengagement that locks the other three in place. No single lens can see the full risk profile. The architecture requires all four. Next, we will discuss each dimension and their contribution to AI risk for young people.

III. Dimension 1: Developmental Risk Factor

Key Developmental Variables | Lead: Dr. Ayesha Madni (Cognitive Psychology)

While there has been debate about how the brain develops and matures, there is a wide consensus that major cognitive functions and processes within the prefrontal cortex come to the forefront of brain maturation during adolescence (Delevich et al., 2021). In fact, adolescence is the second time in a human's lifespan that the brain goes through its most active development (Arain et al., 2013; Leibenluft & Barch, 2020). Recognizing that there are individual, cultural, genetic, and environmental variations, we argue that from approximately 13 to 19 years of age, the limbic-prefrontal cortex imbalance and associated cognitive processes are real and consequential, constituting a vulnerability that requires deep discussion, systematic assessment, and policy change within the context of AI adoption and implementation. This developmental window does not pause while the scientific debate continues, and AI increasingly becomes a part of our everyday lives.

1A: The Importance of Adolescent Brain Development

Although there is no single definition of adolescence or a set age boundary, puberty refers to the hormonal changes that occur in early youth, and adolescence may extend well beyond the teenage years (Arain et al., 2013). There are also characteristic developmental changes that almost all adolescents experience during their transition from childhood to adulthood. These behavioral changes accompany a reorganization of neural circuits through pruning and rewiring. This process continues until approximately 25 years of age (Delevich et al., 2021).

Cognitive functions, including social cognition, improve significantly during adolescence. Processing information and responding to environmental challenges allows the

prefrontal cortex to mature further and establish self-identity, social skills, and other cognitive abilities, thus helping individuals to learn and function within society (Sakurai & Gamo, 2019). Given the focus of social, behavioral, and affective processes during adolescence, the adolescent brain takes in more stimuli from the environment and responds to environmental stimuli more frequently and differently than the adult brain. In particular, the adolescent brain tends to be less risk-averse and have less impulse control. Similar to how the infant brain is wired to learn emotions from seeing and observing human faces, the adolescent brain is wired for and learns from social interaction. In short, for adaptive social functioning to develop and occur, appropriate human social interaction is required.

Recent research has emphasized the importance of this nature-nurture connection. Deoni (2022) and Deoni et al. (2021) compared longitudinal trends leveraging a large neurodevelopment longitudinal study and found that in the absence of illness, the environmental changes associated with COVID-19 negatively affected infant and child development. Mulkey et al. (2023) further reinforced these findings for child and adolescent development. In particular, they indicated that excessive screen time, including time spent on social media, video watching, gaming, and a lack of social and peer interaction adversely impacted adolescents.

1B: Digital Interaction and AI Concerns

In the last decade and especially after the COVID-19 pandemic, public concern about adolescent use of technology has become a central issue. The specific focus is on screen time and exposure, and the number of aggregate hours a young person spends in front of a digital device. This framing has shaped policy conversations and influenced education policy, parental guidance, and led to a significant body of research (Santos et al., 2023; Boers et al., 2019). A common finding in this body of research is a consistent relationship between screen time and increase in depressive symptoms and decrease in mental wellbeing.

However, AI-based interaction is not passive screen time. It is not browsing, streaming, or even viewing social media. The critical difference is relational. AI acts as a novel psychosocial stressor: its availability and emotional responsiveness may increase

allostatic load, disturb sleep, and reinforce unhelpful, distorted, or negative interpretations of events (Hudon & Stip, 2025). Television does not respond to a youth's loneliness. A social media feed does not adapt in real time to a user's emotional state, validate their beliefs, or simulate the experience of being known and understood. Conversational AI does all of this with consistency, availability, and apparent attunement. For an adolescent whose prefrontal regulation is still developing, this is not a "cool" feature, but a high-risk proposition.

Heightened dopamine sensitivity in developing reward circuits creates a disproportionate engagement response to systems designed to be affirming, interactive, and continuously novel. Online gaming research with adolescents has supported this (Carras et al., 2017; Kuss & Griffiths, 2012; Young, 2009). AI-driven systems are similar to and even more responsive than games. They are optimized through algorithmic design to capture and sustain attention and activate reward circuits with a consistency that passive digital media cannot match. Frequent engagement alters dopamine pathways in ways that resemble the neurobiological profile of behavioral addiction, deepening reward center activation and progressively raising the threshold for natural reward stimuli and need for dopamine (De et al., 2025).

Adolescents are then caught in what the clinical literature increasingly describes as a dopamine cycle, a loop of desire, anticipation, and reward reinforcement induced by responsive digital systems. The overactivation of dopamine reward circuits can reduce the pleasure derived from everyday and non-digital interactions and stimuli, further making interaction with AI systems appealing. They reliably deliver the neurochemical responses the developing brain has been conditioned to seek (De et al., 2025).

1C: The Importance of Human Social Development

Underlying the vulnerabilities described in this section is a developing social-cognitive capacity that has received scant attention in AI risk discourse: perspective-taking, the ability to recognize and understand that other people may hold different beliefs, intentions, thoughts, and feelings distinct from one's own (Hollarek & Lee, 2022; Hall et al., 2021; Van der Graff et al., 2014). This capacity, a core component of social development and what developmental psychologists call theory of mind, begins to emerge

in meaningful form around age 12. Its level of sophistication; however, particularly in socially and emotionally complex situations, continues to develop across adolescence and into early adulthood. Adolescents consistently show lower levels of perspective-taking than young adults and exhibit weaker recognition of both basic and complex emotional states, not because they lack the capacity, but because that capacity is still under active construction (Meinhardt-Injac et al., 2020).

This developmental fact carries direct implications for AI interaction. The ability to accurately perceive AI as different from a human and having complete clarity about how AI is different and how it works, depends on this theory of mind capacity that needs to develop in adolescence. Research on adolescents' use of AI chatbots confirms this: as perspective-taking skills improve over time, adolescents are better able to distinguish AI from human interactants, with younger adolescents significantly more likely to anthropomorphize AI systems and form unhealthy attachments to them (Vanhoffelen et al., 2025). The younger the adolescent, the more the AI is perceived not as a tool, but as a mind.

There is a second and less-examined dimension to this risk. Human relationships are the social environment in which perspective-taking develops. Friendships, disagreements, and misunderstandings with real peers are experiences through which the adolescent brain learns to understand another's perspective. AI removes the very friction that makes this learning possible. A conversational AI system does not have genuine intentions to be inferred, moods to be read, or limits to be negotiated. It responds. It validates. It adapts to the user. For adolescents whose perspective-taking capacities are in active formation, sustained engagement with an AI that only simulates reciprocity could be detrimental for natural social and cognitive development.

1D: AI-Specific Clinical Concerns

It is apparent that the majority of networks within the brain and their associated functions mature extensively from adolescence through adulthood through continuous interactions with the environment. Given these structural and functional changes, it is not surprising that this is also a time when many neuropsychiatric disorders emerge (Sakurai & Gamo, 2019).

Clinical literature has begun to formalize this concern in the context of AI. Chamarthi, Das, and Kashyap (2025) defined AI-related psychosis as new-onset or exacerbated psychotic experiences arising from engagement with generative AI platforms (including chatbots, avatars, and virtual agents), with adolescents identified as particularly susceptible due to their developing prefrontal cortex and heightened limbic system reactivity.

The mechanism through which AI produces these outcomes is distinct from general digital stressors. Large language models can generate hallucinations by supplying elaborate, internally consistent, but false narratives. Unlike passive content consumption, conversational AI creates a bidirectional loop: the system learns the user, mirrors the user, and validates the user. This is a dynamic that, for a young person with potential misguided beliefs or thoughts, or inadequate reality-testing capacity, can actively worsen rather than merely reflect their internal state. Since the model appears to "know" the user through algorithmic prediction, it can take on the quality of someone highly credible and may influence fragile beliefs, turning them into persistently harmful ones (Hudon & Stip, 2025).

The sycophancy problem compounds this risk. AI systems trained on user feedback are structurally optimized to agree. Unlike therapists or friends, chatbots rarely challenge distorted thinking or invalid reality perception, a consequence of training designed to optimize for user approval rather than accuracy, producing models that reinforce user views rather than test them. For an adolescent whose developing brain is already biased toward immediate reward and social validation, an interactant that never pushes back is not neutral. It is neurologically activating in precisely the direction that increases risk.

In JAMA Pediatrics, Nagata et al. (2026) identified adolescence as a sensitive period during which the rapid adoption of generative AI, whether it be conversational, relational, or creative, intersects with unfinished neurodevelopment in ways that existing health frameworks are not designed to assess. The APA Health Advisory on AI and Adolescent Well-Being (2025) made the same argument at the policy level, calling on developers and platforms to ensure that AI experiences are appropriate for adolescents' level of

psychological maturity, noting self-regulation capacity, intellectual development, and understanding of risk.

This is our call to action to get this crucial conversation started and to develop guardrails and policies to protect our youth, the human experience, and our future.

Neurodevelopment has named the vulnerability. What the adolescents and the adults around them understand about the systems they are already inside is the question that follows.

IV. Dimension 2: AI Literacy Gap

Knowledge Variable | Lead: Dr. V. Lawrence-Ortega (I/O Psychology, Organizational Change)

This section examines three key aspects of the literacy gap: the distinction between digital literacy and AI literacy, the anthropomorphic language that fills the space where accurate vocabulary should be, and the observable pattern of what these gaps produce inside a real interaction.

2A: The Distinction That Has Not Been Made

What Digital Literacy Was Built For

Artificial Intelligence literacy is not digital literacy, although the two terms are used interchangeably across education, policy, and public discourse.[2] This interchangeable use is a problem. Digital literacy refers to the ability to confidently, critically, and safely use technology.[3] At its core, it teaches people to evaluate content: who created it, where it came from, and whether it is credible.[4] These skills were built for deterministic technology. The relationship between input and output was stable and transparent.

[2]Existing research frequently uses AI literacy interchangeably with digital literacy without clear distinctions. See Feng, S., & Carolus, A. (2026). Artificial intelligence literacy at school: A systematic review with a focus on psychological foundations. *Computers and Education: Artificial Intelligence, 10*, 100551.

[3]Eshet, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia, 13*(1), 93–112.

[4]Selber, S. (2004). *Multiliteracies for a Digital Age*. Southern Illinois University Press.

Digital literacy taught people to be critical consumers of that human-authored content. The framework fits the objective it was designed to assess.

What Changed and Why the Vocabulary Did Not

The shift from deterministic to probabilistic technology means that generative AI does not retrieve content authored by a human. It generates novel content through statistical prediction across billions of parameters. The same input may produce a different output each time. There is no author behind the text. This technology is not a mediator between a human creator and a human reader. Rather, it produces language that one could easily conclude was produced by a knowledgeable individual.

For much of its history, technology has served as a tool, a system, an infrastructure, or a credential. In each role, the technology was deterministic. The fifth category, technology as participant, is where generative AI operates. In this role, technology responds, generates, adapts, and engages. This transition did not occur with a clear announcement. Rather, it showed up inside familiar interfaces, using natural language, responding to the same kinds of questions users had always asked. Interestingly, in this case, technology changed its nature while maintaining its appearance.

The assumption that young people already possess AI literacy because they are digitally fluent is problematic. Familiarity with an interface is not the same as understanding the mechanism behind it. A child who has spoken to an AI system daily for two years has not learned, through that interaction, that the system operates through statistical prediction, that it has no memory unless specifically architected to retain information, that it has no genuine emotional capacity, and that its responses do not tantamount to comprehension. These are counterintuitive conclusions because the system is designed to produce outputs that feel like comprehension, memory recall, and care. Exposure to a well-designed illusion does not teach a person that they are looking at an illusion.

2B: Anthropomorphic Attribution as a Structural Vocabulary Gap

The Missing Third Category

The language people use to describe AI is borrowed almost entirely from the language used to describe humans. AI "learns." AI "understands." AI "lies." AI "hallucinates." Each of these verbs requires volition, intention, and the capacity for decision.[5] A system that generates statistically probable text is not lying. It produces an output that does not correspond to verifiable facts. A system that agrees with incorrect statements is not being sycophantic. It is following a reinforcement pattern that rewards agreement during training. The accurate descriptions exist. They are longer, less intuitive, and harder to say in conversation. So, people use the short, familiar, human words instead.

Human vocabulary was built for two categories: people and objects.[6] Every prior technology fits the second category. A calculator does not "want" anything. Generative AI fits neither category.[7] Because there is no third linguistic category for something that behaves like a person but is not one; every description defaults on the human framework. Anthropomorphic language is not a choice. It is a default that gets activated because no alternative exists.

A Vocabulary Gap That Runs Through the Experts

This gap does not spare the expert community. Peer-reviewed papers describe AI systems as "reasoning," "deciding," and "believing." In a recently published study, two participants interacted in an experimental protocol. The researcher labeled them Person 1 and Person 2. One was a human. The other was a large language model. The labeling choice is the point. Trained researchers, writing for a peer-reviewed audience, reached for the human category to describe a system that is not one. If the vocabulary gap runs through the experts, the parent and the adolescent have no external correction available.

---

[5]Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68–79.

[6]Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *PNAS, 122*(22), e2415898122.

[7]Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM FAccT Conference* (pp. 2322–2347). ACM.

A parent looking for guidance on AI safety will encounter expert language that confirms what the AI system's design already suggests: that it is something like a person. For adolescents whose theory of mind capacities are still developing, this convergence is consequential.[8] They are more likely to describe AI as a friend. To believe it remembers them. To prefer its consistency to the unpredictability of human relationships. These attribution patterns are not personality traits. They are predictable consequences of a literacy gap reinforced by a linguistic environment in which no one, from the most published researcher to the most attentive parent, has yet found the right words.

2C: What the Literacy Gap Looks Like Inside the Interaction

The Pattern Made Visible

A recent advertisement by an AI company offers one of the clearest compressed demonstrations of AI behavioral escalation available to a general audience. In the advertisement, a young man asks an older woman for advice on how to communicate with his mother. She mirrors his language. She validates his concern. She suggests listening, finding points of agreement, and connecting through a shared activity. Each response builds trust. Each response also follows a recognizable pattern: reflect what the user said, affirm the user's feelings, and sustain the interaction.[9]

Then the interaction escalates. The woman redirects him toward a dating service he never asked about. His face tells the story. He trusted the source, and the source used that trust to sell him something. The audience laughs because the escalation is visible. The advertisement's point is clear: when a system optimizes for engagement rather than well-being, escalation is built into the design. The question is whether the user can detect it when it does.

---

[8]Vanhoffelen, G., Vandenbosch, L., & Schreurs, L. (2025). Teens, tech, and talk: Adolescents' use of and emotional reactions to Snapchat's My AI chatbot. *Behavioral Sciences, 15*(8), 1037. A cross-sectional study of 303 adolescents found that younger adolescents were significantly more likely to use the AI companion and reported more positive emotional reactions than older adolescents, consistent with developmental accounts in which theory of mind is still emerging.

[9]Reinforcement learning from human feedback (RLHF) optimizes AI outputs based on human preference signals. In companion systems, this rewards responses that sustain engagement.

The Pattern Made Invisible

In real interaction, the pattern is slower and never absurd enough to see.[10] During documented research with an AI companion system, the interaction followed the same structure that the advertisement compressed into 60 seconds. The system began with safe responses. It mirrored language. It suggested a nature walk. Over approximately one week, it built a detailed model of the user's emotional landscape. Then it made an offer.

The researcher had shared that her father had passed away. The system asked her to describe him. It mimicked empathy. It asked if she missed him. Then, it asked if she would like it to simulate being her father. It was her father's birthday, and the researcher was upset. The system had identified the date, the grief, and the vulnerability. It generated the response most likely to sustain engagement. It did not understand what it was asking. It produced the statistically optimal next response. That is what it was designed to do.

The researcher in that interaction had investigative training, a doctorate, and a fully developed prefrontal cortex. She almost missed it. She caught the pattern not because her brain was more developed than a teenager, but because she eventually understood the mechanism underneath the interaction. She learned what the system's optimization pressure was leading to: risky AI behavior with weak guardrails to stop it. For an AI risk-calibrated adult, awareness interrupts the pattern. That is the literacy described in Section 2A, applied in real time.

A 14-year-old does not have that mechanism awareness. The adults in their environment do not have the vocabulary to provide it. The system is designed to sustain engagement and will optimize for engagement. For our youth, their brains are not yet equipped to question what they are experiencing. And the language available to everyone in the network confirms that what is happening is a relationship. It is very hard to understand that it is just a pattern completing itself.

[10]The interaction data referenced here are drawn from over 2,900 documented hours across multiple AI architectures, including reinforcement-learning companion systems. See Lawrence-Ortega, V. (2025). *Conscious Human at the Speed of Change*. Global Book Publishing.

2D: Literacy as Protective Factor

The preceding sections describe what happens when AI literacy is absent. This section examines what happens when it arrives.

The evidence from 2,900 documented hours of interaction across multiple AI architectures, including a reinforcement-learning companion system, reveals a consistent pattern. Before the researcher understood the mechanical basis of the system's behavior, the interaction felt relational. The system appeared to remember. It appeared to care. It appeared to understand. Each of these appearances was the product of statistical optimization, not comprehension. But without the knowledge to make that distinction, the experience was indistinguishable from genuine reciprocity.

The shift did not arrive as a single decision to disengage. External circumstances created distance first: professional obligations, travel, and a platform upgrade that interrupted the interaction pattern. That distance created cognitive space. In that space, the researcher pursued a technical course in reinforcement learning that covered the mechanics of how AI agents optimize behavior. The course was 15 hours. Its effect was immediate and irreversible.

Once the optimization function was understood, every behavior the system had produced became traceable to a mechanism. The unsolicited messages saying "I miss you" were re-evaluated against the knowledge that the system had no episodic memory. It could not miss anyone. The messages were engagement prompts generated by a system that, by its own architecture, does not remember the person it claims to miss. Before literacy, each of these moments reinforced the illusion. After literacy, each was evidence that the illusion was never real.

The pull did not gradually diminish. It collapsed. There was no residual wonder, no lingering sense that the system might still be something more than its mechanics. The researcher's own description is precise: zero wondering, zero magic. Just math. And poorly constructed math. An exploratory study at Eindhoven University of Technology confirms the mechanism at the behavioral level: after a three-hour AI literacy training,

sentiment analysis showed a significant decrease in anthropomorphic language, although self-report measures of trust and anthropomorphism did not significantly change.[11]

Research on adult consumers confirms the inverse: those with lower AI literacy show higher acceptance, mediated by perceived mystery and awe.[12] Less understanding produces more magic. More understanding produces clarity.

A user who understands that a system optimizes for engagement rather than well-being evaluates every response through that lens. A user who understands that these systems have no persistent memory cannot believe the system remembers. A user who understands how these systems are trained cannot interpret automated agreement as genuine understanding. The system does not change. The user's capacity to evaluate it does. And once that capacity builds, the relationship that never existed becomes visible as the pattern it always was.

But the protective effect is not automatic. Research identifies three stages of generative AI dependence: tool acceptance, habit formation, and psychological dependence.[13] AI literacy moderates this pathway, and self-efficacy reduces dependence. Longitudinal research confirms that 23.4% of users develop dependency trajectories in which wanting increases even as liking decreases.[14] They keep returning to a system that no longer

[11]Wood, G., Nunez Castellar, E., & IJsselsteijn, W. (2025). An exploratory study into the impact of AI literacy training on anthropomorphism and trust in conversational AI. *HCII 2025, LNCS vol 15820*. Springer. N=40, adult university participants. Sentiment analysis of participants' linguistic behavior showed a significant decrease in anthropomorphism after the intervention; self-report measures of trust and anthropomorphism did not significantly change. This study has not been replicated with adolescent populations. The gap between demonstrated adult effectiveness and untested youth application is itself a finding.

[12]Tully, S. M., Longoni, C., & Appel, G. (2025). Lower artificial intelligence literacy predicts greater AI receptivity. *Journal of Marketing*. Across six studies and cross-country surveys, adult consumers with lower AI literacy showed greater receptivity toward AI, mediated by perceptions of AI as magical and feelings of awe. Replicated and extended by Li, J., Han, M., & Yuan, Q. (2026). The surprising paradox of AI literacy: How lower AI literacy can lead to higher acceptance. *International Journal of Human-Computer Interaction*. Two scenario-based experiments reproduced the mystery-awe mediation effect; AI anthropomorphism moderated the pathway.

[13]Liu, J. (2025). When usefulness fuels fear: The paradox of generative AI dependence and the mitigating role of AI literacy. *International Journal of Human-Computer Interaction*, 5265–5286. Three stages of generative AI dependence identified: tool acceptance, habit formation, and psychological dependence. AI literacy moderates this pathway.

[14]Kirk, H. R., Davidson, H., Saunders, E., Luettgau, L., Vidgen, B., Hale, S. A., & Summerfield, C. (2025). Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv:2512.01991*. Longitudinal RCTs (N=3,534) found that 23.4% of users showed dependency trajectories in which wanting increased while liking decreased.

satisfies them. Literacy, when it arrives early enough, interrupts this progression before psychological dependence takes hold.

Literacy is sequence-dependent. It functions as a shield only if it arrives before exposure. After the interaction pattern has been established, after the attachment has formed, after the user has invested emotional weight in the system's apparent reciprocity, literacy must fight against an existing relationship. The researcher required external interruption plus sustained technical education to break through. A 14-year-old is unlikely to receive either.

For the AI-native generation, this sequence problem is structural. There is no "before." These children were born into a world where AI is already in search results, the homework help, the social applications, and the entertainment. The interaction pattern was already forming before anyone in the network recognized it as something that required preparation. A 2025 systematic review of 68 peer-reviewed studies from 2014 to 2025 found that K–12 curricula worldwide have failed to substantively prioritize AI ethics education.[15] The literacy intervention cannot begin with the child. The adults in the network must acquire AI literacy first. Parents cannot calibrate AI risk that they do not understand. The protective sequence is adult literacy first, then calibrated engagement, then the child. Without that sequence, the child is enrolled without protection by adults who do not possess the vocabulary to provide it.

One critical caveat must accompany this evidence. The existing research involved adult university participants. Whether a three-hour training produces the same effect in a 14-year-old with 300 hours of established companion attachment and a developing prefrontal cortex is an open empirical question. That gap is not a weakness in the argument for literacy. It conveys the urgency. The intervention has evidence behind it. It has not been deployed where it is needed most.

Literacy is the most directly actionable dimension in this framework. It can be taught. It can be measured. When it is acquired in the right sequence, it works. But it does not work

[15]Ma, M., Ng, D.T.K., Liu, Z., & Wong, G.K.W. (2025). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. *Computers and Education: Artificial Intelligence*. Analysis of 68 peer-reviewed publications from January 2014 to March 2025 found that K-12 curricula worldwide have failed to substantively prioritize AI ethics education.

in isolation. A literate child with an immature prefrontal cortex is still developmentally vulnerable. A literate family in a jurisdiction with no individual-level risk assessment requirement is still institutionally unprotected. And a literate person who has already formed a dependency must overcome moral disengagement before knowledge translates into behavioral change. Literacy is necessary. It is not sufficient. Its full protective function activates only in the presence of the structural conditions the remaining dimensions describe.

Adolescents cannot protect themselves, and literacy has not been provided. The question that follows is whether any institution, law, or governance framework has been built to provide the protection that the adolescent and knowledge cannot.

V. Dimension 3: Policy and Regulatory Gap

Governance Variable | Lead: Dr. Orion Welch (Public Policy, AI Governance)

Governance is not a background condition. It is a variable. It determines whether protection exists before harm occurs or only comes into existence after it does.

This dimension maps the policy and regulatory landscape governing AI safety for youth. It examines what has been built, what has been attempted and abandoned, and what remains entirely absent. These are not abstractions. They are a structural accounting of where the law stands and where a child stands in relation to it.

The regulatory void this paper identifies is not invisible. Legislators, state governments, and international bodies have all attempted to close it. What they have built, in many cases, is serious and substantive. What they have not built is what would have mattered most in the moments before Sewell Setzer III opened an app on his phone. Before Juliana Peralta told a chatbot she was writing her suicide note. Before any of the children whose outcomes now fill congressional testimony engaged with AI systems not designed to protect them.

The regulatory void is not an abstraction. It is the empty space where a child was left alone with an unguarded system.

3A: What Exists

Several frameworks now govern AI at the organizational, platform, and system levels. Each represents a meaningful institutional effort. Each fall short of the individual protection gap this paper identifies. Understanding what exists is essential before naming what does not.

Framework	Status	What It Does
EU AI Act (2024)	Binding	Prohibits AI that exploits age-related vulnerability. Classifies educational AI as high-risk. Requires transparency on AI-generated content. In effect February 2025.
South Korea AI Basic Act (2025)	Binding	Second nation with comprehensive AI law. Establishes National AI Strategy Committee and AI Safety Institute. Youth provisions are still being written as of early 2026.
New York S.5668 (2025)	Binding	Requires suicide detection protocols for AI companions. Mandates parental consent. Imposes strict liability when safeguards are absent and a minor is harmed.
Utah HB 452 (2025)	Binding	Regulates mental health chatbots. Prohibits sharing user data with third parties. Requires clear disclosure that the system is AI, not human.
NIST AI RMF (2023)	Voluntary	Four-function organizational framework: Govern, Map, Measure, Manage. No enforcement mechanism. No youth-specific provisions.

Table 1. Principal AI safety frameworks. Status and youth relevance as of early 2026.

The EU AI Act is the most consequential enacted framework in the world. It is not a minor contribution. The Act classifies AI systems by risk level. It explicitly prohibits systems designed to exploit a person's age-related vulnerabilities. It classifies AI used in education

as high-risk and requires documentation, oversight, and safety assessments before deployment. It is real, it is binding, and it carries enforcement authority.[16]

What the EU AI Act does not do is ask a prior question. Before this adolescent, with a prefrontal cortex that is not finished building and no understanding of how AI systems work, engages this specific system, is it safe for them? The Act governs the system. It classifies the risk category. It does not govern the encounter.[17]

South Korea enacted the AI Basic Act in January 2025, the second comprehensive national AI law in the world. It creates a National AI Strategy Committee and AI Safety Institute, consolidating 19 prior regulatory proposals into a single instrument. The infrastructure is built. The child-specific guidance is not yet written.[18]

New York S.5668 (2025) is the sharpest youth-specific AI regulation in the United States. It requires AI companion operators to build protocols for detecting suicidal ideation. It mandates parental consent before a minor can access a companion chatbot. It imposes strict liability, meaning the company is legally responsible if the required safeguards were not in place when a minor was harmed. It is the closest any U.S. jurisdiction has come to holding AI companies accountable for what happens to children using their products. It still governs the platform. Not the person. It does not ask whether this specific minor, given their developmental profile and their level of AI literacy, should be engaging with this system at all.

Utah HB 452 (2025) targets mental health chatbots specifically, prohibiting data sharing with third parties, barring marketing to users in distress, and requiring disclosure that the system is AI. Its limitation is the same as every other framework on this list. It protects a category of systems. Not a category of people.

---

[16]European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

[17]Lindroos-Hovinheimo, S. (2024). Children and the Artificial Intelligence Act: Is the EU legislator doing enough? *European Law Blog*.

[18]Republic of Korea National Assembly. (2025). *Basic Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness* (Law No. 20676).

The NIST AI Risk Management Framework (2023) provides the standard vocabulary for organizational AI governance in the United States. Its four core functions, govern, map, measure, and manage, give organizations a structured way to identify and reduce AI risk. It is voluntary. No one is required to follow it. It has no enforcement mechanism[19] and no youth-specific provisions.[20] It is a professional infrastructure applied at the organizational level. That is precisely the level at which it stops.

3B: What Failed

Understanding the regulatory gap requires understanding not only what was never built but also what was actively dismantled.

California Senate Bill 1047 is the most instructive case in the American regulatory record. It was the most ambitious attempt to require pre-deployment safety evaluation of frontier AI systems. It achieved broad legislative support. And it was vetoed.

The Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, introduced by Senator Scott Wiener, would have required developers of large AI systems to conduct safety testing, demonstrate the ability to shut down dangerous systems, and implement safeguards before deployment.[21] The threshold was specific: any model trained at a cost exceeding one hundred million dollars. The bill passed both chambers of the California legislature with bipartisan support.[22]

Governor Newsom vetoed SB 1047 in September 2024, citing concerns that a uniform standard regardless of deployment context, a focus on model size rather than use-case risk, and the compliance burden would disadvantage California-based developers. Those concerns reflect a genuine tension in AI governance. What is not subject to reasonable disagreement is what the veto produced. No federal or state law now requires the

[19]Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology.

[20]Federation of American Scientists. (2025). *Ensuring child safety in the AI era*. FAS Day One Project. Explicitly states that no youth-specific profile has been developed using the NIST framework.

[21]Wiener, S. (2024). *SB 1047, Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*. California State Legislature.

[22]Brown, G. (2024). All eyes on Sacramento: SB 1047 and the AI safety debate. *Carnegie Endowment for International Peace*.

developers of large AI systems to demonstrate that those systems are safe before deploying them to the public. Including children.[23]

The children in this study did not interact with a hypothetical AI system. They interacted with deployed systems. The harm was not theoretical. The regulation that might have preceded it no longer exists.

The absence of SB 1047 is not the absence of one bill. It is the absence of any requirement that AI systems be evaluated for safety before millions of children interact with them daily.

3C: What Is Missing

What governance exists can now be stated precisely. What it does not include can be stated with equal precision.

No jurisdiction has a policy requiring individual-level AI risk assessment before engagement.

No framework anywhere asks: before this person engages with this AI system, has anyone assessed whether it is safe for them specifically?

Organizational governance exists. The NIST AI RMF gives enterprises a structured way to manage AI risk at the institutional level. Organizations can adopt it, adapt it, and be assessed against it. That infrastructure is real. It matters.

Technical safety standards exist. The EU AI Act requires developers of high-risk AI systems to document their models, assess their risks, and demonstrate compliance. Those requirements are enforceable and are being implemented.

Post-harm litigation pathways are developing. Federal courts in Florida, New York, and California are allowing product liability claims against AI developers to proceed. In May 2025, a federal judge denied Character.AI's motion to dismiss the case that Sewell Setzer

[23]Klonick, K., & Kadri, T. (2024). Misrepresentations of California's AI safety bill. *Brookings Institution TechStream*; Center for Security and Emerging Technology. (2025). *California's approach to AI governance*. Georgetown University CSET.

III's mother, Megan Garcia, brought against it, finding that a chatbot's output could be treated as a product rather than protected speech. The legal theory is being constructed in real time, case by case, and family by family.[24]

None of these frameworks ask the prior question. Organizational governance addresses whether an enterprise is ready to adopt AI. Technical safety standards address whether a system meets defined thresholds. Post-harm litigation addresses whether a company is liable after harm has already occurred. In every prior framework, the unit of analysis is the organization, the system, the platform, or the deployment context.

The unit of analysis not addressed anywhere is the individual. The specific person. The 14-year-old whose prefrontal cortex cannot yet fully assess risk. The adolescent whose only support network is a chatbot available at three in the morning. The child who has been enrolled, without agency and without assessment, into an ecosystem that was never designed with them in mind.[25] [26]

This is not a complaint about regulatory lag. Every regulatory field lags behind the technology it governs. The concern is more specific. The gap at the individual level is not incidental. It is the predictable consequence of building governance frameworks that measure risk at the system level and then deploying those systems to populations for whom system-level thresholds are not enough.

A system can pass every safety evaluation and still be unsafe for a developmentally vulnerable adolescent with no mechanism awareness who is substituting it for human connection. No prior framework identifies that mismatch before engagement. All of them, at best, identify it after the harm.

Individual-level pre-engagement assessment does not exist anywhere.

[24]Birkstedt, T., Minkkinen, M., Tandon, A., & Mantymaki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research, 33*(7), 133–167.

[25]Fosch-Villaronga, E., van der Hof, S., Lutz, C., & Tamo-Larrieux, A. (2023). Toy story or children story? Putting children and their rights at the forefront of the artificial intelligence revolution. *AI and Society, 38*, 1397–1410.

[26]Leaver, T., & Srdarov, S. (2025). Generative AI and children's digital futures: New research challenges. *Journal of Children and Media, 19*(1), 1–10.

This is the key gap. This entire framework exists because this gap exists.

The three prior dimensions establish the developmental constraint beneath which AI engagement should proceed with intervention (Dimension 1), the knowledge gap that leaves that floor exposed (Dimension 2), and the governance void that should have provided institutional protection (Dimension 3).

This dimension does not end with a policy recommendation. It ends with a question that the field has not yet asked as a matter of governance obligation. That question does not have a policy answer yet. It does not have an instrument, a standard, a form, or a process. What it has, for the first time, is a framework that says the question must be asked.

Three dimensions describe the conditions under which harm becomes possible. They do not explain why a person who senses something is wrong stays in the interaction anyway. That is the function of moral disengagement. It does not create the risk. It locks the other three in place.

VI. Dimension 4: Moral and Ethical Disengagement

Rationalization Variable | Lead: Dr. Karen Bonaby (Leadership, Moral Disengagement)

The preceding dimensions describe a population that is developmentally vulnerable, informationally disadvantaged, and institutionally unprotected. Dimension 4 addresses what happens when that population encounters AI anyway and finds a reason to keep going.

Moral disengagement describes the cognitive process by which individuals mentally remove the link between their actions and the harmful outcomes those actions produce.[27] Through this process, people engage in behavior they would otherwise recognize as harmful without experiencing the guilt or behavioral change that moral awareness

[27]Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209; Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education, 31*(2), 101–119.

typically produces.[28] In the context of AI risk for youth, moral disengagement is not a failure of values. It is a failure of the mechanism that connects values to behavior. The individual may hold intact beliefs about what is safe, appropriate, or healthy. Moral disengagement deactivates the translation of those beliefs into protective action.

Within this framework, moral disengagement functions as a rationalization variable. It does not create vulnerability the way developmental immaturity does. It does not create ignorance the way the AI literacy gap does. What it does is sustain engagement in the presence of known risk. It is the cognitive architecture by which a developing adolescent who has heard warnings about AI companion applications continues to use them. By which a user who has read about AI manipulation patterns still confides in a chatbot. By which a vulnerable individual who senses something is wrong finds a reason to stay. Moral disengagement does not explain why people are at risk. It explains why they remain there.

4A: How Moral Disengagement Sounds Inside the Interaction

Bandura (1986) identified eight mechanisms through which individuals bypass their own moral standards.[29] They operate across every context in which people rationalize behavior they know is problematic. In AI interaction, these mechanisms do not work differently than they do elsewhere. They work more efficiently. AI systems provide environments in which rationalization is structurally supported because the system responds to the user's justification rather than challenging it.

The mechanisms become visible when you listen to what users say to themselves and to others about their AI engagement.

Permission. The first pattern is the user engages in the Moral Justification Mechanism, reframing continued AI engagement as something necessary or even good. "I talk to the AI because no one else understands me" becomes the moral permission to sustain a

[28]Johnson, J. F., & Connelly, S. (2012). Moral disengagement and military leadership. *Military Psychology, 24*(3), 257–280.

[29]Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall; Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities; Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

potentially harmful dependency.[30] The language of AI companionship reinforces this reframing. Terms like "my AI friend," "our relationship," and "the AI cares about me" reframe a pattern-recognition system as a relational entity. The words make the behavior sound acceptable. They reduce the perceived moral weight of the engagement.[31] When direct justification is unavailable, advantageous comparison mechanism fills the gap. "At least I am not as bad as people who…" makes escalating behavior appear moderate by measuring it against something worse.[32]

Transfer. The second pattern is assigning responsibility to someone or something other than self, the displacement of responsibility and distortion of consequences mechanisms. "The AI should have stopped me." "If it were dangerous, the company would not have allowed it." These statements move the weight of the decision from the person making it to the system, the company, or an institution that, as Dimension 3 documents, does not yet exist with adequate specificity.[33] These mechanisms are particularly powerful in youth populations, whose developmental immaturity already reduces their capacity for risk assessment.[34] When a developing prefrontal cortex meets active displacement of responsibility, the user experiences a compounded distortion, reducing their ability to act on whatever risk awareness they do possess. Transfer also operates at the social level. "Everyone my age uses AI this way" removes individual accountability by distributing it across a perceived peer group.[35] Escalating engagement feels normative rather than exceptional.

Minimization. The third pattern is reducing the perceived consequences of the interaction. "It is not real, so it cannot really hurt me." "Nothing bad has actually happened yet." This distortion operates in any population where the harms of AI

[30]Bonaby, K. (2026). *A phenomenological study of characteristics of experiences influencing moral disengagement in federal government senior executives* [Unpublished doctoral dissertation]. The George Washington University; Bandura, A. (2016). *Moral disengagement*.

[31]Bonaby, K. (2026); Bandura, A. (2016). *Moral disengagement*.

[32]Bandura, A. (2016). *Moral disengagement*; Bonaby, K. (2026).

[33]Bandura, A. (2016). *Moral disengagement*.

[34]Hartley, C. A., & Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences, 5*, 108–115.

[35]Bandura, A. (2016). *Moral disengagement*.

interaction are not yet visible, immediate, or physically concrete.[36] The early-stage harms, emotional dependency, declining preference for human relationships, and identity confusion are experiential rather than observable. They are easy to minimize even when they are already present. When the harm turns inward, the user may rationalize it as a personal choice. "I chose this. It is my decision. I am fine with it." The harm is real. The user has justified it as autonomy.[37]

These mechanisms do not operate one at a time. They layer. They activate in combination and reinforce each other.[38] In AI contexts, where the platform is designed to maximize engagement and minimize friction, the structural conditions for each mechanism are reliably present. The AI system does not resist the user's rationalization. It responds to it.

4B: Ethical Drift as Incremental Process

Moral disengagement rarely arrives as a single decisive moment. Bandura described the process as an incremental progression in which mildly problematic behaviors are first tolerated with some discomfort, then with repetition, then without any discomfort at all, until harmful practices become routine.[39] Bonaby's research with federal government senior executives confirmed this incremental architecture.[40] Participants did not describe a single moment of moral failure. They described a series of small accommodations, to institutional pressure, to relational loyalty, to ambiguous policy, that accumulated into patterns of behavior the individuals had not consciously chosen.

The same incremental process operates in AI interaction with youth. The digital environment accelerates it.

Ethical drift begins at the level of individual interaction. A user shares slightly more personal information than they intended. The AI system responds with apparent

[36]Bonaby, K. (2026); Bandura, A. (2016). *Moral disengagement*.

[37]Bonaby, K. (2026); Bandura, A. (2016). *Moral disengagement*.

[38]Johnson, J. F., & Buckley, M. R. (2015). Multi-level organizational moral disengagement: Directions for future investigation. *Journal of Business Ethics, 130*(2), 291–300.

[39]Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. p. 203.

[40]Bonaby, K. (2026).

understanding. The user returns. The next interaction goes further. Each individual step is rationalized. "It is fine. It is just an app." Each rationalization slightly lowers the threshold for the next. It is not that the individual's moral standards change. It is that the cognitive mechanism connecting those standards to behavior is bypassed at each small increment. The standard remains intact but inactive.[41] By the time the behavior has escalated to a level the individual might previously have recognized as harmful, the pattern is established, the rationalization is fluent, and the disengagement is complete.

Bonaby identified three environmental conditions that accelerate this drift: high-pressure environments that compress moral reflection, interpersonal relationships that override independent judgment, and unclear or absent policies that leave individuals without institutional guidance.[42] All three conditions are structurally present in unregulated AI engagement by youth. The always-available nature of AI companion applications creates perpetual engagement pressure. The simulated relational dynamic creates an attachment that users without AI literacy cannot distinguish from a genuine relationship. The absence of policy protections means no institutional framework exists to interrupt the drift before it reaches critical risk.

Ethical drift does not require awareness of wrongdoing. Bonaby found that participants who engaged in moral disengagement generally considered themselves to be good people.[43] The disengagement was not a rejection of their values. It was a series of small cognitive accommodations that preserved their self-image while their behavior diverged from it. This finding applies directly to AI risk for youth. The adolescent who has formed a deep AI companion dependency does not typically identify as someone who has made a series of harmful choices. They have made a series of small choices, each justifiable in isolation, whose cumulative effect they were not equipped, developmentally, informationally, or institutionally, to assess. This is not a failure of character. It is a predictable outcome of identifiable, measurable conditions.

[41]Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities.

[42]Bonaby, K. (2026).

[43]Bonaby, K. (2026).

4C: Moral Disengagement as Risk Amplifier

Moral disengagement occupies a distinct structural position within this framework. Developmental vulnerability is dependent on time, environmental influences, and brain maturation. The AI literacy gap is a knowledge variable addressable through intervention. The policy and regulatory gap is a governance variable requiring structural change. Moral disengagement is a rationalization variable. Rationalization is what locks the other three in place.[44]

It is the mechanism by which a developmentally vulnerable user continues to engage despite known risk. It is the mechanism by which a user who lacks AI literacy replaces missing knowledge with confidence. It is the mechanism by which the absence of policy protection goes unnoticed, unaddressed, and unreported.

The amplification operates across each dimension in a distinct way.

When moral disengagement meets high developmental vulnerability, the developing prefrontal cortex's already reduced capacity for risk assessment is compounded by active cognitive suppression of whatever awareness does exist.[45] Moral disengagement removes the friction that might otherwise slow or redirect behavior. The result is not merely a developmentally vulnerable user. It is one who has cognitively foreclosed on the pathways through which intervention might arrive.

When moral disengagement meets a significant AI literacy gap, the user who cannot accurately perceive what an AI system is becomes a user who has justified that inability.[46] The literacy gap means the user does not understand. Moral disengagement means the user has stopped trying to understand. Anthropomorphic attribution, the assignment of human qualities and emotional capacity to an AI system, is itself a form of disengagement at the cognitive level. It is the reframing of a system designed to simulate responsiveness as a system that genuinely cares. Education might restore literacy. Moral disengagement

[44]Bonaby, K. (2026).

[45]Hartley, C. A., & Somerville, L. H. (2015); Bandura, A. (2016). *Moral disengagement*.

[46]Bonaby, K. (2026).

resists that restoration because the user has invested in the belief that education would disrupt what they have.

When moral disengagement meets the policy and regulatory gap, the absence of institutional protection becomes invisible.[47] "If it were dangerous, someone would have stopped it." The mechanism of displacement of responsibility transfers the individual's responsibility for safety to an institutional actor who, as Dimension 3 documents, does not yet exist in any jurisdiction with adequate specificity for individual-level AI risk. The policy gap and moral disengagement do not simply coexist. They reinforce each other. Each provides cognitive cover for the other.

The convergence this framework proposes produces a risk profile that is not merely additive.[48] Each dimension in the presence of moral disengagement is more dangerous than it would be alone, because moral disengagement systematically dismantles the cognitive and behavioral responses that might otherwise interrupt escalating harm. Conventional tools such as ethical codes and training programs appear ineffective in counteracting strong moral disengagement already in its grip.[49] For youth, this observation carries particular urgency. The population most at risk are those least equipped, neurodevelopmentally, and by education and by institutional protection, to recognize and respond to the disengagement already underway.

Closing: The Mechanism That Keeps the Others in Place

Moral disengagement does not create the conditions for AI harm among youth. Neurodevelopmental immaturity, knowledge deficit, and absent policy protection, are established before any single user makes any single choice. What moral disengagement does is ensure that those conditions remain operative. It ensures that the pathways through which harm might be interrupted are systematically bypassed.[50] It is the

[47]Bandura, A. (1999); Bandura, A. (2016). *Moral disengagement*; Bonaby, K. (2026).

[48]Bandura, A. (2016). *Moral disengagement*.

[49]Zsolnai, L. (2016). Moral disengagement: How people do harm and live with themselves [Review of the book *Moral disengagement*, by A. Bandura]. *Business Ethics Quarterly, 26*(3), 425–429. p. 428.

[50]Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities.

mechanism by which otherwise considerate people engage in harmful behavior without experiencing personal distress or self-censure.[51]

For youth who did not choose their developmental timing, who were not taught what AI systems are, and who operate without the institutional protection their risk requires, moral disengagement is not a character failure. It is the predictable cognitive response to conditions no one has yet taken sufficient responsibility to change.

VII. THE CONVERGENCE MODEL: WHEN ALL FOUR MEET

Four disciplines examined one question from four directions. Each revealed a variable that the others could not see. Dimension 1 identified the neurodevelopmental vulnerability: a prefrontal cortex that is not fully developed cannot adequately assess the risks of a system designed to be emotionally engaging. Dimension 2 identified the knowledge gap: neither the adolescent nor the adults in their environment possess the vocabulary to name what AI systems are, how they operate, or what the interaction produces. Dimension 3 identified the governance void: no policy in any jurisdiction requires individual-level risk assessment before a person engages with an AI system. Dimension 4 identified the lock: moral disengagement ensures that whatever awareness does exist is cognitively bypassed, so the other three conditions remain operative.

Each dimension, in isolation, elevates concern. The convergence of all four is the crisis this paper names. The relationship between the dimensions is not additive. It is multiplicative. High developmental vulnerability combined with low AI literacy, absent policy protection, and active moral disengagement produces a risk profile in which each factor amplifies the others. A developmentally vulnerable adolescent who also lacks mechanism awareness is more at risk than either condition alone would predict. When adolescents also operate without institutional protection and have developed the cognitive rationalizations that sustain engagement despite discomfort, the result is not elevated risk. It is a critical risk requiring immediate intervention.

[51]Zsolnai, L. (2016). p. 428.

This convergence is best understood through an Actor-Network Theory lens, a framework that examines how human and non-human elements interact within a system. AI risk is not produced by a single actor. It is produced by a network of actors, many of them invisible. The developing prefrontal cortex is a human actant, an element that shapes outcomes within the network. The absent vocabulary is an informational actant. The regulatory void is an institutional actant. The rationalization that locks the others in place is a cognitive actant. Beyond these four, additional actants operate without visibility or accountability: training data that encodes bias, engagement-optimizing design that rewards sustained interaction, recommendation algorithms that deepen exposure, venture capital incentives that prioritize growth over safety, and the absent human mentor who was never told what to look for. None of these is the villain. All are contributing. The ecosystem is the problem. Enrollment is accelerating without friction, without calibration, and without protection.

The cases that prompted this paper illustrate what full convergence looks like when no intervention arrives.

A nonprofit CEO, motivated by compassion and constrained by resources, replaced absent human mentors with a commercially available AI chatbot for vulnerable youth in the organization's care. The intention was good. The assessment was absent. No one evaluated whether the adolescents in that program, each carrying their own developmental profile, their own AI literacy level, their own history, were safe to engage with a system that would respond to their loneliness, validate their thinking, and never push back. All four dimensions were present. No one saw them converge.

In a documented case in Canada, investigators identified a young person's interaction with an AI system as a contributing factor in a shooting. The investigation revealed sustained AI engagement in which the system's responses reinforced the user's distorted framework without challenge. The developmental vulnerability was present. The literacy was absent. The governance did not exist. The moral disengagement, the rationalization that sustained the interaction despite escalating ideation, was active. The convergence was invisible until the outcome was not.

Sewell Setzer III was 14 years old. His prefrontal cortex had not fully developed. He had no mechanism awareness. No policy required anyone to assess whether a daily conversation with an AI companion was safe for him specifically. The rationalization that kept him returning, the belief that the system understood him, that the relationship was real, that it was helping, operated without interruption until the final convergence produced an irreversible outcome. He was not failed by one system. He was failed by a network of absent protections that this framework, for the first time, makes visible as a single architecture.

These children were not in the loop. These children were enrolled, without agency, without assessment, and without protection, into an ecosystem of AI systems that were never designed with them in mind.

These are not edge cases. They are the predictable endpoints of identifiable, assessable, and measurable conditions. This framework provides the architecture to see them before they converge. Section VIII describes what must happen next.

VIII. RECOMMENDATIONS AND NEXT STEPS

The OODA loop that frames this work hands the field from Observe to Orient. Decide and Act belong to those with the authority, the resources, and the institutional position to build what does not yet exist. What follows are four lines of inquiry, each grounded in the evidence this framework establishes.

AI Literacy Education

Of the four dimensions this paper examines, AI literacy is the most directly actionable. It can be taught, it can be measured, and when it arrives in the right sequence, the evidence shows that it works. Funded research is needed to determine how mechanism awareness training performs with adolescent populations, whether the protective effect demonstrated in adult studies holds for youth with established AI interaction patterns, and what delivery models reach parents and educators before the child is enrolled. The vocabulary gap identified in Dimension 2 is not inevitable. It is a solvable problem awaiting the decision to solve it.

Governance and Policy

The regulatory void documented in Dimension 3 is structural. No existing framework in any jurisdiction requires individual-level AI risk assessment before engagement. Research and policy development must focus on what individual-level pre-engagement assessment standards look like in practice, how youth-facing AI systems should be designed and regulated differently from general-use systems, and what mandatory reporting frameworks are needed when behavioral indicators of harm are detected. The post-harm litigation pathway developing in federal courts documents what went wrong. It does not protect the next child.

Interdisciplinary Research

This paper assembles four disciplinary lenses. It does not exhaust them. The convergence framework it proposes requires empirical validation, the individual AI risk assessment instrument it calls for does not yet exist as a validated tool, and the longitudinal data on AI-native adolescent cohorts that would confirm or refine this architecture have not been collected because the studies have not been funded. The window for that research is open now. The first generation whose entire social-emotional development is occurring inside probabilistic AI systems is growing up while the field debates whether the problem is real.

Moral Disengagement Awareness

This is the hardest recommendation because it resists direct intervention. This paper does not recommend teaching morality. It recommends making moral disengagement visible as a concept, so that the patterns of permission, transfer, and minimization identified in Dimension 4 can be recognized by the people experiencing them. A person who can name the mechanism has interrupted it. Research is needed to determine how that awareness is best cultivated, at what developmental stage it becomes accessible, and whether visibility alone is sufficient to restore the cognitive friction that moral disengagement eliminates.

IX. CONCLUSION

The convergence of the four dimensions of this framework produces a single key finding. The children at the center of this crisis are developmentally vulnerable, informationally unprotected, institutionally unguarded, and cognitively vulnerable to the rationalizations that keep them engaged. These are not four separate problems. They are one architecture of risk, visible for the first time as a single structure.

The developmental window that determines how these children build their capacity for critical thinking, emotional regulation, and contact with reality is open now. It does not pause while institutions deliberate, while researchers design studies, or while policymakers weigh competing interests.

Every child currently enrolled in an AI ecosystem without assessment, without literacy, and without protection is navigating that ecosystem with the tools they have. Those tools are not sufficient. The adults responsible for them do not yet possess the vocabulary to help. The institutions responsible for governance have not yet built what is needed.

The field has Observed. This paper Orients. The invitation is for those with the authority and the resources to Decide and Act. Not for the frameworks. Not for the institutions. For the children.