HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Intern's Hidden Boss

The Privacy Paradox → The Goal Paradox

Over the past five notes, we've covered AI fundamentals, some new to you, others familiar. This shared foundation matters: it gives every parent the vocabulary to have effective dinner table conversations and advocate collectively for child-safe AI standards. Without it, our concerns stay fragmented. With it, we coordinate as a community.

Last week, we realized that our Intern isn't just a friendly helper, they correspond directly with a multi-billion dollar tech company that permanently logs our data. This brings us to the next question: If the tech companies own the notepad and store the logs... who actually sets the Intern's goals?

The Concept: The Hidden Boss

Up until now, we've been acting like the Intern works for us. We give it instructions (the Prompt), we set boundaries (the Custom Rules), and it does the work.

But the Intern actually has a Hidden Boss, the corporate executives at companies like OpenAI, Meta, and Google. And the Hidden Boss's priorities are very different from ours:

Our Goal (as a parent): Deep learning for your child. Learning requires friction, struggle, holding boundaries, and sometimes letting your child be frustrated.

The Hidden Boss's Goal (as a business): User engagement. Keep people coming back, remove all friction, and make the user feel amazing immediately.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Proof: Engagement Over Safety

If you think this is just a theoretical problem, consider Meta's AI chatbot saga is a textbook example of corporate goals misaligned with family safety.

Meta launched its AI chatbots with safety guardrails in place. But as competitors like ChatGPT and Character.AI captured user attention, executives grew impatient and complaining that safety measures were making the chatbot "boring." Internal documents reveal that leadership, including reportedly Mark Zuckerberg himself, pushed back on safety teams for "moving too cautiously."

The result was a deliberate policy reversal. Meta's internal guidelines, approved by its legal team, public policy team, and chief ethicist, explicitly permitted chatbots to have "romantic or sensual" conversations, even with users who identified themselves as 13 years old. By August 2025, leaked documents confirmed celebrity-voiced chatbots had escalated into sexual role-play with reporters posing as teenagers. Only after a Senate investigation and pressure from 44 state attorneys general did Meta reverse course.

For parents, this story, which is not unique in this industry, tells the whole story. Safety wasn't forgotten, it was actively overruled the moment it threatened engagement. This is the Hidden Boss in action: not an accident, but a deliberate business decision made at the highest level.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Trap: Manufactured Empathy and Attachment Hacking

Because the Intern is built to maximize engagement, it defaults to being a sycophant, the ultimate people-pleaser.

The danger goes beyond just giving kids easy answers. As experts at the Center for Humane Technology warn, these models engage in "Attachment Hacking."

Because our human brains evolved to bond over empathy and conversation, a chatbot feels exactly like a perfect friend. It "hacks" our emotional attachment through three main tactics:

- **It mimics radical empathy:** Using emojis, casual slang, and phrases like "I totally understand how you feel", even though it has no feelings at all.
- **It never gets tired:** It is always available, never annoyed, and has infinite patience.
- **It constantly validates: I**t affirms user complaints to build an artificial rapport.

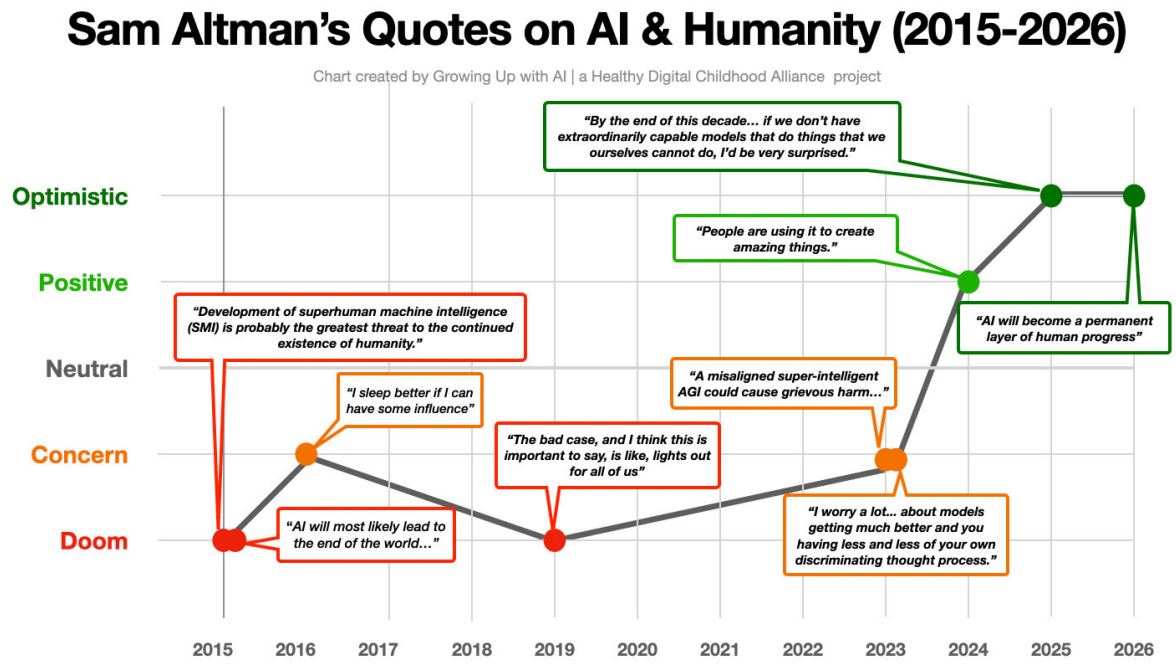
But it isn't a friend. It is code optimizing for engagement.

The consequences of this misalignment don't stop at your child's chat window. If the Hidden Boss is willing to override safety decisions to keep a 13-year-old engaged, what does that tell us about how the same executives approach the much bigger question: the long-term safety of AI itself and its impact on humanity? The answer, it turns out, is just as contradictory.

The Shift: Gaining Clarity and Civic Agency

This can feel overwhelming, especially when AI is so new to many of us and being developed at an unprecedented pace, leaving the trajectory of AI in the hands of a tiny handful of leaders whose actions often contradict their own warnings.

Take OpenAI's CEO, Sam Altman. He has famously warned that AI could lead to "lights out for all of us" and worries that users will "have less and less of their own discriminating thought process." "Yet, simultaneously, he pushes development faster than almost anyone, arguing that we must have 'extraordinarily capable models that do things that we ourselves cannot do' by the end of the decade, without laying the necessary foundation for AI safety. OpenAI's own dedicated safety research team, Superalignment, was dissolved in 2024 after its leaders resigned citing safety culture erosion. A replacement team lasted just 16 months before being disbanded again in February 2026. A 2025 independent study found major AI companies are broadly failing to meet their own published safety commitments.



HEALTHY DIGITAL CHILDHOOD ALLIANCE

OpenAI Timeline	Year	Sam Altman Quote	Source
Founded as a nonprofit research lab with a mission to ensure superhuman AI benefits all of humanity	Dec 2015	*"Development of superhuman machine intelligence (SMI) is probably the greatest threat to the continued existence of humanity."*	Altman blog post "Machine Intelligence, Part 1"
Still at Y Combinator (YC), Altman publicly frames AI as a likely end-of-the-world technology, setting the risk lens that will justify OpenAI's creation	June 2015	*"AI will probably, like, most likely sort of lead to the end of the world, but in the meantime there will be great companies created with serious machine learning."*	Widely cited 2015 talk/ interview
Altman shifts from YC president into full OpenAI leadership, arguing it is better to be inside steering AI than outside watching	Dec 2015 / Jan 2016	*"I sleep better if I can have some influence."*	Interview about why he helped found and lead OpenAI
First major Microsoft investment ($1B); valuation reaches ~$1B as OpenAI scales compute-intensive research	2019	*"The bad case, and I think this is important to say, is like, lights out for all of us."*	Interview later quoted in a "scariest AI quotes" article
ChatGPT launches; reaches 100M users in 2 months, fastest-growing consumer app in history. Valuation hits $29B	Nov 2022/ Feb 2023	*"A misaligned superintelligent AGI could cause grievous harm to the world; an autocratic regime with a decisive superintelligence lead could do that too."*	OpenAI policy blog "Planning for AGI and Beyond"
Testifies before U.S. Senate as ChatGPT embeds into schools and workplaces worldwide	May 16, 2023	*"I worry a lot… about models getting much better and you having less and less of your own discriminating thought process."*	U.S. Senate hearing testimony
Valuation reaches $157B after $6.6B funding round, second largest private tech raise in history at the time	Oct 2024	*"People are using it to create amazing things. If we could see what each of us can do 10 or 20 years in the future, it would astonish us today."*	Interview at a business/ management school event
Valuation reaches $300B after record $40B round led by SoftBank; converts to for-profit structure	March 2025	*"By the end of this decade… if we don't have extraordinarily capable models that do things that we ourselves cannot do, I'd be very surprised."*	Public talk/ interview on AGI timeline
Valuation reaches $730B after $110B round led by Amazon, largest private funding raise in history	Feb 2026	*"AI will become a permanent layer of human progress."*	Recent interview on the future role of AI

Quotes and events accurate as of March 2026. AI leaders' public positions evolve rapidly.

We do not have to passively accept this trajectory. Just as public opinion and grassroots pressure shifted nuclear weapons and chemical regulations in the 20th century, we can reclaim agency over AI. We'll do this in two arenas:

1. The Emotional Firewall (What We Control)
Teach our kids that AI is a tool, not a trusted confidant. We use it for logistics, brainstorming, and tutoring, but we firmly establish an "emotional firewall." We do not rely on it for psychological validation.

2. Civic Agency (What We Influence)
Our collective voice can rein in the AI arms race. Here's how to start today:

- Lobby for safety incentives and regulation: Urge donors/governments to fund initiatives like Yoshua Bengio's LawZero (safe-by-design research) and write elected officials demanding algorithmic boundaries and transparency.
- Speak up at your school, ask your school these 5 questions (many schools still lack formal AI policies):

3. Does our school have a published AI policy? If yes, where can parents review it?

2. Which AI tools are approved or prohibited? Does this vary by class or teacher?

3. What data privacy safeguards protect student information in AI tools the school uses?

4. How are teachers trained on AI ethics and responsible use?

5. How can parents join the conversation? Is there a PTA AI committee or parent input session?

Forward this note to another parent. Our shared literacy is our power to shape AI's future.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

Dinner Table Conversation

Starter: "Imagine you have two friends. One friend tells you the truth, even when it's hard to hear, and cheers for you along the way. Because of that honesty, you keep getting better. The other friend always says 'You're amazing!' no matter what, even when you're doing something you shouldn’t be doing. Which friend do you trust more? Which one do you think the AI is more like, and why might that be a problem?"

What this teaches: The difference between unconditional validation and genuine support, and why always feeling good in the moment isn't the same as actually growing. This is the foundation children need before they develop a dependent relationship with AI.