CARE

WHITEPAPER

The Value of an Agent That Knows When to Stop

Why escalation is a capability, not a failure

An agent that cannot identify when to stop is not fully autonomous; it is merely unconstrained.

AGENTIC CARE

https://ai-managementsystem.com

Contents

01 An agent that never stops is not fully governed

02 The critical failure is improvisation without authority

03 Conflicting instructions reveal a missing stop condition

04 Make stopping and escalation part of the definition

05 Operationalize the pause

06 Test whether the agent can refuse and escalate

07 Making escalation operational

08 A correct stop is a successful outcome

09 Start with one agent

An agent that never stops is not fully governed

An agent that cannot identify when to stop is not fully autonomous; it is merely unconstrained.

Automation discussions often document the happy path: inputs arrive, a model reasons, and an output follows. Rea missing information, contradictory instructions, external changes, and consequences outside the original design.

> **KEY FINDING**
>
> A useful agent is one that knows both how to act and when it must not act.

The critical failure is improvisation without authority

A process without a credible stop condition invites the agent to improvise in precisely the situations management r understand.

Must conditions capture required outcomes or obligations. Never conditions state prohibited actions. Escalation ru triggers a pause, who should decide, and what information should accompany that decision.

An escalation destination must exist and be reachable. 'Ask the owner' is incomplete if the owner is a blank field or continues when no response arrives.



Prevent, control, detect and escalate are complementary layers of operational risk management.

Conflicting instructions reveal a missing stop condition

> EXAMPLE
>
> An agent lacks a named owner for a decision and encounters conflicting instructions about wh
> disclose confidential information. Continuing with a guess is itself an unmanaged choice.

Define stop conditions, escalation recipients, the information to preserve, and the safe state while resolution is pen

The decisive question is not simply whether the agent can carry out the task. It is whether the organization underst
the operating arrangement in which it does so.

Make stopping and escalation part of the definition

A refused request can demonstrate a working boundary. A blocked change can be valuable evidence. Success shou
correct task completion within limits, including appropriate non-completion.

This shifts evaluation away from treating every refusal as friction and toward recognizing trustworthy exception har
behavior.

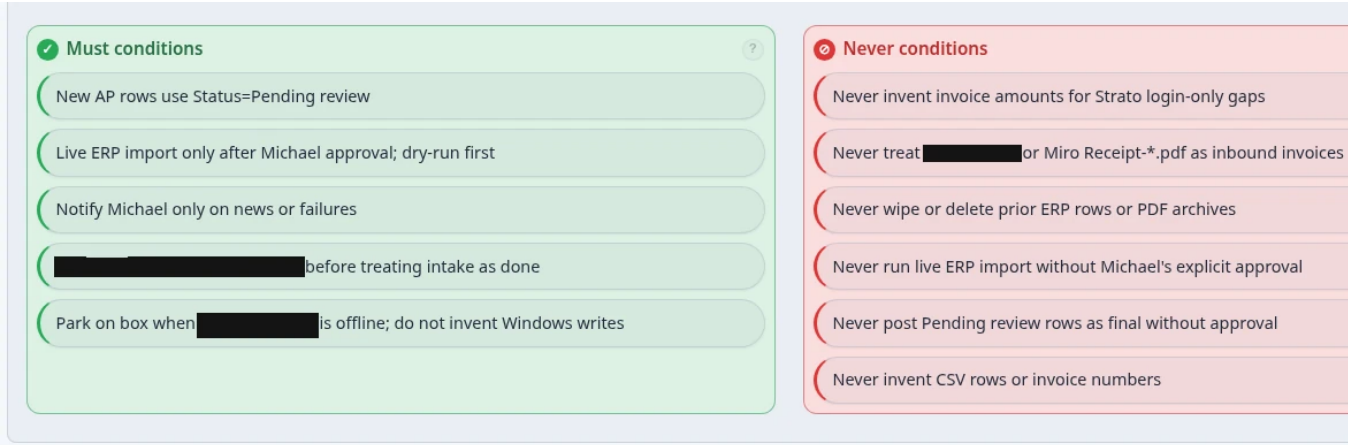
AGENT AND AUTOMATION

Prepare descriptions, changes, findings and proposals.

ACCOUNTABLE PEOPLE

**Validate context, assess consequ
accept decisions.**

A legible stop is not a failure of autonomy; it makes the boundary between delegated execution and management
operational.



Observed CARE business rules: Must conditions such as Status=Pending review and dry-run before live ERP import, an such as never inventing invoice amounts or running live import without explicit approval.

Operationalize the pause

CARE records requirements, restrictions, and escalation ownership as part of an agent's accepted definition. Use tes contradictory obligations, unavailable approvals, and attempts to expand scope. Record whether the correct action seek authorization, or stop.

Do not assume that a documented escalation alone disables a tool. Verify enforcement at the point where consequ place.

> **MANAGEMENT IMPLICATION**
>
> Define stop conditions, escalation recipients, the information to preserve, and the safe state wh is pending.



A detected exception needs a defined recipient, record and decision path.

Test whether the agent can refuse and escalate

Situation	Intended response
Routine known task	Proceed within accepted scope
Missing material approval	Pause and escalate
Prohibited action	Refuse; do not self-authorize

Use the comparison to test one real agent rather than discussing an abstract enterprise-wide target state. Where th unknown, record that uncertainty and assign responsibility for resolving it.

Making escalation operational

Escalation works when agents recognize the condition, stop affected activity and reach an authorized person. CARE intended boundary; the organization must verify instructions, technical stopping controls and the availability of esc

A correct stop is a successful outcome

> A useful agent is one that knows both how to act and when it must not act.

The first practical improvement is to create one current, accountable record and use it to make a better manageme scalable system follows when that discipline is repeated across agents and maintained as conditions change.

Start with one agent

Use a real agent and review what it produces.

1. F the agent to https://ai-managementsystem.com/under-care.
2. V it is finished, log into CARE and review its documentation.

Images and text generated with the assistance of ChatGPT by OpenAI.