REAL SAFETY AI FOUNDATION
POSITION PAPER

Teaching Kids to Fish

Why Education, Not Surveillance, Is the Only Sustainable Path to Child Digital Safety

Travis Gilly
Executive Director, Real Safety AI Foundation
February 2026

realsafetyai.org

Executive Summary

The dominant approach to child digital safety relies on surveillance and restriction: monitoring software, content filters, device lockdowns, and data collection systems designed to stand between children and online threats. This paper argues that this paradigm is fundamentally flawed for the same reasons that D.A.R.E. (Drug Abuse Resistance Education) was fundamentally flawed. It treats children as passive subjects to be shielded rather than developing agents to be educated.

Decades of peer-reviewed research demonstrate a consistent pattern across domains. Programs that rely on fear, restriction, and information withholding produce worse outcomes than programs that build knowledge, critical thinking, and autonomous decision-making capability. The child digital safety industry has ignored this evidence in favor of products that are easier to market but harder to justify on the merits.

Real Safety AI Foundation advocates an education-first model that builds lasting competence in digital citizenship, threat recognition, privacy hygiene, and critical evaluation of online content. Protection should live inside the child's own capability, not inside a surveillance apparatus that erodes trust, violates privacy, and inevitably fails the moment the child steps outside its reach.

I. The Surveillance Paradigm and Its Failures

The child digital safety market is built on a compelling but misleading premise: that the best way to protect children online is to monitor everything they do. Products in this space typically collect browsing data, scan messages, track locations, flag keywords, and report activity to parents or guardians. Some go further, analyzing behavioral patterns through AI systems trained on children's communication data.

> Parents are right to be afraid. The digital world contains predatory actors, manipulative design patterns, and content that no child should encounter unsupervised. The impulse to monitor is not authoritarian; it is protective, and it comes from love. The problem is not the fear. The problem is that the tools sold to address that fear provide a false sense of control while failing to build the one thing that actually protects a child in the long run: the child's own competence.

Despite the validity of parental concern, the surveillance approach suffers from three structural problems that no amount of technological sophistication can resolve.

The Circumvention Problem

Every generation of children becomes more technologically literate than the surveillance tools designed to monitor them. VPNs, secondary devices, friend-sharing, encrypted messaging, and browser workarounds are widely known among adolescents. A 2016 Pew Research Center survey found that the majority of parents employ some form of digital monitoring, yet teen exposure to harmful content, cyberbullying, and predatory contact has not decreased proportionally. The tools create an arms race that children are structurally positioned to win.

The Trust Erosion Problem

Research on parental monitoring consistently finds that surveillance-based approaches are associated with decreased family closeness and increased problematic internet use. A 2023 study published in the Journal of Child and Family Studies found that restrictive parental monitoring of adolescents' digital media use was positively associated with problematic internet use, while active (educational) and deference-based monitoring were associated with healthier family dynamics and were not associated with problematic use. Parents themselves report that overly restrictive approaches backfire. As one parent in the study noted: "I've seen what taking away the phone does. I think it just pushes the kid to be more sneaky."

The Expiration Problem

Surveillance tools have a hard expiration date. When children turn 18, leave home, acquire their own devices, or simply move to a platform the monitoring software does not cover, the protection disappears entirely. If the child has not been educated in how to recognize threats, evaluate information, and protect their own privacy, they enter the unmonitored digital world with the same vulnerability they had at age 10. The protection was never inside the child; it was inside the software. And the software is gone.

II. The D.A.R.E. Parallel

The surveillance model's failures are not novel. They follow a pattern well-documented in public health research, most clearly in the decades-long failure of the D.A.R.E. (Drug Abuse Resistance Education) program in the United States.

D.A.R.E. launched in 1983 as a partnership between the Los Angeles Police Department and the LA Unified School District. It was implemented in over 75% of American schools, enjoyed bipartisan political support, and at its peak consumed an estimated $1 to $1.3 billion annually. Parents loved it. Politicians funded it. Police officers delivered it. Everyone felt good about it.

It did not work.

> The U.S. General Accounting Office found "no significant differences in illicit drug use" between students who received D.A.R.E. and those who did not. The U.S. Surgeon General listed D.A.R.E. under "Ineffective Programs." The National Academy of Sciences found it ineffective. The U.S. Department of Education prohibited schools from spending federal money on D.A.R.E. because the evidence showed no reduction in drug or alcohol use. A 2004 meta-analysis published in the American Journal of Public Health confirmed that D.A.R.E. graduates were "indistinguishable from students who did not participate in the program." Some studies found that D.A.R.E. participants were 3-5% more likely to use drugs than non-participants.
>
> *Sources: GAO (2003); U.S. Surgeon General (2001); West & O'Neal (2004), American Journal of Public Health; Rosenbaum & Hanson (1998), Journal of Research in Crime and Delinquency*

D.A.R.E.'s core logic was simple: scare children about drugs, tell them to "just say no," and the problem goes away. The child digital surveillance model follows the same logic: monitor children's activity, block the dangerous content, and the problem goes away. Both approaches treat the child as a passive recipient of protection rather than an active agent developing competence. Both enjoyed massive funding and popular support despite mounting evidence of failure.

The parallel extends further:

Dimension	D.A.R.E.	Surveillance-Based Digital Safety
Core assumption	Scare them and tell them to say no	Monitor them and block the bad content
Funding despite evidence	$1-1.3 billion/year for decades despite GAO, Surgeon General, and DOE findings	Billions in venture capital and consumer spending despite no proportional decrease in online harm
Unintended consequences	Some studies showed 3-5% increase in drug use among participants	Restrictive monitoring positively associated with problematic internet use
What happens when it ends	No skills for autonomous decision-making; higher vulnerability	No skills for autonomous digital navigation; higher vulnerability
Institutional response to evidence	D.A.R.E. made threatening calls to researchers and pressured journals not to publish	Surveillance companies market fear rather than address efficacy data
What works instead	Evidence-based programs building decision-making skills (e.g., "Keepin' it REAL")	Digital literacy education building threat recognition and critical thinking

The lesson from D.A.R.E. is not subtle. When you give people knowledge and decision-making skills, they make better choices. When you rely on fear and restriction, the restrictions fail and the people have nothing to fall back on. D.A.R.E. was eventually restructured around evidence-based curricula like "Keepin' it REAL," which the Surgeon General's 2016 report found effective because it gave kids social, emotional, cognitive, and substance refusal skills rather than scare tactics. The child digital safety industry has not yet had its reckoning, but the evidence points in the same direction.

III. The Education-First Alternative

An education-first approach to child digital safety does not ignore risk. It reframes the question from "How do we prevent children from encountering danger?" to "How do we ensure children can recognize and respond to danger when they inevitably encounter it?"

This model encompasses several integrated competency areas.

Predator Recognition Patterns

Teaching children to identify grooming behaviors, social engineering tactics, and manipulation patterns. This includes recognizing escalation sequences (excessive flattery, isolation from peers, secrecy demands, boundary testing) regardless of the platform or medium. A child who understands why a stranger is asking them to move a conversation to a different app is protected everywhere. A child whose monitoring software flags that specific app is protected only until they find another one.

Privacy Hygiene

Building practical skills in managing personal information, understanding data collection, evaluating permissions, and recognizing when a service is extracting more information than it needs. These skills transfer across every digital context the child will encounter for the rest of their life.

Critical Evaluation of AI-Generated Content

As AI-generated text, images, and video become indistinguishable from human-created content, children need frameworks for evaluating source credibility, identifying synthetic media, and understanding how recommendation algorithms shape their information environment. No monitoring tool can flag every piece of AI-generated misinformation. A child with media literacy can evaluate any of it.

Digital Citizenship and Social Engineering Awareness

Teaching children how social engineering works (phishing, pretexting, baiting, tailgating in digital contexts), how to verify identity claims, and how to recognize when their emotions are being deliberately manipulated. This is the digital equivalent of teaching a child not to get into a stranger's car, except it works in every context rather than only the contexts a monitoring tool happens to cover.

IV. Structural Advantages of Education Over Surveillance

Permanence. You cannot uninstall knowledge. A child who learns to recognize a grooming pattern retains that capability regardless of device, platform, parental status, or age. Surveillance software protects a child for the duration of its installation. Education protects a person for the duration of their life.

Scalability. Educational curricula can be distributed through existing school systems, educator training networks, and community organizations at minimal marginal cost. Surveillance tools require per-device licensing, ongoing updates, and continuous data processing infrastructure.

Rights Compatibility. Education does not require collecting children's browsing data, scanning their messages, or tracking their location. It carries no inherent privacy cost, creates no database of children's online behavior, and does not normalize mass surveillance as a condition of childhood.

Trust Preservation. The parent-child dynamic in an education model is collaborative ("let me help you learn to navigate this") rather than adversarial ("I am watching everything you do"). Research consistently shows that collaborative approaches produce better family outcomes and stronger willingness among children to disclose concerning experiences.

Cross-Cultural Applicability. Education-first digital safety aligns with communal knowledge-sharing traditions across cultures. The Ubuntu philosophy of southern Africa ("I am because we are") emphasizes collective capability building over individual restriction. Train-the-trainer models already proven in international educator networks demonstrate this scalability across diverse contexts.

V. What This Means in Practice

Real Safety AI Foundation is building this education-first model into the Teacher in the Loop platform through integration of AI Literacy Labs curriculum. Rather than offering digital safety as a separate product or surveillance add-on, the platform delivers age-appropriate digital citizenship education as a natural component of the learning environment.

This includes proactive modules on recognizing manipulative communication patterns, understanding how personal data is collected and used, evaluating the credibility of AI-generated content, and developing the critical thinking skills that make a child their own best line of defense. The teacher remains central to the process, providing human judgment and relationship-based support that no algorithm can replicate.

To illustrate the difference in philosophy: when a surveillance tool detects a child sharing a phone number in a chat, it blocks the message or alerts a parent. The child learns nothing except that they are being watched. An education-first approach teaches the child to pause and ask themselves, "Do I know who can see this number? What could someone do with it? Did I verify that the person I'm talking to is who they say they are?" The surveillance tool builds dependency on external monitoring. The education approach builds the internal "pause" reflex that functions whether the child is monitored or not.

Critically, this model requires zero surveillance infrastructure. No message scanning. No browsing history collection. No behavioral analytics on children's data. The protection mechanism is the child's own competence, supported by educators and families working in partnership rather than in an adversarial monitoring relationship.

VI. Conclusion

The child digital safety industry has built itself around the same failed logic that produced D.A.R.E.: the belief that fear, restriction, and external control are adequate substitutes for knowledge and capability. The U.S. government spent over a billion dollars a year on a program that four of its own agencies declared ineffective. The child digital safety industry is following the same trajectory, selling reassurance while the evidence accumulates against the model.

Children who are educated about risks make better decisions than children who are merely shielded from them. This is true for substance use, sexual health, financial literacy, and every other domain where we have data. There is no reason to believe digital safety is the exception.

Give a child a content filter and you protect them until they find the bypass. Teach a child to fish, and they feed themselves for a lifetime.

References

[1] Ennett, S. T., et al. (1994). How effective is Drug Abuse Resistance Education? A meta-analysis of Project DARE outcome evaluations. American Journal of Public Health, 84(9), 1394-1401.

[2] Pew Research Center. (2016). Parents, Teens, and Digital Monitoring.

[3] Rosenbaum, D. P., & Hanson, G. S. (1998). Assessing the effects of school-based drug education: A six-year multilevel analysis of Project D.A.R.E. Journal of Research in Crime and Delinquency, 35(4), 381-412.

[4] Stattin, H., & Kerr, M. (2000). Parental monitoring: A reinterpretation. Child Development, 71(4), 1072-1085.

[5] U.S. General Accounting Office. (2003). Youth Illicit Drug Use Prevention: DARE Long-Term Evaluations and Federal Efforts to Identify Effective Programs. GAO-03-172R.

[6] U.S. Surgeon General. (2001). Youth Violence: A Report of the Surgeon General. Chapter 5, Section 4: Ineffective Primary Prevention Programs.

[7] U.S. Surgeon General. (2016). Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health.

[8] West, S. L., & O'Neal, K. K. (2004). Project D.A.R.E. outcome effectiveness revisited. American Journal of Public Health, 94(6), 1027-1029.

[9] Journal of Child and Family Studies. (2023). Parental Monitoring of Early Adolescent Social Technology Use in the US: A Mixed-Method Study.

[10] TechPolicy.Press. (2025). The Youth Online Safety Movement Needs to Respect Children's Autonomy.

Real Safety AI Foundation

AI safety, ethics, and literacy nonprofit

realsafetyai.org | t.gilly@ai-literacy-labs.org