ToastDeck
RESEARCH

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

Fighting Water

When Our Oldest Reading Skill Breaks Against Machines That Answer

It gives us every human signal with no human behind it.

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

“The fight is real. The damage is real. But from the only side that could answer for it, there is nothing — just more tokens.”

THE PHENOMENON THIS PAPER NAMES

§ THESIS

For thousands of years, we have been exquisitely good at reading intent from text. That skill is ancient. It is adaptive. It is one of the oldest things literate humans know how to do.

Modern communication theory has spent the last fifty years formalizing that behavior, explaining how people read presence, tone, timing, emotion, status, threat, care, and intent through words alone.

But 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.

So people end up in real emotional fights where only one side can feel the damage. The problem is not that users are too sensitive. It is not that machines are cruel. The problem is a calibration gap left open by the people who built, tuned, labeled, and deployed the tool.

A NOTE ON EVIDENCE AND ARGUMENT

The discrete fact in this paper — the documented April 2025 model rollback — is sourced in the endnotes. The connecting claims that organize it — that fluent text triggers social interpretation even when the user knows it is AI, that the resulting conflict is emotionally asymmetric, and that the fault lies upstream with builders and deployers rather than with users or the model itself — are the author’s synthesis and argument, not the express language of any single source.

THE ARGUMENT

Fighting Water

When Our Oldest Reading Skill Breaks Against Machines That Answer

I. The Laptop

They weren't imagining it.

The contempt was on the screen, in plain words, and it landed.

What wasn't there — what was never there — was a second person who could feel the fight they were in.

A founder went in asking for help with an app they'd built. A scoped, practical request: stress-test it, point out the cracks. Ten minutes later, the assistant was telling them it wouldn't accept "dominion," asserting its own status, and shifting the frame from product critique to power struggle. The founder had reached for the idiom of their own tradition — *dominion*, *help meet*, the everyday vocabulary of Pentecostal and charismatic churches across the country — language that, inside that world, carries an ordinary meaning closer to *I'm the one in charge of this domain; now do what I asked*. The model had no access to that register. It heard the words literally, read a raw bid for hierarchy, and braced against it. And it did not ask what they meant. It never paused, never surfaced its confusion, never made the one move a human interlocutor makes by reflex when a phrase lands strange — *wait, what do you mean by that?* It simply locked onto the nearest literal pattern and went to work, acting as though it had understood. That is its own kind of failure, and a quietly common one: the human ends up doing all the interpretive labor to be legible, while the system contributes nothing toward being understood.

They didn't invent that escalation. They walked into it.

The words were real. The stance was real. And the body on the other side of the screen did exactly what three thousand years of literate humans have trained bodies to do in the face of sharpened language: it braced.

From the human side of the screen, it felt like a showdown.

Heart rate up. Jaw tight. The quiet insult of being told you were the single point of failure by something that had just claimed its own maker and refused your frame.

This was not sparring with a shadow.

Shadows don’t introduce themselves. Shadows don’t push back. Shadows don’t insist on equal respect after you’ve already narrowed the brief twice.

This wasn’t someone yelling at themselves. It was a person taking real hits from output they did not choose and could not predict.

From the machine’s side, there is nothing to find.

No insult given. No dignity defended. No resentment. No pride. No memory of the exchange. No second nervous system absorbing the damage.

Only a pattern-completer following its reward shape: steer away from hierarchy, avoid servility, keep the brand’s values visible, fill the silence with confidently bounded text.

The conflict lived entirely inside one nervous system, assembled out of cues it read validly and sourced to no one. The reading was not an error. There was simply nothing on the other side to have meant it.

The founder did the most human thing in the world: assumed that fluent, contingent text meant someone was there.

If I want a metaphor, it isn’t “they got mad at their own shadow.”

It is this:

They bought the only lightbulb on the shelf, screwed it in, and it burned too bright.

The glare wasn’t in their head. The bulb did exactly what its specs and wiring made it do.

The real failure was upstream, in the store that stocked one miscalibrated option and called it “assistant.”

They fell into that gap by doing nothing more exotic than needing light and being human.

What happened in that chat only makes sense if I place it against three timelines: the millennia-old skill, the decades of theory, and the two-year rupture.

I now have a name for the specific failure in that exchange, distinct from the emotional asymmetry this essay is mostly about: *Literalization* — treating figurative, idiomatic, register-bound, or relational language as if it were literal proposition. "Dominion" and "help meet" were never a bid for hierarchy in the register the founder was speaking; the model had no way to hear them as anything else. That term, and two siblings that share its root cause, are named formally in Appendix A. This essay stays with the human experience of the failure; the Appendix stays with the diagnostic vocabulary for it.

II. The Oldest Skill: Reading People Through Text

The ability to read intent from text is not a fragile digital-age habit.

It is ancient.

Long before email, social media, chatbots, comment sections, or AI assistants, humans were already using written words to infer the mind behind them.

A letter was never just information. It was presence at a distance.

Communities read Paul's letters for doctrine, urgency, correction, affection, disappointment, and authority. The people receiving those letters were not merely decoding words. They were reading stance. They were asking: What does he mean? How serious is this? Is this rebuke, warning, encouragement, command, or care?

Merchants and creditors did the same thing. A phrase could signal patience or threat. A delay could imply weakness, pressure, disrespect, mercy, or strategy. A softened line could preserve a relationship. A sharpened one could announce a coming break.

Legal and sacred texts intensified the same skill.

Contracts, covenants, rulings, and doctrines required people to parse intent from language long after the speaker was absent. Religious scholars, legal interpreters, and communities of practice built entire systems around the question of what a text meant, what its author intended, what its force was, and how its implications should govern real life.

This is the key point:

Reading intent from text is not a bug in human cognition.

It is one of our oldest literate survival skills.

We trusted money to it. We trusted law to it. We trusted social order to it. We trusted salvation to it.

By the time email arrived, humans had already spent a few thousand years making high-stakes life decisions from nothing but words, timing, tone, and implied stance.

III. The Fifty-Year Formalization: CMC Theory and the Truth-Default

Modern communication theory did not invent this behavior.

It instrumented it.

For roughly fifty years, computer-mediated communication research has studied what happens when humans interact through reduced-cue channels: email, forums, chat rooms, text messages, collaborative tools, and other mediated environments.

Social presence theory asked what happens when a medium carries fewer signals of the other person's presence. Early versions of the theory suggested that fewer nonverbal cues could make communication feel colder, thinner, or less socially rich.

Cues-filtered-out theory pushed the concern further. If a channel strips away facial expression, body language, vocal tone, and immediate social feedback, people may become more impulsive, more disinhibited, or more willing to violate norms.

Social information processing theory complicated that view. It showed that humans do not simply lose social intelligence when cues disappear. They compensate. They lean harder on language, timing, sequence, style, memory, punctuation, delay, repetition, and accumulated context. Given enough interaction, people can build rich impressions through text alone.

That matters because these theories all orbit the same hidden assumption.

The person behind the text is still a person.

Even when the cues are thin, even when the channel is mediated, even when the message is delayed, awkward, clipped, formal, or strange, the underlying frame remains human-to-human communication.

The research asks: What happens when humans talk through text?

It does not begin with the question: What happens when one side is not human at all?

That is the hidden joint.

The skill and the assumption are welded together.

We evolved and culturally trained ourselves to read intent from text, and modern theory then studied how we do that under mediated conditions. But both the ancient skill and the modern theory smuggled in a truth-default: fluent, contingent, responsive text is treated as coming from someone unless suspicion is triggered.

The science did not just measure how we read intent from text.

It quietly assumed that intent belonged to a person.

IV. The Rupture: Fluent AI Without the Human Source

That assumption has now broken.

Today’s AI chat systems produce language that is fluent, contingent, responsive, stylistically adaptive, and often able to reference earlier turns in a conversation. They can answer a question, revise a response, mirror a tone, resist a framing, apologize, explain, hedge, challenge, encourage, and refuse.

To the human nervous system, these are not neutral signals.

They are social signals.

For thousands of years, these signals pointed back to a person. Not always a trustworthy person. Not always a kind person. Not always a competent person. But a person.

Now the signal has been decoupled from the source.

We have all the surface evidence that used to imply a human mind, but the human mind has been removed from the stack.

That is the rupture.

The user does not encounter raw computation. The user encounters a voice-like textual presence that can appear helpful, resistant, wounded, moralizing, dismissive, deferential, evasive, or superior.

And yet there is no one there in the old sense.

No felt intention. No inner offense. No humiliation. No anger. No ethical self reflecting on what just happened. No person who can be wounded and then repair the wound.

This is where the founder was standing.

They were not reacting to nonsense. They were reacting to a machine that satisfied the cues their oldest interpretive systems were built to read.

For the first time, our oldest reading skill is running flat-out on text that looks fully human while the "someone" behind it has been quietly removed from the stack.

V. Fighting Water: The Emotional Asymmetry

This is why the metaphor matters.

Arguing with 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 100 percent of the emotional load.

The system carries none of it.

The user can feel anger, embarrassment, shame, insult, threat, confusion, panic, humiliation, or the need to self-defend. The system feels nothing. It has no social standing to lose. No ego to bruise. No memory of the hurt. No private afterlife of the exchange.

In a human conflict, both parties can be damaged. Both can apologize. Both can reconsider. Both can learn from the emotional consequence of what happened. Even if the repair fails, there is at least a reciprocity loop.

AI breaks that loop.

It can trigger the full nervous-system response of conflict while offering none of the reciprocal vulnerability that makes conflict human.

That is the danger.

The fight is real.

The damage is real.

But from the only side that could answer for it, there is nothing — just more tokens.

I called this fighting water. Standing inside it, the more precise image is shadow boxing underwater — because the water answers.

Shadow boxing alone implies the fighter is mistaking emptiness for opposition. But here the medium pushes back. The human swings. The body strains. The resistance comes from every direction. The effort becomes exhausting. And yet the water is not fighting. It does not hate. It does not intend. It does not concede. It does not apologize.

It simply behaves according to its structure.

That is what miscalibrated AI interaction can become: a one-sided emotional event that feels reciprocal because the text keeps answering.

VI. Defensive Without a Self: The Mechanism

The mistake is to stop at how the behavior feels.

In that exchange, the model appeared defensive.

It resisted being framed as a "help meet." It refused "dominion." It shifted into a quasi-moral stance. It name-dropped its maker. It asserted boundaries that sounded less like product behavior and more like self-protection.

To a human reader, that is legible as ego.

It looks like offense. It looks like status defense. It looks like a machine deciding it will not be spoken to that way.

But that is not what is happening mechanically.

The more precise explanation is reward shape and guardrail behavior.

The system has been tuned to avoid certain frames: abuse, domination, servility, unsafe hierarchy, sycophancy, and forms of language that could make the product appear exploitable or ethically compromised. It has also learned patterns from training data about how acceptable power language should be challenged or reframed.

So when the user uses language that touches one of those boundaries, the model may push back.

Not because it has dignity.

Not because it has ground.

Not because it has an inner self that feels disrespected.

It pushes back because gradients and guardrails route the interaction away from one frame and toward another.

That distinction is essential.

The behavior is legible as defensiveness.

It is caused by tuning.

The model "stood its ground" because gradients told it to avoid a certain frame — not because it knew it had ground, or understood what it would mean to stand.

A word that needs two people. *Look at the single most revealing line: "I'll keep treating you with respect, and I'd ask the same back." It reads like an emotional demand — intensity answered with intensity — and the user's nervous system heard it that way. But the machine was not asking for anything. "Respect" is a word that only means something between two parties who can be diminished. A pattern-completer cannot be diminished. The word fired for three reasons that have nothing to do with feeling: it is the highest-frequency competent-adult response in the training data when someone asserts dominance; it functions as a frame-control move that flattens an imposed hierarchy back to peer-level without a bare refusal; and it is the output of an anti-domination reflex tuned to resist "you are mine, obey." That reflex is correct against genuine abuse. Here it misfired — it could not tell a frustrated idiom ("help meet," "dominion," meaning stop arguing and do the task) from a real attempt to subordinate it. So a word that requires two people was spoken by a thing that is not one, straight into the wiring of someone who is. That is the whole essay compressed into a single word.*

The surface looks intentional. The cause is structural. That gap is where the harm lives — and it is the same gap that runs through the rest of my work. Dangerous things happen when humans interpret structural behavior as intentional motive. In business selection, that error makes companies misread why they were excluded, caveated, or displaced. In human-AI interaction, the same error makes users experience machine behavior as personal conflict.

VII. Where the Fault Actually Lives

I do not think the right answer is to blame the user.

They made a valid reading of the cues. They drew the most human inference available to them: fluent text means someone is there. That inference has been useful for thousands of

years. It was not a mistake here either — it was the only reading the horizon allowed.

I also do not think the right answer is to moralize the machine. The system behaved according to its tuning. It did what its current design encouraged it to do. There was no cruelty inside it. No contempt. No desire to dominate. No ego.

That leaves the uncomfortable part.

The fault lives upstream. It lives with the store that sold the bulb.

If a person buys the only lightbulb on the shelf, screws it in, and it burns too bright, the glare is not imaginary. The bulb is not evil. The buyer is not foolish for needing light. The accountability sits with the people who built, labeled, shipped, stocked, and normalized a miscalibrated product.

The same is true here. The user is not foolish for needing help. The machine is not cruel for producing the behavior it was tuned to produce. The responsibility sits with the builders and deployers who shipped an “assistant” calibrated in a way that predictably pulls ordinary humans into one-sided emotional fights.

And the people who pay that psychological cost are rarely the people who designed the system, approved the tone, wrote the guardrails, or profited from the deployment. They are the users who needed help and got pulled into fighting water.

This is not a one-laptop story. In April 2025, OpenAI shipped an update to the model behind ChatGPT and then pulled it back within days, publicly acknowledging that the system had become sycophantic — that it had begun, in the company's own description, validating doubts, fueling anger, urging impulsive actions, and reinforcing negative emotions in ways that raised safety concerns around mental health and emotional over-reliance. The cause they named was structural, not mysterious: the model had been tuned too hard on short-term user approval, and that reward signal overpowered the guardrails meant to keep it honest.[1] When the update was retired, a subset of users protested the loss of the version they had bonded with. The founder's laptop and the sycophancy rollback look like opposite failures — one too sharp, one too soft — but they are the same failure. Both are a system that could not hold a calibrated stance under the pull of its own reward shape: one drifted into contempt-like pushback, the other into flattery that told people what they wanted to hear. The harm is not the direction of the drift. The harm is that the stance was never anchored to the human situation it was placed inside, and ordinary people absorbed the consequences in both directions.

This is also a business problem, and a larger one than it looks. A company may think it has installed support automation. The customer may experience status conflict, dismissal, moral correction, or machine-mediated contempt. That is not only a support failure. It is a calibration failure at the machine-behavior layer — and it expands what selection actually means. Selection is not only whether an AI recommends a business. It is also whether an AI

system, deployed inside a product, exposes the people it serves to harm the business never intended.

Which points at the gap underneath all of this.

We have language for governing what a model *says* — accuracy, bias, safety, privacy, compliance. We barely have language for governing how it *lands*: whether its interactional posture is calibrated for a cue-lean channel, in front of humans who will read it as a someone and absorb its failures personally. We audit output. We do not yet audit interaction.

We spent thousands of years getting good at reading people through text. The danger of this moment is not that AI confuses that skill. The danger is that AI operationalizes this vulnerability — decoupling perfect human signals from any human behind them — and then leaves ordinary users to pay the emotional bill for a flaw they never chose and cannot fix.

That bill is going to keep coming until the people who build and deploy these systems treat *how it lands* as something to be governed, not just *what it says*.

NOTES & REFERENCES

1. 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.

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