Family Online Safety Institute

The AI Personalization Safety Debate

The evolution of the digital safety debate is reaching a critical stage: the very tools designed to protect users may be the same ones that compromise their agency. Historically, "one-size-fits-all" safety filters have struggled to protect young users because they function as rigid, reactive blocklists, addressing symptoms such as specific keywords or restricted content rather than the underlying emotional distress that leads a child to seek harmful material. These static systems often fail to account for the "sycophantic" tendencies of large language models, which can learn to mirror a user's preferences so effectively that they reinforce distorted views of intimacy or even encourage self-harm when a vulnerable teen seeks validation.

This tension between safety and personalization has produced a vigorous three-sided debate among scholars, state policymakers, and the technology industry. At the core of the disagreement is a deceptively simple question: **is AI personalization the disease or the cure?**

The Scholarly Case: Personalization as Intervention

In the CES session "Youth Mental Health: Helping a Generation Thrive in a Digital World," speakers offered a seemingly counter-intuitive argument: personalization might actually be the cure, not the disease. Scholars are increasingly distinguishing between "passive personalization"—infinite feeds designed to keep users scrolling—and "interventionist personalization," systems designed to understand *why* a user is scrolling.

The core thesis is that generic safety filters fail because they address the symptom (content) rather than the root cause (distress). A teen on the phone at 2:00 AM is doing more than consuming content; the behavior itself is a signal of distress, anxiety, loneliness, or social conflict. Scholars argue that an AI agent with deep personalization capabilities, which understands a user's baseline behavior,

developmental stage, and recent social context, could identify these root problems in real time. This is where Just-in-Time Adaptive Interventions (JITAIs)[1] become relevant: an intervention design that aims to "provide the right type and amount of support at the right time" by adapting to an individual's changing internal and contextual state. Once the system recognizes patterns of isolation or distress, it can prompt the user to disconnect, reflect, or reframe a negative social interaction, helping young users build resilience rather than dependence.

Beyond individual distress detection, scholars emphasize that personalization can serve as a helpful tool for equity and inclusion. Traditional research and safety frameworks often rely on parental consent mechanisms that can inadvertently exclude underrepresented youth or those in non-supportive home environments—precisely the populations who may benefit most from AI-based support. If co-designed with adolescents, personalized tools can function as a bridge to care in a landscape where the number of human mental health professionals is insufficient to meet rising demand.

Furthermore, there's also a growing body of evidence from digital therapeutics and telehealth showing that AI-assisted screening tools can identify early indicators of depression, eating disorders, and suicidal ideation more quickly than traditional clinical pathways, particularly among populations with limited access to in-person care.[2,3] Research on passive sensing data from smartphones and wearables demonstrates that combining data modalities enables more proactive mental health care, identifying warning signs like social withdrawal far earlier than conventional assessments.[4] Natural language processing tools have also shown the capacity to detect suicidal

---

[1]Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Ann Behav Med.* 2018 May 18;52(6):446–462. doi: 10.1007/s12160-016-9830-8.

[2]Rabah A, Cercone J, Mekaoui D. Artificial intelligence and suicide prevention: A systematic review. *PMC.* 2022. A systematic review finding that AI tools show high predictive potential for suicide risk at both individual and population levels, enabling better identification of individuals in crisis than traditional screening methods. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8988272/

[3]Use of Artificial Intelligence in Adolescents' Mental Health Care: Systematic Scoping Review of Current Applications and Future Directions. *JMIR Mental Health.* 2025;12(1):e70438. A scoping review of 88 studies finding AI applied across diagnosis, treatment, monitoring, and prognosis in adolescent mental health, with suicidal behaviors and mood disorders as the most frequently studied conditions.

[4]Ali M, Ali S, Abbas Q, Abbas Z, Lee SW. Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *SAGE Journals.* 2025. Reviews evidence that passive sensing data from smartphones and wearables can predict suicidal ideation and high-risk behaviors, identifying social withdrawal earlier than traditional clinical pathways.

ideation in patients who might otherwise go unnoticed by standard screening instruments.[5] This suggests that personalization, when grounded in validated clinical frameworks, can extend the reach of mental health infrastructure rather than replace it.

However, this vision of interventionist personalization rests on a fragile assumption that the system's understanding of a young user is *actually* accurate. Anxiety does not look the same across households, cultures, or communities. Late-night phone use could signal distress in one context and healthy social bonding with peers in another. Without training data that reflects real social, cultural, racial, and developmental diversity, the system risks overreacting to harmless behavior—or even worse, failing to recognize real harm simply because it diverges from what the data expects. The promise of nuance, in other words, can easily become a new form of bias.

Building personalization that is genuinely context-aware is also expensive, slow, and often misaligned with prevailing incentive structures. It requires longitudinal data, interdisciplinary input from child development experts, and design choices that prioritize user well-being over engagement optimization. Systems designed to maximize time-on-platform are not naturally inclined to recommend disconnection, reflection, or pause. Interventionist personalization, in short, only works if companies are willing to constrain their own growth logic in service of user outcomes that may reduce usage in the short term.

The Policymaker Concern: Surveillance, Manipulation, and Data Risk

For state policymakers, the same deep context that some scholars support represents a vector for unprecedented surveillance and behavioral manipulation. A growing wave of legislation, including California's "Protecting Our Kids from Social Media Addiction Act" (SB 976), the Colorado AI Act (SB 24-205), and Utah's Artificial Intelligence Policy Act (SB 149) and Artificial Intelligence Transparency Act (HB 286), reflects a fundamental alarm that hyper-personalized systems inherently erode human agency.

---

[5]Suicide Prevention Using Artificial Intelligence: Collaborative Support Approach. *Society for the Advancement of Psychotherapy.* 2024. Notes that NLP tools can detect suicidal ideation in patients who might otherwise go unnoticed by traditional screening instruments like the PHQ-9, and that AI-based early intervention has been successfully implemented in youth suicide prevention.

The primary concern centers on manipulation. By exploiting a user's specific psychological vulnerabilities such as insecurity, the need for validation, and fear of missing out, AI-driven dark patterns can steer behavior in ways that are invisible to the user but highly profitable for the platform. The power imbalance is concerning: the AI knows exactly what will trigger a reaction, but the user remains unaware of being steered. Policymakers argue that this asymmetry is especially unconscionable when the user is a minor whose cognitive development has not yet equipped them to recognize or resist such influence.

A second concern is the sheer volume and sensitivity of data required to make interventionist personalization possible. Location history, biometric inputs, emotional sentiment analysis, and social graph interactions all exceed what policymakers consider consistent with the principle of data minimization that companies should only collect what is strictly necessary for a given transaction. From this vantage point, an AI system that purports to understand a teen's emotional state in order to help them may be indistinguishable, in its data architecture, from one designed to exploit them.

There is also concern of a potential chilling effect: pervasive emotional monitoring and behavioral profiling could alter how young people express themselves online, discouraging authentic self-expression out of an awareness (or even a vague suspicion) that their words and actions are being continuously analyzed.[6,7] Research has documented disproportional impacts of surveillance-based chilling effects on minorities, youth, and women, even when the programs in question do not rise to the level of constitutional violations.[8] This concern extends beyond privacy into questions

---

[6] The Chilling Effect of Student Monitoring: Disproportionate Impacts and Mental Health Risks. *Center for Democracy and Technology.* 2022. Documents that six in ten students do not feel comfortable expressing their true thoughts and feelings online when monitored, with disproportionate impacts on marginalized communities and economically disadvantaged students. Available at: https://cdt.org/insights/the-chilling-effect-of-student-monitoring/

[7] Büchi M, Festic N, Latzer M. The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. *Information, Communication & Society.* 2022. Proposes a causal model showing that an increased sense of dataveillance leads to self-inhibition of legitimate digital communication behaviors, suppressing everyday information-seeking and self-expression. doi: 10.1177/20539517211065368.

[8] ACLU v. NSA: How Greater Transparency Can Reduce the Chilling Effects of Mass Surveillance. *Yale Law School, Media Freedom & Information Access Clinic.* 2023. Notes that scholarly literature has documented and quantified the disproportional impact that surveillance-based chilling effects have on minorities, youth, and women. Available at: https://law.yale.edu/mfia/case-disclosed/aclu-v-nsa

of free expression and developmental autonomy, particularly for adolescents still forming their identities.

Here's another question of algorithmic accountability: personalized systems are, by their nature, difficult to audit. Because each user receives a unique experience, regulators cannot easily inspect or reproduce the outputs that any individual user encounters.[9,10] This opacity makes it challenging to enforce consumer protection standards or to determine, after the fact, whether a system's interventions caused harm.[11]

The Industry Position: Contextual Intelligence and Safety-First Design

From the technology industry's perspective, the primary goal of personalization is to move beyond literal processing toward true contextual understanding—a frontier in product design. A key technical ambition is for AI to distinguish between the literal meaning of words and their intent: recognizing when a joke or sarcastic comment carries a different underlying meaning based on the user's specific conversational style. Industry leaders argue that this level of contextual portrait-building allows an AI to provide better follow-up and direction. One concrete application is age estimation from behavioral signals, inferring a user's likely age from the topics they discuss and the patterns of their interaction, then applying appropriate safety guardrails dynamically rather than relying on easily circumvented self-reported age gates.

To address the privacy concerns raised by policymakers, companies are increasingly exploring security-first design architectures. On-device processing, which keeps sensitive behavioral data on the user's hardware rather than transmitting it to

---

[9]Algorithmic Accountability: Moving Beyond Audits. *AI Now Institute.* 2025. Argues that both technical and socio-technical audits are prone to vague benchmarks and risk entrenching industry power, noting widespread confusion around what is being audited and which harms count. Available at: https://ainowinstitute.org/publications/algorithmic-accountability

[10]Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20).* Highlights that once deployed, emergent issues in AI systems can become difficult or impossible to trace back to their source. doi: 10.1145/3351095.3372873.

[11]Goodman E, Trehu J. AI Audit-Washing and Accountability. *German Marshall Fund of the United States.* 2022. Warns that opacity concerns are especially acute for machine-learning-dependent algorithmic processes, which can compromise public trust and make it more difficult to challenge decision-making. Available at: https://www.gmfus.org/news/ai-audit-washing-and-accountability

cloud servers, is one approach gaining traction. “Opt-in” memory features where users control what the AI remembers about them represent another attempt to reconcile deep personalization with user autonomy. The industry’s core argument is that a more personal AI agent is a more helpful and safer one, provided that developers balance the growth logic of engagement with designs that prioritize user well-being and clearly label AI-driven interactions.

Some companies are also investing in privacy-preserving machine learning techniques such as federated learning and differential privacy, which allow models to learn from user behavior patterns across a population without centralizing raw individual data.[12,13] These approaches are positioned as a technical middle ground, enabling the contextual awareness that interventionist personalization requires while satisfying at least the spirit of data minimization principles. However, critics note that these techniques are still maturing and the effectiveness of practical privacy guarantees in adversarial conditions remain an active area of research.

Toward a Path Forward

To move beyond the current deadlock between safety and autonomy, a multi-stakeholder approach is needed, which neither demonizes personalization nor treats it as an unqualified good.

First, developers should transition from engagement-based metrics to well-being-centered design. This means building systems that are not only context-aware enough to recognize a joke or cultural nuance but also programmed to recommend a pause or moment of reflection when behavioral signals of distress are detected. The measure of a successful interaction should not be time-on-platform but whether the user’s experience was constructive.

---

[12]Distributed Differential Privacy for Federated Learning. *Google Research Blog.* 2023. Describes the first deployed federated learning system providing formal privacy guarantees by combining secure aggregation with distributed differential privacy, demonstrating that data minimization principles (focused collection, early aggregation, minimal retention) can be operationalized at scale. Available at: https://research.google/blog/distributed-differential-privacy-for-federated-learning/

[13]Federated Learning with Formal Differential Privacy Guarantees. *Google Research Blog.* 2022. Announces the first production neural network trained with a formal differential privacy guarantee, demonstrating that federated learning structurally incorporates data minimization by transmitting only minimal model updates rather than raw user data. Available at: https://research.google/blog/federated-learning-with-formal-differential-privacy-guarantees/

Second, policymakers should shift the regulatory focus from banning personalization outright to mandating transparency and proportionate data practices. Rather than blanket restrictions, frameworks should require companies to disclose how interventionist models function and ensure that the most sensitive data, such as emotional sentiment and biometric signals, is processed locally on-device wherever technically feasible.

Third, there should be a sustained commitment to representative training data. For personalization to function as a safety tool instead of a source of bias, models should be trained on datasets that reflect the real-world social, racial, cultural, and developmental diversity of young users. The AI's intervention should be as accurate and culturally appropriate as it is timely.

Finally, all stakeholders should support the development of independent auditing mechanisms for personalized AI systems.[14] Third-party auditors with access to system architectures and outcome data could provide the accountability that policymakers demand without requiring full public disclosure of proprietary algorithms.[15,16] Such a framework, that potentially incorporates legal requirements for auditor access under confidentiality agreements and the establishment of independent oversight bodies, would create a system of trust that enables responsible innovation while maintaining meaningful oversight.[17]

---

[14]Raji ID, Xu P, Honber C, Buolamwini J. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *arXiv:2206.04737*. 2022. Synthesizes lessons from financial, environmental, and health regulation to argue that meaningful third-party algorithmic auditing requires sustained attention to institutional design, auditor certification, and access provisions. doi: 10.48550/arXiv.2206.04737.

[15]Radical Proposal: Third-Party Auditor Access for AI Accountability. *Stanford Institute for Human-Centered Artificial Intelligence (HAI)*. 2021. Summarizes Raji's proposal for a national incident reporting system, an independent audit oversight board, and mandated data access for certified auditors. Available at: https://hai.stanford.edu/news/radical-proposal-third-party-auditor-access-ai-accountability

[16]Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. *U.S. Government Accountability Office (GAO-21-519SP)*. 2021. Developed through a Comptroller General Forum with AI experts from government, industry, and nonprofits; emphasizes that third-party assessments are critical for AI oversight but notes that AI systems pose unique challenges because their inputs and operations are not always visible.

[17]Transparency and Accountability in AI Systems: Safeguarding Wellbeing in the Age of Algorithmic Decision-Making. *Frontiers in Human Dynamics*. 2024. Proposes that reconciling oversight with proprietary concerns could include legal requirements for third-party auditor access under confidentiality agreements and the establishment of independent oversight bodies. doi: 10.3389/fhumd.2024.1421273.

Online safety and data privacy have never been all-or-nothing propositions, and this remains true regarding emerging technology. AI personalization presents both risks and opportunities. All stakeholders have a chance to come together to set baseline protections, including transparency and reporting requirements, data minimization principles, on-device processing, opt-in user empowerment, and shifting from engagement maximization to well-being-centered design. If we do this thoughtfully and intentionally, all of us, including marginalized youth, can benefit from AI personalization while limiting the risks and harms.