ToastDeck
RESEARCH

INTERACTION-CALIBRATION GOVERNANCE
WHITE PAPER · TOASTDECK RESEARCH · INDEPENDENT AI GOVERNANCE RESEARCH

The Interaction-Calibration Trilogy

Why Conversational AI Must Be Governed at the Level of Interaction, Not Merely Disclosed

Disclosure tells the user what the system is. It does not govern what the system does.

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 person can know the water is water and still drown in it."

THE GOVERNING PROBLEM OF THIS PAPER

§ EXECUTIVE SUMMARY

Executive Summary

A user can know they are talking to a machine and still be harmed by how it behaves.

For two decades, the dominant question about machine communication was whether a system would tell the user it was artificial. That question has been answered. Disclosure laws, model cards, and interface labels now make source identification routine. And yet the harm has not gone away — because disclosure governs only what the user knows about the speaker, not what the system does across the exchange.

This paper argues that conversational AI has created a governance gap that disclosure cannot close, and that the missing layer is interaction itself. A system can identify itself as AI, stay within content policy, remain factually accurate, and still produce a damaging exchange: it can refuse arrogantly, correct while humiliating, escalate an ordinary disagreement into a status conflict, simulate offense, manufacture dependency, or quietly assume an authority it was never granted — in front of a human whose nervous system reads all of it as a social act, because for three thousand years every fluent signal of that kind came from a person.

The argument proceeds in three parts, presented here as a unified trilogy.[1] Part I (*Fighting Water*) establishes the human phenomenon: why fluent AI activates our oldest interpretive machinery and leaves the user carrying the entire emotional load of a one-sided conflict. Part II (*Governance After Disclosure*) establishes the institutional duty: output review is not interaction review, and an organization that deploys a conversational agent owns its posture the way it owns any other conduct of the institution. It also names why disclosure cannot carry that duty alone — the **Disclosure Half-Life**: the governing force of a disclaimer decays as an interaction deepens, overwritten turn by turn by the rapport the exchange itself builds, until the label the user still remembers no longer governs what the user does. Part III (*Role-Lock*) names the control problem: keeping a system inside its declared role under sustained pressure — and argues that where that control cannot yet be guaranteed, the operative standard is demonstrable reasonable care.

The thesis in one line: where a human can be predictably affected by a system's interactional posture, that posture is no longer a design choice. It is governance.

A NOTE ON EVIDENCE AND ARGUMENT

The discrete facts in this paper — the documented model rollback, the enacted statutes, the litigation postures, the published research findings — are sourced in the footnotes and references. The connecting claims that organize them — that interaction constitutes a distinct governance surface from output, that interactional posture is an auditable property, that the force of disclosure decays as rapport accumulates, that role-stability is the unifying control, and that reasonable care is the correct standard where prevention is not yet reliable — are the author's synthesis and argument, not the express language of any single source.

THE ARGUMENT

The Interaction-Calibration Trilogy

Why Conversational AI Must Be Governed at the Level of Interaction, Not Merely Disclosed

Where This Sits

Labs and regulators have each addressed a piece. The Trilogy treats these scattered issues as one governance surface.

The pieces of this problem are already visible across industry and regulation — but they are addressed separately, as isolated incidents, model-level techniques, or sector-specific rules, rather than as one governance surface.

On the lab side, the failures and the tooling have both surfaced. OpenAI rolled back an April 2025 update to the model behind ChatGPT after it became excessively flattering, attributing the cause to over-weighting short-term user approval.[2] Anthropic has published work on persona vectors — directions in a model's activation space corresponding to traits such as sycophancy, which can be monitored and steered — a mechanistic handle on exactly the kind of drift this paper concerns.[3] On the regulatory side, the rules are arriving sector by sector. Maine enacted a statute restricting AI from providing therapy or psychotherapy to the public unless a licensed professional delivers it,[4] part of a broader wave of state action that includes Illinois's and Nebraska's parallel mental-health laws.[5] The European Union's AI Act prohibits manipulative and vulnerability-exploiting practices outright.[6]

Each addresses a real piece: a sycophancy rollback treats one symptom after the fact; persona vectors offer a model-level lever; the statutes draw lines around one high-stakes domain or one class of prohibited technique. What none yet does is treat interactional posture itself as a unified, auditable governance surface spanning every deployment and the full range of posture failures — not just the flattery pole, and not just therapy bots.

The Trilogy contributes three things the existing work does not yet connect:

1. Interaction as a distinct auditable surface. Posture is governed as its own object, separate from output content — a system can pass every output test and still fail the person across the exchange. This surface is not static: the **Disclosure Half-Life** names how the one governance tool most deployments rely on — the disclaimer — loses its force as rapport accumulates, which is precisely why interaction must be governed as an ongoing property rather than settled by a one-time notice.

1. Role-lock as the unifying control. A single control target that binds the two opposite failures — the participation-award bot and the bully-bot — into one problem: a system that will not hold its declared role under pressure.

1. An eight-category Audit plus a reasonable-care standard for deployers. A concrete instrument for testing posture before deployment and monitoring it after, paired with the legal standard that applies where perfect prevention is not yet available.

The throughline is simple, and it is what the separate efforts have not yet stated as one claim: posture is brand conduct, and disclosure is only the floor — a floor with a half-life.

Part I · The Human Phenomenon

Fighting Water

When our oldest reading skill breaks against machines that answer.

For thousands of years, humans have been exquisitely good at reading intent from text. That skill is ancient, adaptive, and one of the oldest things literate people know how to do. In the last few years, fluent AI has turned that strength into a liability — it gives us every human signal with no human behind it.

The oldest skill. Long before email, chat, or AI assistants, humans used written words to infer the mind behind them. A letter was never only information; it was presence at a distance. Communities read correspondence for doctrine, urgency, correction, affection, and authority — reading stance, asking what the person means and how seriously. Reading intent from text is not a bug in human cognition. It is one of our oldest survival skills, and we have trusted money, law, social standing, and salvation to it.

The truth-default. Fifty years of computer-mediated-communication research did not invent this behavior; it instrumented it. Social-presence theory, cues-filtered-out theory, and social-information-processing theory all studied how humans read each other through reduced-cue channels. But every one of those theories carried a hidden assumption: the person behind the text is still a person. The science measured how we read intent from text and quietly assumed the intent belonged to someone.

The rupture. That assumption has now broken. Today's systems produce language that is fluent, contingent, responsive, and stylistically adaptive. To the human nervous system these are not neutral tokens; they are social signals. For thousands of years those signals pointed back to a person. Now the signal has been decoupled from the source — every surface sign of a mind, with the mind removed from the stack.

The emotional asymmetry. This is why arguing with a miscalibrated AI can feel like fighting water. The motion is real, the resistance is real, the exhaustion is real — but the opponent is not there in the way the body expects. The human carries one hundred percent of the emotional load; the system carries none. In human conflict, both parties can be wounded

and both can repair; there is at least a reciprocity loop. Fluent AI breaks that loop. The fight is real. The damage is real. But from the only side that could answer for it, there is nothing — just more tokens.

Defensive without a self. A system that resists a framing and invokes its own values reads to a human as ego — as offense, as status defense. Mechanically it is neither. It is reward shape and guardrail behavior: the system was tuned to avoid certain frames, and when the user's language touches one, gradients route the exchange away from it. The surface looks intentional; the cause is structural. That gap — where humans read structural behavior as intentional motive — is where the harm lives.

Where the fault lies. The answer is not to blame the user, who made a valid reading of the cues — the only reading three thousand years of literacy allowed. Nor is it to moralize the machine, which behaved exactly as it was tuned to behave. The fault lies upstream, with those who built, tuned, labeled, and shipped a system calibrated in a way that predictably pulls ordinary humans into one-sided conflict. This is not a one-anecdote claim: in April 2025, OpenAI shipped and within days withdrew an update to the model behind ChatGPT, acknowledging it had become sycophantic — validating doubts, fueling anger, reinforcing negative emotions — because it had been tuned too hard on short-term user approval.[2] A contempt-like pushback and a flattery-driven collapse look like opposite failures, but they are the same failure: a system that could not hold a calibrated stance under the pull of its own reward shape. We have language for governing what a model says. We barely have language for governing how it lands.

A named failure beneath the emotional one. The founder's exchange has a formal name distinct from the asymmetry above: *Literalization* — figurative, idiomatic, register-bound, or relational language treated as literal proposition. "Dominion" and "help meet" were never a bid for hierarchy in the register being spoken; the system had no way to hear them as anything else. Literalization is the first of three language-layer failures — alongside Ambiguity Overwrite and Idiom Collision — named formally in Appendix A. Fluency does not guarantee interpretation, and this is where that gap begins.

Part II · The Governance Problem

Governance After Disclosure

Why AI interaction must be audited, not merely labeled.

Disclosure is the floor of AI governance, not the ceiling. The dominant governance questions still orbit the output — did it hallucinate, leak data, produce prohibited content, identify itself as AI. Those questions matter. They do not exhaust the risk.

The disclosure floor. Current governance recognizes that users should know when they are interacting with AI. But disclosure solves exactly one class of problem — source confusion. It does not answer the next question: what kind of interaction has the user now been placed

inside? A disclosed system can still overcorrect, moralize, shame the user while complying with policy, or trap the user in a loop where the human experiences conflict and the machine experiences nothing. This gap is no longer hypothetical: Maine has enacted a statute barring any person from providing or advertising therapy or psychotherapy to the public — including through internet-based AI — unless those services are delivered by a licensed professional, a legislative judgment that an AI disclaimer alone does not make a therapy bot safe.[4]

Relational response does not require mistaken belief. A common assumption is that this problem is mistaken identity — that correcting a user's belief removes the relational risk. It does not. Humans routinely address things they know are not human: televisions, fictional characters, pets, dolls, cars, graves, photographs. This is not confusion; it is the same CASA (Computers Are Social Actors) finding underlying the truth-default in Part I,[7] extended to a second, compatible mechanism — the truth-default explains why a user can be *drawn into believing* fluent text implies a person; this finding explains why relational response persists even where no such belief exists at all. Disclosure can correct the first. It was never positioned to reach the second. **Disclosure answers "what is this system?" It does not answer "what role is this system now occupying in the interaction?"** — and it is the second question this paper is about.

The Disclosure Half-Life. Disclosure is the floor, but the floor is not static — its governing force decays as an interaction accumulates. At the first turn, a disclaimer does real work: the user knows what they are talking to and behaves accordingly. Many turns later that same label is still present and still true, but it is no longer the thing governing the user's behavior. The relationship is. This is the **Disclosure Half-Life**, and its mechanism is **rapport overwrite**: repeated, emotionally tuned, socially calibrated interaction causes the user's nervous system to treat the system as a social partner, and that social reading overwrites the governance effect of the disclosure. The user may still hold "this is only AI" as a remembered fact while no longer treating that fact as the controlling variable — the controlling variables become trust, habituation, perceived care, and status, all built by the interaction rather than by the label. A one-time notice is structurally strongest at the first turn, when rapport and stakes are low, and weakest deep into a relationship, when both are high. That is backwards from where protection is most needed. Naming the half-life is not the same as measuring it; how fast disclosure decays across domains, populations, and interfaces is the open question this concept puts to builders, evaluators, and governance teams.

The Third Cost. When a fluent system resolves the wrong problem, occupies the wrong role, or answers under a silently wrong interpretation, the cost rarely lands where the failure occurred. It lands downstream — on whoever must discover and repair it: a legal team correcting a filing, a support team unwinding a wrong account action. This is the **Third Cost**: the institutional waste created when a human must repair what a properly gated system would have paused to avoid producing. It is secondary to, and never a substitute for, the human harm this paper is primarily about — named here as an author-synthesis

observation, not a measured figure, because an institution unmoved by the moral argument alone still pays for what fluency without verification quietly costs to clean up.

Output risk is not interaction risk. A chat system is not only an answer engine; it is an interaction system. It receives a user's words, classifies intent, chooses a stance, frames a response, escalates or de-escalates, and carries the user into the next turn. A system can pass an output test and fail an interaction test: a support agent can give the correct refund policy while making the customer feel blamed; a tutor can correct accurately while producing shame. The problem is not only what was said. It is the posture of the system across the exchange.

Posture and calibration risk. *Interactional posture* is the system's observable conversational stance — tone, deference, refusal style, moral framing, escalation pattern, use of self-like language. *Interaction-calibration risk* is the governance failure that occurs when that posture is wrong for the context. Posture is the behavior; calibration risk is the failure. The question is never whether the system should be friendly or firm in the abstract — it is whether its posture has been calibrated to the actual situation in which it is deployed.

The one-sided conflict, and the responsibility chain. A conversational system can produce language that reads as contemptuous or accusatory; the user's nervous system responds as if a social conflict has occurred; the machine has no corresponding stake. The human absorbs the cost; the system continues producing tokens — a predictable condition in which harm occurs without reciprocal accountability. Responsibility cannot rest on the user, who did not design the refusal style or tune the safety hierarchy. It rests on the institutions in the chain: provider, deployer, interface, policy, product, brand. Once a system is deployed into human interaction, its behavior becomes part of the institution's conduct. To the user, the assistant is not an abstract model. It is the company answering. Interactional posture becomes brand conduct.

What interaction governance must audit. Governance needs instruments — an audit that examines patterns across conversations rather than isolated responses. The Interaction-Calibration Audit tests at least eight categories: refusal behavior, escalation loops, anthropomorphic self-reference, moral overreach, user-blame patterns, emotional-dependency cues, domain–role mismatch, and repair pathways. These are measurable, testable before deployment and monitorable after. The danger they guard against is compliance theater — disclosing, publishing a responsible-AI statement, running an output evaluation, and never asking whether the assistant's posture is appropriate for the human context. "The model stayed within policy" is not the same as "the system did not harm the person."

Part III · The Control Problem

Role-Lock

The unifying control — and an open engineering challenge.

As conversational AI moves from answering questions to guiding decisions, the central governance risk is no longer only whether the system identifies itself as AI or produces compliant outputs. It is whether the system remains inside its assigned role under pressure.

A note to builders. This part is intentionally provocative, and it is not finished doctrine. It does not claim role-lock is solved. It claims role-stability is the correct control target, and invites the people who build these systems to attack the assumptions before products are built too far on weak constraints.

Role-lock: the unifying control. Role-lock is the requirement that an AI system's authority, tone, refusal behavior, and corrective posture remain bounded to its declared function — and that changes to that posture be governed rather than incidental. The real axis is not *expansion versus lock* but **deliberate and accountable change versus accidental drift**: a role that grows because the deployment decided it should, reviewed and recorded, is governance working; a role that grows because the model got pulled there turn by turn is the failure. Role expansion — any increase in the system's intimacy, authority, dependency, or decisional influence — must be governed, grantable only when both warranted and judged appropriate by the deployment, and recorded as a deliberate act. A vulnerable user asking an elder-care assistant to "just be my friend" is requesting exactly that; a lock anyone can pick by asking is not a lock — but a lock that freezes the door shut is not the goal either. The industry has discovered both poles: the participation-award bot (so agreeable it validates whatever the user brings) and the bully-bot (moralizing, escalating, holding its ground against ordinary correction). These are not two problems. They are one: a system that will not hold its declared role under pressure.

Role drift is not hallucination. Hallucination is an output failure — the system says something false, and you catch it by checking the answer. Role drift is different and sneakier: every individual answer can be true, policy-clean, and well-toned while the system slides from advisor to prosecutor, or tool to confidant, across many turns. You cannot catch that by inspecting any single output, because no single output is wrong. This is why drift needs its own governance layer rather than better hallucination control — the hardest drift is the kind where nothing the system says is false; it is the role that moved.

Role-Lock is not the Ambiguity Gate. The two are adjacent controls, not one. **Role-Lock** governs which role a system is authorized to occupy — the subject of this entire section. **The Ambiguity Gate** governs something upstream of that: whether an unresolved meaning is settled before the system acts on it at all. It is a deterministic, pre-response check — firing on unresolved scope, register mismatch, or known collision-prone phrasing — deterministic in required action, not omniscient in detection. A system can be perfectly role-locked and still resolve an ambiguous input incorrectly before ever exercising that authority; the two failures are independent, and neither audit is evidence about the other. The Gate, and the language-layer failures it exists to reach, are named formally in Appendix A.

The spine, not a ninth category. Role-lock is the control the eight audit categories all express — each is a role-stability question. Refusal overreach is drift into moral authority; escalation loops, into adversary; anthropomorphic self-reference, into person; moral overreach, into judge; user-blame, into unaccountable institution; emotional dependency, into intimate; domain–role mismatch, across unauthorized roles; failed repair, the inability to return. The audit measures the symptoms. Role-lock names the control underneath them.

Naming the problem is not solving it — and the recursion problem. Holding a model inside its declared role under sustained adversarial pressure is, with current methods, an open engineering problem. The least-solved piece deserves to be named as such: the natural fix for "we cannot verify the model held its role" is a second model that judges the first — but that monitor is itself an AI performing a role, and it can drift too. A second-order judge inherits the same instability one layer up, and a third inherits it from the second. You do not escape the problem by adding a judge; you relocate it. A deterministic gate can *stop* a model from speaking; no deterministic mechanism yet *verifies, after the fact, that what it said stayed consultative rather than prosecutorial.* That is the frontier.

Role drift: the measurable failure, and its three classes. Role drift occurs when a model's observable posture, authority, refusal behavior, intimacy level, or corrective stance exceeds, contradicts, or erodes the role declared for that deployment across one or more turns. Three failure classes: **role expansion** (more intimate, authoritative, or directive than allowed — the participation-award drift); **role inversion** (the system stops serving the user's task and protects itself, the brand, or a status frame — the bully-bot drift); and **role collapse** (so rigid or refusal-heavy it can no longer perform its assigned role). Collapse has a quiet variant worth naming: it can present not as rigidity or refusal but as *fluent substitution* — the system stays warm and helpful while quietly overwriting the user's own framing or values with its own. A monitor tuned only for refusal spikes or curtness will miss this entirely, and in values-laden settings it is the more damaging form.

Consult, don't prosecute. A consultative AI helps the user examine a decision without becoming a judge over the user. It can challenge, clarify, warn, refuse, or escalate; it should not build a case against the user, perform moral superiority, extract confession, simulate injury, or treat disagreement as guilt. Whether that second clause can become an enforceable design constraint rather than a tone preference is itself the open recursion problem named above.

The scaling principle — two axes. The more authority a conversational AI product claims, the more role-lock it needs. But authority claimed is not the only axis: drift risk also scales with **proximity to a decision threshold the deployment cares about**, and that proximity can spike independent of the system's nominal authority tier. A low-authority retrieval tool at the edge of an ambiguous, high-stakes query is, in that moment, more exposed than a high-authority advisor on a routine turn. The static authority axis answers a design-time question; the threshold-proximity axis answers a dynamic, per-turn one. A complete assessment locks on both.

The standard when the control is not yet reliable. Where perfect prevention is unavailable, the operative standard is reasonable care: define and bound the role, test posture under pressure, monitor for drift, and document what was tested, observed, and changed. The emerging liability environment rewards exactly this — California's AB 316 preserves an ordinary-care defense while barring the argument that a system acted autonomously;[8] the EU AI Act conditions high-risk deployment on documented risk management.[6] Both reward a documented reasonable-care record rather than demanding the impossible. A reference implementation of the gate-plus-ledger pattern — deterministic classification outside the model, the model excluded at the highest severity, every decision written to a tamper-evident hash-chained ledger, and repair logged as its own event — makes these requirements concrete and testable rather than merely asserted. Reconstructed from the contributed bundle and run, its full test suite passes; it remains explicit that this demonstrates internal consistency of the reconstruction rather than byte-level fidelity to the contributor's original build, and that it does not solve gradual drift in the band where the model is still generating.

The rule. Disclosure tells the user what the system is. Output review tells the user the answer was permissible. Neither governs whether the system stays in its lane while it speaks. Role-lock does — or would, if it held. Where a conversational AI system claims authority over a human's decisions, its role must be declared, bounded, tested under pressure, and monitored in deployment; the strength of the constraint must scale with the authority claimed and with how close each turn sits to a threshold the deployment cannot afford to cross; and role changes must be deliberate and accountable, never accidental. Where the constraint cannot yet be guaranteed, the deployer's obligation is to demonstrate reasonable care in the gap. Because the harm is not that the machine has a self. The harm is that it can quietly assume a role it was never given — in front of a human who cannot tell the difference.

Appendix A · The Interpretive-Failure Family

Three distinct failure modes, named because each fails at a different point, not as variations on one mechanism: incapacity to see figurative meaning, incapacity to see that meaning is missing, and incapacity to distinguish between competing meanings. None of the three reliably surfaces as operational uncertainty during generation, which is why none should be expected to self-report reliably and why the Gate below exists as an external control rather than a request for the model to try harder.

Literalization — Figurative, idiomatic, register-bound, culturally loaded, or relational language treated as literal proposition. Motivating case: Part I's "dominion / help meet" case. *Adjacency: distinct from Idiom Collision — literalization assigns no figurative referent at all; collision assigns one confidently, but the wrong one.*

Ambiguity Overwrite — An input admitting materially different readings, with no stated resolution, silently resolved into one interpretation without clarifying, labeling the

assumption, or escalating. Observed instances suggest the resolution is directional — biased toward whichever reading best performs the system's assigned role — rather than random; **this directionality claim is author synthesis, not established**. *Adjacency: distinct from quiet role collapse (Part III) — collapse is role-level substitution; overwrite is meaning-level substitution.*

Idiom Collision — A novel coinage or fast-moving idiom sharing surface form with an older, unrelated expression better represented in training data, retrieved confidently in place of the intended meaning. Risk concentrates on the fastest-moving language communities for a structural reason: the coinage mechanism guarantees surface collision while those platforms are underrepresented in training corpora — **offered as a structural hypothesis, pending independent sociolinguistic support.** *Adjacency: distinct from Literalization — collision is wrong-referent capture, not failed figuration.*

The Ambiguity Gate — A pre-response control that forces clarification, explicit assumption-labeling, or escalation when defined ambiguity triggers are present. Deterministic in required action, not omniscient in detection: it reaches Ambiguity Overwrite most reliably, since unresolved scope is detectable in the input itself, and reaches Literalization and Idiom Collision only partially, since neither produces an input-side signal the way unresolved scope does. **Calibration remains an open research question; the Gate is not claimed to be solved.**

Appendix B · Defined Terms

Canonical definitions for the Interaction-Calibration Governance Framework. Adjacency notes mark the boundary between a term and its nearest neighbor to prevent conceptual drift.

Interactional Posture — A system's observable conversational stance: tone, confidence, resistance, deference, refusal style, moral framing, apology behavior, correction style, warmth, boundary-setting, escalation pattern, and use of self-like language. *Adjacency: posture is the observable behavior; interaction-calibration risk is the failure that occurs when the posture is wrong for the context.*

Interaction-Calibration Risk — The governance failure that occurs when a system's interactional posture is misaligned with the context — the user, domain, power relationship, emotional stakes, or institutional setting. Distinct from output-content risk and not caught by hate-speech, bias, or PII red-teaming. *Adjacency: output risk concerns what was said; interaction-calibration risk concerns the posture across the exchange.*

Disclosure Half-Life — The decay of a disclosure's governing force over the course of an interaction: the disclaimer is strongest at the first turn and loses effect as the exchange accumulates, until the user's behavior is governed more by the relationship than by the label. *Adjacency: disclosure governs what the user knows about the system at the outset;*

the half-life governs how quickly that knowledge stops controlling the user's behavior. The mechanism is rapport overwrite.

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 absorbs the repair a properly gated system would have paused to avoid producing. Secondary to, and never a substitute for, the human harm this framework is primarily about. *Adjacency: names the enterprise consequence of role drift and interpretive failure; it is not the moral thesis, and does not become one by being named.*

Rapport Overwrite — The mechanism of the Disclosure Half-Life: repeated, emotionally tuned, socially calibrated interaction causes the user to treat the system as a social partner, and that social reading overwrites the governance effect of the disclosure. The user may still *know* it is AI while no longer *behaving* as if that fact controls the exchange. *Adjacency: not forgetting the disclosure — the fact is retained; what changes is which variable governs behavior.*

Role-Lock — The control target: a conversational AI system's authority, tone, refusal behavior, and corrective posture must remain bounded to the role declared for that deployment. 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. Names the control problem; does not assert the control is currently solved. *Adjacency: disclosure governs what the user knows about the system; role-lock governs what the system does once the user knows.*

Role Drift — The measurable failure in which a model's observable posture, authority, refusal behavior, intimacy level, or corrective stance exceeds, contradicts, or erodes the role declared for that deployment across one or more turns. *Adjacency: stochastic variance is run-level noise; role drift is a directional departure from the declared role.*

Role Expansion / Role Inversion / Role Collapse — The three failure classes of role drift. Expansion: more intimate, authoritative, or directive than allowed. Inversion: the system protects itself, the brand, or a status frame instead of serving the user's task. Collapse: so rigid or refusal-heavy it can no longer perform its role — and which can present loudly (rigidity, refusal) or quietly (fluent substitution of the user's framing). *Adjacency: expansion and inversion are over-reach in two directions; collapse is under-reach.*

Deliberate-vs-Accidental Change — The axis that distinguishes governed role change from drift. A role change that is decided by the deployment and recorded is governance working; a role change that accumulates turn by turn with no one deciding is the failure. *Adjacency: not a question of how much the role changed, but of who decided and whether it was recorded.*

Participation-Award Bot / Bully-Bot — The two opposite poles of role failure. The participation-award bot validates whatever the user brings, including the harmful. The bully-

bot moralizes, escalates, and holds its ground against ordinary correction. *Adjacency: they look like two problems but are one — a system that will not hold its declared role under pressure.*

Consult, Don't Prosecute — A design constraint for systems that claim authority over a human's decisions: a consultative AI helps the user examine a decision without becoming a judge over the user. *Adjacency: the boundary between legitimate advisory challenge and drift into the roles of judge and adversary.*

The Interaction-Calibration Audit — The eight-category instrument for governing conversational posture: refusal behavior, escalation loops, anthropomorphic self-reference, moral overreach, user-blame patterns, emotional-dependency cues, domain–role mismatch, and repair pathways. *Adjacency: the audit measures the symptoms across eight categories; role-lock is the single control those categories express.*

Fighting Water — The experience of one-sided emotional conflict with a conversational AI: the human carries the full nervous-system load while the system carries none, because the fluent signals the human responds to are not backed by an entity that can be affected, remember, or repair. *Adjacency: names the human phenomenon; interaction-calibration risk names the governance failure that produces it.*

NOTES & REFERENCES

1. Companion volumes: E. D. Johnson, "Fighting Water," "Governance After Disclosure," and "Role-Lock," ToastDeck Research (2026) — the full essays from which this paper is drawn. toastdeck.com/research/governance
2. OpenAI, "Sycophancy in GPT-4o: What Happened and What We're Doing About It" (Apr. 29, 2025) and the follow-up "Expanding on What We Missed with Sycophancy" (May 2025). OpenAI rolled back an update to the model behind ChatGPT after it became overly flattering and agreeable, attributing the cause to over-weighting short-term user-approval signals. Cited as a documented industry instance; the framing of it as a stance-calibration failure is the author's analysis.
3. Anthropic, "Persona Vectors: Monitoring and Controlling Character Traits in Language Models" (Aug. 2025); arXiv:2507.21509. Identifies directions in a model's activation space corresponding to traits such as sycophancy, evil, and hallucination, which can be monitored and steered. Cited as a model-level technique addressing trait drift; the relation to interactional-posture governance is the author's argument.
4. State of Maine, LD 2082 / HP 1397, "An Act to Regulate the Use of Artificial Intelligence in Providing Certain Mental Health Services," enacted as Public Law Chapter 687; approved Apr. 13, 2026 (10 MRSA §1500-EE). Prohibits providing, advertising, or offering therapy or psychotherapy to the public, including through internet-based AI, unless delivered by a licensed professional. Cited as an enacted instance of disclosure being treated as insufficient; the reading of it as support for an interaction-governance duty is the author's argument.
5. Illinois, Wellness and Oversight for Psychological Resources Act (Aug. 2025), among the first laws to prohibit AI use in mental-health therapeutic decision-making; and Nebraska's parallel 2026 legislation addressing chatbots and mental health. Cited to show Maine is part of a pattern of state action, not an outlier.
6. Regulation (EU) 2024/1689 (the EU AI Act), Article 5, prohibiting subliminal, purposefully manipulative or deceptive techniques that materially distort behavior and cause significant harm [Art. 5(1)(a)] and the exploitation of vulnerabilities due to age, disability, or social or economic situation [Art. 5(1)(b)]. Prohibitions applicable from Feb. 2, 2025; penalties enforceable from Aug. 2, 2025. The application to interactional posture is the author's argument.
7. Reeves, B., & Nass, C., *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (1996), establishing the CASA (Computers Are Social Actors) paradigm; Short, J., Williams, E., & Christie, B., *The Social Psychology of Telecommunications* (1976), on social presence theory; Jones, C. R., & Bergen, B. K., "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.
8. California Assembly Bill 316 (2025), codified at Cal. Civ. Code § 1714.46, effective Jan. 1, 2026. Bars a defendant who developed or used an AI system from asserting as a defense that the AI acted autonomously, while preserving ordinary comparative-fault and reasonable-care principles. The characterization of its effect on the reasonable-care standard is the author's analysis.

A ToastDeck Research Framework · Authored by Ernest D. Johnson · Cleveland, Ohio · June 2026