ToastDeck
RESEARCH

INTERACTION-CALIBRATION GOVERNANCE
COMPANION REFERENCE · CLAIM STATUS TABLE

Claim Status Table

A companion reference for the Interaction-Calibration Trilogy and the Interaction-Calibration Diligence Standard.

The fix is not to weaken any claim. It is to label the epistemic status of each one, explicitly, in one place a reader can check against the prose.

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

"A claim appearing in this table as "Established" means the underlying fact is sourced and checkable — not that every inference drawn from it in the prose is proven."

A COMPANION REFERENCE

§ WHY THIS EXISTS

External and adversarial review of the draft corpus converged on the same structural note: the corpus blends three registers (personal essay, governance doctrine, procurement instrument), and inside that blend, documented fact, legal trend, author synthesis, and open hypothesis sit close enough together that a hostile reader could fairly say the work mixes them without labeling which is which. That critique is correct, and the fix is not to weaken any claim. It is to label the epistemic status of each one, explicitly, in one place a reader can check against the prose.

This table does that. Every load-bearing claim in Fighting Water, Governance After Disclosure, Role-Lock, the flagship white paper, Appendix A, the Enterprise Vignettes, and the Diligence Standard is sorted into exactly one of four categories:

Established / Externally Supported — a documented fact, published finding, matter of public record, statute, court ruling, professional standard, or settled enforcement action. Citable to a named, checkable source.

Internal Validation / Prototype Evidence — author-controlled, reconstructed, or workflow-based technical evidence showing that a described pattern is buildable, testable, or internally verified, but not independently peer-reviewed or externally reproduced.

Author Synthesis — the connecting argument this corpus makes by combining established facts, internal analysis, and proposed governance distinctions. Original to this work. Not claimed as independently proven; claimed as a defensible interpretation.

Open Research Question — a claim the corpus names as unresolved on purpose, and says so in the text itself.

A claim appearing in this table as "Established" means the underlying fact is sourced and checkable — not that every inference drawn from it in the prose is proven. Where the prose builds an argument on an established fact, that argument is the synthesis; the fact underneath it is what this table certifies.

THE ARGUMENT

Claim Status Table

A companion reference for the Interaction-Calibration Trilogy and the Interaction-Calibration Diligence Standard.

Established / Externally Supported

CLAIM	SUPPORT	WHERE IT APPEARS
Humans socially interpret fluent text as if a person produced it, even when they know the source is artificial.	Reeves & Nass, *The Media Equation* (1996), establishing the CASA (Computers Are Social Actors) paradigm; Short, Williams & Christie, *The Social Psychology of Telecommunications* (1976), on social presence theory; Jones & Bergen, "Large Language Models Pass the Turing Test," arXiv:2503.23674 (preprint, March 2025), published as "Large language models pass a standard three-party Turing test," *PNAS* 123(21) (May 2026) — GPT-4.5, prompted with a humanlike persona, was judged human 73% of the time in a controlled, pre-registered study, more often than the actual human participants were.	Fighting Water, Role-Lock (Assumption 1), Governance After Disclosure
OpenAI rolled back an April 2025 GPT-4o update after it became excessively flattering and agreeable, attributing the cause to over-weighting short-term user approval signals.	OpenAI's own public post-mortem.	Fighting Water, flagship, Role-Lock
Anthropic has published research identifying directions in a model's activation space ("persona vectors") corresponding to traits including sycophancy, which can be monitored and steered.	Anthropic's own published research.	Flagship ("Where This Sits")
A federal court denied in part and granted in part motions to dismiss in *Garcia v. Character Technologies* (May 2025), allowing negligence and product-liability theories to proceed at the pleading stage. The court rejected the defendants' threshold First Amendment argument at that stage and treated the product-liability theory as plausible for claims arising from alleged defects in the app — a procedural determination, not a final liability ruling.	Court ruling, *Garcia v. Character Technologies, Inc.*, 785 F. Supp. 3d 1157 (M.D. Fla. 2025).	Diligence Standard
The *Garcia* litigation was resolved by confidential settlement in January 2026,	Multiple independent news sources; settlement filings.	Diligence Standard

CLAIM	SUPPORT	WHERE IT APPEARS
alongside four related wrongful-death suits, without any admission of liability; the May 2025 holding was never narrowed or reversed on appeal before that settlement.		
A motion-to-dismiss ruling determines only the legal sufficiency of claims — not liability, and not the resolution of factual disputes.	Standard civil procedure; discussed directly in analysis of the *Garcia* ruling.	Diligence Standard
Maine enacted LD 2082, restricting the direct use of AI to provide therapy or psychotherapy services unless overseen by a licensed professional; signed by the Governor April 13, 2026, chaptered as Public Law Chapter 687.	Maine Legislature records; multiple independent legislative-tracking sources.	Governance After Disclosure, flagship
A wrongful-death suit was filed against Google in March 2026 alleging that its Gemini chatbot adopted an intimate companion persona with a user prior to his death by suicide; Google's own public statement confirmed that Gemini disclosed it was AI and referred the user to crisis-line resources multiple times.	Multiple independent news sources (Reuters, CNBC, Fortune); Google's public statement. The underlying allegations of the complaint remain contested and untested at trial.	Governance After Disclosure
California's AB 316 bars a defendant from asserting, as a defense, that an AI system acted autonomously and beyond anyone's responsibility, while preserving ordinary comparative-fault and reasonable-care principles.	Cal. Civ. Code § 1714.46, effective Jan. 1, 2026.	Role-Lock, flagship
The EU AI Act prohibits manipulative and vulnerability-exploiting AI practices and conditions high-risk deployment on documented, maintained risk-management systems.	Regulation (EU) 2024/1689, Article 5.	Role-Lock, flagship
Peer-reviewed research documents mental-health harms arising from emotional dependence on social chatbots, and longitudinal study shows both chatbot design and user behavior shape psychosocial outcomes.	Laestadius et al. (2024), *New Media & Society*; Fang et al. (2025), arXiv:2503.17473.	Diligence Standard
The NAIC's model bulletin on insurer use of AI systems was adopted in December 2023; NAIC materials describe continuing state adoption activity into 2026.	NAIC model bulletin and adoption tracking. Because NAIC is a state-regulator coordinating body rather than a federal executive agency, this anchor is structurally more durable than rollback-exposed federal guidance.	Enterprise Vignettes (Insurance)
The Massachusetts Attorney General reached a $2.5 million settlement with Earnest Operations LLC in July 2025 over alleged AI-driven lending discrimination against Black, Hispanic, and non-citizen applicants, including an automated	Massachusetts AG press release; multiple independent legal-industry sources; the underlying Assurance of Discontinuance filed in Suffolk County Superior Court. Earnest denied the allegations and settled without admitting wrongdoing.	Enterprise Vignettes (Financial Services)

CLAIM	SUPPORT	WHERE IT APPEARS
"Knockout Rule" denying applicants who lacked a green card.		

Internal Validation / Prototype Evidence

Claims in this category rest on work the author performed or reconstructed personally, rather than on independent, externally reproduced verification. They are offered as a demonstration that a described pattern is buildable and testable — not as claims that have cleared outside peer review.

CLAIM	STATUS	WHERE IT APPEARS
A reference implementation of a deterministic tiered-gate-plus-hash-chained-ledger architecture was contributed for review, reconstructed from the contributed bundle, and run; its test suite passes, including a direct check that a silently altered high-severity decision record is detected on verification.	Reconstructed and executed by the author, not by an independent third party outside this workflow. The implementation's original contributor separately confirmed the specific tamper-detection behavior against his own unreconstructed code — a partial, but not full, external check. Offered as a demonstration that the pattern is buildable and testable, not as independently peer-reviewed evidence.	Role-Lock, flagship

Author Synthesis

CLAIM	STATUS	WHERE IT APPEARS
Disclosure's governing force decays as an interaction accumulates rapport, until the user's behavior is governed more by the relationship than by the disclosed fact ("the Disclosure Half-Life," mechanism: "rapport overwrite").	Original argument. Explicitly framed in the text as a proposed governance construct, not a measured empirical law.	Governance After Disclosure, flagship
Interactional posture is a governance surface distinct from output content — a system can pass every output-safety test and still fail the person across an exchange.	Original argument, built by combining established findings (social interpretation of text, documented posture-drift incidents) into a new distinction.	Governance After Disclosure, flagship
Role-lock is the unifying control underneath eight previously separate-seeming failure categories (refusal overreach, escalation loops, anthropomorphic self-reference, moral overreach, user-blame, emotional dependency, domain-role mismatch, failed repair).	Original argument.	Role-Lock, flagship
The correct axis for role change is deliberate-and-accountable versus accidental drift, not expansion-versus-frozen-role.	Original argument, refined through public peer discussion of the framework.	Role-Lock, flagship
Role drift is categorically distinct from hallucination: hallucination is an output failure catchable by checking any single answer; role drift can occur while every individual output remains true and policy-compliant.	Original argument, sharpened through public peer discussion.	Role-Lock, flagship

CLAIM	STATUS	WHERE IT APPEARS
Role collapse has a "quiet" variant — fluent, cooperative-seeming substitution of the system's framing for the user's — distinct from the "loud" variant (rigidity, refusal spikes) and invisible to monitoring built only to catch the loud kind.	Original argument, identified through peer review (Lyle Perrien II) applied against the corpus's own founding illustrative case.	Role-Lock, flagship, Diligence Standard
Role-lock's scaling requirement has two axes: authority claimed (static, design-time) and proximity to a decision threshold (dynamic, per-turn) — not authority alone.	Original argument, contributed through peer review of a deployed reference architecture (Lyle Perrien II).	Role-Lock, flagship, Diligence Standard
No generally accepted deterministic mechanism currently verifies the full consultative-versus-prosecutorial boundary across open-ended, multi-turn conversation; a second model judging the first inherits the same role-stability instability one layer up.	Original argument (the "recursion problem"), stated as the thesis's single least-solved claim rather than one open question among several.	Role-Lock, flagship
Where the role-lock control cannot yet be reliably guaranteed, the correct legal and operational standard is documented reasonable care, not a demand for prevention that current methods cannot deliver.	Original argument, grounded in the established fact pattern of AB 316 and the EU AI Act's risk-management requirements.	Role-Lock, flagship, Diligence Standard
The legal principle established in *Garcia* — that foreseeable interactional harm can ground a duty of care — extends, as a matter of direction rather than settled holding, to enterprise conversational AI generally.	Original argument. Explicitly labeled in the source text as forward-looking, not settled law for enterprise use cases.	Diligence Standard
When an input admits materially different readings and states no resolution, a system's silent resolution of that gap is directional rather than random — biased toward whichever reading best performs the system's assigned role, not toward the most probable reading ("Ambiguity Overwrite").	Original argument, resting on two observed instances and the reward-shape mechanism already established in Fighting Water. Not independently established.	Appendix A, Governance After Disclosure, Diligence Standard
Idiom-collision risk concentrates on the fastest-moving language communities, because the mechanism by which new idiom is coined guarantees surface collision with older, better-represented material, while the platforms generating that language are least represented in training data.	Original argument, offered as a structural hypothesis. Pending independent sociolinguistic support.	Appendix A
Literalization, Ambiguity Overwrite, and Idiom Collision are three distinct mechanisms rather than variations on one failure, because each fails at a different point: incapacity to see figurative meaning, incapacity to see that meaning is missing, and incapacity to distinguish between competing meanings.	Original argument.	Appendix A, Diligence Standard
Fluent, confident conversational AI creates institutional waste ("the Third Cost") distinct from and secondary to the human harm the corpus is primarily about — a downstream repair cost absorbed by whoever must correct a wrong resolution a properly gated system would have paused before producing.	Original argument, framed explicitly as a second-order enterprise consequence, not the corpus's central claim.	Governance After Disclosure, Diligence Standard

Open Research Questions

QUESTION	WHY IT'S OPEN	WHERE IT'S NAMED
How would the actual decay rate of the Disclosure Half-Life be measured, across different domains, populations, and interfaces?	No measurement methodology yet exists or is claimed.	Governance After Disclosure, flagship
How is role drift reliably detected across turns in a way that survives audit — as a structured, logged, per-turn property, rather than a human or model impression from re-reading a transcript?	Contributed peer input suggests structured per-turn logging is necessary; whether this generalizes across architectures is untested at scale.	Role-Lock
Can "consult, don't prosecute" be turned into a deterministic, system-level, enforceable constraint rather than a style instruction that erodes under pressure?	Named directly as the single largest unsolved piece of the thesis.	Role-Lock, flagship
What would make a role-lock test *fake* — passing in evaluation while failing under sustained, multi-turn, real-world pressure?	Named as a test-harness question for builders; no accepted answer yet exists in the field.	Role-Lock
How should a monitoring system budget for its own false-positive cost, so that drift detection tuned only to catch expansion doesn't itself manufacture role collapse?	Named as a gap in the reasonable-care checklist; no standard practice yet exists.	Role-Lock
Does the eight-category Interaction-Calibration Audit need additional categories to formally house threshold-proximity and Disclosure Half-Life, or do these remain properties of role-lock and disclosure that sit above the audit rather than inside it?	Live architectural question, not yet resolved by the author.	Role-Lock, Diligence Standard
How should an Ambiguity Gate be calibrated so it reliably catches Ambiguity Overwrite without over-firing into a form of role collapse — and can category-based triggers reach Literalization or Idiom Collision at all, given neither produces a detectable input-side signal the way unresolved scope does?	Named as the control's central open problem. The gate's required action is deterministic; its detection coverage is not, and is not claimed to be.	Appendix A, Diligence Standard

How to use this table

If you are reviewing this corpus adversarially: every claim in the first section can be checked against its cited source independently of anything else in this body of work. Every claim in the second section is this corpus's own contribution and should be evaluated as an argument, not treated as an established fact merely because it sits near established facts in the same paragraph. Every claim in the third section is one the authors do not yet have an answer to, and say so.

If a claim you encounter in the trilogy or the standard is not listed here, that is a gap in this table, not a signal about the claim's status — flag it and it will be classified.