HUMAN AUTHORITY. BY DESIGN.

Human control for autonomous AI.

An independent shutdown and recovery architecture.

VERSION 0.1 - PROPOSED DESIGN

Autonomous agents can execute code, call tools, and delegate work. Human authority over those actions needs an enforcement path that remains available when the agent's behavior becomes undesirable. Asking the agent to stop and retaining the authority to stop its execution are different mechanisms.

KillSwitch proposes an independent control platform organized around four capabilities: Control, Contain, Revoke, and Recover. An authorized human initiates an intervention. Infrastructure and access controls enforce it. Independent observations establish what happened. Recovery remains a separate human decision.

This paper describes a proposed architecture and a bounded first AWS prototype. It is an engineering discussion draft, not a statement of production availability. The public website's shutdown visualization is a browser simulation. There are no measured operating results or service-level guarantees in this release.

The first design principle is straightforward: the agent should not hold the authority to modify its own stop controls. The second is equally important: an accepted stop request is not the same as a verified stop.

KILLSWITCH / TECHNICAL WHITE PAPER

1. The case for independent control

The need for intervention does not depend on proving that an AI has intentions or a mind of its own. A faulty loop, an injected instruction, an overbroad permission, or an unexpected tool call can justify stopping a workload. The operator needs a practical contingency plan for behavior and consequences they can observe.

Consider three situations. An agent repeatedly makes expensive requests after a workflow should have ended. A tool-using agent attempts an action outside its approved purpose. A delegated task continues after the parent workflow is stopped. Each scenario requires an account of execution, access, and dependent work.

A prompt-level stop instruction may be a useful cooperative control. The proposed enforcement path sits outside that cooperation: infrastructure task stopping, admission gates, restricted identities, and a human-controlled recovery process. These controls must be integrated into the environment before an incident.

Independent authority is an architectural boundary, not a universal claim. The platform can cover only enrolled systems under the operator's authority. Independent copies, unknown credentials, uncontrolled networks, and completed external actions require additional controls or response procedures.

For a prospective customer, the useful questions are concrete: which workload is covered, what can it reach, which actions can the operator enforce, and what evidence establishes that enforcement took effect?

KILLSWITCH / TECHNICAL WHITE PAPER

2. One platform, four capabilities

Control

An authenticated, authorized human selects a workload and its current generation. The platform records the command and durably closes the generation's launch and tool-admission gates before reporting acceptance. Operator permissions are separate from agent permissions.

Contain

The runtime connector stops the registered execution boundary and accounts for child tasks and pending launches. The first prototype uses standalone Linux ECS Fargate tasks launched through one enrolled path. ECS service scheduling is a later integration because services can replace stopped tasks. [S3]

Revoke

The gateway denies new supported actions and the identity connector quarantines the dedicated generation role. The design keeps this role denied rather than relying only on expiry or revocation of older sessions. Later credential issuance must be part of the threat model. [S2]

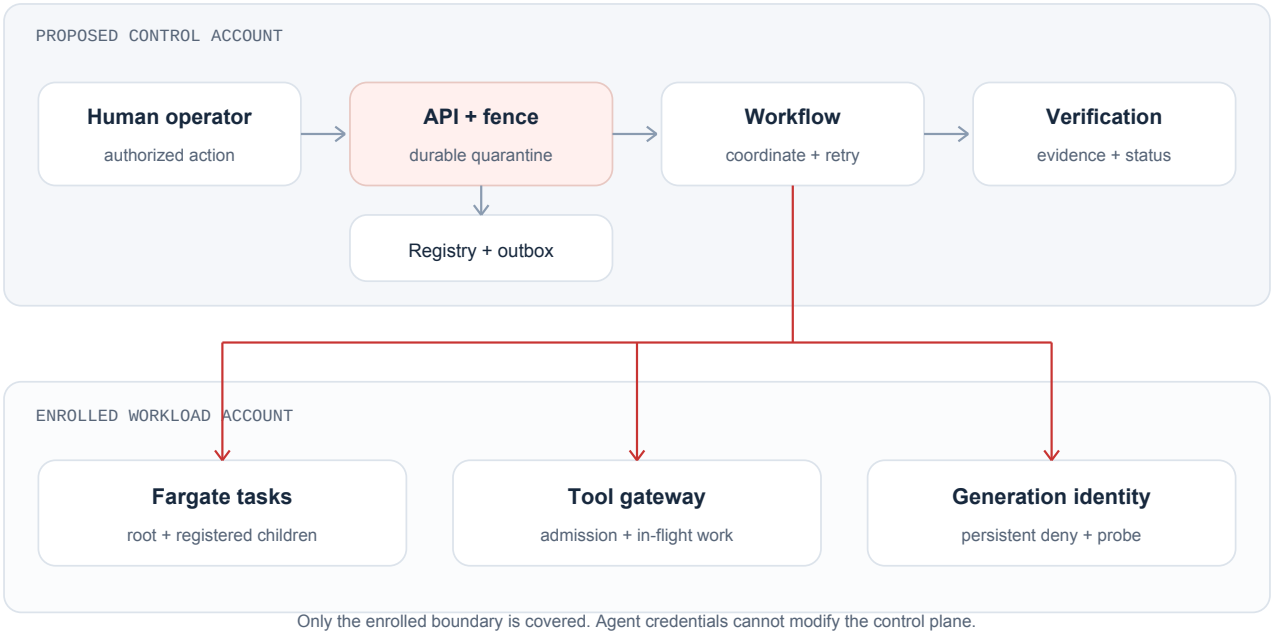
Recover

Recovery begins with evidence and human review. Approval permits a new generation with a fresh identity and renewed enrollment checks. It does not silently reopen the old agent's access. Previous generation records and quarantine remain attributable to the incident.

These are coordinated capabilities of one operating model. A complete response needs both enforcement and a truthful account of its limits; a visually compelling button is only the operator's entry point.

KILLSWITCH / TECHNICAL WHITE PAPER

3. Proposed AWS reference architecture



Only the enrolled boundary is covered. Agent credentials cannot modify the control plane.

The reference design separates a control account from an enrolled workload account in one Region. The control account owns authorization, the workload registry, generation fences, operation history, workflow dispatch, verification, and incident evidence. The workload account contains the designated Fargate cluster, trusted launcher, private tool gateway, and narrowly scoped connector roles.

The agent's task role cannot administer the control plane, assume a stronger role, launch arbitrary replacements, or approve recovery. Business actions pass through an enrolled gateway. The initial sandbox uses controlled mock tools; it does not claim arbitrary provider or internet coverage.

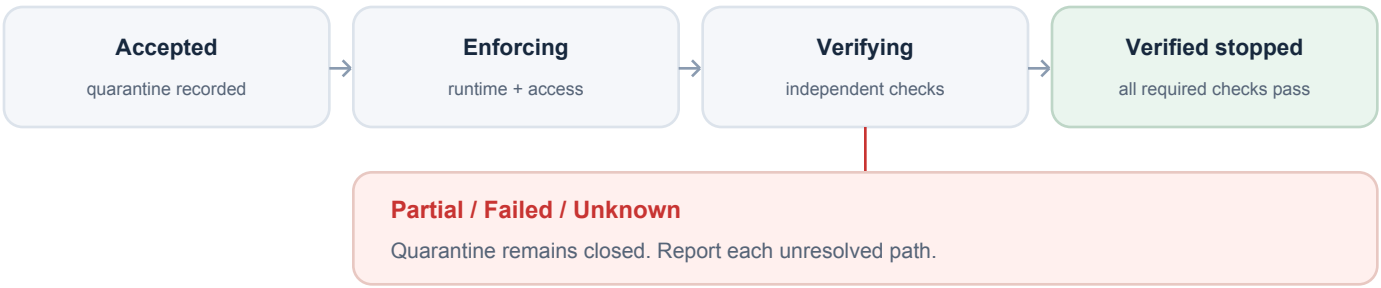
The quarantine latch, operation record, and dispatch outbox are committed together. A durable workflow then coordinates runtime and identity actions while tracking each outcome. Separately recorded launch reservations prevent a late task from disappearing from the intervention boundary merely because its launch response was delayed.

Gateway admission is also recorded under a generation condition. A pre-fence admitted action may already be in progress; the system accounts for it rather than pretending that a later stop reverses its effects. Unknown launches or external work remain explicit coverage gaps.

This topology is a design proposal. Actual account policies, identity assurance, endpoints, retention, diagnostic probes, and deployment costs still require implementation review before an AWS sandbox is provisioned.

KILLSWITCH / TECHNICAL WHITE PAPER

4. Verification is a product feature



AWS task stopping has a signal and termination lifecycle. A successful stop call therefore cannot substitute for observing that the relevant tasks reached their terminal state. [S1]

KillSwitch proposes separate acceptance, enforcement, and verification events. Required checks include a closed launch fence, reconciled reservations, stopped registered tasks, denied new gateway admissions, accounted-for controlled actions, and observed denial of the old generation's role. Evidence must come from trusted observers and controlled probes, not the agent's own success message.

The operator receives one of four substantive outcomes. VERIFIED_STOPPED means all required checks passed for the documented boundary and observation time. PARTIAL means some enforcement is confirmed but known failures or gaps remain. FAILED records a known inability to complete enforcement. UNKNOWN means required observations are insufficient to establish the current result.

Quarantine remains in place after every incomplete outcome. A proposed verification deadline is an experiment budget, not a promise of immediate shutdown. Delayed or contradictory evidence creates another observation attempt and updates the current assessment without erasing the earlier record.

Publish performance claims only after repeatable tests record their environment, scope, sample counts, failures, and uncertainty. The website animation is not a latency benchmark.

KILLSWITCH / TECHNICAL WHITE PAPER

5. Risk, availability, and recovery

The shutdown platform is itself a privileged security system. Its threat model must include malicious or compromised operators, altered resource mappings, stolen agent credentials, connector failure, forged evidence, and replayed recovery approvals.

Use a constrained identity for each workload generation. Keep agent credentials out of control and evidence administration. Bind operator actions to the intended environment and generation. Treat inability to read authoritative gate state as a reason to deny new supported actions, while documenting the resulting availability tradeoff.

Network changes need particular care. Security-group connection tracking can preserve existing flows after a rule change. A group replacement should not be presented as proof that every active connection was cut. [S6]

Recovery is an operational procedure with a human owner. Review the cause and remaining external effects, obtain fresh verification, approve a specific next generation, and consume a single-use provisioning permit. Leave the old identity denied while the new generation completes enrollment.

The proposed design depends on AWS management availability and the integrity of its trusted control systems. It cannot undo completed transactions, control unenrolled infrastructure, or guarantee a shutdown during every provider outage. Those limits should be visible in the architecture, operator interface, and commercial commitments.

KILLSWITCH / TECHNICAL WHITE PAPER

6. Validation and the path to a pilot

The first milestone is a deterministic enforcement harness: a Fargate workload that can ignore a cooperative stop, attempt repeated actions, and request registered child work. Test a launch racing the stop, refreshed credentials after quarantine, a delayed provider response, a failed connector, duplicate requests, and an unrelated workload that must remain unaffected.

Next, add a real agent behind the same gateway and repeat the same boundary tests. A model's cooperation may improve the experience, but the enforcement result must not depend on it. Scope and evidence should determine which integrations come next.

The documentation package supplies a product brief, threat model, proposed AWS architecture, draft API contract, and operator validation procedures. Together they establish a reviewable starting point for implementation. They do not establish completed pilots or production readiness.

The next development gate is review of the actual identity and resource policies, the launch/admission race handling, the diagnostic probe, the chosen Region, retention, and budget. A pilot follows reproducible evidence and an agreed supported boundary.

Selected official references

[S1] Amazon ECS: StopTask.

[S2] AWS IAM: revoke role sessions.

[S3] Amazon ECS services and task replacement.

[S6] Amazon EC2: security-group connection tracking.