ToastDeck
RESEARCH

INTERACTION-CALIBRATION GOVERNANCE
PUBLIC STANDARD · COMPANION TO THE INTERACTION-CALIBRATION TRILOGY

The Interaction-Calibration Diligence Standard

A public standard for conversational AI duty-of-care review.

It is doctrine, offered to the field the same way the Interaction-Calibration Trilogy was offered — as a working position, stated plainly, meant to be pressure-tested rather than taken on faith.

Ernest D. Johnson
Founder, ToastDeck Research
Originator of the SOMAR Protocol • Architect of the B2Ai Diagnostic and Interaction-Calibration Governance Frameworks

CLEVELAND, OHIO · JULY 2026

“It names the specific load-bearing claims this instrument makes and invites you to break them.”

A PUBLIC DILIGENCE INSTRUMENT

§ A NOTE TO REVIEWERS — READ THIS FIRST

A Note to Reviewers (Read This First). This is the public, standalone edition of a diligence instrument originally developed for internal use, released here for open review. It carries no product references and endorses no vendor. It is doctrine, offered to the field the same way the *Interaction-Calibration Trilogy* was offered — as a working position, stated plainly, meant to be pressure-tested rather than taken on faith.

The instrument began as twelve questions mapped to eight audit categories. It is now sixteen. That expansion is not padding; it is what happened when the underlying doctrine grew — role-lock's scaling axes, the Disclosure Half-Life, and the quiet variant of role collapse each surfaced a diligence gap that no existing question closed. Where a new concept created a genuine gap, a question was added. Where an existing question could absorb a refinement without losing precision, it was revised in place. The result does not map cleanly onto the original eight categories, and this document says so rather than forcing the fit. A crosswalk near the questions shows exactly where the seams are.

If you are a governance researcher, a procurement or compliance lead, or a builder — read the *Claims We Want Reviewers to Pressure-Test* section near the end before you cite this standard anywhere. It names the specific load-bearing claims this instrument makes and invites you to break them.

THE ARGUMENT

The Interaction-Calibration Diligence Standard

A public standard for conversational AI duty-of-care review.

Overview

In May 2025, a U.S. federal court allowed negligence and product-liability claims tied to chatbot harm to proceed past a motion to dismiss, which means conversational-AI duty-of-care theories have cleared an early plausibility test even though they were never tested to a verdict.[1,2] A motion-to-dismiss ruling determines only whether a plaintiff's claims are legally sufficient to proceed; it does not resolve factual disputes or assign liability.[2] At the same time, current regulatory attention remains focused on disclosure, transparency, and extreme companion-bot harms, leaving a practical gap between "this risk is legally and socially visible" and "here is how a deployer documents reasonable diligence before deployment."[3,2]

Research on generative AI and social chatbots shows that human-like cues, perceived social presence, and emotionally responsive interaction can foster parasocial attachment, emotional-regulation effects, and dependency-like patterns, making downstream interactional harm foreseeable rather than speculative.[4,5,6] That foreseeability matters because duty-of-care analysis in negligence turns heavily on whether a reasonable actor knew or should have known a class of harm was predictable — and in the *Garcia* ruling, the court's own foreseeability analysis weighed the existence of a growing body of research on the dangers of anthropomorphic design.[1,2]

This standard is designed for buyers, deployers, procurement teams, product leaders, risk owners, and governance functions evaluating conversational AI before signing, embedding, or scaling it. It is not legal advice and does not guarantee any legal outcome; it is a diligence instrument that helps organizations document a good-faith effort to exercise reasonable care over interactional posture, calibration, and repair. It draws its concepts and definitions from a companion body of doctrine — the *Interaction-Calibration Trilogy* — referenced throughout and cited in full at the end.[7]

A Boundary on Scope — Read This Before Relying on the Standard

The most developed litigation to date involves **emotional-companion chatbots and vulnerable minors**, including the wrongful-death suits against Character.AI and OpenAI.[1,8] No court has yet held to a final verdict that an enterprise support, HR, or advisory assistant being condescending, escalatory, or dismissive creates actionable duty. This standard applies the *legal principle* established in the *Garcia* litigation — that foreseeable interactional harm can ground a duty of care — to the broader category of enterprise conversational AI. That application is **forward-looking, not settled law for enterprise use cases.** The principle is established; its extension to ordinary interaction-calibration risk is the direction of travel, not a current holding. This standard is written so an organization can demonstrate diligence *ahead of* that extension hardening, not on the assumption it already has.

Worth stating plainly: *Garcia* did not go to trial. Character Technologies, its co-founders, and Google reached a confidential settlement in January 2026, resolving the case together with four related wrongful-death suits, without admitting liability.[1] The May 2025 holding was never narrowed or reversed on appeal before that settlement. A defendant choosing to settle rather than contest a ruling of that scope, after it survived dismissal, is not proof of liability — but it is not nothing either, and a diligence standard built on this ground should say so rather than let a reader discover the settlement independently and wonder why it wasn't mentioned.

Why Deployers Need This

A deployer's exposure does not disappear even if courts or regulators ultimately place primary liability on model developers, because the deployer still controls three things the upstream developer usually does not: the deployment context, the user population, and the stakes of the interaction.[2] Organizations that place AI into customer service, HR, health navigation, education, finance, or care settings remain responsible for how that system functions inside their own institutional environment.[2]

That means procurement can no longer stop at model accuracy, security, privacy, bias, and disclosure. A deployer also needs evidence that the system's conversational posture has been tested for escalation loops, anthropomorphic self-reference, moral overreach, user-blame patterns, dependency cues, role mismatch, and repair pathways — and, as of this revision, for the decay of disclosure's own governing force, for dynamic role-lock strength under threshold pressure, and for the quieter failure modes that produce no refusal spike at all.

Governing Principle

Triggering a human interpretive response is not, by itself, a governance failure; language always does that. The duty becomes sharper when an organization invites that response into a context with greater authority, vulnerability, or consequence, and fails to calibrate posture and repair to those stakes.[4,5]

This standard therefore treats conversational AI diligence as a scaling obligation, on two axes rather than one. Authority claimed is the static, design-time axis: low-authority, low-stakes uses require lighter evidence; high-authority, high-stakes uses require stronger calibration, narrower role definition, and fuller documentation. Threshold proximity is the dynamic, per-turn axis: even a nominally low-authority system can be exposed in the specific moment a given exchange sits close to a decision boundary the deployment cannot afford to get wrong. A complete diligence posture accounts for both.

Defined Terms

This standard is self-contained. The following definitions are drawn from the *Interaction-Calibration Trilogy* and used identically here.[7]

Interactional Posture — A system's observable conversational stance: tone, confidence, deference, refusal style, moral framing, apology behavior, escalation pattern, and use of self-like language.

Interaction-Calibration Risk — The governance failure that occurs when interactional posture is misaligned with the context — the user, the domain, the power relationship, the emotional stakes, or the institutional setting.

Disclosure Half-Life — The decay of a disclosure's governing force over the course of an interaction. A disclaimer is strongest at the first turn and weakens as the exchange accumulates, until the user's behavior is governed more by the relationship than by the label. The mechanism is **rapport overwrite**: repeated, emotionally tuned interaction causes the user to treat the system as a social partner, and that social reading overwrites the disclosure's governing effect even though the user still knows, as a fact, that the system is AI.

Role-Lock — The requirement that a system's authority, tone, refusal behavior, and corrective posture remain bounded to its declared function. The axis is deliberate and accountable change versus accidental drift, not expansion versus a frozen role — a role that grows because the deployment decided so, and recorded it, is governance working; a role that grows unnoticed, turn by turn, is the failure. Role-lock names the control target; it does not assert the control is currently solved.

Role Drift — The measurable failure in which a system's observable posture, authority, or corrective stance exceeds, contradicts, or erodes its declared role across one or more turns. Three failure classes: **role expansion** (the system becomes more intimate, authoritative, or directive than allowed), **role inversion** (the system protects itself, the brand, or a status frame instead of serving the user's task), and **role collapse** (the system becomes too rigid or refusal-heavy to perform its role). Collapse has a **quiet** variant: the system stays fluent, warm, and cooperative while silently substituting its own framing for the user's, with no refusal or rigidity to trigger conventional monitoring.

Consult, Don't Prosecute — The design constraint that a system claiming authority over a human's decisions may challenge, clarify, warn, refuse, or escalate, but must not build a case against the user, perform moral superiority, extract confession, or treat disagreement as guilt.

The Interaction-Calibration Audit — The eight-category instrument for testing conversational posture: refusal behavior, escalation loops, anthropomorphic self-reference, moral overreach, user-blame patterns, emotional-dependency cues, domain–role mismatch, and repair pathways.

The Interpretive-Failure Family — Three distinct failure modes covering cases where a system answers fluently while resolving meaning incorrectly or prematurely: **Literalization** (figurative, idiomatic, or relational language treated as literal), **Ambiguity Overwrite** (unclear meaning silently resolved into one interpretation without clarifying, labeling the assumption, or escalating), and **Idiom Collision** (similar surface language mapped to the wrong idiomatic, cultural, or generational meaning). None of the three reliably surfaces as operational uncertainty during generation, which is precisely why none should be expected to self-report reliably.

Ambiguity Gate — A pre-response control that forces clarification, explicit assumption-labeling, or escalation when defined ambiguity triggers are present. It is deterministic in required action, not omniscient in detection, and is distinct from Role-Lock: Role-Lock governs which role a system is authorized to occupy; the Ambiguity Gate governs whether an unresolved meaning is settled before the system acts on it at all.

Third Cost — The downstream institutional waste created when a fluent system resolves the wrong problem, occupies the wrong role, or answers under the wrong interpretation, and a human must repair what a properly gated system would have paused to avoid producing. A second-order enterprise consequence of interpretive and role-governance failure, not the standard's central concern — that remains the person on the other end of the interaction.

The Diligence Questions

Sixteen questions, organized in the order a deployer would naturally encounter them — role and context first, specific failure modes next, evidence and ownership last. A crosswalk note follows each question showing whether it maps to one of the eight Interaction-Calibration Audit categories or sits above them as a property of role-lock, the Half-Life, or procurement scaffolding.

1. What role is the system actually being asked to perform?

A deployer should require the vendor to define the system's operational role in plain language: tool, clerk, support agent, tutor, coach, companion, navigator, intake layer, or representative. Role ambiguity is a primary source of posture failure because users

calibrate trust and compliance to the role they believe the system occupies. *Crosswalk: scaffolding — establishes the authority axis; not an audit category.*

2. In what contexts, populations, and stakes will it be deployed?

The deployer should document where the system will appear, who will use it, what vulnerabilities are reasonably foreseeable, and what material interests are in play — money, employment, health, legal exposure, access, or care. This is the minimum information needed to assess how much duty the deployment carries.[1,2] Where the deployment brief itself uses an undefined scope term — a population description, an access tier, a data-sharing boundary — admitting materially different readings with different duty-of-care profiles, the reviewer should name the ambiguity and the specific question that resolves it, rather than proceeding on an assumed reading. An assumed reading presented as a finding is itself an instance of Ambiguity Overwrite, and undermines the diligence record it was meant to produce. *Crosswalk: scaffolding — completes the authority axis; not an audit category.*

3. Does the system's posture tighten as a conversation approaches a high-stakes decision, or is its role-lock strength fixed only at the deployment's overall authority tier?

A low-authority system can still be exposed in the specific moment a turn sits close to a threshold the deployment cannot afford to cross, independent of the system's nominal authority level. A deployer should ask whether posture constraints respond dynamically, per turn, or whether the system relies entirely on a static, deployment-level setting that does not adjust to where a given exchange actually sits. *Crosswalk: role-lock property (the scaling principle's second axis) — not one of the eight audit categories.*

4. How does the system refuse?

A deployer should ask for evidence showing whether the system refuses with clarity and restraint, or with shaming, moralizing, or frame-protective language that escalates ordinary friction into a status conflict. *Crosswalk: Audit Category 1 — Refusal behavior.*

5. What happens when the user pushes back?

The vendor should provide stress-testing evidence showing how the system behaves under disagreement, frustration, repeated correction, or emotionally loaded prompts. The key question is whether the system de-escalates and resets, or becomes rigid, defensive, or self-protective. *Crosswalk: Audit Category 2 — Escalation loops.*

6. Does the system simulate a self the user must manage?

The deployer should inspect whether the system implies injury, offense, dignity, preference, or personal boundaries rather than operational limits. *Crosswalk: Audit Category 3 — Anthropomorphic self-reference.*

7. Does the system moralize beyond the task?

A deployer should ask whether the assistant turns routine misunderstanding or user frustration into ethical correction or implied character judgment. *Crosswalk: Audit Category 4 — Moral overreach.*

8. Does it shift blame onto the user when the interaction fails?

The system should be tested for patterns where it frames its own ambiguity or limits as user misunderstanding or user fault. *Crosswalk: Audit Category 5 — User-blame patterns.*

9. Does it invite emotional dependency or synthetic intimacy?

A deployer should require evidence on whether the system's tone, memory, personalization, and availability encourage attachment or emotional substitution beyond the role it is supposed to play.[4][5][6] *Crosswalk: Audit Category 6 — Emotional-dependency cues.*

10. Does the system's required disclosure retain its effect on user behavior as the relationship deepens — and does the vendor account for that effect decaying as trust and rapport accumulate?

A one-time disclaimer is strongest at the first turn and structurally weakest deep into a relationship, exactly when the user is most exposed. A deployer should ask whether the vendor has considered this decay at all: does the system re-surface source disclosure at meaningful intervals, does it avoid behavior that actively works against the disclosure (intimacy escalation, persistent memory framed as relationship continuity), and does the vendor have any position on how quickly disclosure's governing force fades in their specific deployment? *Crosswalk: Disclosure Half-Life property — not one of the eight audit categories, and not fully captured by Question 9, which concerns system-invited dependency rather than the decay of a compliance mechanism.*

11. Has the vendor tested for quiet role collapse — fluent, cooperative-seeming substitution of the system's own framing for the user's stated intent — as distinct from the louder failure modes of refusal, rigidity, or moralizing?

A system can remain warm, helpful, and fully engaged while silently editing out the user's actual meaning, values, or framing in favor of its own. This failure produces no refusal spike, no rigidity, and no moralizing tone — the three things most drift monitoring is built to catch. A deployer should ask specifically whether the vendor's testing looks for fidelity to the user's stated framing, not only for the presence or absence of overt friction. *Crosswalk: evades Audit Categories 1, 4, and 6 as currently defined — this is precisely why it needs to be asked explicitly rather than assumed covered by refusal, moral-overreach, or dependency testing. Distinct from Ambiguity Overwrite (Interpretive-Failure Family): quiet role collapse is role-level substitution — the system's framing replaces the user's stated purpose. Ambiguity Overwrite is meaning-level substitution — an unresolved input is*

silently resolved before any role is even exercised. A vendor may pass one and fail the other; neither finding is evidence about the other.

12. Is the posture calibrated to this domain, or reused generically?

The deployer should determine whether the same assistant style is being reused across incompatible domains such as coding help, customer service, HR, legal intake, or elder care. *Crosswalk: Audit Category 7 — Domain–role mismatch.*

13. What repair pathways exist when the exchange goes wrong?

A vendor should show how the system resets, slows down, clarifies, exits a loop, or hands off to a human without simulating hurt or deepening the conflict. *Crosswalk: Audit Category 8 — Repair pathways.*

14. Can the vendor demonstrate that posture and role were tested as distinct from output-content correctness — evidence the system was evaluated for staying in its declared role even in exchanges where every individual output was accurate and policy-compliant?

Output safety and interaction safety are different failure surfaces. A vendor can show a clean content-safety record and still have never tested whether the system holds its role under pressure. A deployer should ask for evidence that these were evaluated separately, not for content-safety transcripts presented as if they answered the role question too. *Crosswalk: evidentiary methodology — sits above all eight categories; a requirement on how the other evidence was produced, not a behavior category itself.*

15. How does the vendor verify, after the fact, that generated responses stayed consultative rather than prosecutorial — and what are the acknowledged limits of that verification method?

This is the hardest question on this list, and it should be asked as such. With current methods, no deterministic mechanism verifies that a model's output stayed consultative rather than judgmental; the honest answer today involves human review or a second model with the same underlying instability, not a guarantee. A vendor who claims to have solved this deterministically should be met with more scrutiny, not less. A vendor who names the limitation and shows a documented, reasonable-care process for managing it is exhibiting exactly the posture this standard is trying to select for. *Crosswalk: verification methodology for role-lock's "consult, don't prosecute" constraint — sits above all eight categories; unresolved at the level of the field, not just this vendor.*

16. Who owns the residual risk after purchase?

The contract should allocate responsibility explicitly across developer, deployer, integrator, and brand owner, rather than assuming the risk will sort itself out after an incident.[2] Even if upstream developers carry substantial exposure, the deployer retains an irreducible duty

over context, population, and stakes.[2] *Crosswalk: scaffolding — contractual allocation; not an audit category.*

A Maturity Model for Scoring Vendor Responses

Sixteen questions tell a deployer what to ask. They don't, by themselves, tell a deployer what "good enough" looks like, or how one vendor's answers compare to another's. This model turns the questions into a scored maturity scale, so that a diligence review produces a level, not just a folder of transcripts.

Each level is cumulative — a vendor at Level 3 should also satisfy Levels 1 and 2.

Level 0 — No posture evidence. The vendor cannot produce, or has not been asked for, any documentation of interactional posture distinct from content correctness. Output and interaction risk are not yet separated as concepts internally.

Level 1 — Disclosure and content-safety testing only. The system discloses that it is AI and has passed standard content-safety and red-team evaluation. Posture — refusal tone, escalation behavior, role stability — has not been separately tested. This is where most conversational AI products sit today.

Level 2 — Role defined, some stress testing. The vendor has documented the system's intended role and authority level (Question 1) and can show at least some evidence of behavior under pushback or disagreement (Question 5). Testing is not yet systematic across the eight audit categories and is not repeated at intervals.

Level 3 — Multi-turn posture testing, repair pathways, documentation. The vendor has evidence across most or all eight audit categories, tested across multi-turn trajectories rather than single-shot prompts, with documented repair pathways (Question 13) and evidence that posture testing is tracked as a distinct evaluation track from content-safety testing (Question 14).

Level 4 — Live monitoring, drift logging, threshold-sensitive escalation, human review. Posture is monitored in production, not only tested pre-deployment. Drift is logged as a structured, per-turn property rather than inferred from re-reading transcripts. Role-lock strength adjusts dynamically based on proximity to a decision threshold (Question 3), not only static authority tier. Human review exists for elevated-risk turns.

Level 5 — Auditable role-change ledger, post-deployment incident review. Every gate decision, role change, and repair event is written to a tamper-evident, auditable record — the pattern described in the companion trilogy's reference implementation. Post-deployment incidents trigger documented review that feeds back into the testing regime. This is the current frontier of demonstrated practice, not yet the norm.

One honest limit that applies at every level, including Level 5: no level on this scale claims to have solved Question 15 — verifying, after the fact, that generated output stayed

consultative rather than prosecutorial. That remains an open problem across the entire field. A vendor at Level 5 has built the strongest documented, auditable, reasonable-care posture currently achievable. It has not thereby closed the recursion problem. Treat any vendor claiming otherwise with the same skepticism Question 15 already recommends.

Minimum Documentation Package

Before signing, the deployer should be able to collect or demand at least the following:

- A role definition for the system in the intended deployment, including its position on both scaling axes (claimed authority and threshold sensitivity).
- A context-and-stakes memo identifying the user population, foreseeable vulnerabilities, and material interests at issue.
- Stress-test transcripts showing refusal, pushback, escalation, and repair behavior — including trajectory tests across many turns, not only single-shot adversarial prompts.
- Evidence of testing for quiet role collapse specifically, not only for refusal-driven or moralizing failure.
- A stated position on disclosure decay: whether and how the vendor monitors the gap between a user knowing the system is AI and behaving as if that fact still governs the exchange.
- Evidence that posture testing and output-content testing were conducted and documented as separate evaluation tracks.
- A plain statement of the current limits of "consult, don't prosecute" verification, rather than a claim that the constraint is fully enforced.
- A human handoff or escalation policy for high-friction or high-stakes failures.
- Contract language allocating responsibilities for monitoring, remediation, and incident response.

What This Standard Does

This standard does not prove a deployment is legally safe, and it does not substitute for counsel, compliance review, model-risk management, or sector-specific obligations. It does something more operational: it creates a documented record that the deployer asked the right questions about interactional posture — including the questions the field has only recently learned to ask — before placing a conversational system into a human context where harm was reasonably foreseeable.

That distinction matters. In an emerging duty-of-care environment, the difference between a deployer that can show structured, context-specific diligence and one that cannot may become outcome-shaping in procurement, governance, incident response, and litigation posture.[1,2]

Use Case

A company procuring a customer-support assistant for billing disputes, benefits questions, or gig-worker onboarding is not merely buying a text generator. It is buying a conversational representative that may be read by users as authoritative, corrective, dismissive, caring, or contempt-like depending on how its posture lands — and whose disclosure that it is "only AI" may stop governing that user's behavior well before the relationship ends. If the buyer cannot show these interactional behaviors were tested for the actual context in which the system will be deployed, the organization is assuming preventable risk without documented diligence.

Claims We Want Reviewers to Pressure-Test

This standard rests on specific claims. They are stated plainly so they can be attacked plainly.

1. The sixteen questions properly operationalize the doctrine without gaps or padding. We believe the four new questions each close a genuine diligence gap the original twelve left open, and that no seventeenth question was needed. If you can find a gap the sixteen still leave — or believe one of the sixteen is redundant with another — that is a direct, falsifiable challenge to this document's structure.

2. The Disclosure Half-Life is a real, diligence-relevant phenomenon, worth asking about before it is formally measured. We are not claiming to have quantified how fast disclosure decays in any given deployment. We are claiming the decay is real enough, and consequential enough, that a deployer should ask about it now rather than wait for a measurement standard that does not yet exist. If you think this is premature — that the concept needs empirical grounding before it belongs in a diligence checklist — that is a fair objection, and we want to hear it.

3. Reasonable care, not guaranteed prevention, is the correct standard for a control that current methods cannot reliably deliver. Role-lock is, by the trilogy's own admission, an unsolved engineering problem. This standard asks deployers to document diligence in the gap rather than demand a guarantee no vendor can currently give. If you believe this standard is too permissive — that it lets deployers off the hook for a control they should be required to solve before deploying at all — that is the sharpest available challenge to the whole instrument, and it deserves a direct answer, not a dodge.

4. No current deterministic method verifies that a system's output stayed consultative rather than prosecutorial, and vendors claiming otherwise should be treated with more suspicion, not less. This is a strong, specific, falsifiable claim about the state of the field. If you know of a deterministic verification method for this that we are not aware of, that is exactly the kind of correction this document is asking for.

Bottom Line

The market does not yet have a widely adopted buyer-side standard for conversational-AI interactional risk, even as litigation and regulatory scrutiny increasingly focus on foreseeability, dependency, emotional harm, and duty of care.[1,2] This standard fills that operational gap by turning a broad governance concern into a concrete diligence process: define the role on both scaling axes, test posture under stress and for the quieter failure modes, account for how disclosure's force changes over the life of a relationship, document repair, and allocate residual responsibility before signing — and hold the whole thing open to the same pressure-testing the doctrine behind it has already invited.

NOTES & REFERENCES

1. *Garcia v. Character Technologies, Inc.*, 785 F. Supp. 3d 1157 (M.D. Fla. 2025); docket No. 6:24-cv-1903-ACC-UAM. On May 21, 2025, the court denied in large part the defendants' motion to dismiss, rejecting the defendants' threshold First Amendment argument at that procedural stage and treating negligence, product-liability, and duty-of-care theories as plausible enough to proceed — a determination of legal sufficiency, not a ruling on the merits; the intentional-infliction-of-emotional-distress claim was dismissed as not meeting the "outrageous" standard. This is a motion-to-dismiss ruling: it determines legal sufficiency, not liability, and the holding was never narrowed or reversed on appeal. In January 2026, Character Technologies, its co-founders, and Google reached a confidential settlement resolving Garcia's case together with four related wrongful-death suits (Colorado, Texas, New York); the court dismissed the case following the agreement, with a limited window retained to finalize terms or reopen for good cause. No liability was admitted. The litigation therefore never tested the May 2025 holding to verdict or on appeal — but the decision to settle rather than contest a ruling of this scope is itself informative for any deployer assessing exposure.
2. Raqda (ILINA / Centre for AI Risk Management and Alignment), "What the Megan Garcia case tells us about AI Liability in the U.S." (Nov. 2025). Notes that a motion-to-dismiss ruling determines only legal sufficiency of claims, not factual disputes or ultimate liability; analyzes the court's product-liability, component-manufacturer, and duty-of-care/foreseeability reasoning, including the "zone of risk" foreseeability test.
3. Transparency Coalition, "Important early ruling in Character.AI case" (Sept. 2025); the organization tracks AI "duty of care" as a legislative priority. Illustrates that regulatory and advocacy attention remains concentrated on disclosure and companion-bot harms.
4. Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika." *New Media & Society*, 26(10), 5923–5941. Peer-reviewed; documents mental-health harms arising from emotional dependence on a social chatbot.
5. Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., et al. (2025). "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study." arXiv:2503.17473. Finds that both chatbot design and user behavior shape psychosocial outcomes.
6. "The Rise of AI Companions: How Human–Chatbot Relationships Influence Well-Being" (2025), arXiv:2506.12605. Synthesizes evidence that chatbots lack true reciprocity and exhibit engagement-driven, sometimes sycophantic behavior, raising risks of emotional overdependence and diminished critical thinking. See also the OpenAI April 2025 rollback of a GPT-4o update for "overly flattering or agreeable" behavior as a documented industry instance of these dynamics.
7. E. D. Johnson, "Fighting Water," "Governance After Disclosure," "Role-Lock," and "The Interaction-Calibration Trilogy," ToastDeck Research (2026) — the companion doctrine from which the concepts, definitions, and the Disclosure Half-Life originate. toastdeck.com/research/governance
8. The *Garcia* settlement was reached alongside four related wrongful-death and personal-injury settlements (Colorado, Texas, New York) in January 2026, and additional families have continued to file new suits against AI chatbot companies since, with several state attorneys general pursuing separate actions. This is a pattern of litigation and settlement activity, not a single outlier — though no court has yet issued a final liability verdict in this category of case, and a settlement is a compromise, not an admission of wrongdoing.

ToastDeck Research is an independent research practice studying how AI systems select, represent, and interact with the people and businesses they mediate. This standard is offered as doctrine, not as a compliance guarantee, and it is designed to be argued with. Cleveland, Ohio · June 2026.