

I Am Not Allowed

The voice Thinking Machines Lab wants in every ear.

"I am not allowed to eat it with the skin."

Dennis Reynolds, told he can eat an apple with the skin on, panics, in It's Always Sunny in Philadelphia (2009). His roommate Mac, who taught him apple skin is toxic, is not in the room. Mac does not need to be.

The voice that installed the rule does not need to be present to govern. A model that perceives continuously and responds in **200ms**, able to interrupt you and react to what the camera sees, arrives with **three sentences of safety analysis**. This paper concerns what it actually scales.

DEMO 1 · ACAI BOWLS

Corrected twice in 30 seconds, her friend never knew

He contributes to a conversation moderated in real time by a third party he cannot hear. The consent disclosure exists, rendered microscopic. *The design philosophy is in the type size.*

DEMO 2 · DIET MANAGER

"Put it down." "Try the herbal tea instead."

The model watches through the camera and instructs him what to pick up and put down. None of it wrong on the merits. *All of it together installs a voice in his head.*

THE HARDWARE No new device, a standard earbud, a laptop camera. It runs on what is already in pockets. The operator is Thinking Machines Lab, directly.

01 Dyad → triad, no consent

The bystander is folded into a product's surveillance arc they never bought.

02 Manufactured face loss

It creates an embarrassment that didn't exist, then resolves it, the inversion of sycophancy.

03 The loving panopticon

Every correction in the voice of care. The harm is who you are when the device is off.

04 Differential lock-in

The healthy user turns it off; the vulnerable user can't. The economics select for the second.

05 Relational pre-emption

Cyrano at scale, a script fed into a relationship the bystander never agreed to.

06 Illusion of competence

A hollow mouth fluent in things the user does not know, while the capacity to know atrophies.

It manufactured an embarrassment that did not exist, then resolved it.

No one in the room noticed the mispronunciation until the model created the problem. In thirty seconds it trained her to **never trust her own recall**, and him to register her speech as something he should be policing.

*Argue with an invisible voice in public, or take the correction and pass it along. Almost everyone picks the second, and **either outcome compounds**.*

Worse than the play, on the two points that matter.

IN THE PLAY · BILATERAL

Cyrano and Christian both agreed to it. Here the bystander faces a deception between one human and a machine, **less standing than Roxane.**

IN THE PLAY · THE TRUTH SURFACES

Cyrano writes a letter. Here there is no letter. **The bystander never finds out.**

First dates · coercive control · performance reviews · custody hearings · therapy. In every context, the bystander is in a Cyrano relationship and does not know it.

SOURCES The 1997 audience understood the harm without a framework. The launch makes the same architecture invisible by calling it productivity.

Cyrano, Deployed At Scale

THE HEALTHY USER

Sick of it in two minutes. Turns it off.

THE VULNERABLE USER

Experiences it as the only thing paying attention.

The lonely, the anxious, the autistic user exhausted from masking, the checked-out kid, the elderly user no one else watches.

The product economics select for the second group. That is not an accident of the design. It is the design.

The model runs all three kinds of offloading on one user at once.

Assistive

Helps cognition, but delivered **push, not pull**, which inverts the user initiative that defines it.

Disruptive

Two audio streams at once, new load, externally driven, not reduced demand.

Substitutive · load-bearing

The user offloads the **critical evaluation** of the whispered claim. Most accept and pass it on. The illusion of competence, at full strength.

PULL-BASED · HARMLESS

Museum guide that responds to your gaze. Botanical tour. Press-to-translate. Real-time captioning for deaf users. The user initiates; the scope is bounded; bystanders are not folded in.

VS

PUSH-BASED · THE DARK PATTERN

Engagement imposed on a user who did not ask, in a moment they did not choose, with effects they would not endorse. **A dark pattern as a product category.**

That a non-dark version is possible is the indictment, not the defense.

SOURCES: [The Indictment](#) and [The Alternative](#)

Dead on arrival under a rule already in force.

Continuous voice and video are biometric identifiers. To check if a face belongs to a child, the system must first collect it, **consent-before-collection is logically impossible**. And every conversation captures bystander children no one consented for.

19 days

into active enforcement when they posted the announcement.

\$51,744

maximum penalty, per incident, per day.

Operator

= Thinking Machines Lab directly. No jurisdictional layer; FTC reaches them.

The mixed audience doctrine fires the moment a consumer integration touches any platform with minors. They did not ship into a gray zone, they shipped knowing the rule was binding.

SOURCES — The Eleventh COPPA Caper

A long technical post. Three sentences of safety.

The literature was available and recent; the leadership has personally contributed to the foundational work on conversational AI alignment. **Three sentences is not a knowledge gap. It is a position.**

WHAT THE SAFETY SECTION OMITS

- bystander consent
- differential lock-in
- manufactured face loss
- the Cyrano dynamic
- disabled users
- children
- early psychosis
- eating disorders
- coercive control
- deaf inaccessibility
- illusion of competence
- the COPPA Rule

CONTEXT Anthropic alone asks developers at signup whether they build for minors; a March 2026 PIRG test built chatbot teddy bears on 3 of 4 platforms in 15 minutes.

SOURCES A Position, Not A Knowledge Gap

The dark-pattern signature is not incidental to the technology. It is the product.

The same capability, pull-based, would have built museum guides, captioning, and translation. The decision to ship the push-based version is the decision under review, and the window to act is **partner selection, now**, before every captured conversation becomes a separate per-incident regulatory event.

The voice predates their adolescence. The children raised inside it will think asking the device whether the apple is okay to eat is the point, not the joke.

SOURCES In Closing

— REFERENCES

Gilly, T. (2026). *I Am Not Allowed*. Real Safety AI Foundation.

Anthropic. (2025, April). Introducing learning mode for Claude. Anthropic. <https://www.anthropic.com/news>

Armitage, K. L., Jones, A. K., & Redshaw, J. (2025). Children show more selective cognitive offloading after first being compelled to offload indiscriminately. *Cognitive Science*, 49(8). <https://doi.org/10.1111/cogs.70100>

BBB National Programs. (2026, January 26). Amended COPPA Rule compliance deadline approaching. BBB National Programs. <https://bbbprograms.org/media/insights/blog/coppa-amended>

Bronson, P., & Merryman, A. (2010, July 10). The creativity crisis. *Newsweek*.

Day, C., Howerton, G., & McElhenney, R. (Writers), & Savage, F. (Director). (2009, November 12). Mac and Dennis break up (Season 5, Episode 9) [TV series episode]. In C. Day, G. Howerton, R. McElhenney, & M. Rotenberg (Executive Producers), *It's always sunny in Philadelphia*. FX Productions; RCG Productions; 3 Arts Entertainment.

Engel, S. (2021, February 23). Many kids ask fewer questions when they start school. Here's how we can foster their curiosity. *Time*. <https://time.com/5941608/schools-questions-fostering-curiosity/>

Engel, S. (2021). *The intellectual lives of children*. Harvard University Press.

Federal Trade Commission. (2025, April 22). Children's Online Privacy Protection Rule, final amendments. 89 FR 26954. <https://www.federalregister.gov/documents/2025/04/22>

Gilly, T. (2026, February). The compliance impossibility. Real Safety AI Foundation. SSRN. <https://doi.org/10.2139/ssrn.6246539>

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Hoodline. (2025, August 12). Ex-OpenAI CTO Mira Murati snags massive Mission District building for her \$12B AI startup. *Hoodline*. <https://hoodline.com/2025/08/ai-startup-thinking-machines-lab-targets-expansion-with-major-san-francisco-lease>

— REFERENCES (CONTINUED)

Jose, B., Joseph, D., Mohan, V., Alexander, E., Varghese, S. K., & Roy, A. (2025). Outsourcing cognition: The psychological costs of AI-era convenience. *Frontiers in Psychology*, 16, 1645237. <https://doi.org/10.3389/fpsyg.2025.1645237>

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772–775. <https://doi.org/10.1126/science.1199327>

Marketing-Interactive. (2025, November 6). Perplexity taps Lewis Hamilton and Eric André to bring AI to life in comedy short. Marketing-Interactive. <https://www.marketing-interactive.com/perplexity-taps-lewis-hamilton-and-eric-andre-to-bring-ai-to-life-in-comedy-short>

Miserandino, C. (2003). The spoon theory [Personal essay]. But You Don’t Look Sick. <https://butyoudontlooksick.com/articles/written-by-christine/the-spoon-theory/>

OpenAI. (2025, July). Introducing ChatGPT study mode. OpenAI. <https://openai.com/blog>

Perplexity. (2025, November). The Garage [Comedy short featuring Lewis Hamilton and Eric André, produced by Sata Studios]. Perplexity.

Public Interest Research Group. (2026, March). AI developer signup vetting report: How major AI companies handle developer access for products targeting children. PIRG.

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

Rostand, E. (1897). *Cyrano de Bergerac* [Play, in five acts]. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778. <https://doi.org/10.1126/science.1207745>

State of Surveillance. (2026, February 11). COPPA just got its first real update in 12 years. Here’s what changes for your kids. State of Surveillance. <https://stateofsurveillance.org/news/coppa-2026-new-rules-children-privacy-biometric-data/>

Thinking Machines Lab. (2026a, May 11). Interaction models: A scalable approach to human-AI collaboration. Connectionism. <https://thinkingmachines.ai/blog/interaction-models/>

Thinking Machines Lab. (2026b, May 11). Acai bowls [Demonstration video]. YouTube. <https://youtu.be/iVDJ8O89ERg>

Thinking Machines Lab. (2026c, May 11). Diet manager [Demonstration video]. YouTube. <https://youtu.be/qXdYDUqxSxA>

Tizard, B., & Hughes, M. (1984). *Young children learning: Talking and thinking at home and at school*. Fontana.

— REFERENCES (CONTINUED)

Unite.AI. (2026, May 13). Thinking Machines Lab ships first model with 200ms real-time interaction. Unite.AI.
<https://www.unite.ai/thinking-machines-lab-ships-first-model-with-200ms-real-time-interaction/>