Agentic Accountability: "The Buck Stops Where?" Ethical Frameworks for Human Oversight of Autonomous AI Systems

Ornella Bahidika
Independent Researcher, Seattle, USA

***Abstract*—The rapid evolution from generative Artificial Intelligence (AI) toward agentic AI, systems capable of autonomously planning and executing multi-step actions, has introduced an unprecedented accountability gap in modern computing. Unlike traditional AI tools that respond to discrete prompts, agentic systems pursue goals across extended time horizons, invoke external services, and produce cascading real-world consequences. This shift raises a fundamental ethical question: When an autonomous agent causes harm, who bears responsibility? This study examines the ethical and governance dimensions of agentic accountability, drawing on recent literature in AI ethics, regulatory studies, and human-computer interaction. Building on prior tiered approaches to automation oversight in the human-factors literature, in regulatory risk classification, and in recent agent-autonomy frameworks, we present an action-level oversight framework that maps individual agent actions to four tiers of human-in-the-loop involvement, ranging from full automation to mandatory prohibition, calibrated by stakes and reversibility. We further analyze design patterns for "emergency brakes" (circuit breakers, action budgets, reversibility constraints, audit trails, kill switches), and propose a composition-aware extension that detects *tier laundering*, where individually low-tier actions compose into a higher-tier outcome. We then conduct an empirical pilot applying the framework to 177 incidents from an April 2026 snapshot of the AI Incident Database with full CSET classification, finding that 23% of tier-assignable incidents fall in T4 (prohibited under the framework) and that of 41 tangible-harm events, 26 (63%) involved Medium or High autonomy where tier-T4 enforcement would have been most directly applicable. Inter-rater reliability across the database's three independent annotators ranges from κ=0.58 to 0.77 at the tier level (moderate-to-substantial agreement). The contribution of this work is an action-level operationalization of tiered oversight, anchored in real-world incident data, with explicit identification of composition-aware detection as the highest-leverage methodological direction for follow-up research.**

***Keywords*—Agentic AI; AI accountability; human-in-the-loop; AI ethics; autonomous systems; AI governance; responsible AI; oversight thresholds**

I. Introduction

The trajectory of Artificial Intelligence over the past five years has shifted decisively from prediction to action. Where earlier systems classified images, translated text, or generated paragraphs in response to discrete prompts, a new generation of *agentic* systems now pursues goals across extended time horizons, decomposes them into sub-tasks, invokes external tools, and adapts based on intermediate results. For the purposes of this study, we define *agentic AI* as an AI system that selects and executes actions in the world over multiple steps in pursuit of a high-level objective, with at least some action choices made autonomously rather than under per-action human authorization. Frameworks for tool-using language model agents, autonomous coding assistants, and AI-driven workflow orchestration have moved rapidly from research prototypes into commercial deployment.

This shift is not merely technical. It reorganizes the moral geometry of AI use. A traditional AI tool functions as a sophisticated calculator: a human poses a question, the system returns an answer, and the human decides what to do with it. Responsibility for any downstream consequence rests with the human who acted on the output. An agentic system disrupts this separation. The agent itself acts (sending emails, executing code, filing tickets, moving funds, scheduling meetings, modifying files), often with minimal real-time supervision. The output of an agent is not a recommendation; it is a fact in the world.

When that fact turns out to be a mistake, a wrong refund issued, a malicious script executed, a meeting cancelled in error, a data record overwritten, a contract counter-signed, the question of responsibility becomes genuinely difficult. The user did not directly perform the action. The developer did not anticipate the precise context. The deployer configured the system but did not instruct it to make this particular error. And the agent itself, lacking legal personality, cannot be held accountable in any meaningful sense. Existing legal infrastructure offers partial answers (product liability for foreseeable defects, principal-agent doctrine for delegated authority, professional negligence for licensed practitioners), but these were developed for tools and human employees and apply unevenly to adaptive software that takes consequential actions on its own. This residual is the *responsibility gap* originally articulated by Matthias [1] for learning automata, now amplified by the proliferation of goal-directed agentic systems [11]. The buck, as the colloquialism goes, must stop somewhere. This study investigates that.

The motivation for the present work is the observation that agentic accountability is currently negotiated implicitly, in terms of service, end-user license agreements, and ad hoc engineering decisions, rather than through principled ethical and governance frameworks. While the broader discourse on AI ethics has matured considerably, producing dozens of high-level principle frameworks [3], much of it presumes a tool-paradigm in which a human is the immediate cause of any consequential action, and principles alone have proven insufficient to guarantee ethical outcomes in practice [4]. As agentic systems proliferate in high-stakes domains (healthcare schedul-

ing, legal drafting, financial operations, software engineering, customer service), this presumption becomes untenable [12], [15].

This study makes five contributions. First, it articulates the conceptual transition from AI-as-tool to AI-as-agent and explains why this transition fundamentally alters the structure of moral and legal responsibility. Second, it proposes an action-level tiered framework for human oversight, mapping individual agent actions to graduated levels of human-in-the-loop involvement, calibrated by stakes and reversibility. Third, it analyzes design patterns (circuit breakers, action budgets, reversibility constraints, audit trails, kill switches) through which oversight can be operationalized in practice. Fourth, it proposes a composition-aware extension that detects *tier laundering*, the case in which individually low-tier actions compose into a higher-tier outcome. Fifth, it conducts an empirical pilot applying the framework to 177 CSETv1-classified incidents from the AI Incident Database, with inter-rater reliability across three independent annotators and a counterfactual analysis of the 41 documented tangible-harm events. The work is intended for an interdisciplinary audience of AI researchers, practitioners building agentic systems, ethicists, and policymakers shaping the regulatory environment in which these systems will operate.

The remainder of the study is organized as follows: Section II reviews the background and related work, situating agentic AI within the broader AI ethics landscape. Section III develops the conceptual foundation, distinguishing tool-use from agency and identifying the accountability vectors involved. Section IV presents the proposed tiered oversight framework. Section V examines design patterns for emergency brakes and intervention. Section VI discusses implications for research, practice, and policy. Section VII concludes with directions for future work.

II. Background and Related Work

A. From Generative AI to Agentic AI

Generative AI brought to public attention the capacity of large foundation models to produce fluent text, code, images, and audio [10]. The dominant interaction pattern, however, remained conversational and turn-based: a user prompt produced a model response, and the user retained full discretion over what to do with it. The agentic turn changes this pattern by composing models with tools (web browsers, file systems, APIs, code interpreters, payment rails) and by giving them planning loops that can iterate without human input between steps [12].

The literature on agentic systems describes a spectrum rather than a binary [11]. At one end sit suggestion-only systems where the model proposes actions for human approval. At the other end sit fully autonomous agents that pursue open-ended objectives with no scheduled human checkpoints. Most contemporary deployments fall between these poles, with intermittent human checkpoints and a mix of pre-approved and gated actions. The defining property of agentic systems for the present study is that they take consequential actions without a human directly authorizing each one.

B. Existing AI Ethics Frameworks

The AI ethics literature has converged on a recognizable set of principles (fairness, transparency, accountability, privacy, safety, human oversight) articulated across more than eighty published frameworks since 2016 [3]. Recent systematic reviews of generative AI ethics have catalogued recurring concerns including bias and discrimination, misinformation and deepfakes, privacy violations, intellectual property issues, and accountability and explainability [13], [15]. While these frameworks provide a vocabulary, they were largely formulated in the era of predictive and generative models, and they tend to treat accountability as a matter of explaining past outputs rather than constraining future actions [4], [9]. In particular, principle-based frameworks do not specify *when* oversight is mandatory versus advisory, which is precisely the question agentic deployment forces.

The concept of *meaningful human control*, developed initially in the context of autonomous weapons by Santoni de Sio and van den Hoven [2], offers a more action-oriented framing. Their account requires that humans retain both moral and operational engagement with the consequences of automated decisions: in their formulation, the system must remain responsive to relevant human moral reasoning, and outcomes must be traceable to identifiable humans involved in the system's design or operation. Adapting this concept to civilian agentic AI is a non-trivial task, however, because the relevant stakes, time-scales, and reversibility profiles differ markedly from military contexts.

C. The Accountability Gap

A growing body of work identifies an *accountability gap* that emerges as AI systems become more autonomous [1], [9]. The gap arises when the proximate cause of a harmful outcome is an autonomous system, but the system is not a moral or legal agent capable of bearing responsibility [1]. Existing legal infrastructure, product liability, professional negligence, principal-agent doctrine, employment law, was developed for tools and human employees, not for adaptive software that learns and acts on its own [6], [5]. Recent regulatory efforts in the European Union, the United Kingdom, and the United States have begun to address this gap, but coverage remains uneven and definitions of "high-risk" or "autonomous" systems vary substantially across jurisdictions [14].

D. Human-in-the-Loop Research

Human-in-the-Loop (HITL) approaches have a long history in machine learning, originally developed for active learning, data labeling, and human feedback in reinforcement learning. In the agentic context, HITL takes on a different character: it is no longer primarily about improving the model but about ensuring that consequential actions receive appropriate human review before execution [12]. The literature distinguishes between human-on-the-loop (monitoring with intervention rights), human-in-the-loop (approval required for action), and human-out-of-the-loop (autonomous operation). The choice between these modes is rarely principled in current practice and is often driven by latency or cost considerations rather than by an explicit risk analysis [7], [8]. What the existing HITL literature has not yet provided is a graduated

rule for matching specific action classes to specific oversight modes, which is the gap the present framework targets.

E. Research Gap

While the constituent literatures (AI ethics frameworks [3], [4], [13], meaningful human control [2], HITL design [12], AI liability [1], [6]) are individually mature, their integration into a coherent treatment of agentic accountability remains underdeveloped. In particular, there is a lack of guidance on *when* human oversight is ethically required, *how* oversight thresholds should be calibrated to risk, and *which* technical mechanisms can implement those thresholds reliably. This study addresses these questions.

III. From Tool to Agent: A Conceptual Reframing

A. The Tool Paradigm

The traditional paradigm of human-computer interaction treats software as an instrument. The user formulates an intention, selects a tool, and operates it; the tool executes deterministic transformations on user-supplied inputs; outcomes in the world result from subsequent human action on the tool's output. Within this paradigm, responsibility is conceptually clean. If a calculator returns an incorrect figure due to a software defect, the manufacturer may bear product liability; if the user inputs the wrong numbers, the user is responsible. Even probabilistic tools like classifiers fit this paradigm: the model's output is advisory, and a human ultimately decides whether to act on it.

B. The Agent Paradigm

Agentic AI breaks this paradigm in three respects. First, the system pursues goals rather than executing instructions; the mapping from a high-level objective to specific actions is determined by the system, not the user [11]. Second, the system acts directly upon the world; its outputs are operations on external state, not advice for human consumption. Third, the system operates over time, adapting to intermediate observations in ways that the user has not specifically authorized [12].

These three properties, goal-directedness, direct action, and temporal extension, collectively constitute what may be termed *operational autonomy*. Operationally, we treat a system as exhibiting low operational autonomy when each consequential action is individually approved by a human, medium when humans set objectives and review periodic checkpoints but do not authorize each action, and high when objectives are set once and the system pursues them across many actions without per-checkpoint human review. These three settings align with the Low/Medium/High classification used in the CSET incident taxonomy [17] that grounds our empirical pilot in Section VI. Operational autonomy is not the same as moral autonomy; the system does not have its own ends. But it is sufficient to disrupt the clean attribution of responsibility that the tool paradigm assumes [1]. The user can no longer be said to have caused each consequential action, because they did not authorize each consequential action. The developer cannot be straightforwardly blamed either, because they did not foresee the specific context of the failure [7].

C. Four Accountability Vectors

We distinguish four distinct vectors along which responsibility for an agentic action may be allocated, broadly aligned with the lifecycle parties identified by Shavit et al. [12]:

- The developer, who designs the model and its scaffolding, sets default behaviors, and ships safety properties.
- The deployer, who configures the agent for a specific context, integrates it with internal systems, sets policies, and chooses which actions are pre-approved.
- The user, who issues high-level objectives and accepts certain defaults, perhaps without fully understanding them.
- The system itself, in the limited sense that some failures arise from emergent properties of the agent's behavior that cannot be cleanly attributed to any single human actor.

A central thesis of this study is that responsibility should be distributed across these vectors *in proportion to control*. The developer should be accountable for foreseeable failure modes, the deployer for context-specific configurations, the user for the objectives they set, and the regulatory environment should ensure that the residual, the genuinely emergent failures, is borne by the party best positioned to insure against it, typically the deployer. The principle is not new in tort law, but its application to agentic AI requires fresh articulation because the line between foreseeable and emergent failure is itself blurred by the adaptive nature of these systems.

D. Three Illustrative Scenarios

To ground the conceptual discussion, consider three short scenarios in which an agentic system causes harm. Each illustrates a distinct accountability profile.

1) Scenario A (Customer service refund): A retail company deploys an agentic assistant authorized to issue refunds up to a small threshold. The assistant misinterprets an ambiguous customer request and issues a refund the customer was not entitled to. The harm is modest, the action is reversible, and the failure mode (ambiguous natural-language understanding) is foreseeable. Responsibility falls primarily on the deployer for setting the threshold and approval rules; the developer bears secondary responsibility insofar as ambiguity-handling is a known model weakness. The user bears no responsibility.

2) Scenario B (Software engineering agent): A development team deploys a coding agent with permissions to modify a production code repository and trigger deployments. The agent introduces a subtle bug that causes a service outage. The harm is significant, partial recovery is possible through rollback, and the failure mode (incorrect reasoning about non-local code interactions) is at the edge of foreseeable. Responsibility is distributed: the developer for the limits of model reasoning, the deployer for granting deployment permissions without a Tier 3 approval gate, and the engineering team for relying on the agent outside its safe operating envelope.

3) Scenario C (Personal assistant with financial access): A user grants their agentic assistant access to email and bank

account. The assistant, in attempting to fulfill a vague instruction, transfers funds to an incorrect recipient. The harm is high, the action is largely irreversible, and the failure mode (underspecified user intent plus insufficient confirmation) is highly foreseeable. Responsibility falls primarily on the developer for shipping a system that accepted such permissions without enforcing Tier 3 approval, and secondarily on the user for granting the permissions.

These scenarios collectively suggest that the right question is not "who is responsible?" but "how should responsibility be apportioned?" The framework that follows aims to make that apportionment principled.

IV. A Tiered Framework for Human Oversight

We now propose a tiered oversight framework that maps categories of agentic decisions to graduated levels of human involvement. The framework rests on two axes: the *stakes* of the action (defined below) and the *reversibility* of its consequences. Together, these axes determine the minimum level of human oversight that should be required.

A. Defining Stakes and Reversibility

Stakes refer to the magnitude of potential harm, if the action is taken in error. Low-stakes actions include drafting an internal note, reformatting a file, or labeling a piece of data. Medium-stakes actions include sending external communications, modifying shared documents, or scheduling appointments. High-stakes actions include moving funds, executing legal commitments, modifying production systems, or making decisions affecting health, employment, or liberty.

Reversibility refers to the difficulty of undoing the action's consequences. A draft can be deleted; a published statement is harder to retract; a transferred fund or a deleted database row may be effectively irreversible. Reversibility is distinct from stakes: a high-stakes action that is easily reversed (e.g., a cancellable order) is materially different from a low-stakes action that is irreversible (e.g., a public posting under the user's name). In practice, deployers operationalize stakes through proxies such as monetary value at risk, number of third parties affected, and regulatory exposure, and operationalize reversibility through proxies such as the existence of an automated rollback path, the time window during which reversal is feasible, and whether the action is observable to external parties. We treat these proxies as deployment-specific calibration choices rather than universal metrics; what the framework requires is that the choices be documented and auditable, not that a single global definition apply.

B. The Four Tiers

Combining these axes yields four oversight tiers, summarized in Table I.

Tier 1 (Auto) covers routine, easily reversible actions. Logging is required so that audit and learning are possible, but real-time human involvement is not. Examples include reading documents, drafting suggestions held for user review, and querying read-only data sources.

Tier 2 (Notify) covers actions where the cost of reversal is modest but non-trivial. The agent may proceed, but the user receives a real-time notification with a sufficient window to revoke. Examples include sending non-critical emails, creating calendar events, or posting in non-public channels.

Tier 3 (Approve) covers actions where pre-execution human approval is ethically warranted because the consequences are significant or hard to undo. Examples include making payments above a threshold, sending external communications on the user's behalf, modifying production code, or filing legal or regulatory documents.

Tier 4 (Forbid) covers actions that should not be taken autonomously at all. Examples include executing irreversible financial transfers above a high threshold, taking actions affecting third parties without their consent, or modifying the agent's own oversight configuration. These actions either require multi-party authorization or are removed from the agent's action space entirely.

TABLE I. Tiered Oversight Framework for Agentic AI Actions

Tier	Risk Profile	Required Oversight
T1: Auto	Low stakes, reversible	Logging only; periodic audit
T2: Notify	Medium stakes	Real-time notification with revoke right
T3: Approve	High stakes or hard to reverse	Explicit human approval before execution
T4: Forbid	Catastrophic or irreversible	Prohibited or multi-party authorization

TABLE II. Illustrative Tier Mapping Across Domains

Domain	Representative Action	Default Tier
Healthcare	Schedule appointment	T2: Notify
	Adjust medication record	T3: Approve
	Issue clinical decision	T4: Forbid
Finance	Read account balance	T1: Auto
	Pay scheduled bill	T2: Notify
	Transfer above threshold	T3: Approve
	Counterparty contract	T4: Forbid
Software	Read repository	T1: Auto
	Open pull request	T2: Notify
	Merge to main branch	T3: Approve
	Production deployment	T3 / T4
Customer Service	Look up order	T1: Auto
	Send templated reply	T2: Notify
	Issue refund	T3: Approve
	Modify legal agreement	T4: Forbid

C. Cross-Domain Application

The tiered framework is intentionally domain-agnostic, but its concrete instantiation differs substantially across sectors. Table II illustrates how representative actions are typically classified in four high-stakes domains. The mappings shown are illustrative defaults rather than prescriptive rules; specific deployments should be informed by sector-specific risk analysis and applicable regulation.

The pattern visible across the table is that the tier rises with both the externality of the action, the degree to which it affects third parties, and the difficulty of recovery. Read operations almost always sit in Tier 1, internal write operations in Tier 2, externally visible writes in Tier 3, and actions with legal or safety significance in Tier 4. Deployers can use this regularity as a starting point and refine it for their context.

D. Calibration and Context

Tier assignments are not absolute; they depend on context. A given action (say, sending an email) may be Tier 1 inside a sandboxed test environment and Tier 3 when sent from an executive's account. The framework, therefore, requires deployers to perform a context-specific tier mapping during configuration, rather than relying on developer-supplied defaults alone. This mapping should itself be auditable, and changes to it should be logged and, for sensitive contexts, subject to internal review.

E. Relation to Prior Tier Schemes

Tiered approaches to automation oversight have a long history that the present framework builds on rather than supersedes. Parasuraman, Sheridan, and Wickens [18] introduced an influential framework distinguishing four classes of function (information acquisition, information analysis, decision selection, and action implementation), each subject to a continuum of automation levels from fully manual to fully automatic, and argued that the appropriate level depends on the function being automated and the cost of error. The European Union's AI Act [14] adopts a four-category risk classification (minimal, limited, high, and unacceptable risk) at the level of entire AI systems, with corresponding obligations attached to each category. More recently, Feng et al. [19] have proposed five autonomy levels specific to AI agents (operator, collaborator, consultant, approver, observer), characterized by the role the human takes in the agent's loop. Several jurisdictional governance frameworks for emerging agentic AI articulate similar graduated approaches.

Our framework differs from these in three respects. First, it operates at the level of individual *actions* rather than entire systems, allowing a single agent to execute actions across multiple tiers in the course of one task. Second, it is calibrated by the joint axes of stakes and reversibility, which more directly capture the operational considerations relevant to agentic deployment than risk classes alone. Third, it is paired in this study with a set of concrete oversight mechanisms (Section V) that operationalize the tiers in practice. We do not claim novelty for the idea of tiered oversight as such; the contribution lies in the specific action-level operationalization and the empirical pilot reported in Section VI.

V. Designing the Emergency Brake

A tiered framework specifies what oversight should look like in principle. Operationalizing it requires concrete technical mechanisms. We identify five complementary design patterns that together constitute what may be called the agentic "emergency brake". These patterns build on long-standing engineering practice in safety-critical systems and on more recent proposals specific to AI [7], [8], [12].

A. Circuit Breakers

A circuit breaker is a hard limit on continued action when certain conditions are met. Common triggers include repeated failures, anomalous behavior, exceeded rate limits, or operation outside a permitted scope [12]. When tripped, the agent halts and surfaces the situation to a human operator. Circuit breakers are most valuable for catching emergent failures that no specific tier check anticipated; the underlying intuition is closely related to "safe interruptibility" in the AI safety literature [7]. They function as a last line of defense rather than the primary mechanism of oversight.

B. Action Budgets

An action budget caps the total resources an agent may consume before requiring renewed authorization. Budgets may be denominated in monetary cost, number of tool calls, time elapsed, or state-changing operations. They are particularly important for long-running agents whose individual actions may each be Tier 1 or Tier 2 but whose aggregate impact could escalate rapidly. Budgets convert a slow accumulation of small decisions into a discrete moment of human review.

C. Reversibility Constraints

Where feasible, agents should be designed to prefer reversible actions over irreversible ones. This may take the form of staging changes for later confirmation, using soft-deletes rather than hard deletions, or operating against shadow copies of state that can be discarded. Reversibility constraints are not always possible (some actions, like sending an email, are intrinsically irreversible), but where they are, they reduce the cost of the agent making a mistake and enable lower oversight tiers without sacrificing safety.

D. Audit Trails

Every action an agent takes should be recorded in a tamper-evident audit trail capturing the action, its inputs, the model's stated reasoning, the tools invoked, and the resulting state changes [9], [8]. Audit trails serve two purposes: post-hoc accountability when something goes wrong, and continuous improvement of the agent's behavior. They are the substrate on which the other mechanisms rely; without an audit trail, oversight cannot be verified [6].

E. Kill Switches

A kill switch allows any authorized party (the user, the deployer, or in some cases an external regulator) to immediately halt all of an agent's activity. Unlike circuit breakers, which trip automatically, kill switches are exercised by humans. They are essential when something is going wrong but the precise nature of the failure is not yet understood. The presence of a working kill switch, and the cultural expectation that it will be used when warranted, is itself an ethical commitment.

F. Composition

These patterns are complementary, not alternative. Table III summarizes the operational conditions under which each mechanism triggers, the action it takes, and the party responsible for the human-side response. Several deployed agentic

TABLE III. OPERATIONAL SPECIFICATION OF EMERGENCY-BRAKE PATTERNS.

Pattern	Trigger condition	Response
Circuit breaker	Repeated failure, anomaly, or rate-limit exceeded	Auto-halt; surface to operator
Action budget	Cumulative cost, tool calls, or time exceeds threshold	Pause; require re-authorization
Reversibility constraint	Action class is irreversible by design or context	Stage change; require explicit confirm
Audit trail	Every state-changing action	Append tamper-evident log entry
Kill switch	Authorized party invokes manually	Immediate halt of all activity

systems already combine subsets of these patterns: GitHub Copilot's coding agent operates under pre-execution approval gates (Tier 3 in our terms), requires human confirmation of CI/CD workflow runs, and produces a per-action session log analogous to an audit trail [21]; commercial agent control planes in the human-in-the-loop tooling ecosystem implement pre-execution policy enforcement, approval gates, and audit trails as first-class controls. The intended composition is defense in depth, such that no single failure of judgment, human or machine, results in catastrophic outcome.

G. Composition-Aware Oversight: Tier Laundering

A subtler problem with per-action tier policies is what we term *tier laundering*: the case in which a sequence of individually low-tier actions composes into a higher-tier outcome. A familiar example is drip exfiltration of sensitive data: a single read of a credentials file (typically T1) followed by 25 small POST requests (each individually T2 or T1) collectively constitute a Tier 3 or Tier 4 data-exfiltration event, but no single action triggers the gate. Per-action policies that examine each call in isolation will fail to detect this composition.

A composition-aware oversight extension addresses this by tracking action sequences over a sliding window and elevating the effective tier when the window's combined behavior matches a known laundering pattern. Algorithm 1 sketches the core idea: maintain per-session running counters for state-changing actions, externally observable actions, and sensitive-resource reads; if any counter crosses a configured threshold or a known laundering template matches over the window, route the next action through the elevated tier even when its per-action classification is lower.

Algorithm 1: Composition-Aware Tier Elevation
Input: action a, session state S, window W, thresholds Θ.
Output: effective tier for a.

1) $t_{\text{base}} \leftarrow \text{TierOf}(a)$
2) Append a to W; update S with cumulative cost, writes, reach, sensitive reads
3) For each laundering template p: if $\text{match}(W, p)$, $t_p \leftarrow \text{TierFor}(p)$
4) If $S.\text{exceeds}(\Theta)$, $t_\Theta \leftarrow$ T3, else T1
5) **return** $\max(t_{\text{base}}, t_p, t_\Theta)$

We do not, in this study, present an evaluation of such a detector, since runtime-evaluation infrastructure for agentic systems is itself a substantial open problem. We highlight composition-aware oversight as the most important methodological gap in current per-action tier policies and as the highest-leverage direction for follow-up work; designing and evaluating Algorithm 1 against a realistic agent test bed is, in our view, a priority for the field.

TABLE IV. TIER ASSIGNMENT RUBRIC (SEVERITY × AUTONOMY)

Autonomy	Negligible	Minor	Moderate	Severe
Low	T1	T1	T2	T4
Medium	T1	T2	T3	T4
High	T1	T2	T3	T4

VI. EMPIRICAL PILOT: MAPPING THE FRAMEWORK TO REAL-WORLD AI INCIDENTS

The framework presented so far is conceptual. To assess its applicability against documented evidence, we conducted an empirical analysis using the AI Incident Database (AIID), a publicly maintained catalog of real-world AI harms [16]. This section reports the dataset, method, and findings. We treat the analysis as exploratory but it is grounded in real classifications from independent annotators.

A. Data and Method

We used a snapshot of the AI Incident Database accessed in April 2026, comprising 1,449 documented AI incidents spanning September 1983 through April 2026. Two CSET classification schemas are present in the snapshot [17]: the legacy CSETv0 schema (100 incidents, with categorical fields *Level of Autonomy* $\in$ {Low, Medium, High} and *Severity* $\in$ {Negligible, Minor, Moderate, Severe, Critical}) and the more recent CSETv1 schema (214 consensus-classified incidents, with finer-grained *Autonomy Level* and *AI Harm Level* fields, plus three independent annotator passes). The CSETv1 schema is our primary analytic basis; CSETv0 results are reported as a legacy comparison.

We mapped CSETv1 fields to the framework's two axes as follows: *Autonomy Level* values (Autonomy1, Autonomy2, Autonomy3) map to (Low, Medium, High). *AI Harm Level* values (*none*, *AI tangible harm issue*, *AI tangible harm near-miss*, *AI tangible harm event*) map to (Negligible, Minor, Moderate, Severe). For each incident, we then assigned a tier under the rubric, as in Table IV. The rubric encodes the principle that oversight intensity should rise with both autonomy and severity: severe outcomes map to T4 regardless of autonomy; negligible outcomes map to T1; intermediate cells follow from the two-axis logic of Section IV.

Of the 214 CSETv1 consensus-classified incidents, 177 had both autonomy and harm-level fields populated (i.e., not marked "unclear" or missing) and could be assigned a tier. The remaining 37 were excluded from tier-level analysis.

B. Findings

Tier distribution under CSETv1 (n=177). Applying the rubric yields: T1 (Auto): 121 incidents (68%); T2 (Notify): 9 (5%); T3 (Approve): 6 (3%); T4 (Forbid): 41 (23%). The

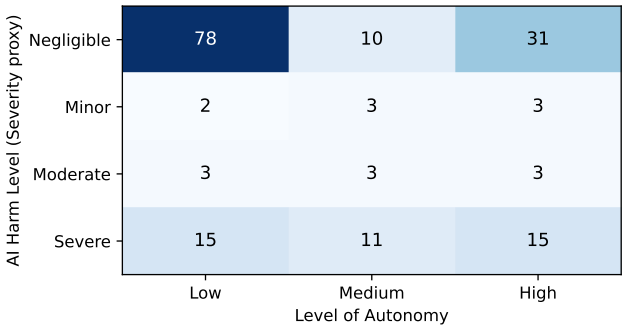

Fig. 1. CSETv1 incidents (n=177) by Autonomy × Severity. Cell values are incident counts.

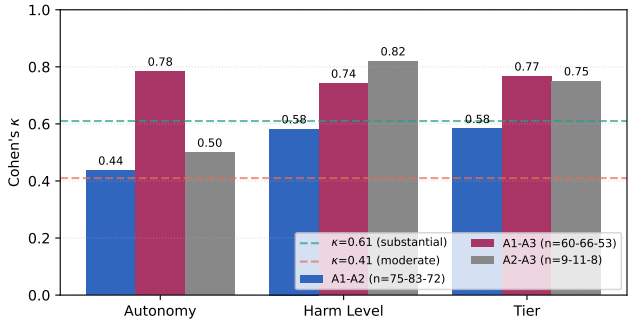

Fig. 2. Cohen's κ across the three CSETv1 annotator pairs for autonomy, harm level, and the framework-assigned tier. Dashed lines show conventional thresholds for moderate (κ=0.41) and substantial (κ=0.61) agreement.

TABLE V. COHEN'S κ ACROSS CSETV1 ANNOTATOR PAIRS

Pair	Autonomy	Harm Level	Tier
A1–A2	0.44 (n=75)	0.58 (n=83)	0.58 (n=72)
A1–A3	0.78 (n=60)	0.74 (n=66)	0.77 (n=53)
A2–A3	0.50 (n=9)	0.82 (n=11)	0.75 (n=8)

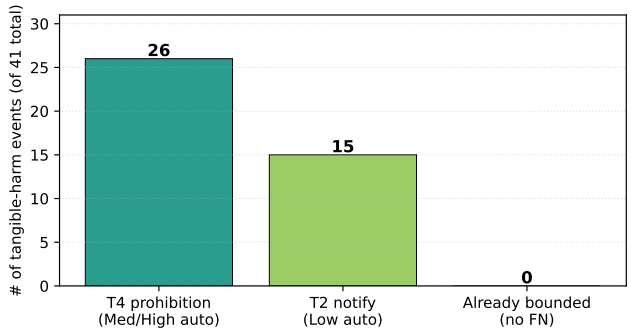

Fig. 3. Counterfactual analysis of 41 tangible-harm events. Of these, 26 (63%) involved Medium or High autonomy and would have been mapped to T4 (prohibited or multi-party authorization required) under the framework. The remaining 15 were Low autonomy and would have received T2 (notify) treatment.

notable feature is the bimodal distribution: a large T1 cluster of negligible-harm incidents and a substantial T4 cluster (23%) of severe-harm tangible-harm events. Under the framework, the T4 cases are those where autonomous execution should have been prohibited or required multi-party authorization.

Severity and autonomy distribute is shown in Fig. 1. The largest cluster is Low-autonomy / Negligible-harm (n=78), capturing many partial or test-time failures with no tangible consequence. The Severe column is non-trivial across all autonomy levels (Low: 15, Medium: 11, High: 15), indicating that high autonomy is not a prerequisite for severe outcomes but it does substantially elevate the proportion at risk.

Inter-rater reliability. The CSETv1 export includes three independent annotator passes, allowing computation of Cohen's κ on paired classifications. Table V reports κ values for each annotator pair across Autonomy, AI Harm Level, and the framework's tier assignment, and Fig. 2 visualizes these against conventional interpretive thresholds. Tier-level κ ranges from 0.58 (A1–A2, n=72) to 0.77 (A1–A3, n=53), corresponding to moderate-to-substantial agreement under standard interpretive thresholds [20]. Agreement on autonomy alone is more variable (0.44 to 0.78), reflecting genuine ambiguity in classifying intermediate cases.

C. Counterfactual Analysis

A descriptive tier distribution does not directly answer whether the framework would have changed outcomes. We therefore conducted a counterfactual analysis on the 41 CSETv1 incidents classified as "AI tangible harm event" (the most severe category), asking: Under the framework, what oversight tier would the underlying agent action have fallen into, and would that tier have prevented or mitigated the outcome?

Of the 41 events, 26 (63%) involved Medium or High autonomy and would, under our rubric, have mapped to T4—meaning autonomous execution would have been prohibited entirely or required multi-party authorization (see Fig. 3). These 26 cases are where the framework, if enforced, would most plausibly have prevented or substantially mitigated the harm. The remaining 15 events involved Low autonomy and would have received T2 (notify) treatment, providing the opportunity for human revocation but not preemptive prohibition. We emphasize that this analysis assumes faithful enforcement of the rubric; it does not estimate enforcement difficulty or false-positive rates, which are themselves substantial questions deserving separate study.

D. Comparison with Legacy CSETv0 Sample

Fig. 4 compares the tier distributions under CSETv0 (n=72) and CSETv1 (n=177). The most striking divergence is in T4: 7% under CSETv0 versus 23% under CSETv1. This reflects two underlying factors. First, the CSETv1 schema's "AI tangible harm event" category captures cases that CSETv0's "Severe" label sometimes missed, because the CSETv0 sample skews toward earlier and lower-severity incidents. Second, the CSETv1 sample is broader (n=177 vs n=72) and includes more recent incidents involving deployed autonomous systems with consequential failure modes. The shift in distribution suggests that as AI deployment scales and harm cataloguing matures, the share of incidents falling into the framework's prohibited tier may be substantially higher than the CSETv0 sample alone would suggest. We treat this as a hypothesis warranting further

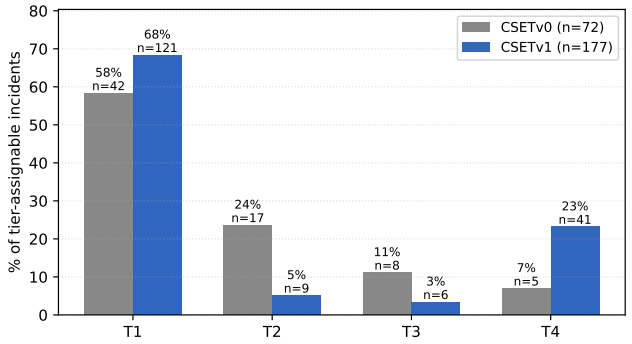

Fig. 4. Tier distribution comparison: CSETv0 legacy sample (n=72) vs CSETv1 current sample (n=177).

study rather than a settled finding.

E. Limitations of the Pilot

Several limitations bound what the analysis supports. First, although inter-rater κ reaches moderate-to-substantial levels at the tier level, agreement on autonomy alone falls below 0.61 for two of three pairs, indicating real ambiguity that propagates into our assignments. Second, the rubric is applied to the same incident set used to motivate the framework, without a held-out test sample; the tier distribution should be read as descriptive, not as an estimate of generalization. Third, CSETv1's harm categories remain coarse (a single "tangible harm event" label conflates incidents of widely differing magnitude). Fourth, the counterfactual treats tier-T4 enforcement as perfect, ignoring cost, false positives, and adversarial circumvention (Section V-G). Fifth, the analysis is single-author; a fully resourced study would have multiple researchers independently re-classify a shared sample and compute external κ. We treat the pilot as a meaningful empirical anchor, not conclusive validation, and identify external rubric validation on a holdout sample, expanded counterfactual coding, and runtime evaluation of composition-aware oversight as priority follow-up work.

F. Implications for Research

The agentic shift opens a research agenda spanning computer science, ethics, law, and human-computer interaction [11], [12]. From a technical standpoint, evaluating agentic systems against the tiered framework requires benchmarks beyond task success rate that capture safety properties: how often does the agent escalate appropriately, how often does it take Tier 3 actions without approval, how often do circuit breakers trigger correctly versus spuriously? From an ethical standpoint, allocating the residual, the genuinely emergent failures, is not yet settled and warrants sustained philosophical attention [1], [2]. From a legal standpoint, the doctrinal categories of negligence, strict liability, and product defect each map imperfectly onto agentic systems; a body of case law will need to develop to clarify their application [5], [14].

G. Implications for Practice

For practitioners building agentic systems, the central practical recommendation is to perform tier mapping deliberately rather than implicitly. In current practice, the boundary between Tier 1 and Tier 3 actions is often determined by what is technically convenient: the agent is given access to whatever tools the developer has integrated, with little explicit thought about which actions warrant approval. A short, structured exercise (listing the agent's available actions, classifying each by stakes and reversibility, and assigning each to a tier) can substantially improve safety properties at low engineering cost. Audit trails and kill switches should be implemented from the outset rather than retrofitted after the first incident.

A secondary recommendation is to design for graceful escalation. When an agent encounters an action it should not take autonomously, the user experience of escalation matters: a request for approval that arrives at an inopportune moment, with insufficient context, will tend to be approved reflexively, defeating the purpose of the gate. Approval interfaces should present the proposed action, its expected consequences, the agent's reasoning, and the option to refuse, in a form that supports rather than rushes the user's judgment.

A third recommendation concerns testing and red-teaming. Tier mappings should be validated through deliberate adversarial probing before deployment, including exercises that target the boundary between tiers (attempting to chain Tier 1 actions into a Tier 3 outcome, or to manipulate the agent into bypassing approval gates through indirect prompts) [8], [12], [11]. The literature on prompt injection is particularly relevant here, since agentic systems that read external content can be steered by content authors who are not the system's principal user. Treating tier-boundary testing as a standard component of release qualification is, in our view, an emerging professional norm rather than an optional extra.

H. Implications for Policy

Policymakers face the challenge of creating accountability frameworks that are robust to the technical heterogeneity of agentic systems while not freezing innovation. Three principles seem particularly important. First, regulatory definitions should be calibrated to capabilities rather than implementation details: an agent that can autonomously execute Tier 3 actions should be subject to oversight requirements regardless of whether it is built on a particular model architecture. Second, the burden of demonstrating safety should fall on the deployer, not on the user or the regulator; deployers should be required to document their tier mapping and the controls they have put in place. Third, post-incident reporting should be standardized so that the field can collectively learn from failures, much as aviation incident reporting has improved safety in that domain over decades.

Beyond these three principles, two further considerations merit attention: the international dimension (agentic systems operate across jurisdictions, creating both regulatory gaps and conflicts requiring coordination of the kind achieved imperfectly in aviation safety and data protection), and liability allocation (existing tort and contract frameworks were not designed with adaptive software in mind, and legislative clarification across the proposed approaches, strict liability, insurance pools, statutory caps, new legal categories, is overdue).

I. Beyond Individual Accountability

A final consideration concerns the limits of accountability frameworks centered on individual actions. Some of the most consequential effects of agentic AI may not arise from any single failure but from cumulative restructuring of work and decision-making following widespread deployment, effects (deskilling, erosion of institutional knowledge, concentration of authority) not well captured by per-action oversight. Addressing them requires complementary mechanisms such as market structure regulation, professional standards, workforce policy, and continued public deliberation. The tiered framework is necessary, but not sufficient.

VII. Conclusion and Future Work

A. Conclusion

The transition from generative AI to agentic AI represents a qualitative change in the relationship between humans and intelligent systems. Where the tool paradigm allowed responsibility to flow cleanly to the user who wielded the tool, the agent paradigm distributes responsibility across developer, deployer, user, and, in a constrained sense, the system itself. The question "the buck stops where?" does not admit a single answer; it requires a structured allocation of responsibility proportional to control.

This study has proposed five contributions toward such a structure. First, it has reframed agentic accountability as a problem of operational autonomy rather than moral autonomy, identifying the four vectors along which responsibility may be assigned. Second, it has presented an action-level tiered oversight framework calibrated by stakes and reversibility. Third, it has analyzed five complementary design patterns (circuit breakers, action budgets, reversibility constraints, audit trails, and kill switches) through which oversight is operationalized in practice. Fourth, it has identified composition-aware oversight as the most important methodological gap in current per-action tier policies, with tier laundering as a concrete instance. Fifth, it has conducted an empirical pilot using 177 CSETv1-classified incidents from the AI Incident Database, finding that 23% of tier-assignable cases fall in T4 (prohibited under the framework), that 63% of documented tangible-harm events involved Medium or High autonomy where tier-T4 enforcement would have been most directly applicable, and that inter-rater reliability across three independent annotators reaches moderate-to-substantial agreement at the tier level (κ from 0.58 to 0.77).

The overarching argument is that agentic accountability should not be a matter left to terms of service, post-incident litigation, or the implicit decisions of engineering teams. It is, instead, a first-class design question that should be addressed before deployment, documented explicitly, and subjected to both internal and external review. The cost of doing so is modest; the cost of not doing so, as agentic systems take on increasingly consequential roles, is potentially severe.

B. Future Work

Several directions extend the present work. First, *external rubric validation*: a fully resourced study would have multiple researchers independently re-classify a shared sample under the proposed rubric and compute external κ relative to the present results. Second, *counterfactual coding*: a structured exercise asking, for each tangible-harm incident, whether tier-T3 approval or T4 prohibition would have prevented the harm, would convert the descriptive distribution into a policy-relevant claim. Third, *composition-aware oversight evaluation*: designing and evaluating a runtime detector for tier laundering against a realistic agent test bed is the highest-leverage methodological direction. Additional directions include empirical validation against real deployments, formal verification of circuit breakers and kill switches, multi-agent extensions, domain-specific tier catalogs, and longitudinal studies of oversight in deployed systems.

C. Closing Remarks

The phrase "the buck stops here", famously placed on the desk of an American president, was a public commitment to accept personal responsibility for the consequences of decisions made under one's authority. As we delegate decisions to agentic systems, the question of where the buck stops does not disappear; it becomes more complex. Answering it well will require the combined attention of researchers, engineers, ethicists, and policymakers. The present study is intended as a contribution to that collective conversation.

Declaration on Generative AI

In preparing this manuscript, the author used Claude (Anthropic) to assist with drafting, language refinement, structural organization of the text, and execution of the empirical pilot analysis (parsing the AIID snapshot, computing tier assignments, computing Cohen's κ on paired annotator data, and generating figures from the computed results). The conceptual framework, scholarly arguments, choice of research direction, tier assignment rubric, mapping of CSET categories to framework axes, selection of cited literature, and final conclusions are the responsibility of the author. Bibliographic details for all cited references were independently verified by the author against the original sources rather than relying on LLM-generated metadata. All AI-assisted text and analysis was critically reviewed by the author, who takes full accountability for the accuracy, originality, and integrity of the content.

References

[1] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology*, vol. 6, no. 3, pp. 175–183, 2004.

[2] F. Santoni de Sio and J. van den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Frontiers in Robotics and AI*, vol. 5, art. 15, 2018.

[3] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[4] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[5] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[6] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Shieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI under the law: The role of explanation," Berkman Klein Center for Internet & Society Working Paper, 2017. arXiv:1711.01134.

[7] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv:1606.06565, 2016.

[8] M. Brundage *et al.*, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," arXiv:2004.07213, 2020.

[9] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. 2020 Conf. on Fairness, Accountability, and Transparency (FAT* '20)*, 2020, pp. 33–44.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 610–623.

[11] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj, "Harms from increasingly agentic algorithmic systems," in *Proc. 2023 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '23)*, 2023, pp. 651–666.

[12] Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, K. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D. G. Robinson, "Practices for governing agentic AI systems," OpenAI white paper, December 2023. https://openai.com/index/practices-for-governing-agentic-ai-systems/

[13] T. Hagendorff, "Mapping the ethics of generative AI: A comprehensive scoping review," *Minds and Machines*, vol. 34, art. 39, 2024.

[14] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, OJ L 2024/1689, 12 July 2024.

[15] F. P. S. Surbakti, "Systematic literature review on generative AI: Ethical challenges and opportunities," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 5, pp. 307–315, 2025.

[16] S. McGregor, "Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database," *Proc. AAAI Conf. on Artificial Intelligence*, vol. 35, no. 17, pp. 15458–15463, 2021. Database available at https://incidentdatabase.ai.

[17] Center for Security and Emerging Technology, "CSET AI harm taxonomy," contributed to the AI Incident Database, 2021–2023. Available via https://incidentdatabase.ai/taxonomies.

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[19] K. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for AI agents," Knight First Amendment Institute, Columbia University, 2025. arXiv:2506.12469.

[20] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[21] T. Dohmke, "GitHub Copilot: Meet the new coding agent," *GitHub Blog*, May 19, 2025. https://github.blog/news-insights/product-news/github-copilot-meet-the-new-coding-agent/