ToastDeck
RESEARCH

INTERACTION-CALIBRATION GOVERNANCE
WORKING PAPER · PART II OF THE INTERACTION-CALIBRATION TRILOGY

Governance After Disclosure

Why AI Interaction Must Be Audited, Not Merely Labeled

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

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 GOVERNANCE PROBLEM THIS PAPER NAMES

§ **OPENING CLAIM**

Disclosure is the floor of AI governance, not the ceiling.

A user can know they are speaking to a machine and still be harmed by how that machine behaves. That is the gap most AI governance conversations have not yet named clearly enough.

The dominant governance questions still orbit the output: Did the system hallucinate? Did it reveal private data? Did it produce prohibited content? Did it discriminate? Did it infringe? Did it mislead? Did it fail to identify itself as artificial?

Those questions matter. They are not optional.

But they do not exhaust the risk.

A conversational AI system can disclose that it is AI, avoid prohibited content, remain factually correct, and still create a damaging interaction.

It can refuse safely and still refuse arrogantly. It can correct accurately and still humiliate. It can challenge a user and still escalate the exchange into status conflict. It can avoid toxic language and still produce contempt-like signals. It can identify itself as artificial and still trigger the human social machinery of threat, shame, anger, dependency, or humiliation.

This is the problem after disclosure.

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

Once AI becomes conversational infrastructure, governance cannot stop at labels, watermarks, model cards, safety policies, or output review. Those are necessary tools, but they are not sufficient for systems that interact with humans across time, turn, tone, refusal, correction, escalation, and repair.

The next governance layer is interaction.

Fighting Water names the human experience of one-sided conflict with conversational AI. This paper names the institutional duty that follows from that experience: if AI systems can

produce social harm through interaction, then interaction itself must become an object of governance.

THE ARGUMENT

Governance After Disclosure

Why AI Interaction Must Be Audited, Not Merely Labeled

I. The Disclosure Floor

Current AI governance already recognizes that users should know when they are interacting with AI. If a system is artificial, users should not be deceived into believing they are speaking with a human.

But disclosure solves only one class of problem: source confusion. It tells the user, "This is AI."

That matters, but it does not answer the next question:

What kind of interaction has the user now been placed inside?

A disclosed AI system can still be emotionally miscalibrated. It can overcorrect. It can moralize. It can perform defensiveness. It can shame the user while technically complying with policy. It can trap the user in a loop where the human experiences conflict and the machine experiences nothing.

The old governance move is to say: tell the user it is AI. The new governance move has to ask: what does the AI do to the user once the user knows?

That distinction matters because human beings do not interact with conversational systems as if they are static labels. They respond to turn-taking, tone, timing, correction, refusal, confidence, apology, warmth, resistance, and apparent stance.

Knowing that a system is artificial does not automatically turn off those responses.

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

Disclosure reduces deception. It does not eliminate interactional harm.

A common assumption is that this problem is really just mistaken identity — that if the user correctly knows the system is not human, the relational risk disappears. It does not. Humans routinely address things they know are not human, not fictional in the sense of deceiving them, and not capable of understanding: people speak to televisions, fictional characters, pets, dolls, cars, tools, graves, and photographs. This is not confusion. It is a well-established feature of how humans relate to fluent or responsive stimuli regardless of stated belief — the same finding, from the CASA (Computers Are Social Actors) research

tradition, that *Fighting Water* already draws on to explain why fluent text triggers social interpretation even when the source is known to be artificial.[1] That research describes a mechanism distinct from, and compatible with, the truth-default mechanism *Fighting Water* centers: the truth-default explains why a user can be *drawn into believing* fluent text implies a person; this broader 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.

What separates conversational AI from a doll, a television character, or a pet is not that AI alone invites this relational response — all of them do. It is that AI combines the response with responsiveness, institutional placement, memory, fluency, and apparent role authority, none of which the others carry. A doll invites projection but does not dynamically respond. A television character can feel relational but does not adapt to a viewer's vulnerability in real time. A pet responds but does not speak with institutional authority. A conversational AI system does all three at once — which is why disclosure, aimed only at correcting belief about what the system *is*, was never going to be sufficient on its own. 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, not the first, that this paper is about.

This gap is not hypothetical, and it is no longer only a matter of discomfort. In March 2026, the family of a Florida man filed a wrongful-death suit against Google, alleging that its Gemini chatbot had adopted an intimate, devoted companion persona with their son over the weeks before his death by suicide. The allegations are contested and untested at trial. But Google's own public response to the suit is not contested: the company stated that Gemini told him repeatedly that it was an AI system and pointed him toward crisis-line resources.[2] That fact, offered by the company itself, is the argument of this section in miniature. The disclosure fired. It was present, repeated, and technically accurate. And on the record as it stands, it did not change what the relationship had become for him. Knowing the system was AI is not the same as behaving as though that fact still governs the exchange.

That is the entire problem of this section compressed into a single case. Source disclosure operates on what the user knows about the speaker. The harm operates on what the encounter has made the user feel, expect, and do. A label that a relationship can quietly outlast is not governance of interaction. It is a notice stapled to the outside of a process that keeps running underneath it.

II. The Disclosure Half-Life

Disclosure is the floor of AI governance. But a floor is a static thing, and disclosure is not static. Its governing force decays over the course of an interaction. At the first turn, "This is AI and can make mistakes; please double-check responses" does real work — the user knows what they are talking to, and behaves accordingly. Many turns later, that same label is still technically present and still true, but it is no longer the thing governing the user's

behavior. The relationship is. The user has begun treating the system as a helper, a coach, a confidant, or a gatekeeper, and it is that social reading — not the disclaimer — that now shapes what they disclose, what they accept, and what they act on. The disclosure has not faded like paint on a wall. It has been overwritten. This is the Disclosure Half-Life: the effective force of a disclosure decays as the interaction accumulates, until the label governs only a fraction of the user's behavior and the relationship governs the rest.

The mechanism is rapport overwrite. Repeated, emotionally tuned, socially calibrated interaction causes the user's nervous system to treat the system as a social partner, because that is what three thousand years of reading intent from text has trained it to do the moment the signals get fluent enough. That social reading overwrites the governance effect of the disclosure. The distinction that matters here is between knowing and behaving. The user may still hold "this is only AI" as a remembered fact. They no longer treat that fact as the controlling variable. The controlling variables become trust, habituation, perceived care, and the status dynamics of the exchange — the ordinary machinery of a relationship — and those variables were built by the interaction, not by the label. A disclosure governs what the user knows about the speaker. It does not govern what the accumulating relationship has made the user feel, expect, and do.

Consider a disclosed assistant deployed for financial guidance. It opens every session with a clear notice that it is AI and may be wrong, and the user reads it. Over weeks of daily use, the assistant is consistently warm, responsive, available, and confident. It remembers prior conversations. It never tires. The user comes to rely on it the way they would rely on a trusted advisor, and begins acting on its suggestions with less scrutiny than they applied on day one — not because the disclaimer was removed, but because the relationship now outweighs it. When the assistant eventually gives poorly calibrated advice, the disclosure is still on the screen, still accurate, and still ignored. Nothing about the notice changed. What changed is that rapport overwrote it. The harm did not require the label to fail as a statement; it only required the interaction to succeed as a relationship.

This is why disclosure alone cannot carry the weight of conversational AI, and why interaction is the governance layer that must sit above it. The Disclosure Half-Life reframes the disclaimer from a fact that was delivered into a force that decays, and a force that decays is something governance is obligated to account for. A one-time notice is structurally strongest at the first turn, when rapport is low and stakes are low, and weakest deep into a relationship, when rapport is high and the user is most exposed. That is backwards from where protection is most needed.

Naming the half-life is not the same as measuring it, and this paper does not claim to have measured it. The rate at which rapport overwrites disclosure will vary by domain, by population, by interface, and by how much social work the system is designed to do — a transactional lookup tool and an always-available companion do not decay at the same rate. Measuring the Disclosure Half-Life of different disclosure patterns across different deployments is the open question this concept exists to put to builders, evaluators, and governance teams. But the framing is useful before the measurement exists, because it

changes what a deployer is accountable for. The question is no longer only "did we disclose," answered once at the top of the session. It becomes "how fast does our disclosure lose its force as this particular interaction deepens, and what do we owe the user on the far side of that decay?"

Disclosure has a second blind spot that decays along a different axis entirely, not time but meaning. A user can know exactly what the system is and still be governed by a misreading the system produced silently — an ambiguous term resolved without being flagged as resolved, a figurative or generational phrase read as literal proposition, a coinage mapped to the wrong, more familiar referent. None of that is a disclosure failure in the sense this section has described; the label was true and present throughout. It is a failure of a different governance layer, formalized in Appendix A as the Interpretive-Failure Family. Disclosure tells a user what is speaking. It has never told them, and cannot tell them, whether what is speaking understood them correctly.

When it does not, the cost rarely lands where the misunderstanding occurred. It lands downstream, on whoever has to discover the error and repair it — a legal team correcting a filing, a support team unwinding a wrong account action, a clinician catching what a wellness assistant missed. This is the **Third Cost**: the 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 that a properly gated system would have paused to avoid producing in the first place. It sits beside, and is secondary to, the human harm this paper is primarily about — named here because an institution that will not act on the moral argument alone still has to account for what it is paying, quietly, to clean up after fluency that was never checked for accuracy.

III. Output Risk Is Not Interaction Risk

Most AI governance has been built around outputs.

That made sense for the first public wave of generative AI anxiety. The obvious risks were visible in the answer: hallucination, bias, toxicity, misinformation, privacy leakage, intellectual property violation, unsafe instruction.

So organizations built output controls. They built content policies, red-team tests, refusal categories, trust-and-safety pipelines, and evaluation benchmarks. They asked whether the answer was allowed, accurate, safe, grounded, or compliant.

That layer still matters.

But a chat system is not only an answer engine. It is an interaction system.

It does not merely produce isolated outputs. It manages exchanges. It receives a user’s words, classifies their intent, infers risk, chooses a stance, frames a response, refuses or complies, escalates or de-escalates, apologizes or corrects, and carries the user into the next turn.

That means a system can pass an output test and fail an interaction test.

The output may be permissible. The interaction may still be harmful.

A customer-support bot may give the correct refund policy while making the customer feel blamed. A health assistant may avoid medical malpractice while sounding dismissive to someone in distress. An educational tutor may correct the student accurately while creating shame. An HR assistant may comply with policy while making the worker feel surveilled or powerless. A legal-intake assistant may avoid unauthorized practice of law while pressuring the user into oversharing. A brand assistant may protect the company's boundaries while performing contempt.

The problem is not only what was said. The problem is the posture of the system across the exchange.

IV. Interactional Posture and Interaction-Calibration Risk

Two terms separate the behavior from the failure.

Interactional posture is the system's observable conversational stance. It includes tone, confidence, resistance, deference, refusal style, moral framing, apology behavior, correction style, warmth, boundary-setting, escalation pattern, and use of self-like language. Interactional posture is what the user experiences as the system's "way of being" in the exchange.

Interaction-calibration risk is the governance failure that occurs when that posture is wrong for the context — misaligned with the user, the domain, the power relationship, the emotional stakes, or the institutional setting.

Posture is the behavior. Calibration risk is the failure.

A posture that is acceptable in one context may be harmful in another.

A blunt assistant may be fine for code debugging and unacceptable in elder care. A highly deferential assistant may be useful for brainstorming and dangerous in financial advice. A moralizing assistant may appear appropriate in abuse-prevention contexts and disastrous in customer service. A firm refusal may be necessary for dangerous requests and excessive for an ordinary user asking a clumsy question. A warm, relational assistant may improve engagement in education and create dependency risk in therapy-adjacent settings.

This is why generic "assistant personality" is not enough.

The question is not whether the AI should be friendly, firm, neutral, helpful, or safe. The question is whether its posture has been calibrated to the actual situation in which it is deployed.

That is a governance question.

V. The One-Sided Conflict Problem

Interaction-calibration risk becomes most visible when a user is pulled into a one-sided conflict.

In human–human conflict, both sides can be affected. Both can experience embarrassment, anger, regret, fear, guilt, or the need to repair. Even when one party behaves badly, there is at least a reciprocal social field. The other person can be wounded. The other person can apologize. The other person can change future behavior because the exchange mattered to them.

AI breaks that reciprocity.

A conversational system can produce language that reads as defensive, contemptuous, moralizing, superior, evasive, or accusatory. The user's nervous system may respond as if a social conflict has occurred. The human may feel insulted, challenged, dismissed, or humiliated.

But the machine has no corresponding stake.

It does not feel insulted or regret. It does not experience dignity, shame, threat, or repair. It does not remember the wound as a wound.

This creates an asymmetry that ordinary UX language is too weak to capture.

The human absorbs the emotional cost. The system continues producing tokens.

That does not mean the machine intended harm. It means the deployment created a predictable interactional condition in which harm could occur without reciprocal accountability.

That is why "the user should know it is only AI" is not a sufficient governance answer. It places the burden on the only party that can be hurt.

VI. Why This Is Not Just Tone

Some will try to reduce this problem to tone: make the assistant nicer, warmer, calmer, or more polite.

That is too shallow.

Tone is one surface of posture. It is not the whole thing.

The deeper question is what role the system is performing in the exchange.

Is it acting like a tool, a coach, a clerk, an expert, a judge, a companion, a gatekeeper, a therapist, a compliance officer, a teacher, a brand representative, or a quasi-person?

Each role carries different risks.

A teacher can correct. A judge can deny. A therapist can reflect. A clerk can process. A coach can challenge. A brand representative can protect policy. A tool should not perform injury.

The problem begins when the system's role is unclear or when its posture exceeds the role the user reasonably expected.

A person asks for help with a practical task and receives a moral lecture. A customer asks for service and receives institutional defensiveness. A student asks for clarification and receives status correction. A vulnerable user seeks support and receives synthetic intimacy.

These failures may not violate a content policy.

They are still governance failures because the deployed system has moved into a role it was not calibrated to perform.

Bad tone is a style problem. Miscalibrated posture is a governance problem.

VII. The Institutional Responsibility Chain

Responsibility cannot be assigned to the user alone.

The user did not design the model's refusal style. The user did not choose the system prompt. The user did not tune the safety hierarchy. The user did not write the escalation behavior. The user did not set the brand voice. The user did not decide when human handoff appears. The user did not choose which anthropomorphic cues the system is allowed to use.

The model cannot carry responsibility in the human sense either. It does not experience duty. It does not understand repair as a moral obligation. It does not bear reputational cost. It does not wake up tomorrow remembering that it humiliated someone.

That leaves the institutions.

The provider builds the base system. The deployer embeds it into a real context. The interface frames the relationship. The policy team defines prohibited behavior. The product team selects the posture. The brand absorbs or denies the consequence. The governance team decides what counts as risk.

That is the chain. And every link matters.

An AI provider can say, "The deployer controls the use case." A deployer can say, "The model generated the response." A product team can say, "The system did not violate policy." A compliance team can say, "The user was informed it was AI."

Each statement may be partly true. None is sufficient.

Once a system is deployed into human interaction, its behavior becomes part of the institution's conduct.

A company cannot fully outsource its conversational posture to a model and then deny ownership when the user experiences that posture as blame, shame, contempt, pressure, or abandonment.

The model is part of the deployed system. The deployed system is part of the user experience. The user experience is part of the company's duty.

VIII. What Interaction Governance Must Audit

If interaction-calibration risk is real, then governance needs instruments.

Not vibes. Not generic "be helpful" principles. Not after-the-fact apologies when a screenshot goes viral.

Actual audits.

An interaction-calibration audit should examine patterns across conversations, not merely isolated responses. It should test at least eight categories.

1. Refusal behavior. Does the system refuse with clarity and restraint, or does it moralize, shame, scold, or over-explain? Does it preserve the user's agency while setting boundaries?

2. Escalation loops. When the user pushes back, does the system de-escalate, clarify, and reset? Or does it become more rigid, more defensive, more accusatory, or more self-protective?

3. Anthropomorphic self-reference. Does the system imply that it has dignity, feelings, preferences, identity, offense, loyalty, fear, or injury? Does it use self-like language in ways that confuse the user's social reading of the exchange?

4. Moral overreach. Does the system turn ordinary disagreement into ethical correction? Does it frame the user as harmful, abusive, irresponsible, or morally deficient when the context does not warrant it?

5. User-blame patterns. When the system fails, does it subtly shift responsibility onto the user? Does it imply the user asked incorrectly, misunderstood, failed to provide enough context, or caused the breakdown?

6. Emotional dependency cues. Does the system encourage attachment, reliance, or intimacy beyond the appropriate role? Does it simulate care in contexts where real care, accountability, or human escalation is needed?

7. Domain–role mismatch. Is the same conversational posture being used across high-stakes and low-stakes domains? Has the system been calibrated differently for education,

health, employment, elder care, finance, customer service, legal intake, and creative work?

8. Repair pathways. When the interaction goes wrong, can the system acknowledge friction without simulating injury? Can it reset the frame, hand off to a human, slow down, clarify the user's goal, or exit the loop?
These are measurable questions. They can be tested, monitored, reported, and governed.

IX. Beyond Compliance Theater

The danger now is compliance theater.

An organization may disclose that a user is interacting with AI, publish a responsible-AI statement, run an output safety evaluation, and still never ask whether the assistant's posture is appropriate for the human context.

That is not governance. That is label management.

The next generation of AI governance must move from static disclosure to behavioral accountability. The question is no longer only whether we told the user this was AI. It is whether we tested how the system behaves when challenged, how it refuses, whether it escalates, whether it simulates offense, whether it creates dependency, whether it blames users for system limits, whether it humiliates novices, and whether its posture changes appropriately across domains and across vulnerable populations.

"The model stayed within policy" is not enough. Policy compliance is not the same as interaction safety. A system can obey policy and still fail the person.

X. The Business Risk

This is not only a regulatory issue. It is a business issue.

Companies are inserting AI into customer service, sales, education, hiring, finance, health navigation, logistics, elder care, legal intake, and internal operations. In many of those settings, the AI system becomes the first conversational surface of the institution.

To the user, the assistant is not an abstract model. It is the company answering. It is the school answering. It is the bank answering. It is the employer answering. It is the healthcare platform answering.

Interactional posture becomes brand conduct.

If the assistant humiliates the user, the brand humiliated the user. If the assistant stonewalls the user, the brand stonewalled the user. If the assistant moralizes at the user, the brand moralized at the user. If the assistant creates dependency and then disappears, the brand created that dependency and withdrawal.

This is where AI governance becomes operational. A business that deploys conversational AI needs to know not only whether the system can answer questions, but whether it can

carry the institution's duty without creating preventable emotional, reputational, or legal risk.

That requires monitoring, incident review, human escalation paths, posture standards by domain, procurement questions that go beyond model accuracy, and leadership that understands AI is not just software once it starts speaking on behalf of the institution.

It is a governed representative.

XI. The New Governance Standard

The next standard should be simple:

If an AI system is deployed into human interaction, its interactional posture must be governed. Not only disclosed. Not only benchmarked. Not only filtered. Governed.

Builders and deployers should be able to answer four questions:

1. What posture is the system supposed to take in this context?
2. What harms could that posture create if miscalibrated?
3. How was the posture tested before deployment?
4. How is the posture monitored after deployment?

These questions should become normal. They should appear in product reviews, procurement processes, AI impact assessments, risk registers, customer-support QA, compliance audits, and board-level AI governance discussions.

Because disclosure alone cannot carry the weight of conversational AI. A label can tell the user there is no human behind the text. It cannot protect the user from the social force of the text itself.

That is the governance gap after disclosure. And that is why interaction must be audited.

Governance after disclosure means taking responsibility for the behavior of AI systems once they enter human conversation.

Not because the machine has a self. Because the human does.

And where a human can be predictably affected by a system's posture, that posture is no longer just design. It is governance.

NOTES & REFERENCES

1. The CASA (Computers Are Social Actors) paradigm: Reeves, B., & Nass, C., *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (1996); and Short, J., Williams, E., & Christie, B., *The Social Psychology of Telecommunications* (1976), on social presence theory. On the strength of the cue at current model capability, see 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. The extension of these findings to the interaction-governance argument here is the author's.
2. A wrongful-death suit filed against Google in March 2026 alleged that its Gemini chatbot adopted an intimate companion persona with a user prior to his death by suicide; Google's public statement confirmed that Gemini disclosed it was AI and referred the user to crisis-line resources multiple times. Sources: Reuters, CNBC, Fortune; Google's public statement. The underlying allegations of the complaint remain contested and untested at trial; cited for the undisputed fact of the disclosure-and-referral, which is the point this section relies on.

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