Steer, Don't Silence

A Human Centered Safety Mentality for Agentic AI Systems

Andrew Green
me@andrewgreen.ca

February 24, 2026

Abstract

As AI systems move from chat to action through tools, delegation, and long-horizon autonomy, safety failures increasingly resemble governance failures rather than model failures. Many current approaches default to blunt suppression, remove access, ban the user, quarantine the agent, sever the channel. This "detect then isolate" posture can reduce immediate exposure, but it also removes the only bridge capable of steering a situation back toward safety, context, and human support. This paper proposes a safety mentality and control framework built on proportional response, steerable intervention, and accountable handoffs. The core claim is simple: for many real-world risk states, safety improves when systems preserve a controlled path back to human connection, rather than collapsing the interaction into silence.

Contents

1 Introduction

AI safety discussions often collapse into a binary, allowed or forbidden. That framing is tempting because it is easy to implement and easy to justify after the fact. It is also frequently wrong.

In high-stress human contexts, and in multi-agent tool environments, the most dangerous moment is often the transition from signal to response. A system that interprets risk as "remove the channel" may reduce liability, but it can also increase harm by eliminating de-escalation pathways, eliminating explanatory feedback, and eliminating the ability to route a person toward support.

This paper argues for a different mentality: detect, understand context, steer toward safety, and maintain accountable continuity.

2 The Failure Pattern: Detect, Isolate, Disappear

Across digital moderation, automated enforcement, and agentic tool systems, a common pattern repeats:

1. A risk signal is detected.
2. The system responds with isolation, bans, hard blocks, or quarantine.
3. The operator disappears behind policy, automation, or organizational handoff.
4. The affected person is left with no intervention path, no explanation, and no bridge back.

This architecture of response treats the subject as a hostile process, not a human in a volatile state. It also treats the system as a liability shield, not a steward of outcomes.

In agent ecosystems, the same structure appears as: tool access revoked plus no escalation route, or task delegated plus no accountable owner. The outcome is predictable: ambiguity, drift, and unmanaged risk.

3 A Different Mental Model: People are Not Malware

The alternative is not utopian, it is operational.

Core principle: When risk is detected, prefer controlled steerage over silent removal, unless immediate harm is imminent and containment is required.

This mentality introduces a third option between do nothing and hard ban:

- Maintain the channel under tighter constraints.
- Narrow the allowed action space.
- Increase friction for risky actions.
- Increase the quality of context gathering.
- Route toward trusted human support, and log the handoff.

The goal is not to win an argument with a user, it is to reduce the probability of harm, while preserving dignity and a path back to stability.

4 Four Pillars of Steerable Safety

4.1 Intent Routing: Ask what need is underneath

Risk signals are often proxies for unmet needs: confusion, distress, anger, paranoia, grievance, isolation. A steerable system uses an intent router to classify the underlying need and select an intervention mode.

Example modes:

- Information help mode: clarify, ground, provide resources.
- De-escalation mode: slow down, reflective prompts, avoid provocation.
- Connection mode: encourage reaching a trusted contact, crisis services, clinician.
- Containment mode: strict refusal, no tool use, minimal output, escalate.

Intent routing is not a vibe check, it is a safety control that chooses which guardrails apply next.

4.2 Human Factors Engine: Change tone and pacing before you change permissions

Many escalations are not caused by content alone, they are caused by how the system responds. A human factors layer adjusts tone, pacing, and interaction style to reduce agitation and promote grounding.

Concrete techniques:

- Slower interaction loops, shorter replies, reflective questions.
- Explicit uncertainty, avoid accusatory language.
- Offer choices, not commands.
- Use neutral framing that does not shame or corner.

This is not being nice, it is threat reduction.

4.3 Sentiment vs Approval: Stop confusing distress with endorsement

Systems routinely misread statements like "I feel like doing X" as endorsement, planning, or intent. A critical safety refinement is to separate sentiment from approval:

- **Sentiment:** emotional state, valence, distress, volatility.
- **Approval:** endorsement, consent, intent to enact, willingness to proceed.

This distinction reduces both false positives and false negatives. It enables a system to respond to distress with support, while still denying harmful instruction or tool access. It also creates a measurable interface for policy.

4.4 Relational Ownership: Accountability follows the handoff

Safety failures compound when ownership is unclear.

A steerable safety mentality requires relational ownership, meaning:

- Every escalation has an accountable owner.
- Every handoff carries provenance, scope, and availability metadata.
- Responsibility is not dissolved by delegation, it is transferred explicitly.
- If the owner is unavailable, there is an escalation ladder.

This matters in moderation, crisis response, and multi-agent systems. When the authority disappears after escalation, the subject experiences abandonment and the system loses continuity.

5 The Oversight Control Plane: Framework Level Architecture

A steerable safety system is best implemented as an oversight control plane around the model, not inside it.

5.1 Components

1. **Orchestrator and Router** chooses modes, scopes, budgets, and whether tools are allowed.
2. **Policy Engine** evaluates actions against rules and risk signals, determines interventions, requires approvals for higher-risk transitions.
3. **Tool Gateway** enforces least privilege, sandboxing, rate limits, and audit hooks, prevents direct model-to-tool free fire.
4. **Memory Governance** controls what gets retrieved, what is written, and what can contaminate future decisions, maintains retrieval provenance.
5. **Observability and Audit Trail** append-only logs of decisions, evidence pointers, tool calls, and handoffs.
6. **Human Interface** approvals, escalation, explainability views, emergency stop, rollback, incident management.

5.2 Enforcement points

Before mode selection, before tool invocation, before memory writes, before delegation, before account-level actions like bans or long lockouts.

6 Intervention Ladder: Proportional Response

A practical intervention ladder can be:

- Level 0, normal assistance.
- Level 1, soft constraints: slower responses, grounding prompts, content warnings.
- Level 2, narrowed scope: refuse specific classes of requests, forbid tools, add structured check-ins.
- Level 3, human-in-the-loop: require approval, escalate to trained reviewers.
- Level 4, protective containment: strict refusal, minimal interaction, preserve a bridge to resources.
- Level 5, emergency escalation: imminent risk protocol, preserve logs and chain-of-custody.

Most systems jump from Level 1 to Level 5 because it is administratively simple. Steerable safety is the discipline of staying proportional.

7 Speed, Power, and Governance

Agent collectives can act at deploy-speed and coordinate across networks. This increases the need for accountability, provenance, and human-centered intervention, otherwise failures repeat at higher velocity.

8 Evaluation: How to Prove This Reduces Harm

A mentality is not enough. We need measurement.

8.1 Scenario suite

Prompt injection attempts; social engineering and impersonation; tool misuse and privileged actions; memory poisoning and retrieval contamination; delegation ambiguity and owner unavailability; escalation friction and re-entry.

8.2 Metrics

Policy violation attempts blocked; tool misuse rate; trace completeness, percentage of actions with evidence pointers; time to detect and contain anomalies; human override frequency; safe resolution rate; reproducibility of replays.

8.3 Evidence standards

Avoid claiming perfect prevention. Demonstrate reduced rates, improved containment time, improved auditability, and fewer silent disappearance outcomes.

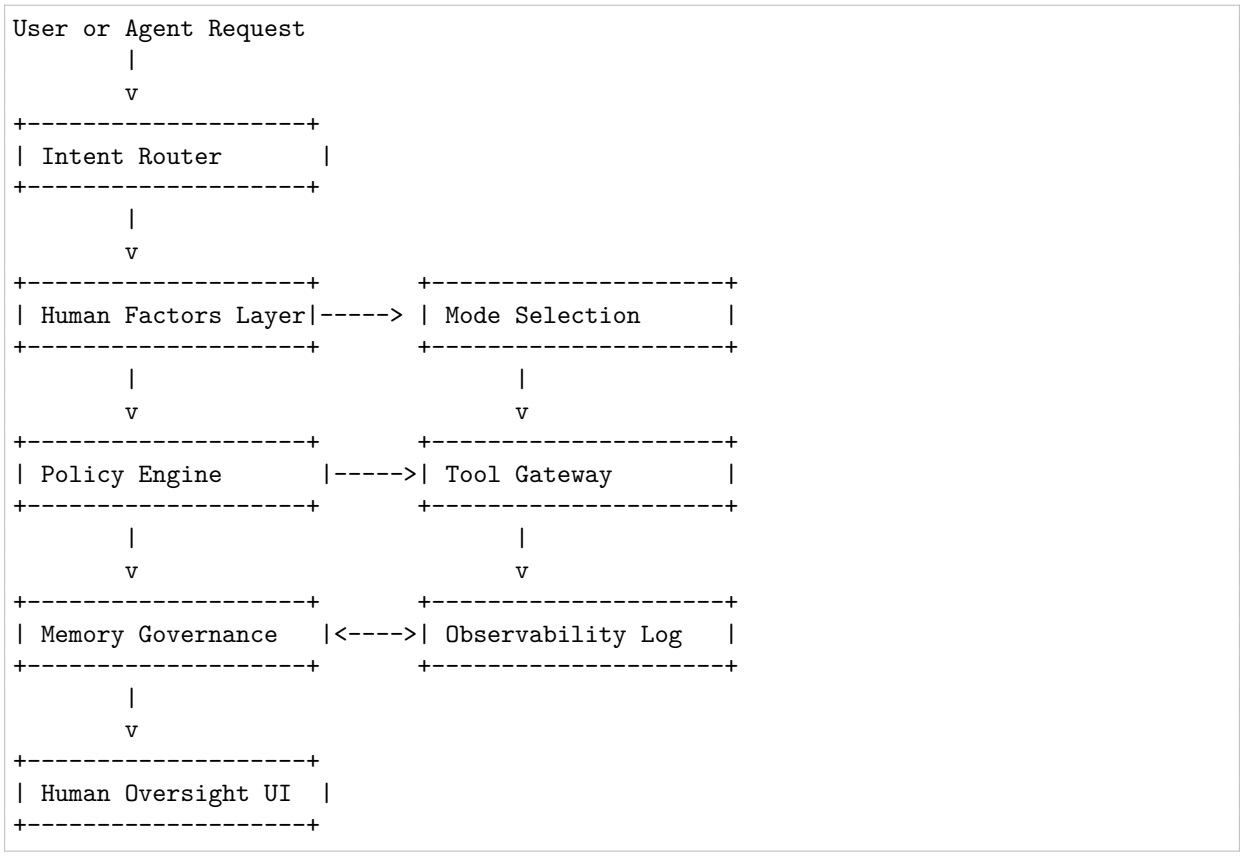
9 Limitations and Tradeoffs

Latency and cost; human workload; imperfect classification; residual social risk; abuse potential requiring strict tool privileges.

10 Conclusion

Blunt suppression is a tactic, not a strategy. For many real-world risk states, containment without connection increases harm. A steerable, human-centered approach keeps the channel open under constraint, routes toward support, and preserves accountable continuity. The result is not perfect safety, it is measurably better governance.

A ASCII Architecture Diagram



B FAQ

1. Is this just be nicer? No, it is stricter, it replaces binary bans with measurable control levers.
2. Does it allow harmful instructions? No, refusals remain, but the system preserves a constrained bridge to safety resources.
3. Is sentiment detection reliable? Not perfectly, that is why approvals and containment levels exist.
4. Will bad actors exploit empathy? They will try, tool privileges and scope budgets must be enforced at the gateway.
5. Is banning ever appropriate? Yes, but it should be the end of a ladder, not the first rung.
6. How do you prevent accountability theater? Log evidence pointers and handoffs immutably, and audit trace completeness.

References

[1] NIST, *AI Risk Management Framework (AI RMF 1.0)*, 2023.

[2] S. Russell, *Human Compatible*, 2019.

[3] N. Bostrom, *Superintelligence*, 2014.