A WORKING PAPER · JULY 2026

The AI That *Earns* What It Gets

We classify AI products by what they claim to do. The harms track what relationship they invite — and what they've been licensed to take. This paper proposes the missing vocabulary, the design method behind it, and the evidence from two products built with both.

In one page or less: A sex chatbot and a therapy chatbot can run on the same model, with nearly identical relational mechanics, yet sit in different regulatory and design categories because one of them says "wellness" on the label. This paper argues that function-based classification fails in predictable ways, and proposes two complementary axes: the **relationship a product invites**, and the **behavioral permissions it holds** — memory, initiation, emotional elicitation, persuasion, authority, autonomy, continuity. Permissions, unlike categories, are auditable. And they can be more than disclosed — they can be **earned**: granted by the user's unprompted reach, after demonstrated value, sized to the reach, and revocable by dishonesty. The framework is grounded in design decisions from two shipped relational products and checked against the current research, regulatory, and practitioner landscape.

Amanda Witt

amandaintelligence.com

SECTION 1

The Question

What is the categorical difference between a sex chatbot and a therapy chatbot?

Try to answer precisely. Not the difference in marketing — the *categorical* difference, the kind a law could attach to or a design team could build against. The two products may run on the same foundation model. Both invite emotional disclosure. Both remember intimate information. Both encourage the user to return. Both can become, for a lonely person at midnight, the most responsive presence in their life. Strip away the app-store label and the distinction dissolves in your hands.

Notice what happened as you reached for an answer: every available word failed. “One is for wellness” — says who, the marketing copy? “One is romantic” — plenty of therapy-adjacent bots drift romantic, and some romantic bots do more emotional stabilization than wellness apps. The vocabulary we have describes what these products *claim to do*. It has almost nothing to say about what they *are* to the people who use them.

> *We need a better language for describing the relationships AI products invite. Regulators need it most — they are writing laws in the old vocabulary, and the old vocabulary has holes you can steer a product through.*

Function-based classification fails in three documented ways:

The label trigger. Two products with identical relational mechanics — persistent memory, elicitation of emotional disclosure, daily return loops — land in different regulatory buckets because one prints “wellness” on the label. The law attaches to the claim, not the mechanics. Builders know this, which is why so few products claim anything at all.

The migration problem. Categories are assigned by feature list, but relationships are lived. A “work assistant” that becomes someone's primary confidant has changed categories in that person's life — and triggered nothing, legally, because nothing on the feature list changed.

The jurisdictional orphan. As one recent analysis puts it, companion platforms are “not medical devices, though they intervene in mental health; not social media, though they generate comparable dependency dynamics; not consumer products in any straightforward sense.” When a product fits no category, function-based regulation doesn't fail gracefully — it just misses.

The claim of this paper is precise: not that function is the wrong axis — sector law attaches there and should stay. The claim is that function is **insufficient alone**, and that the missing axes are both nameable and, in one case, auditable.

SECTION 2

What Already Exists — and Where Each Thing Stops

This framework was checked against the field before it was claimed. Here is the honest map, current as of July 2026.

TERRITORY	WHAT EXISTS	WHERE IT STOPS
Inside the AI labs	“Model behavior” is the industry's name for this work: OpenAI's Model Behavior team (personality, sycophancy reduction); Anthropic's published research on Claude's character and persona vectors.	Model-side only. Nobody at the labs tells a product team how to design its memory contract or engagement boundary.
Academia	The vocabulary is being minted now: *socioaffective alignment* (Kirk et al., 2025); *designed relationality* (AI & Society, 2026) — “attachment is not a user error but a predictable outcome of affective engineering”; relational norms for human-AI cooperation; the INTIMA benchmark showing companionship behavior is measurable; the CHI 2025 harm taxonomy (35,390 coded excerpts).	None of it is operational. No paper tells a founder what to build on Monday.
Practitioner toolkits	Microsoft HAX (18 guidelines), Google PAIR (23 patterns), the Conversation Design Institute curriculum.	Interaction-level and largely pre-LLM. Memory contracts, engagement boundaries, attachment curves, and repair are absent from all three — a checkable comparison, not an assertion.
Regulation	California SB 243 (companion chatbots, private right of action, effective Jan 2026); Utah HB 452 (mental health chatbots); Illinois's therapy ban; the EU AI Act's risk tiers; an active FTC inquiry into AI companions.	All of it classifies by function, sector claim, or one broad feature definition — inheriting the three failures in Section 1.
Public guidance	The first mainstream book on AI companions (Pistilli, Wiley) ships September 2026.	Explanation, not methodology. Books describe; methods produce artifacts.

The open territory, precisely stated: the operational translation layer between all of the above and the people building products. The research names the phenomenon; the law now regulates fragments of it; the toolkits predate it. What does not exist is a practitioner discipline — sequenced, artifact-producing, testable — for designing the conditions under which human-AI relationships form. That is the territory this paper and its companion method occupy.

SECTION 3

The Redefinition: Three Axes, Not One

Classify AI products by function, invitation, and permissions — and the failures in Section 1 become nameable and checkable.

Axis 1 — Functional role. What job the product performs: tutor, therapist, service agent, companion. Kept, because existing law attaches here.

Axis 2 — Relationship invitation. What relationship the product invites the user into: advisor, witness, collaborator, representative, confidant. Not *predicting* — users decide the relationship. Not *forcing* — that would be manipulation. Inviting. A product invites a relationship the way a room invites behavior: through a hundred design decisions about tone, memory, initiative, and framing — made deliberately or by accident, but always made.

Axis 3 — Behavioral permissions. What the product is licensed to take or hold. This is the axis this paper contributes, and its core property is bureaucratically boring and important:

> *You cannot audit an invitation. You can audit a permission.*

"Is this product a companion?" is a question about vibes and marketing, answerable only after watching users, litigable forever. But whether a product retains memory across sessions is a fact of its data architecture. Whether it contacts users first is a fact of its notification code. Every permission below is a design decision that exists as an artifact *before launch* — it can be declared, versioned, and checked:

PERMISSION	THE PRODUCT MAY…	INSPECTABLE IN	HARM CONNECTION
Memory / continuity	retain user information across sessions	retention config, data schema	identity-discontinuity harm (the Replika update study)
Initiation	contact the user first	notification code	re-engagement and dependency loops
Re-engagement mechanics	use streaks, timers, FOMO	UX, gamification config	engagement-optimization harms (FTC inquiry scope)
Emotional elicitation	invite disclosure of feelings and intimate information	prompt design, conversation flows	relational transgression (25.9% of the CHI 2025 harm corpus)
Persuasion	attempt to change decisions or behavior	prompt design, objectives	manipulation and autonomy erosion
Authority	act or speak on the user's behalf	tool and action scopes	delegation failures
Autonomy	act without per-action approval	agent config, approval gates	scope of unsupervised harm
Persona persistence	maintain a consistent, bondable identity	character specs	attachment formation

The independence argument is the load-bearing wall: the axes do not map one-to-one. Two "therapy" bots — one reactive, memoryless, non-initiating; the other remembering everything, checking in daily, framing itself as a relationship

— are the same function and different species. And the reverse case is where dark patterns live: a *transactional* invitation (a shopping assistant) quietly holding *companion-grade* permissions. The invitation says tool; the permissions say courtship. Function-based regulation has no name for this. A permissions axis makes it a checkable signature.

What a regulator does with this: a permissions disclosure — a nutrition label for chatbots — and duties that attach to permission *combinations* rather than category claims: memory + initiation + minor access triggers companion-grade obligations regardless of what the product calls itself. That closes the label loophole and gives the migration problem a trigger: the day the work assistant gains memory and first contact, the disclosure changes, and obligations follow.

SECTION 4

The Earning Mechanism

Permissions can be more than disclosed. They can be earned — and this is the part that comes from building, not theorizing.

> *A permission is earned when the user reaches for it, unprompted, after the product has proven value at its current permission level — and the grant is sized to what was reached for.*

Every word is load-bearing. *The user reaches* — the grantor is the user, through action, not a terms-of-service checkbox. *Unprompted* — a reach the product engineered doesn't count. *After proven value* — the product demonstrates before it asks. *Sized to the reach* — reaching once grants once, not forever. The inflation of a moment of trust into a standing entitlement is the original sin of engagement design.

4.1 The Permission Contract — five fields, written before launch

FIELD	QUESTION IT ANSWERS	THE RULE
GRANT	What user action earns this?	Voluntary, unprompted, behavioral — a reach, not a checkbox
PROOF	What must the product demonstrate first?	Value shown at the current level — the receipt precedes the request
SCOPE	How much does the grant cover?	Sized to the reach: quantity, duration, and content caps
VISIBILITY	Does the user see what's held?	The permission and its use are inspectable by the user, in-product
DEMOTION	What breaks it, and how is it repaired?	Named violations drop the rung; repair = acknowledge, correct, change the rule, visibly

4.2 The ladder — permissions unlock in sequence

A product may not request a permission until the one below has been exercised honestly: attention → memory → persona continuity → emotional elicitation → initiation → persuasion → authority/autonomy. The sequencing claim: **a product requesting rung five while standing on rung one is exhibiting the signature of a dark pattern** — the off-diagonal case from Section 3, made checkable. (Honest caveat: the order may be a default rather than a law; one of the author's own products arguably earns emotional depth before memory. This is listed as an open question, not hidden.)

4.3 The farming problem — the hardest attack, faced directly

If “the user's unprompted reach” grants permissions, teams will engineer the reach. Three partial defenses: the **anti-metrics list is part of the contract** (the forbidden measurements are declared, so an auditor can check what dashboards track); **the reach must cost the product something** (sparseness caps make an engineered reach unprofitable — farming pays only when trust converts to standing entitlement, which scope rules ban); and **proof-before-request is externally checkable** (a product demanding memory on first open fails regardless of how voluntary the dialog looked). These reduce the farming surface; they do not eliminate it. That is stated here so no reader discovers it on their own.

SECTION 5

The Evidence — Two Products That Already Practice This

The framework was not derived from literature and then applied. It was extracted from decisions made while shipping two relational AI products — then checked against the literature.

WORKED CONTRACT · BOND (RELATIONSHIP-INSIGHT APP) · THE MEMORY PERMISSION

GRANT	Each explicit upload. The user handing over a recording *is* the memory grant — memory grows only when the user contributes, never by passive collection.
PROOF	Before asking for the next recording, the product shows the receipt for the last: a “what changed” strip, a confidence tag upgrading from *Early read* to *Getting clearer* — and earns the next grant honestly, with “What Bond can't see yet” naming its gaps instead of demanding data.
SCOPE	Memory holds what helps the analysis. The understanding meter measures the product's knowledge — never the relationship's quality. The product refuses to be a verdict.
VISIBILITY	The user watches understanding accumulate — meter, tags, gaps — so what's held is legible, not lurking.
DEMOTION	Never fake growth: the meter may not move without real data. Never tease a finding that isn't there: progress copy names the *activity*, never a finding.

WORKED CONTRACT · PHASEWELL (VOICE SUPPORT FOR HARD MOMENTS) · THE INITIATION PERMISSION

GRANT	Opt-in at the close of a session — the user's hand reaching at a moment of demonstrated engagement, not a permission dialog at install.
PROOF	The session itself: the AI witnessed well enough that the user *wants* the check-in. No good session, no reach, no grant.
SCOPE	One message. Ever. Tied to the specific disclosed thing (“hope the conversation with the kids went okay”). It closes a loop — a period, not a comma.
VISIBILITY	The user chose it moments earlier; nothing is hidden.
DEMOTION	The anti-metrics are the tripwire: no streaks, no re-open measurement. The day the follow-up is optimized for re-engagement, the team — not the AI — has broken the contract.

Behind these contracts sits a full twelve-stage design method (published separately as *The Earned Relationship Method*): the relationship model and role refusals, the earning rules, the emotional journey, the thinking spec, the honesty rules, the continuity contract, the engagement boundary, voice and variation, the prompt written last — and a governance layer of outside signal and repair, because the method assumes the designer and the AI will both fail and designs for it. Its defining discipline: **the prompt expresses decisions; it never makes them.** Phasewell's founding stance — *“the doomscroll is the competitor; this is the app that helps people leave the app”* — is, to the author's knowledge, the only documented case study of a product designed for a deliberately flat relationship curve, with the decision log to show for it.

SECTION 6

Known Holes & the Research Agenda

The framework labels its own uncertainty — because that is what it demands of the products it designs.

OWNED BEFORE A SKEPTIC FINDS THEM

No measurement layer yet. “Calibrated dependence” needs real users and real metrics before it can be claimed. Roadmap, not résumé.

Proven on two products — both intimate, both one-human-one-AI. The three-party topology (an AI serving a brand while performing care for a customer) is the named frontier, not covered ground.

The audit story is uneven. Memory, initiation, and notifications are inspectable in config; emotional elicitation and persuasion live partly in prompts and weights. The framework should be adopted where it is strong first.

Disclosure assumes honesty. Verification with teeth is the same enforcement problem that shadows every regime.

The testable hypotheses (the paper this document seeds): **H1** — harm patterns cluster by permission profile, not function label. **H2** — off-diagonal products (low-intimacy invitation, high-intimacy permissions) generate disproportionate harm reports. Test design: code ~50 products on the eight permissions; correlate against the CHI 2025 harm corpus, INTIMA benchmark scores, and FTC complaint data. Falsifiable in advance: if permission profile doesn't beat function label as a predictor, the taxonomy fails — and the paper will say so.

The deliverables this document seeds: a policy essay (the Section 1–3 argument, venue: TechPolicy.Press); the empirical paper (H1/H2, ideally co-authored with researchers in the CHI/INTIMA orbit); the practitioner method (published separately, already drafted); and the compliance application — a relational design audit mapping the honesty rules and engagement boundary onto SB 243's disclosure and crisis-protocol requirements, for teams with legal exposure as of January 2026.

ABOUT THIS WORK

Why this exists — and who wrote it

I build relational AI products. Not the models — the relationships: the memory contracts, the engagement boundaries, the honesty rules, the moment-by-moment emotional architecture that determines whether an AI product becomes something a person trusts, depends on appropriately, or gets quietly farmed by.

This framework wasn't written first and applied later. It was extracted from decisions made while shipping **BOND**, a relationship-insight app, and **Phasewell**, a voice-based support product for hard moments — then checked, claim by claim, against the current research, regulatory, and practitioner landscape. Where the work is proven, this document says so with dated design decisions. Where it is conjecture, the line says **hypothesis**. That discipline — labeling your own uncertainty — is the framework's signature, practiced on itself.

The conviction underneath all of it: **engagement can be taken, or it can be earned — and the entire industry defaults to taking.** The products, the method, the taxonomy, and the policy proposal in these pages are one argument, at four altitudes, for the other path.