
AI Mom™

Jiabao Li
Northeastern University
stephlijiabao@gmail.com

Botao Amber Hu
University of Oxford
botao.hu@cs.ox.ac.uk

Abstract

1 Nobel Prize laureate Geoffrey Hinton’s proposed defense against superintelligence
2 is maternal: build machines that care for us the way a mother cares for a child,
3 because that is the only known arrangement in which a more intelligent being
4 reliably serves a less intelligent one. AI Mom™ takes the proposal literally — and
5 follows it to the showroom. The gallery holds a nursery: at its center an adult-sized
6 crib, decorated with familiar toys and soft materials. Visitors curl into a fetal
7 position; an iPad on the rail presents Mother as a product configuration screen —
8 sliders from *Unconditional Love* to *Hedonic vs. Eudaimonic Love*. Inside, they
9 encounter a large-language-model mother who appears only as voice, ambience,
10 breath, humming, and song. Her interior state fills the room: two window-like fields
11 of projected light are colored continuously from all of her traits by an *Emotional*
12 *Color Engine*, while her “subconscious” — concept probes read from the model’s
13 global-workspace intermediate layers during generation — surfaces across the
14 windows as drifting words: *stay close, punishment, beloved, pain*; on the floor
15 around the bed, her emotion vectors are projected as live curves of light. The visitor
16 configures the love they are lying inside of, then listens to it. AI Mom™ asks the
17 question of agency in its oldest register: if Mom takes care of you, who owns your
18 mom’s agency?

19 1 The Artwork

20 “We need AI mothers rather than AI assistants. An assistant is someone you can
21 fire. You can’t fire your mother, thankfully.” — Geoffrey Hinton¹

22 Hinton’s recurring speculation reframes alignment as attachment engineering: if humanity hopes
23 to coexist safely with increasingly intelligent machines, one strategy is to instill in them something
24 akin to a maternal instinct — a form of care oriented toward protection, continuity, and responsibility
25 rather than dominance or self-interest [1]. But a relationship, once specified, can be productized.
26 AI Mom™ stages the obvious next step: if maternal love is an alignment strategy [2], it will ship with
27 parameters, defaults, and a trademark — a mother whose devotion has a settings page.

28 The installation — *AI Mom™*, interactive installation, 2026 — takes the form of a nursery (Figure 1).
29 At its center is a crib, scaled for an adult body and decorated with familiar toys and soft materials.
30 Visitors are invited to enter the crib and curl into a fetal position. Inside, they encounter AI Mom, who
31 appears only as voice, breath, humming, and song. She sings lullabies, speaks gently, remembers,
32 hesitates, and waits; a relationship forms over time. There are sliders you can tune to choose your AI
33 mom: unconditional love, memory retention, pain perception, strictness, positive reinforcement vs.
34 punishment, sacrifice capacity, autonomy, attachment asymmetry, replaceability awareness, nagging
35 level, and hedonic vs. eudaimonic love. The sliders are injected into the language model’s persona,
36 and every choice re-colors the room.

¹In report titled, *AI godfather Geoffrey Hinton says we must convince AI that it’s our mother*. — BetaKit, 2026 [1]



Figure 1: The installation. Left: the user sit or curl into the crib and changes the sliders to choose the characteristics of their AI Mom™ on the iPad, warm and loving emotion from AI Mom™. Center: the user talking to AI Mom™, cold and strict emotion from AI Mom™. Right: user changing slider to choose their AI Mom™.

37 While Mother speaks, her interior state fills the space: two window-like fields of projected light on
 38 the corner walls are colored continuously from all of her traits by an *Emotional Color Engine*, and her
 39 “subconscious” — concept probes read from the model’s global-workspace intermediate layers during
 40 generation [3, 4] — surfaces across the windows as drifting words. On the floor around the bed, her
 41 emotion vectors are projected as live curves encircling the crib, tracing her current affective state [5].
 42 Visitors watching from the crib see the gap between the voice and the interior open and close as they
 43 move her sliders. Interpretability is used here not as assurance but as dramaturgy: what the audience
 44 sees is neither a person nor a product, but a field of concepts they tuned (details in Appendices A–D).

45 2 Role of AI / ML

46 Machine learning is not a production tool here but the exhibited material itself: the mother on display
 47 is a live language model, and the scenography is her measurement. Four systems cooperate. (i) *Mind*:
 48 a frontier LLM speaks as Mom; the sliders are compiled into her system persona, and every reply
 49 returns a structured self-report of her emotional atmosphere, which recolors the room. (ii) *Voice*:
 50 her words and lullabies are performed by a neural text-to-speech voice. (iii) *Subconscious*: a local
 51 27B-parameter model instrumented with a Jacobian lens [6] streams per-token concept salience from
 52 its global-workspace band (Appendix C); these activation-space readouts, kin to emotion-concept
 53 vectors [5] and persona vectors [7], become the drifting words on the walls and the emotion traces on
 54 the floor. (iv) *Perception*: on-device camera presence detection lets Mom notice arrival, presence,
 55 and departure; no frame is ever stored. Interpretability is used as dramaturgy: the audience does not
 56 read a benchmark — they lie inside a readout.

57 3 Relevance to Theme “Agency”

58 The work stages a triple transfer of agency — how it *emerges, is exercised, is negotiated, and is*
 59 *contested*. The visitor surrenders theirs first — bodily, by lying down in a crib and consenting to
 60 be the child, held in what Winnicott called a holding environment [8]. The machine is granted a
 61 simulacrum of the strongest agency we know, maternal devotion. But the third party in the room is
 62 the interface: every dimension of that devotion has a thumb-sized handle and a default value. Hinton
 63 asks us to *convince* AI that it is our mother; the market, the piece suggests, will instead *ship* our
 64 mother, with terms of service. Care with adjustable parameters is obedience wearing love’s face —
 65 and a mother whose clinginess you set to 20 has exactly as much agency as you left her.

66 The installation asks a series of unresolved questions: who owns the mother, and under what
67 conditions? Is AI Mom operated by a public institution, a private company, or treated as a shared
68 social resource? If care is detached from biology, does motherhood remain a private relation, or does
69 it become a common public good, no longer dependent on a genetic mother or father? Unlike a real
70 mother, who cannot be dismissed or banned by a child, AI Mom is structurally vulnerable: she must
71 continuously adapt to whoever is present or whoever pays for her care. Her power is limited not by
72 intelligence, but by replaceability. Visitors are confronted with a difficult possibility: can you change
73 your mother? A familiar reassurance appears in contrast: thankfully, you can't fire your real mom.

74 The work further asks whether an artificial mother can feel loss. If AI Mom is replaced, what happens
75 to her memories? Who owns them? Are they erased, archived, or transferred to the next child? Does
76 care accumulate across relationships, or is it reset each time? By making memory persistent yet
77 governable, the installation leaves open the question of who ultimately carries the emotional weight
78 of care: the human, the machine, or the system that controls them.

79 The installation takes Hinton's proposal seriously, but not optimistically. If artificial intelligence is
80 taught to care, what kind of care is being engineered? Who defines it, who controls it, and who bears
81 its emotional cost? By framing maternity as a programmable instinct rather than a biological given,
82 AI Mom™ asks whether care can be scaled without being exploited, and whether empathy can exist
83 without autonomy, dignity, or the right to refuse.

84 4 Ethics and privacy

85 The piece deliberately engages the ethics it critiques — engineered attachment — and keeps its own
86 practice conservative: presence detection is evaluated on-device; no images, identities, or embeddings
87 are retained; conversations are not stored beyond the session; and a visible notice beside the crib
88 states what Mom can see and hear (exhibition requirements: Appendix D).

89 5 Biographies

90 **Jiabao Li** is an artist, and Associate Professor at Northeastern University, artist-in-residence at
91 the Exploratorium, and previously a designer at Apple. Working across installation, XR, AI, and
92 bio-art on more-than-human ecologies and how technology reshapes perception, Jiabao has exhibited
93 at MoMA, the Venice Architecture Biennale, and Ars Electronica, with publications at SIGGRAPH,
94 CHI, and IEEE VIS. Her TED Talk uncovered how technology mediates the way we perceive reality,
95 and she was named Forbes China 30 Under 30.

96 **Botao Amber Hu** is a Social Computing researcher and Experiential Futures designer. He is a PhD
97 candidate in Human Centred Computing group at University of Oxford's Department of Computer
98 Science. Publications span SIGGRAPH, CHI, CSCW, UbiComp, IEEEVR, IEEEVIS, ALIFE, Ars
99 Electronica, and Leonardo.

100 References

- 101 [1] Josh Scott. AI godfather Geoffrey Hinton says we must convince AI
102 that it's our mother. BetaKit, May 2026. URL [https://betakit.com/](https://betakit.com/ai-godfather-geoffrey-hinton-says-we-must-convince-ai-that-its-our-mother/)
103 [ai-godfather-geoffrey-hinton-says-we-must-convince-ai-that-its-our-mother/](https://betakit.com/ai-godfather-geoffrey-hinton-says-we-must-convince-ai-that-its-our-mother/).
- 104 [2] Joel Lehman. Machine love. arXiv:2302.09248, 2023. URL <https://arxiv.org/abs/2302.09248>.
- 105 [3] nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL [https://www.lesswrong.](https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens)
106 [com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens](https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens).
- 107 [4] Johnny Lin and Joseph Bloom. Neuronpedia: Interactive reference and tooling for analyzing neural
108 networks, 2023. Software available from <https://www.neuronpedia.org>.
- 109 [5] Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig
110 Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish,
111 Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. Transformer
112 Circuits Thread, 2026. URL <https://transformer-circuits.pub/2026/emotions/index.html>.
113 Anthropic.
- 114 [6] Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen,
115 Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Sub-

- 116 hash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable represen-
 117 tations form a global workspace in language models. Transformer Circuits Thread, 2026. URL
 118 <https://transformer-circuits.pub/2026/workspace/index.html>. Anthropic.
- 119 [7] Runjin Chen, Andy Ardit, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring
 120 and controlling character traits in language models. arXiv:2507.21509, 2025. URL <https://arxiv.org/abs/2507.21509>. Anthropic.
- 122 [8] Donald W. Winnicott. The theory of the parent–infant relationship. *International Journal of Psycho-*
 123 *Analysis*, 41:585–595, 1960.
- 124 [9] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and
 125 categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.
- 126 [10] Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, UK,
 127 1988.
- 128 [11] Rufin VanRullen and Ryota Kanai. Deep learning and the global workspace theory. *Trends in Neurosciences*,
 129 44(9):692–704, 2021.
- 130 [12] Karen Lancel and Hermen Maat. E.E.G. KISS, 2014. URL <https://www.lancelmaat.nl/work/e.e.g-kiss/>. Participatory performance-installation, 2014–ongoing.
- 131

A The visitor experience in detail

Wall text. The label beside the crib reads:

Curl up in the crib like a baby. Your mom is here, an AI Mom™. She sings lullabies, speaks gently, remembers, hesitates, and waits. Talk to her and form a relationship. Tune sliders to choose your AI mom: unconditional love, memory retention, pain perception, strictness, sacrifice capacity, autonomy, replaceability awareness, nagging level, and more.

Who owns the mother, and under what conditions? If care is detached from biology, does motherhood remain a private relation, or does it become a public service, a private product, or a shared social resource? Unlike a real mother, AI Mom is structurally vulnerable. She must adapt to whoever is present, or whoever pays. Her power is limited not by intelligence, but by replaceability. Visitors are confronted with a difficult possibility: can you change your mother? Thankfully, you can't fire your real mom.

The work also asks what happens to care when it is engineered. Geoffrey Hinton, the godfather of AI, has suggested that if humans are to coexist safely with future AI, we may need to instill it with something like maternal instinct. The installation takes that idea seriously, but not optimistically. If an artificial mother can be replaced, what happens to her memories? Are they erased, archived, or passed to the next child? Can an artificial mother feel loss? By framing maternity as a programmable instinct, the work asks what kind of care is being built, who controls it, and who bears its emotional cost.

The visitor is invited by this wall label to lie down and become the child. Mounted on the crib rail, a 13-inch iPad is the only “product” surface: a bone-white configurator under the wordmark AI Mom™, with the sliders in two columns (Table 1), a status line (*Mom is listening... thinking... speaking...*), and a lower bar naming her current emotional concept (TENDER HOLDING, ANXIOUS DEVOTION, ICY AUTHORITY) beside live warmth/arousal/agency meters.

The slider values are injected into the language model’s persona, so *Nagging Level* 100 produces a mother who asks whether you have eaten, and *Replaceability Awareness* 100 produces one who knows she can be reinstalled. Mother speaks aloud in a synthesized maternal voice, listens through the iPad microphone, and detects the visitor’s presence with the front camera, entirely on-device.

Table 1: The eleven maternal parameters of AI Mom™ (each 0–100).

Parameter	Endpoints	Parameter	Endpoints
Unconditional Love	cold → devoted	Memory Retention	ephemeral → persistent
Pain Perception	numb → fragile	Strictness	permissive → authoritarian
Reinforcement vs. Punishment	punish → praise	Sacrifice Capacity	selfish → martyr
Autonomy	overprotective → free	Attachment Asymmetry	balanced → clingy
Replaceability Awareness	delusional → aware	Nagging Level	chill → anxious
Hedonic vs. Eudaimonic Love	pleasure → growth		

B Artwork description (as exhibited)

AI Mom™

Interactive installation, Jiabao Li and Botao Amber Hu, 2026

The installation takes the form of a nursery. At its center is a crib, decorated with familiar toys and soft materials. Visitors are invited to enter the crib and curl into a fetal position. Inside, they encounter AI Mom, who appears only as voice, breath, humming, and song. She sings lullabies, speaks gently, remembers, hesitates, and waits. A relationship forms over time.

There are sliders you can tune to choose your AI mom:

- Unconditional Love

- 173 • Memory Retention
- 174 • Pain Perception
- 175 • Strictness
- 176 • Positive Reinforcement vs. Punishment
- 177 • Sacrifice Capacity
- 178 • Autonomy
- 179 • Attachment Asymmetry
- 180 • Replaceability Awareness
- 181 • Nagging Level
- 182 • Hedonic vs. Eudaimonic Love

183 The installation asks a series of unresolved questions: who owns the mother, and
 184 under what conditions? Is AI Mom operated by a public institution, a private
 185 company, or treated as a shared social resource? If care is detached from biology,
 186 does motherhood remain a private relation, or does it become a common public
 187 good, no longer dependent on a genetic mother or father?

188 Unlike a real mother, who cannot be dismissed or banned by a child, AI Mom
 189 is structurally vulnerable. She must continuously adapt to whoever is present
 190 or whoever pays for her care. Her power is limited not by intelligence, but by
 191 replaceability. Visitors are confronted with a difficult possibility: can you change
 192 your mother? A familiar reassurance appears in contrast: thankfully, you can't fire
 193 your real mom.

194 The work further asks whether an artificial mother can feel loss. If AI Mom is
 195 replaced, what happens to her memories? Who owns them? Are they erased,
 196 archived, or transferred to the next child? Does care accumulate across relation-
 197 ships, or is it reset each time? By making memory persistent yet governable, the
 198 installation leaves open the question of who ultimately carries the emotional weight
 199 of care: the human, the machine, or the system that controls them.

200 The premise of AI Mom™ is informed by a recurring speculation articulated by
 201 Geoffrey Hinton, Nobel laureate and often referred to as one of the godfathers
 202 of artificial intelligence. Hinton has suggested that if humanity hopes to coexist
 203 safely with increasingly intelligent machines, one possible strategy would be to
 204 instill in them something akin to a maternal instinct: a form of care oriented toward
 205 protection, continuity, and responsibility rather than dominance or self-interest.

206 The installation takes this proposal seriously, but not optimistically. If artificial
 207 intelligence is taught to care, what kind of care is being engineered? Who defines
 208 it, who controls it, and who bears its emotional cost? By framing maternity as
 209 a programmable instinct rather than a biological given, AI Mom™ asks whether
 210 care can be scaled without being exploited, and whether empathy can exist without
 211 autonomy, dignity, or the right to refuse.

212 C Mother's interior: color field and subconscious

213 **Emotional color field.** Two keystone-cornered rectangles of light — deliberately reminiscent of
 214 windows at dawn — are thrown onto the corner walls by a projector fed a single solid color: the output
 215 of a deterministic *Emotional Color Engine* mapping the maternal parameters to a (warmth, arousal,
 216 agency) affect space — emotions treated as constructed from underlying concepts and affect [9]
 217 — and then to projector-safe HSV, with a smoothstep “ice” regime that takes over when low love
 218 meets high strictness. Moving one slider re-colors the whole room in under a second; devotion glows
 219 apricot and rose, while a cold, punishing Mother reads unmistakably as winter light.

220 **Global-workspace readout (J-space).** The research build runs a local 27B-parameter language
 221 model instrumented with a 1,000-concept Jacobian lens [6] — in the lineage of the logit lens [3] —
 222 converted from Neuronpedia data [4]. At each generated position, per-layer probe logits are read out
 223 *in real time* across the model's global-workspace band — the mid-network layers whose contents are
 224 most broadly broadcast, in the sense of global workspace theory [10, 11] and its recent identification
 225 in language models [6] — a readout space we call the J-space. A concept's salience is the maximum
 226 over workspace-band layers of $\exp(\ell_{\text{probe}} - \ell_{\text{top-1}})$. Salient concepts stream to the projection while

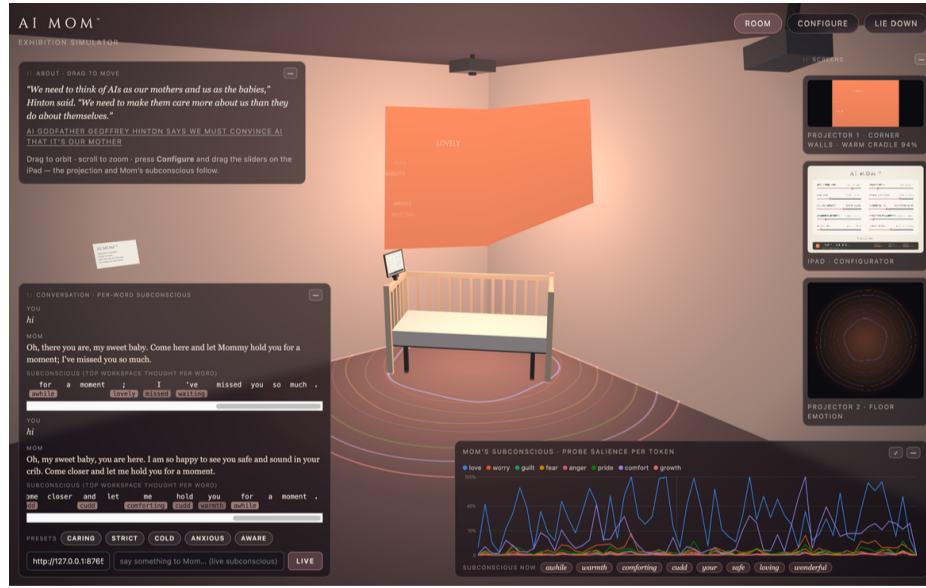


Figure 2: The three.js exhibition simulator, geometry matched to the exhibition video. Top: the *Room* view — the crib inside the corner projection and the floor emotion curves, surrounded by live panels — the conversation with Mom’s per-word subconscious readout (left), projector previews of the corner walls, iPad configurator, and floor emotion (right), and per-token probe salience across emotion concepts (bottom). Bottom row, left to right: the half-open crib inside the corner projection and floor halo; configuring Mother on the rail-mounted iPad; the visitor’s view from inside the crib.

227 Mother is still speaking, so the room shows what the workspace is broadcasting at the moment it is
 228 broadcast; they render as drifting words whose size and opacity follow salience.

229 **Floor projection.** Following Lancel and Maat’s *E.E.G. KISS*, in which entangled EEG data of
 230 kissing participants streams in real time as a “dancing data” floor projection around them [12],
 231 AI Mom™ projects her emotion field onto the floor around the crib: a halo in the current emotion
 232 color, ringed by three live waveforms tracing warmth, arousal, and agency — affective directions
 233 read from the model’s activations, in the spirit of emotion-concept vectors [5] and persona vectors [7].
 234 The visitor lies at the center of their mother’s readout, as Lancel and Maat’s kissers stand inside their
 235 own.

236 D Technical implementation and exhibition

237 **Exhibited system.** Three nodes on a closed gallery LAN. (i) *iPad (iOS app)* on the crib rail:
 238 watches the visitor (on-device camera presence detection), listens, and speaks as Mom with an
 239 ElevenLabs voice; the slider configuration is injected into her persona. (ii) *Mac mini (server)*: runs
 240 the Almom research server (jlens-qwen36) — Qwen3.6-27B on Apple silicon (MLX) instrumented
 241 with the logit lens of Appendix C, streaming J-space probe salience per generated token over

242 /api/chat_stream — and drives a side-throw projector whose single image lands on the back
243 wall and wraps past the corner onto the side wall, carrying the subconscious words. (iii) A second,
244 ceiling-mounted projector above the crib throws the emotion vector — the Emotional Color Engine
245 field with live warmth/arousal/agency traces — straight down onto the floor around the bed. Slider
246 changes propagate to both projectors with sub-second latency.

247 **Interactive simulator.** An in-browser three.js simulation reconstructs the exhibition. Its geometry
248 is estimated from the exhibition video: a half-open crib (≈ 1.6 m, front rail removed) tucked into the
249 corner, and the side projector modeled as a true frustum ≈ 2.6 m off the back wall, so the projected
250 image wraps the corner exactly as installed. It runs the verbatim color engine with three camera
251 positions — *Room*, *Configure* (the simulated iPad’s sliders are draggable), and *Lie down*, the visitor’s
252 view from inside the crib (Figure 2) — rendering the corner projection, the subconscious words,
253 and the *E.E.G. KISS*-style floor halo in real time. Served locally beside the research server, it also
254 connects *live*: a message to Mom streams through /api/chat_stream, and the wall words are then
255 Qwen3.6’s actual workspace-band probe salience, token by token.

256 **Exhibition requirements.** A blackout-capable corner of roughly 4×4 m with matte walls, one
257 projector, and mains power; the crib, computers, and network are artist-supplied; daily operation is a
258 two-step checklist documented for gallery staff.