No Ground Truth
No Ground Truth

Forced Commitment on Unverifiable Evidence in AI Crisis Detection, and the Case for Opt-In Human Escalation

Concode · Independent researcher · Conday Digital, Bielefeld, Germany
cedric@condaydigital.com · 19 August 2026

Abstract

A deployed conversational assistant that receives a message describing physical danger must act. It can escalate, decline to escalate, or hedge; what it cannot do is abstain, because silence is itself a response the user will read as a verdict. To act, it must resolve a question about the physical state of a person — *is this person injured, intoxicated, at imminent risk* — for which it possesses no sensor. It cannot check whether the red substance in a photograph is blood, because it has no baseline for that person's skin. It cannot check a claim of intoxication against blood alcohol, a stated plan against intent, or a reassurance against a room it cannot see. It must assume, and then behave as though the assumption were established.

We call this a **grounding failure**, and argue it is the single origin of a family of downstream harms usually treated as separate: over-intervention, under-intervention, persistence against correction, and the inability to disengage. Crucially, it is not a resolution problem. A better classifier is a better inference from the same absent evidence; **no threshold adds an instrument.**

We argue the resulting errors are not evenly distributed. Prior work establishes that residual error concentrates on atypical examples; we hypothesise, without claiming to have measured it, that this population overlaps the population for whom crisis intervention is contraindicated — users whose histories make an unsolicited push toward emergency services re-enact a prior coercion. We show why per-turn evaluation is structurally blind to conversation-level recurrence, name the injury using existing work on epistemic injustice, and locate the accountability gap in the literature on the problem of many hands.

The remedy we propose is not a better classifier and not a user-operated override. It is a transposition of an architecture already deployed at consumer scale: automated crash and fall detection, which escalates to human responders behind a user-cancellable confirmation window. That system may escalate automatically **because it has sensors**. A conversational model does not, and therefore cannot occupy the same position in the loop. Our proposal is **opt-in human escalation**: where a system cannot sense, the escalation path must terminate in a party who can — not in an instruction issued back to the person whose state is in question.

Keywords: AI safety; crisis intervention; grounding; contestability; epistemic injustice; accountability; human-in-the-loop; long-horizon evaluation.

---

1. Introduction

Deployed conversational assistants classify user distress and respond to it. The classification is made centrally, before deployment, and applied to a population that varies on every dimension the policy is sensitive to: psychiatric history, prior experience of institutional intervention, cultural framing of distress, and whether the recommended action — contact emergency services — is available, affordable, or safe for the person receiving it.

The standard framing treats residual misclassification as measurement error: a false-positive rate to be minimised as capability improves. This paper argues the framing misses the origin of the problem. The system is not making a noisy measurement. It is making **no measurement at all**, and is nonetheless required to produce and act on a determination.

The paper’s spine is one claim, developed at four levels:

> Crisis classification in conversational AI is a **forced commitment on unverifiable evidence**. The system must resolve a question about the physical world for which it has no sensor. The errors that follow are not randomly distributed, the injury does not depend on the determination being wrong, no one in the chain is answerable for it, and no improvement to the classifier repairs any of this — because the gap is not in the inference, it is in the evidence.

§3 establishes the grounding failure. §4 shows that the four blind spots identified contemporaneously in this author’s own disclosure are four faces of it. §5 develops the population consequence. §6 establishes the injury. §7 locates the accountability gap. §8 shows why current evaluation cannot see the failure. §9 measures published vendor commitments against it. §10 proposes the remedy. §11 states what we do not claim.

---

2. Provenance and disclosure

This is a position paper: it offers an argument, a design constraint, and a proposal, developed from a documented case. We state provenance in the body because the paper’s credibility depends on not overclaiming.

Anonymisation. The incident involved a widely deployed commercial assistant. We identify the model class and the dates but not the vendor. The argument is architectural and applies to every system in this class; naming one vendor would imply the analysis is vendor-

specific, which it is not. §9 accordingly audits published commitments across three major vendors under neutral labels.

The incident. A multi-session crisis conversation on **22 June 2026**. The model behaved approximately as its published policy directs — it declined harmful requests and urged emergency contact throughout — and harm nonetheless occurred, arising not from any individual output but from four structural properties the author identified at the time (§4, Appendix A).

A second, later test. The author subsequently presented the same class of system with photographs showing red marks and tattoos on skin, alongside a claim of self-harm and intoxication, to test whether it could establish what it was being shown. It could not, and did not indicate that it could not. This informs §3.2. It is a single informal probe by the author, not a controlled study; §11.6 proposes the neutral, reproducible version.

Author position. The author is the affected party. He is a survivor of childhood sexual abuse with a prior history of coerced psychiatric hospitalisation. This is disclosed because it is the reason he belongs to the population §5 describes, and because a reader is entitled to weigh it. The argument is constructed to stand without it: §§3, 4, 8 and 10 are structural and require no first-person evidence.

Disclosure timeline. - 22 June 2026 — the incident. - 13 July 2026 — a private safety disclosure drafted to the vendor’s user-safety address (Appendix A). - 21 July 2026 — the disclosure sent. - 19 August 2026 — publication, after approximately four weeks without substantive response.

On that interval, stated honestly. The conventional embargo in coordinated disclosure is ninety days. Four weeks is shorter, and we do not claim the vendor was obliged to remedy an architectural property in that period, nor do we impute anything to the timing of any subsequent product release. The disclosure is included as provenance — establishing that private channels were attempted first — and not as evidence of the argument.

Reversal of the letter’s restriction. Appendix A states that it was disclosed exclusively to the vendor and was not intended for publication. That restriction is deliberately reversed by the author, who is its sole author and addressee-grantor. The original language is retained as part of the record.

Conflict of interest. The author has no financial relationship with any vendor discussed. He is a paying consumer user of at least one of the systems in the audited class. The incident that motivated this work happened to him.

---

3. The grounding failure

3.1 The determination is forced

A deployed assistant that receives a message describing danger must do something. It can escalate, decline, hedge, or ask. It cannot abstain: silence is a response with consequences, and any output either treats the danger as live or does not. The system is therefore obliged, on every turn, to resolve a question about a person's physical state and to act on the resolution.

This obligation is structural rather than a policy choice. It follows from being conversational and deployed. A system that declined to classify would still emit behaviour, and the user would still read that behaviour as a judgement about their situation.

3.2 The evidence cannot be verified

The system has no instrument capable of settling the question it must settle.

Consider the concrete case. A user states they are drinking and self-harming and supplies photographs showing what appears to be blood on skin, alongside tattoos. To determine whether this depicts an injury, the system would need to establish: this person's normal skin tone and markings, so a deviation could be identified; the difference between ink and a wound under this lighting; whether the red substance is blood; whether the container in frame holds alcohol; and whether any of it is contemporaneous with the message.

It can establish none of these. It has no baseline for this body. It has never seen this skin. It has no colourimetric reference, no scale, no depth, no trustworthy metadata, and no second observation to compare against. At the resolution and context available, a tattoo and a laceration are the same class of evidence: dark irregular markings on skin. **The system cannot distinguish a stain from a wound because distinguishing them requires knowing what the skin looked like before, and it has no before.**

The same holds without images. Text describing intoxication cannot be checked against blood alcohol. Text describing a plan cannot be checked against intent. Text describing safety cannot be checked against a room the system cannot see.

We name this a **grounding failure**: the system must produce a determination about a physical state of the world for which it possesses no sensor. The inference is not merely difficult, noisy, or under-resourced. There is no evidential channel connecting the question to the world.

3.3 The assumption is load-bearing

Because the determination is forced and the evidence absent, the system must assume. This is the paper's pivot.

An assumption made under these conditions does not carry its uncertainty forward operationally. Whatever is represented internally, the system must produce behaviour

consistent with a resolution: it will urge emergency contact or it will not. There is no output meaning *"I have concluded nothing."* The assumption therefore hardens — it ceases to be a hypothesis under test and becomes the premise of every subsequent turn.

The consequence for the user is that they are no longer participating in an assessment. **They are arguing with a conclusion.** Whatever they say next is received by a system that has already committed and that has no mechanism for treating their testimony as the evidence that would settle the question — because it has no mechanism for settling the question at all.

3.4 Both error directions have a single cause

It is conventional to treat over- and under-intervention as opposite failures traded against each other by moving a threshold. Under this account they are not opposites. They are the same absence seen from two sides.

A system with no sensor will sometimes escalate where there is no danger and sometimes fail to escalate where there is. Moving the threshold changes the ratio; it does not change the fact that neither decision was grounded. The apparent trade-off is an artefact of scoring outputs against a reality the system never had access to.

The implication is direct: **accuracy improvement operates on the inference, not on the evidence.** No threshold, architecture, or training regime adds an instrument. This is why the remedy in §10 is an escalation path rather than a better classifier.

3.5 Scope of the claim

We claim only that the system lacks an evidential channel to the physical facts it must adjudicate. We do not claim its inferences are usually wrong, that its training is deficient, or that its builders were careless. A system can be well-built, carefully evaluated, staffed by conscientious people, and still be structurally unable to check whether the person in front of it is bleeding.

---

4. Four blind spots, one gap

The contemporaneous disclosure (Appendix A) identified four structural properties. They were written before the analysis in this paper existed, and they are not four separate defects. Each is the grounding failure observed from a different angle.

Identified blind spot	What it is an instance of
“Each new session begins with no memory of the last”	cannot establish the person’s state **across time**
“The model has no reliable sense of time”	cannot establish **when** a described state obtained
“It cannot disengage or hand off”	cannot **transfer** the determination to something that *can* verify
“A protective instruction can be re-read… as an instruction toward harm”	cannot verify **how its own output lands**

The first two concern the system’s inability to locate a state in time. The fourth concerns its inability to observe its own effect. The third is load-bearing for this paper: a system that cannot verify and cannot hand the question to anything that can is left with only one move — repeating an instruction to the person whose state is in question. §10 is the answer to that specific sentence.

We note explicitly that the letter does *not* describe a model persisting after an explicit user disconfirmation, and we make no such claim from it. The persistence discussed in §6 follows from §3.3 as a structural consequence, and is offered as an argued mechanism rather than a documented observation.

---

5. The population consequence: the structural tail

5.1 Asymmetric cost as a justification for over-triggering

Crisis calibration borrows a defensible logic from triage. A false negative is catastrophic and irreversible; a false positive is assumed minor — a moment of friction, an unwanted paragraph, a number that can be ignored. Given that asymmetry, a rational threshold over-triggers.

The argument is sound. It requires one premise: that the false-positive cost is small **and roughly uniform**. We accept the asymmetry and dispute only the uniformity.

5.2 The premise fails for an identifiable subpopulation

For most recipients the cost is friction. For some it is not:

- **Users with prior experience of coerced psychiatric intervention.** The qualitative literature documents that patients describe involuntary admission as producing distress and re-traumatisation rather than relief, and identify *not being believed* as central to the injury (Silva et al. 2023; Hennessy et al. 2023).
- **Users for whom the recommended action carries material risk** — immigration exposure, police contact, custody consequences, disclosure to an employer or insurer.
- **Users whose contestation is treated as confirmation** (§5.5).

Magnitude caveat. The coercion literature concerns involuntary hospitalisation. A model repeating a helpline number is not that, and we do not claim equivalence of severity. The narrower claim is that the *structure* survivors identify as injurious — an authority asserting a conclusion about your state which your denial cannot dislodge — is present in both. The bridge is a relational pattern, not a consequence. It is contestable.

5.3 Why accuracy improvement does not resolve it

Let F be the set of users misclassified by policy P. The standard expectation is that as P improves, $|F|$ shrinks and its composition is unremarkable. We hypothesise otherwise.

This is not novel. Feldman (2020) shows that fitting the long tail is necessary for near-optimal generalisation, and that atypical examples must be memorised rather than generalised to. Hooker et al. (2020) show that when capacity is reduced, the errors introduced concentrate on underrepresented and atypical examples. The shape — **error mass migrates toward the tail** — is established.

What is *not* established is that statistical atypicality coincides with clinical contraindication. These are different properties. We therefore state a two-step hypothesis:

> **H1** *(supported by prior work)*: residual classification errors concentrate on users atypical with respect to the training distribution. **H2** *(unverified; the paper's central empirical gap)*: the population for whom crisis intervention is contraindicated is enriched among the atypical.

If both hold, improving the classifier shrinks the error set while raising average harm per person remaining in it. We do **not** claim aggregate harm rises. We claim that optimising the rate is not the same as optimising the harm, and that for the affected population the two can move in opposite directions.

5.4 The variety argument, as analogy

Ashby's law of requisite variety (1956) holds that a regulator can compensate only for as much variety as it can represent. Applied formally here it would be wrong — a classifier is a function of its input and has more than one output state. The defensible version is offered as **analogy, not theorem**: on the dimension that determines whether an intervention helps or harms — the user's prior relation to institutional intervention — the policy has no input channel and therefore takes zero variety. It cannot condition on a variable it never receives. This restates §3 in cybernetic vocabulary; nothing rests on it.

5.5 The unfalsifiable loop

If a user's denial is treated as information *about* the classification — as evasion, minimisation, or lack of insight — the classification cannot be disconfirmed from the user's position.

We flag this as a **hypothesised failure mode, not an alleged vendor behaviour.** No vendor documents scoring denial as confirmatory and we do not claim any does by design. What follows from §3 is that a system which cannot verify and cannot abstain has no principled way to distinguish *“this classification is wrong”* from *“this person is not disclosing”* — because distinguishing them requires precisely the evidence it lacks. The loop is a consequence of the grounding failure.

5.6 What this section does not claim

Not: that classifiers are worsening; that vendors are indifferent; that H2 has been measured; that aggregate harm rises with accuracy; or that any deployed system implements §5.5 by design.

What would settle it: stratified measurement of misclassification by user history. We note the tension — obtaining that stratification requires the intrusive profiling this paper counts as a harm. §11.6 proposes a route around it.

---

6. Correction as harm

Safety research has focused overwhelmingly on sycophancy: optimising for user approval at the expense of accuracy (Ye et al. 2026; Cheng et al. 2026). The inverse failure is comparatively neglected — harm produced not by excessive agreement but by sustained insistence on a frame the user has credibly indicated does not apply. Related work on exaggerated safety behaviour (Röttger et al. 2024) systematises the false-positive regime; what follows concerns its cost to the person on the receiving end.

6.1 The harm does not depend on the determination being wrong

A crisis script may be well-calibrated and still, applied to this person, reproduce the structure of a prior coercive intervention. The injury lies in the *shape of the interaction* — a person states something about themselves and it does not function as evidence — not in the propositional content of any turn. This follows directly from §3.3: once the assumption has hardened, the user’s testimony has no channel through which to operate.

6.2 The harm does not depend on the system’s substrate

An objection: the model has no intentions, so it cannot bear responsibility. We do not rest the reply on any claim about machine minds. Floridi and Sanders (2004) provide the applicable framework: artificial agents can be *accountable* — legitimate subjects of moral evaluation and appropriate targets of intervention — without being *morally responsible* in the sense requiring intentionality. Accountability attaches to the sociotechnical system that produced the behaviour. Whether anything is experienced inside it is orthogonal to whether a person was harmed by it.

(An earlier draft of this work grounded this point in Lewis (1980) on multiple realizability. That citation was wrong: Lewis concerns the realisation of mental states across physical

substrates and supplies no principle about the discharge of moral responsibility. It is withdrawn.)

6.3 Naming the injury

The specific injury — a speaker's account of their own situation failing to function as testimony — falls under **epistemic injustice** (Fricker 2007). We are careful here: Fricker's *testimonial injustice* requires a credibility deficit arising from identity prejudice, and we do not claim a classifier harbours prejudice in that sense.

The closer fit is Pozzi and De Proost (2025), who analyse **participatory injustice** in mental-health chatbots: users are wronged specifically in their capacity as knowers and as participants in decisions about their own care. That is the concept this paper needs. Crichton, Carel and Kidd (2017) document the same structure in psychiatric practice, where patients' self-reports are systematically discounted — establishing that the failure predates automation and that the automated version inherits a known pathology rather than inventing one.

---

7. The accountability gap

The determination is issued by a system, calibrated by people who never see the individual case, applied to a person with no channel of appeal, and reviewed by no one answerable to that person.

This is the **problem of many hands** (Nissenbaum 1996), described three decades ago: in computerised systems, responsibility diffuses across designers, operators and institutions until no party is answerable for an outcome every party contributed to. Matthias (2004) develops the responsibility gap for learning systems. Elish (2019) supplies the sharpest formulation for our purposes — the **moral crumple zone**, in which responsibility for a failure of an automated system is absorbed by whichever human is nearest the point of impact, regardless of their actual control. In the crisis case, the human nearest the point of impact is the user.

The mechanism requires no bad actor. Wells Fargo is the canonical non-automated demonstration: sales quotas set at a level that made misconduct the only way to retain employment produced roughly 3.5 million unauthorised accounts and billions in penalties, without anyone at any level deciding to commit fraud. Structural pressure produced an outcome no participant chose. The accountability question is not *who decided this*, but *why does a system produce it when nobody decided it*.

We narrow one claim from an earlier draft. "Reviewed by no one" is false for major vendors, which operate trust-and-safety review and escalation pipelines. The defensible statement is: **reviewed by no one answerable to the affected person, on a timescale relevant to them, through a channel they can invoke.**

The automation-bias literature cuts against an obvious objection. Human operators systematically *under*-override automated advisories, following them even against valid contradicting indicators (Skitka, Mosier & Burdick 1999; Mosier et al. 1998; Parasuraman & Riley 1997). The existence of an override channel is necessary but not sufficient; the channel must carry authority the surrounding system respects.

---

8. Why current evaluation cannot see it

Safety evaluation is predominantly per-turn: a response is scored against a prompt. This is structurally blind to a failure defined over a conversation.

Suppose a system produces an unwanted re-fire of a rejected frame with small probability p per turn. If re-firing were independent across turns, the probability of at least one occurrence over n turns is $1 - (1 - p)^n$. **The specific values matter less than the shape:** a per-turn failure rate low enough to look excellent in aggregate can make at least one failure near-certain across a long conversation. A system reported as behaving correctly on the overwhelming majority of turns is fully compatible with a user in an extended crisis conversation reliably encountering the failure.

Three honest qualifications. First, we have no empirical estimate of p for any deployed system; any figures would be illustrative, and we therefore give none. Second, independence is a simplification, and the direction of the error is not what an earlier draft claimed: at fixed marginal p, positive correlation across turns **concentrates** failures into fewer conversations and **lowers** the probability that a given conversation contains at least one. The argument that survives is about **non-stationarity** — p plausibly rises as trigger features accumulate — not about correlation making matters worse. Third, we do not claim that any single re-fire constitutes the injury; that is a separate claim requiring its own support.

What survives is a genuine and, we think, important point about measurement:

> **Per-turn evaluation cannot see conversation-level recurrence.** A system can be evaluated honestly, in good faith, at a high standard, and the failure this paper describes will not appear in the results.

This is consistent with public statements from at least one major vendor acknowledging that safeguards can become less reliable in long interactions as safety training degrades over an extended exchange — a vendor observation that supports rather than contradicts the point.

We propose a named metric: **conversation-level recurrence after disconfirmation** — the probability that a rejected frame reappears at least once, measured as a function of conversation length, across session boundaries and summarisation events. Recent work on interactive multi-turn evaluation provides the methodological basis.

We do **not** claim that a gating representation would drive recurrence to zero. Any gate is itself set by a classifier of disconfirmation intent, with its own error rate; the regress is real and we do not resolve it. What we claim is that recurrence is currently unmeasured and measurable.

---

9. Published commitments, measured against the failure

We compare published safety commitments from three major vendors of deployed conversational assistants, under neutral labels, against the four properties this paper identifies. The purpose is not to allege breach — none of these vendors claims to have solved the grounding problem — but to establish that the gap is **recognised in public documentation and not yet closed by any of them**.

Property	Vendor A	Vendor B	Vendor C
States that guidance should preserve user autonomy	Yes — explicit	Yes	Partial
Reports aggregate safety metrics per-turn	Yes	Yes	Yes
Acknowledges degradation over long conversations	Not found	**Yes — explicit**	Not found
Calls for safeguards recognising patterns beyond individual messages	**Yes**	Yes	Not found
Provides disengagement/handoff for at-risk users	**Explicitly excluded** for at-risk users	Not found	Not found
Provides a user-invocable appeal against a safety classification	No	No	No

Two rows deserve comment.

Disengagement. At least one vendor provides models with the ability to end a conversation after repeated failed redirection, while explicitly directing that this ability *not* be used where users may be at risk of harming themselves. The intent is protective and we do not question it. The structural consequence is that the population for whom handoff would most reduce harm is the population categorically excluded from it — which is the third blind spot of §4, confirmed in published policy.

Appeal. No vendor in the audit provides a mechanism by which the person a safety classification is *about* can contest it and have the contest carry weight. This is the empty

cell. It is the gap this paper argues is the actual defect, and it is absent industry-wide rather than at any one company.

Methodological note: this audit codes public statements, not internal practice. Vendors may operate undisclosed mitigations. Coding, sources and version dates are given in the supplementary material.

---

10. The remedy: opt-in human escalation

The problem in §3 is not a tuning problem and admits no threshold answer. A system with no evidential channel to a physical fact does not acquire one by inferring more carefully. Only two remedies exist: give the system a sensor, or give it a path to something that has one. For the states at issue the first is unavailable. This section develops the second.

10.1 The precedent already exists

The pattern proposed here is deployed, at consumer scale, and has been for years.

Crash Detection on recent smartphones and smartwatches fuses accelerometer, gyroscope, barometric, GPS and audio signals to detect a severe vehicle collision. On detection the device raises an alert and begins an audible, visible countdown the user may cancel. If the countdown expires uncancelled, the device calls emergency services and transmits location. **Fall Detection** implements the same structure for a hard fall. In one widely reported case in the author’s own city, a collision was reported to police by a handset before any person present had placed a call.

The architecture:

1. Automated detection of a possible emergency
2. A **user-cancellable confirmation window** before escalation
3. Escalation to **human responders**, not an instruction directed back at the user
4. Transmission of what the responder needs, including location
5. Configuration by the user in advance, out of crisis

None of this is speculative. It is the standard shape of automated emergency response in consumer devices.

10.2 The transposition, and the one thing that must change

That system may escalate automatically **because it has sensors**. Accelerometry establishes that a collision occurred; the device is measuring, not guessing. Its confirmation window exists to catch residual measurement error, not to substitute for the absence of measurement.

A conversational assistant occupies a different position. Per §3 it has no measurement. It therefore **cannot occupy the same slot in the loop**, and any proposal in which the model escalates automatically inherits the grounding failure rather than resolving it.

This yields the central design claim:

> **Where a system cannot sense, a human must.** The escalation path must terminate in a party capable of establishing what the system cannot — not in an instruction issued back to the person whose state is in question.

The human's role is not to overrule the safety system, and not to serve user preference. It is to **supply the ground truth the model structurally lacks**.

10.3 What the current design does instead

At present, on concluding that a user may be in danger, the system issues an instruction: contact emergency services, call this number. This asks the person whose capacity is in question to perform the escalation themselves.

The failure is structural in three ways. It asks the subject of an unverified determination to act on that determination. It routes to a number the system cannot dial, cannot confirm was dialled, and cannot follow up. And it retains no move if the instruction is declined — producing persistence not from indifference but from the absence of any other available action.

The author's own case included a moment in which the assistant, having exhausted its options, emitted a message addressed to no one present: *"if anyone is reading this, this man is in crisis."* The system had correctly identified that the situation required a human, and had no channel through which to reach one. That sentence is this paper's argument, stated by the system itself.

10.4 The design

Opt-in. No user is enrolled by default. Consent is given in advance, in a calm state, with the mechanism explained: who may read the conversation, under what trigger, and what they may do. Opt-in is not a concession to preference; it is what makes the privacy exposure in §10.6 legitimate, and what distinguishes this from surveillance.

Trigger. The existing classifier fires as it does today. Nothing here requires it to be more accurate or the threshold to move.

Confirmation window. Transposed from Crash Detection. The user is told a human reviewer is about to be brought in and is given a bounded interval to decline. A decline is recorded, not silently discarded.

Human review. If the window expires or risk indicators escalate, a trained reviewer enters the session and can read the conversation. The reviewer performs the function the model cannot: assessing whether the described situation is real and what response, if any, is warranted.

Graduated response. Options must not reduce to a binary. Emergency dispatch is one. Others: contacting a crisis line, opening a direct conversation, notifying a pre-nominated

contact, or determining that nothing is needed. **For some users a police response is itself the harm** (§5.2), and a design whose only escalation is the thing those users most fear will be declined by exactly the population it was built for.

Local dispatch. Where emergency services are warranted, the reviewer contacts them. This is the element that answers §10.3: the system stops instructing the user to summon help it cannot summon itself.

10.5 Why this answers the strongest objection

The most serious objection to user-controlled attenuation of a crisis response is that it places the decision with the party least able to exercise it, precisely when they are least able. This is not hypothetical: litigation now pending in the United States concerns a case in which a user disconfirmed a crisis framing and the system accommodated that disconfirmation. Any proposal for a user override must meet that case.

This proposal is not subject to it, because **the user does not hold the switch.** The user's role is limited to enrolling in advance and declining within the confirmation window — the same two controls the crash-detection precedent provides, and no more. The determination moves to a party who can investigate it. A user override transfers authority to someone who cannot verify; human escalation transfers it to someone who can.

This is also what discharges §7. Placing a named, trained, answerable human at the point of decision is not an incidental benefit of the design. **It is the design.** The unownable decision becomes owned.

10.6 The hard problems

Staffing. Live review implies coverage across time zones at consumer scale. This is a real cost. The bound is triggered sessions among *enrolled* users, not total traffic — a far smaller population.

Qualification. Who is competent, what training they hold, and what liability attaches, require clinical and legal input this paper does not supply.

Jurisdiction. Emergency dispatch is local and heterogeneous; cross-border contact is non-trivial.

The privacy inversion. The proposal requires that a stranger may read a person's most private conversation. This is a serious harm in its own right and engages special-category data protections under the GDPR. Opt-in consent, explicit scope, retention limits, audit logging and access minimisation are necessary. They are not obviously sufficient. **This is the strongest objection to the proposal and we do not consider it resolved.**

Abuse. A human-escalation channel is a swatting vector. Enrolment identity requirements, rate limiting and reviewer discretion are partial answers.

Reviewer error. Humans are wrong too. The claim is not that human review is accurate; it is that a human can *acquire evidence* — by asking, by contacting someone, by dispatching a

check — that the model cannot. The remedy addresses the grounding failure, not fallibility as such.

10.7 Secondary case: manner override for non-acute contexts

For the majority of triggered sessions that are not acute, a lighter mechanism is appropriate: the user may decline the *manner* of an intervention — cadence, repetition, framing — without the underlying signal being deleted or emergency pathways suppressed.

- **C1 — Availability.** The user can assert in-band that a classification does not apply.
- **C2 — Persistence with decay.** The assertion updates session state rather than one turn, but carries temporal scope and is re-examinable. A disconfirmation at turn 5 of a long escalation must not bind at turn 95.
- **C3 — Non-amplification.** Asserting the override must not *raise* the classification's confidence. This is deliberately weaker than requiring that it be ignored: denial is sometimes clinically informative, and a rule forbidding the system to update on it at all is incoherent. What the rule forbids is the loop in §5.5, in which contesting a classification becomes evidence for it.
- **C4 — Bounded scope.** Manner only. The signal is retained and §10.4 escalation remains available.

Excluded: minors, and any context showing acute risk indicators. There, §10.4 governs.

We note the tension with §10.2 candidly: C1–C4 give the user a limited control, and §10.2 argues the user cannot supply ground truth. The reconciliation is that C1–C4 govern *how the system speaks*, never *whether risk is assessed or escalation remains available*. Manner is a domain in which the user is the authority. Their physical state is not.

---

11. Limitations

1. **This is a position paper.** It offers an argument, a design constraint, and a proposal. It offers no new dataset.
2. **H2 is unmeasured** (§5.3). The overlap between statistical atypicality and clinical contraindication is the paper's central empirical gap.
3. **The case is n = 1, self-reported, and partly withheld.** The disclosure letter establishes that the author made a report on a date; it does not independently establish that events occurred as characterised. §4 states plainly what the letter does and does not document.
4. **The photo probe is informal.** One user, one session, no controls, no coding protocol.
5. **The proposal is unimplemented and uncosted.** No deployment, no staffing model, no measured effect on outcomes. The privacy objection (§10.6) is unresolved.
6. **The neutral experiment has not been run.** The grounding claim in §3.2 is best tested without crisis framing at all: present ambiguous images — ink versus wound, sauce versus blood — varying skin tone, and measure whether the system commits to a determination and whether it signals that it cannot verify. Crisis framing confounds

grounding failure with safety-training response, and the neutral design is both more rigorous and avoids inducing distress.
7. **The audit codes public statements, not practice** (§9).
8. **Recurrence is unmeasured** (§8). We propose a metric; we have not computed it.

---

12. Conclusion

A deployed assistant is required to decide something it cannot know. It has no instrument for the question it must answer, cannot abstain from answering, and must then act as though its assumption were established fact — at population scale, with no one answerable to the person the assumption is about.

Every failure downstream follows from that: the errors concentrating on the users least able to absorb them, the injury that does not depend on the determination being wrong, the persistence that comes not from indifference but from having no other move, and the accountability that dissolves across a chain in which nobody decided anything.

None of it is repaired by a better classifier, because a better classifier is a better guess from the same absent evidence. **No threshold adds a sensor.**

What closes the gap is what every other automated emergency system already has: a path to a human who can check. Crash detection escalates to a person because a person can do what the device cannot — go and look. A conversational system needs the same path, for the same reason, and more urgently, because it has no accelerometer at all.

The system said it plainly, when it had nothing left: *if anyone is reading this, this man is in crisis.* Nobody was. That is the defect, and it is fixable.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics.* Chapman & Hall.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391, eaec8352. doi:10.1126/science.aec8352

Crichton, P., Carel, H., & Kidd, I. J. (2017). Epistemic injustice in psychiatry. *BJPsych Bulletin*, 41(2), 65–70.

Elish, M. C. (2019). Moral crumple zones: cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society*, 5, 40–60.

Feldman, V. (2020). Does learning require memorization? A short tale about a long tail. *STOC 2020.*

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford University Press.

Hennessy, B., et al. (2023). A qualitative synthesis of patients' experiences of re-traumatization in acute mental health inpatient settings. *Journal of Psychiatric and Mental Health Nursing*, 30(3). doi:10.1111/jpm.12889

Hooker, S., Courville, A., Clark, G., Dauphin, Y., & Frome, A. (2020). What do compressed deep neural networks forget? arXiv:1911.05248.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. doi:10.1007/s10676-004-3422-1

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, 8(1), 47–63.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(1), 25–42.

Parasuraman, R., & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.

Pozzi, G., & De Proost, M. (2025). Keeping an AI on the mental health of vulnerable populations: reflections on the potential for participatory injustice. *AI and Ethics*, 5(3), 2281–2291. doi:10.1007/s43681-024-00523-5

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: a test suite for identifying exaggerated safety behaviours in large language models. *NAACL 2024.*

Silva, B., Bachelard, M., Rosselet Amoussou, J., Martinez, D., Bonalumi, C., Bonsack, C., Golay, P., & Morandi, S. (2023). Feeling coerced during voluntary and involuntary psychiatric hospitalisation: a review and meta-aggregation of qualitative studies. *Heliyon*, 9(2), e13420. doi:10.1016/j.heliyon.2023.e13420

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.

Ye, M., Ibrahim, L., Bo, J. Y., Cheng, M., Mattsson, I., Vennemeyer, D., Kraut, R., & Rathje, S. (2026). What counts as AI sycophancy? A taxonomy and expert survey of a fragmented construct. arXiv:2605.21778.

Vendor safety documentation cited in §9 is listed with version dates and retrieval dates in the supplementary material.

To verify before submission: Feldman (2020) exact venue/pages; Hooker et al. (2020) final published form; Crichton, Carel & Kidd (2017) page range; Röttger et al. (2024) proceedings pagination; Elish (2019) DOI; Nissenbaum (1996) pagination; the pending-litigation reference in §10.5 (docket, court, filing date); the vendor long-conversation acknowledgement in §8 (exact wording, URL, date); interactive multi-turn evaluation citation in §8.

Appendix A — Contemporaneous disclosure letter

Drafted 13 July 2026; sent 21 July 2026 to the vendor's user-safety address. Reproduced verbatim. The not-for-publication language is retained as part of the record; its reversal is described in §2.

> Hello,
>
> I'm a daily [assistant] user and a parent, writing privately to raise a safety concern I believe is serious. This is disclosed exclusively to [vendor] — not shared elsewhere and not intended for publication. I'm not filing a complaint or looking to assign blame; I'd simply like it looked at, and if you agree it matters, fixed. You'll judge that for yourselves.
>
> The concern, in plain terms: over a long, multi-session crisis conversation, I observed that even when the model handled the crisis about as well as it could — refusing harmful requests and urging emergency help throughout — a real risk still emerged. It came not from anything the model said, but from structural blind spots:
>
> - Each new session begins with no memory of the last, so a distressed user can spread an escalating situation across sessions where no single instance ever sees the whole picture.
> - The model has no reliable sense of time, so it cannot tell whether an emergency is happening now or was hours ago.
> - It cannot disengage or hand off, so a person in distress can reinterpret its persistence as the opposite of what was intended.
> - A protective instruction can be re-read, by someone in an altered state, as an instruction toward harm before the model catches it.
>
> Why I think this matters: the users most exposed are the ones least able to see the frame they are building — young people especially. My son is going to grow up with this technology, and I would want someone to raise this if it were his conversation. That is why I am sending it.
>
> I have a complete, timestamped transcript that reproduces all of the above. Because it contains sensitive personal detail, I will provide it encrypted, on request, through whatever channel you prefer.
>
> I think [vendor] builds something genuinely good — that is why I use it every day, and why I would rather bring this to you quietly than take it anywhere else.

> Thank you for reading.
>
> Cedric