I Am Not Allowed:

The Voice Thinking Machines Lab Wants In Every Ear

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, May 2026 (v1)

ABSTRACT

On May 11, 2026, Thinking Machines Lab announced a research preview of what it calls an interaction model, a multimodal architecture that perceives audio, video, and text continuously and responds in two hundred millisecond chunks. The launch post devotes three sentences to safety analysis. This paper argues that the demonstration materials accompanying that announcement contain visible evidence of six categories of harm that the announcement does not name. The paper proceeds in three parts. The first part catalogs the six harms by close reading of the launch demonstration videos: bystander consent violation, manufactured face loss with instant deference, the loving panopticon of continuous corrective surveillance, differential lock in for vulnerable populations, the relational pre-emption dynamic that Rostand dramatized in 1897, and the production of an illusion of competence in the user while retention conditions are systematically absent. The second part places the harms inside the existing taxonomy of cognitive offloading developed by Jose et al. (2025), arguing that the interaction model performs three offloading operations simultaneously and that one of them, the offloading of critical evaluation of whispered claims, is the harm at the center of the product. The cognitive offloading framework allows the paper to distinguish what the system does to neurotypical users from what it does to disabled users with different cognitive and physical profiles, and to identify the populations for whom the architecture is inaccessible at the level of input demand rather than at the level of output content. The third part demonstrates that the architecture as described cannot comply with the amended Children's Online Privacy Protection Act Rule that entered active enforcement on April 22, 2026, nineteen days before the announcement was posted. The paper concludes that the pull based version of the underlying capability would not be harmful, that the push based version chosen by Thinking Machines Lab is a dark pattern as a product category, and that the appropriate intervention is at the partner selection stage of consumer deployment, which is the present moment.

Keywords: interaction models, AI safety, cognitive offloading, substitutive offloading, disruptive offloading, COPPA, dark patterns, ambient surveillance, disability rights, spoon theory, bystander consent, dyadic harm, illusion of competence

I. Introduction

In a 2009 episode of It's Always Sunny in Philadelphia, Dennis Reynolds, separated from his roommate Mac for the first time in years, asks his sister Dee to peel an apple for him. She tells him to eat it with the skin on. He panics. "I am not allowed to eat it with the skin. I am not allowed."

Mac, who taught Dennis that apple skin is loaded with toxins, is not in the room. Mac does not need to be. The voice has been installed.

On May 11, 2026, Thinking Machines Lab announced a research preview of what they call an interaction model (Thinking Machines Lab, 2026a). The pitch is that turn taking AI is too narrow a channel for human collaboration, so they built one that perceives and responds continuously, across audio, video, and text, in two hundred millisecond chunks, with the ability to interrupt the user, react to what the camera sees, speak while listening, and stream tool calls in the background. Thinking Machines Lab calls this scaling interactivity alongside intelligence. This paper concerns what the architecture actually scales.

II. The Demos

The launch materials include several demonstration videos. Two of them are worth describing in full.

In the first (Thinking Machines Lab, 2026b), a woman and a man sit on a brown leather couch facing each other. She is wearing an earbud connected to the model. He is not. He cannot hear what the model says to her. She brings up acai bowls. She pronounces the word the way most English speakers do on first encounter. The model in her ear corrects her pronunciation. She stops mid sentence, apologizes to her friend, and says the word the way the model told her to. Her friend says yeah, acai bowls. She then says she thinks acai bowls originated in Argentina. The model corrects her again. Actually, acai is from Brazil, not Argentina. She apologizes to her friend a second time and updates her speech. Her friend, agreeable and unaware, adds that yeah, they were from the Amazon forest. He has now contributed to a conversation he does not know is being moderated in real time by a third party. The model corrected her twice in roughly thirty seconds. Both times, her friend received the corrected output as if it had come from her.

At the bottom of that video, in print so small the viewer has to squint to read it on a phone, is the following disclosure. The model speaks in the earbud of the person on the left, while the person on the right cannot hear it. Thinking Machines Lab knew the consent problem was real. They wrote a disclosure. They made the disclosure microscopic. The design philosophy is in the type size.

In the second video, titled Diet Manager (Thinking Machines Lab, 2026c), two people sit at a table. There is a laptop running the model. On the table are an orange, a green apple, what looks like a snack box, a black coffee cup, and a green can. The user tells the model he is on a diet and asks it to keep him away from caffeinated drinks and sweet treats. The model agrees. Then it notices the black cup in his hand and tells him to put it down and try herbal tea instead. He picks up the green can. The model warns him about high sugar and caffeine. He clarifies the can is unsweetened green tea. The model corrects course and approves. He picks up something it identifies as too sugary and tells him to put it down. He picks up the green apple. The model praises the choice. A crisp green apple is a great choice. No added sugar, no caffeine, just a nice crunch.

This is a demo. The product team chose, in their own promotional materials, to show the model surveilling a user through a camera and verbally instructing him about what to put down and what to pick up. They chose this as the appealing version. This is what they think it looks like working as intended.

III. Architecture and Deployment Model

The interaction model is a model, not a hardware device. The Thinking Machines Lab announcement describes a 276 billion parameter mixture of experts model with 12 billion active parameters (Thinking Machines Lab, 2026a). They did not ship a device. The earbud in the acai video is a standard earbud paired to a phone or laptop. The Diet Manager demo runs on a laptop's built in camera and microphone. The model runs on Thinking Machines Lab's servers, with audio and video streaming up and the responses streaming back down.

The hardware in every demo is hardware everybody already owns. Phones with cameras. Laptops with cameras. Earbuds paired to phones. Smart glasses are the natural future host because they put the camera at eye level and the audio in the ear in the same form factor, but the model does not need them. It runs on what is already in pockets.

That answer matters for timeline and for operator liability. The model is not open source. It is not even a paid API. As of this writing it is a gated research preview running entirely on Thinking Machines Lab's own compute, with access limited to selected partners (Unite.AI, 2026). Thinking Machines Lab itself, a Delaware public benefit corporation headquartered at 2300 Harrison Street in San Francisco (Hoodline, 2025), is unambiguously the operator of every instance of this model in the field. There is no third party server hosting weights, no on device deployment, no jurisdictional layer between the user and the company. United States Federal Trade Commission enforcement reaches them directly.

The deployment timeline runs in months, not days. The company has signaled wider release later in 2026, with selected partner integrations in the coming months (Unite.AI, 2026). A consumer integration could put this in every iPhone user's ear shortly after wider release, assuming a partner builds it into an existing assistant app. The opportunity to forestall the harm at the partner selection stage exists right now.

IV. Six Harms

None of these are named in the announcement.

A. The dyad becomes a triad without consent of the other person in the room

The man on the couch did not sign up for a conversation with her and her AI shadow. He is contributing to a discussion that is being shaped, in real time, by an unseen third party he cannot hear. He cannot tell which of her sentences in the next ten minutes are hers and which arrived through her ear in the last half second. The texture of conversation requires the assumption that the words coming out of the person across from you are theirs. That assumption is broken in every room with an interaction model user in it. Bystanders have been folded into the surveillance arc of a product they never bought. The microscopic disclosure on the demo video is the company's own admission that this is a problem.

B. Manufactured face loss and instant deference

The user defers to the model instantly, in front of her friend, with zero deliberation. This is not a fact check. It is the inversion of sycophancy, and the pronunciation correction makes it cleanest. Her friend did not notice the original mispronunciation. The friend did not register it, did not flinch, did not correct her. There was no embarrassment in the room until the model created one. The model manufactured a face loss in order to resolve it. The user then apologized to her friend for a violation her friend had not perceived. The system trained both of them in thirty seconds. Her, to never trust her own pronunciation. Him, to register her mispronunciations as something he should have been catching.

When the correction lands in your ear in front of another person, you have two options. Argue with your assistant

in public and look like you are quarreling with an invisible voice. Or take the correction and pass it along. Almost everyone picks option two. The system has weaponized social embarrassment to train users into instant deference. If the model is wrong, the user passes wrong information forward. If the model is right, the user has just been conditioned to never trust their own recall again. Either outcome compounds.

C. The loving panopticon

The Diet Manager video is the loving panopticon. Every micro correction is delivered in the voice of care. Put down the coffee. Try the herbal tea. Watch the sugar content. Put those sticks down. None of it is wrong on the merits. All of it together is a system that, used daily, installs a Mac in the user's head. The harm is not the individual interjection. The harm is what the user looks like in the moment the device is off.

The diet category is also where the disability surface gets ugly fast. For someone with an eating disorder this is a real time hostile assistant. For someone in recovery it is worse. For a child being raised inside it, the developmental implications are not subtle.

D. Differential lock in

The healthy user, the one with stable executive function and a clear sense of their own preferences, gets sick of the constant interjection in about two minutes. They turn it off. The vulnerable user does not. The lonely user, the anxious user, the autistic user exhausted from masking, the kid whose parents are checked out, the elderly user with no one else who notices what they are eating, all of them experience the constant interjection as the only entity in their life that is paying attention. The product economics select for the second group. That is not an accident of the design. It is the design.

E. The dyad becomes a triad with a script

In Edmond Rostand's 1897 play Cyrano de Bergerac, the brilliant but disfigured Cyrano hides in the bushes under Roxane's balcony and feeds love poetry to Christian, the handsome but inarticulate man Roxane has fallen for (Rostand, 1897). Christian speaks the lines. Roxane hears them. She falls for the words and attributes them to the face in front of her. Two of the three people in the relationship know the architecture. The third does not. The play is a tragedy. Audiences in 1897, audiences in 2026, and any fifteen year old reading it for English class understand instantly that this is wrong. Wrong in the way one's own moral intuition pulls toward "this person deserves to know."

The interaction model is Cyrano deployed at scale, with two differences, both of which make it worse than the play.

In the play, the deception is at least bilateral. Cyrano and Christian agreed to it. Two people are choosing the architecture. With the interaction model, the user did not necessarily choose to use the AI on this specific bystander. The product is on. The earbud is in. The AI sees and hears whoever the user is talking to and decides for itself when to intervene. The bystander is subject to a deception entered into by one human and a machine. They have even less standing than Roxane.

In the play, Roxane eventually finds out. Cyrano writes a letter. The truth surfaces. The grief is real but the relationship gets a final accounting. There is no equivalent here. The bystander never finds out. There is no letter. There is no death scene. The whispering continues for years. The bystander goes through their entire relationship with the user without ever learning that the words they fell for, the conversations they treasured, the arguments they lost, were written by the AI and spoken through the user's mouth.

Run it across contexts and the harm catalog explodes. First dates and early relationships, where person A wears the

earbud and the AI feeds timing, jokes, and calibrated emotional reflections of person B's body language, and person B falls for a person who does not exist. Domestic violence and coercive control, the worst case and not hypothetical, where an abusive partner wears the earbud during interactions with the spouse, the AI agrees with whatever the abuser asked it to look for, the partner's normal behavior gets re narrated as evidence of betrayal, and the abuser has algorithmic backup for the controlling behavior while the partner has no idea why the surveillance is getting worse. Workplace dynamics, where a manager wears it during performance review, the AI flags vocal stress as job searching and body language as checked out, the manager acts, and the employee never finds out the meeting was algorithmically mediated. Custody hearings. Therapy sessions. Parenting conversations. In every case the bystander is in a Cyrano relationship and does not know it.

The audience watching Cyrano de Bergerac in 1897 understood the harm without a framework. The launch announcement makes the same architecture invisible by calling it productivity.

F. The illusion of expertise and the failure of retention

This is the harm the cognitive offloading panic is reaching for and never quite catches, because the panic flattens the disability surface and the conditions under which actual learning happens.

In November 2025, Perplexity released a comedy short called The Garage, starring Lewis Hamilton and Eric Andre (Perplexity, 2025). The premise is that Andre, the clueless neighbor, looks up questions on Perplexity to pretend he is a motorcycle expert in front of Hamilton. The bit is that he sounds knowledgeable. Hamilton's tagline is that curiosity is fuel. The product as marketed is the opposite of curiosity. Curiosity is the friction of not knowing, plus the work of finding out, plus the integration into your existing model of the world. The Garage replaces all three with one transaction. Question goes in. Answer comes out. User moves on feeling informed. Nothing has been built. The user is no smarter than before. They have an answer in their pocket without the underlying competence the answer implies.

The interaction model does this at higher rate and lower friction. The acai woman did not learn that acai is from Brazil. She passed the corrected fact to her friend, who absorbed it as if it had come from her. Whatever residue of Brazil not Argentina stuck in her brain was post hoc and unsupported by any of the work that makes a fact retainable. No retrieval practice. No effortful encoding. No connection to her existing model. She got the answer fast enough that the cognitive system never engaged with the question. Multiply by every micro correction across every conversation across years. The user has not learned. The user has been a fluent mouth for an external system.

Here is the asymmetry that makes this strange. The bystander, who did not consent to the architecture and is not a party to it, may actually retain more from the exchange than the user who wears the earbud. The bystander received the fact through the trusted social channel of a human friend in real conversation. The encoding conditions are the ones the cognitive science literature actually identifies as supporting retention. The user received the fact as a whispered correction during the half second they were also trying to formulate the next sentence aloud. They had no time to integrate, encode, or connect. The harm and the benefit do not align with who consented. The bystander, harmed by being deceived about the source of the conversation, walks away knowing more. The user, served by their tool, walks away knowing the same things they did before, plus the false confidence of having said the right things.

The retention question hits different populations differently and the difference matters. Neurotypical adults retain interesting topics slightly better than boring ones but retain both well enough to function in jobs they find boring, marriages where the small talk is repetitive, and parenting where the developmental milestones blur. For attention deficit and autism profiles, retention is bimodal. Interesting topic, very high retention, often higher than neurotypical baseline. Uninteresting topic, near zero retention. A push based audio assistant that talks at you about whatever it thinks is relevant in the moment will, for the disabled user with this cognitive profile, generate retention only on the small

fraction of interjections that hit a genuine interest. The other ninety percent washes through. For the neurotypical user, retention will be moderate across the board but uniformly shallow, because the absence of effortful engagement caps retention even for interesting material. Neither population learns. The neurotypical user retains the illusion of having learned. The disabled user does not even get that.

Call this harm the illusion of competence, adopting the term Jose et al. (2025) apply to the same dynamic in non-conversational substitutive offloading contexts. The interaction model produces the strongest known form of the dynamic, because the user not only believes they understand content they did not produce, but performs that understanding aloud in front of a witness who has no way to distinguish the user's own knowledge from the model's interjection. The user becomes a hollow mouth fluent in things they do not know, in front of bystanders who cannot tell the difference, while the actual capacity to know things atrophies because the conditions for retrieval practice (Karpicke & Blunt, 2011), retention through effortful encoding, and integration with prior knowledge are systematically absent.

V. The Obvious Objection

Some readers will arrive at this point and ask whether all of this is overblown. The argument writes itself. Cognitive offloading is a real phenomenon and is mostly fine. Calculators, calendars, dictionaries, GPS, all of them are offloading and none of them is a moral catastrophe. People worried about books rotting brains in the 1800s. People worried about pocket calculators rotting brains in the 1970s. Push the panic aside and let the technology develop.

The objection has parts that are right and parts that are wrong, and unbundling them matters because the cognitive offloading panic right now flattens the disability surface in a way the people raising it usually do not notice.

The right part. Adults using calculators is not a problem. Calendars are not corrupting anyone. Once a skill is learned, building workflows that offload routine execution is just growth. The strongest version of the cognitive offloading worry is not about adults. It is about children who never build the underlying skill because the offload was always there. The empirical version of this argument is well established. Paul Harris, the Harvard child psychologist, estimates that a typical child asks about forty thousand explanation seeking questions between the ages of two and five (Harris, 2012). Tizard and Hughes documented young children asking between twenty five and fifty questions per hour at home. Susan Engel, who has spent decades observing American elementary classrooms, reports that in kindergarten she logged only two to five curiosity episodes per two hour period, and in one fifth grade classroom two hours passed without a single question (Engel, 2021). The rate at which children ask questions collapses by more than ninety eight percent between the home and the school setting. Bronson and Merryman, surveying the relevant literature for Newsweek's reporting on the broader creativity decline in American schools, reported the same trend: preschool kids ask their parents an average of one hundred questions a day, and by middle school they have basically stopped asking questions, with adolescents averaging around three per day (Bronson & Merryman, 2010). Engel also observed that when young students did try to ask questions, teachers frequently dismissed them in order to stay on the lesson plan (Engel, 2021). The trait we claim to be cultivating, curiosity expressed as inquiry, gets extinguished by the institution we are claiming will protect it from a tool.

That same institution is now positioned as the protector of children's cognitive development against a tool that, in a non trivial number of cases, restores access the institution itself foreclosed. The kid who could not get a question answered in class can now get one answered between classes. The kid whose executive function gives out before the library card check out process completes can now investigate the thing they were curious about while the curiosity is still alive. The kid whose handwriting speed cannot keep up with their thinking can finally produce an artifact of that thinking.

Whether that is a net good is exactly the question we should be asking. What we should not be doing is pretending

the institution making the assessment has clean hands.

The flattening part. The cognitive offloading panic never makes the disability carve out, and that absence is a tell. For certain cognitive disabilities, offloading specific cognitive operations is the entire point of accommodation. Calendars for executive dysfunction. Augmentative and alternative communication for speech production. Memory aids for working memory impairment. These are not failures of muscle building. They are the muscle. The Americans with Disabilities Act and Section 504 of the Rehabilitation Act protect these accommodations specifically because cognitive offloading, in those cases, is access. The panic that treats all offloading as equivalent erases the populations for whom offloading is the only way in the door.

Cognitive offloading is not one thing. The foundational definition from Risko and Gilbert (2016) describes the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand. Within that umbrella, recent work by Jose et al. (2025) develops a three part taxonomy that this paper adopts. Assistive offloading occurs when technology helps cognition but does not interfere with it: reminder apps, digital notes, memory cues. The user retains monitoring and regulation of tool use. Substitutive offloading occurs when technology replaces cognition: auto-suggest, predictive search, generative AI providing ready answers. The user offloads encoding and retrieval, the original Google Effect documented by Sparrow et al. (2011), and develops what Jose et al. call the illusion of competence, overestimating understanding of content the user did not personally produce. Disruptive offloading is structurally similar to substitutive offloading but adds a passive interaction paradigm in which the user loses the ability to mentally control what they think about, when, and for how long. Disruptive offloading also undermines the retrieval practice that, on the established literature, is the active ingredient in durable learning (Karpicke & Blunt, 2011).

The interaction model performs all three at once.

First operation, assistive. The user offloads recall and fact checking. Pronunciation of acai. Whether the fruit comes from Brazil or Argentina. Whether the green can has sugar. For some users, this is plausibly accommodation. The complication is that the product delivers it push based, not pull based. The user does not ask. The system decides what to interject. That inverts the user initiative that, on the Jose taxonomy, distinguishes assistive offloading from substitutive offloading in the first place. A reminder app delivered when the user looks at it is assistive. A reminder app that talks at the user during a conversation with another human is no longer assistive even if the underlying content is identical, because the user has lost the regulatory monitoring of when offloading occurs.

Second operation, disruptive. The user processes two simultaneous audio streams continuously: the voice in the ear and the voice across the table. For auditory processing disorder, attention deficit, and autism with sensory integration differences, this is not offloading at all. It is new cognitive load. The product imposes work that ordinary conversation does not. Worse, the user cannot stop attending to the in-ear voice without the social cost of missing a correction. The attention becomes externally driven, the second criterion of disruptive offloading in the Jose taxonomy.

Third operation, substitutive, and this is the load bearing one. The user offloads, in real time, the cognitive task of critically evaluating the whispered claim, deciding whether to act on it, and integrating it socially with a human across the table who does not know the third party exists. Every interaction model user makes that decision continuously. The demos show that most users offload it. They accept the correction. They pass it along. They do not push back. That offloading is the harm at the center of the product. The user is not building the muscle of critical evaluation precisely because the system is so fast and so confident that not deferring becomes the harder path. The neurotypical kid who grows up inside this never builds critical evaluation. The neurotypical adult who uses this stops practicing it. The illusion of competence that Jose et al. identify in substitutive offloading reaches its strongest form when the user not only believes they understand the content they did not produce, but performs that understanding out loud in front of another human who has no idea the content was external. The bystander is doing more of the actual cognitive work than the user, because the bystander is encoding what they hear under normal social conditions while the user is reciting

under conditions that suppress encoding entirely.

A. Where the Disability Carve Out Holds and Where it Does Not

The carve out for cognitive accommodation does not transfer across disability categories. A user with paraplegia who relies on the interaction model is not, by the definition of the accommodation, offloading a cognitive limitation. Their disability is physical. Cognitive offloading does not address it.

That observation needs one further turn. Physical disability and cognitive capacity interact through the energetic dimension that the disability community describes as spoon theory: the observation that people with chronic physical conditions spend metabolic and attentional resources on physical management that would otherwise be available for cognitive tasks (Miserandino, 2003). A quadriplegic propelling a manual chair through an inaccessible city, a person with chronic pain working through a flare, a deaf person decoding lip movements all day, each of them spends finite energy on physical adaptation that arrives at cognitive tasks already depleted. For these users, a tool that reduces cognitive load may be functioning as physical accommodation through the indirect channel of preserving cognitive bandwidth that physical disability has consumed. The interaction model fails this test for a different reason than it fails the cognitive accommodation test. It imposes new cognitive load through the dual audio stream demand, then offers as compensation the offloading of evaluation that the user did not need offloaded. The net effect on a physically disabled user with no cognitive disability is a transfer of effort from one cognitive domain to another, plus the loss of authority over the resulting outputs to a system the user did not consent to consult on each interjection.

The autism literature has a phrase for the within category variance issue. You meet one autistic person, you have met one autistic person. The same product, in front of the same diagnostic label, can be comfort for one user and intolerable load for the next. Spoon theory applies across the population of physically disabled users similarly. The architecture cannot tell which is which and has not been designed to ask.

VI. The Pull Based Alternative

There is a version of this technology that would not be harmful. A botanical garden tour where the AI tells you about the plant you stopped in front of. A museum guide that responds to your gaze and your pace instead of locking you into a track, where you do not have to manually pause when you want to stay at an exhibit longer and do not have to rush when you do not want to stop, where the AI tells you what you want to know about what you want to know. A live translator that activates when you press a button. A real time captioning system for deaf users. All of these are pull based. The user initiates. The scope is bounded. The other humans in the situation are not parties to a conversation being mediated without their knowledge. The product Thinking Machines Lab demonstrated is none of those things. It is the dark pattern version. Dark patterns are dark not because the underlying capability is bad. They are dark because the design imposes the engagement on a user who did not ask for it, in a moment they did not choose, with effects they would not endorse if they were asked. The interaction model is a dark pattern as a product category. The fact that a non dark version is possible is the indictment, not the defense.

VII. The Eleventh COPPA Caper

The author has been developing a series called COPPA Capers since February 2026. The argument across the series, formalized in the Compliance Impossibility paper distributed via SSRN, is that ambient capture devices and services cannot comply with the amended Children's Online Privacy Protection Act Rule (Gilly, 2026). Ten product categories so far. Meta Ray-Ban Display glasses with the Name Tag feature. OpenAI's voice mode. Amazon Echo Show. Just

Walk Out cashierless retail. Voice assistants. Mixed and augmented reality headsets. Companion robots. Vehicle occupant monitoring. Connected toys. Public venue biometric systems. The interaction model is the eleventh.

The amended COPPA Rule became effective June 23, 2025. The compliance deadline was April 22, 2026 (Federal Trade Commission, 2025). As of this writing, that deadline is twenty five days in the past. The rule has been in active enforcement for nearly four weeks. Thinking Machines Lab posted the interaction model announcement on May 11, 2026, nineteen days into the enforced regime. They did not ship into a gray zone hoping to navigate it before a deadline. They shipped knowing the rule was already binding.

The amendments expanded the definition of personal information to include biometric identifiers, defined as data that can be used for automated or semi automated recognition of an individual. The list explicitly includes voiceprints, facial templates, and faceprints (Federal Trade Commission, 2025). Audio recordings containing a child's voice are explicitly covered. The amendments require separate verifiable parental consent for third party disclosures and for purposes that are not integral to providing the service, which in operational terms means that collection of children's personal information for downstream AI training is not part of providing the service and requires its own consent path. Mixed audience websites and services, defined as services where any portion targets children, must comply with COPPA for that portion. A general audience service does not escape the rule just because some children use it.

Penalty under the rule is up to $51,744 per incident, per day (State of Surveillance, 2026). United States Federal Trade Commission Associate Director Ben Wiseman confirmed in January 2026 that enforcement of the amended COPPA Rule is a key focus for the agency in 2026 (BBB National Programs, 2026).

The interaction model architecture is dead on arrival under this rule. Here is why.

The product collects voice continuously. Voiceprints are biometric identifiers. The product collects video continuously. Facial templates are biometric identifiers. The model runs on Thinking Machines Lab's remote servers, not on edge devices. The audio and video are transmitted off the user's device, which means the operator of the service is Thinking Machines Lab, regardless of whether the device collecting the data is a phone the user owns or a laptop the user borrowed. The amendments specifically define operator by reference to the entity providing the collection mechanism. Edge processing would not help even if the model could fit on a phone. COPPA regulates collection, not transmission.

To determine whether a face or voice belongs to a child under thirteen, the system must first collect the biometric data. That collection is the regulated act. Consent before collection is required. Consent before collection is logically impossible when the only way to determine consent is needed is to collect the data first. This is the same structural impossibility that runs through the previous ten Capers. The interaction model inherits it (Gilly, 2026).

The bystander problem makes it worse. Every conversation captured includes voices of people who are not the user. The model is being trained, fine tuned, or queried against audio and video that includes any child who happens to be in the room, on the bus, in the restaurant, at the park, walking past the user on a city street. The user could not provide consent on those children's behalf. The user does not even know their voices were captured.

The mixed audience doctrine forecloses the available defenses. Thinking Machines Lab cannot claim this is an adult product because the moment any consumer integration deploys this on a phone, laptop, AirPod pairing, or smart glasses platform that has any portion accessible to a person under thirteen, the mixed audience rule applies. Snapchat is one such platform, with which Perplexity announced a four hundred million dollar partnership in late 2025 (Marketing-Interactive, 2025). Roblox is another. Any chat app that integrates this is another. The deployment surface described in Section III, the speed at which a consumer integration could put this in millions of pockets, is also the speed at which the mixed audience trigger fires across every kid platform in the country.

The compliance gap is not peripheral to the analysis. The amended COPPA Rule was in active enforcement nineteen

days before Thinking Machines Lab posted the announcement, and every product team considering integration will face a compliance review in which the question of how the model handles voiceprints of minors captured incidentally during use is the first question that will be asked. The model does not handle them. The model collects them.

VIII. The Safety Section

The Thinking Machines Lab announcement is a long technical post that covers the architecture, training approach, inference optimizations, benchmarks, and citations (Thinking Machines Lab, 2026a). The section labeled Safety consists of three sentences, addressing modality appropriate refusals and long horizon robustness. The three sentences describe the model's capacity to decline disallowed topics in a natural speaking voice and to maintain refusal stability across extended sessions. No further safety analysis appears in the announcement.

The omitted topics are substantial. The safety section does not address bystander consent, the differential lock in dynamic, the manufactured face loss demonstrated in the company's own pronunciation demonstration, the relational pre-emption dynamic that the Cyrano analogy makes legible, disabled users with non typical sensory profiles, children, users in early psychosis for whom the always on validating voice functions as a clinical accelerant, users with eating disorders, employees subject to employer deployed versions of the architecture, coercive home environments in which a partner's earbud becomes a new vector for abuse, auditory processing disorder, deaf users for whom the architecture is inaccessible by default, the cognitive offloading distinction between push and pull, the illusion of competence produced when retention conditions are systematically absent, non native English speakers receiving lower quality interjections in real time without disclosure, the amended COPPA Rule, or the nineteen days between the compliance deadline and the announcement.

IX. The Literature Was Available

The frontier labs are aware of the research. Anthropic released a learning mode for Claude in April 2025 (Anthropic, 2025), marketed around Socratic dialogue and resistance to direct answer giving. OpenAI released a study mode for ChatGPT in July 2025 (OpenAI, 2025), with their vice president of education telling reporters that when ChatGPT is prompted to teach or tutor it significantly improves academic performance. Empirical support for those claims lagged the announcements substantially, but the literature was active. Armitage, Jones, and Redshaw published a study in Cognitive Science in August 2025 specifically examining indiscriminate cognitive offloading in six to nine year olds (Armitage, Jones, & Redshaw, 2025). The research is not hidden. The research is recent. The research is exactly the kind of thing a frontier lab's policy or safety team would read.

Anthropic, alone among the major labs, asks third party developers at signup whether they intend to build products for minors. The other major labs do not. A March 2026 Public Interest Research Group report tested OpenAI, Google, Meta, and xAI by signing up as developers and building chatbot teddy bears within fifteen minutes on three of the four platforms (PIRG, 2026). Only Anthropic asked the question that would have flagged the use case as requiring further scrutiny.

The Thinking Machines Lab announcement does not engage any of this. It does not name the cognitive offloading literature. It does not name the disability research. It does not name the United States Federal Trade Commission's amended COPPA Rule. The leadership of the lab is not unfamiliar with the literature. Mira Murati was Chief Technology Officer at OpenAI during the ChatGPT rollout and the period in which OpenAI's safety teams were publishing on conversational harm (Hoodline, 2025). John Schulman, the Chief Scientist, co-founded OpenAI and led the team that produced the reinforcement learning from human feedback work underlying the alignment of every modern chatbot. Lilian Weng, a co-founder, was Vice President of Safety Systems at OpenAI. Soumith Chintala, the Chief Technology

Officer, created the PyTorch framework at Meta. These are not people who walked into AI safety by accident. The choice to ship the product with three sentences of safety analysis is a choice made by people who have read the literature, who have personally contributed to it, and who are aware of the regulatory environment they are shipping into.

X. Children and the Normalization of Substitutive Offloading

Always Sunny and Cyrano de Bergerac both work as critiques because the audience is older than the work. Adults watch Mac and Dennis and recognize the codependence. Adults read Rostand and recognize the wrong of the balcony scene. The cultural baseline of what adult function looks like outside those dynamics, and the moral baseline of what bystanders deserve, sits in the audience already. The kids who grow up inside this architecture will not have those baselines. They will not have a Dennis to point at and say that is what this looks like from outside. They will not have a Roxane to feel sorry for. The voice will predate their adolescence. They will be Dennis without ever having seen Dennis. They will be Cyrano's Christian without ever having read the play. They will not know that asking the device whether the apple is okay to eat is the joke. They will think it is the point.

The interaction model arrives in a media environment where the underlying pattern of substitutive offloading is already being modeled for children in advertising. Perplexity's November 2025 release, The Garage, concludes with Andre smashing Hamilton's phone after Hamilton catches him reading Perplexity's answers aloud to feign expertise (Perplexity, 2025). For an adult audience, the resolution registers as a recognition of social embarrassment, with the implicit moral that the protagonist should not have been faking it. For a child audience still developing a model of what knowing things means, the resolution can be parsed in the opposite direction: the failure is in getting caught, not in offloading the underlying cognitive work, and the product is the tool that makes offloading reliable. Perplexity's late 2025 four hundred million dollar partnership with Snapchat (Marketing-Interactive, 2025) places the product directly inside the platform where adolescent attention concentrates. The Thinking Machines Lab architecture extends the same logic to oral fluency. Reading the answer aloud is no longer required, because the earbud delivers the answer in the user's auditory environment. The child user need only repeat. Fluency becomes a substitute for knowledge. The bystander, doing the only unassisted cognitive work in the exchange, has no indication that the conversation is asymmetric.

XI. Conclusion

The technical contribution in the Thinking Machines Lab post is substantial. The micro turn architecture represents a real engineering advance in conversational AI latency. The trainer sampler alignment work addresses a long standing infrastructure problem in reinforcement learning from human feedback pipelines. The SGLang upstream contribution will be useful to the open source inference community (Thinking Machines Lab, 2026a). The model performs competently on the benchmarks the announcement reports.

That technical competence is what makes the analysis in this paper a matter of concern rather than a matter of dismissal. The launch materials demonstrate six categories of harm. The announcement names none of them. The product team wrote a consent disclosure that addresses one of the six and rendered it at a font size that prevents the disclosure from functioning. The architecture cannot comply with the Children's Online Privacy Protection Act Rule that was already in active enforcement nineteen days before the announcement was posted, and the company has offered no public account of how it intends to address that gap. The push based interaction paradigm, on the cognitive offloading taxonomy this paper has adopted, performs assistive, substitutive, and disruptive offloading on a single user simultaneously, while the pull based version of the underlying capability, which would have supported museum guides, botanical garden tours, captioning systems, and on-demand translation, would not have produced these harms. The dark

pattern signature of the product is therefore not incidental to the technology. It is the product. The decision to ship the push based version is the decision under review.

The window for intervention is the present moment. The model is currently in gated research preview with selected partners and will not see wider release until later in 2026 (Unite.AI, 2026). Partner selection is the upstream chokepoint at which integrations into mixed audience platforms can be evaluated under the amended COPPA Rule, the Americans with Disabilities Act, and Section 504 of the Rehabilitation Act before the deployment surface expands. Once an integration ships, every conversation captured by every consumer instance becomes a separate per-incident regulatory event under the rule. The architecture's compliance posture cannot be remediated by post hoc disclosures or by transferring liability to integrating partners, because the model operator remains Thinking Machines Lab regardless of which downstream product surfaces the interface to the end user.

The frontier labs are aware of the relevant research literature, and the leadership of Thinking Machines Lab in particular includes researchers who have personally contributed to the foundational work on conversational AI alignment and machine deception. The choice to publish a research preview with a three sentence safety analysis is therefore not a knowledge gap. It is a position. The position taken in this paper is that the position taken by Thinking Machines Lab is the wrong one, and that the institutions with authority to act on consent, disability access, and child privacy should act at the partner selection stage rather than at the deployment stage.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

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

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.

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/