CARE

WHITEPAPER

The Hidden Risks of Agents That Follow Every Rule

Why permission compliance does not guarantee business safety

The agent may be obeying its instructions perfectly while the instructions describe an unsafe operating arrangement.

AGENTIC CARE

https://ai-managementsystem.com

Contents

01 An obedient agent can still be dangerous

02 The operating arrangement can be the risk

03 When an innocent name hides financial authority

04 Assess the process, not just the individual agent

05 The human remains the risk owner

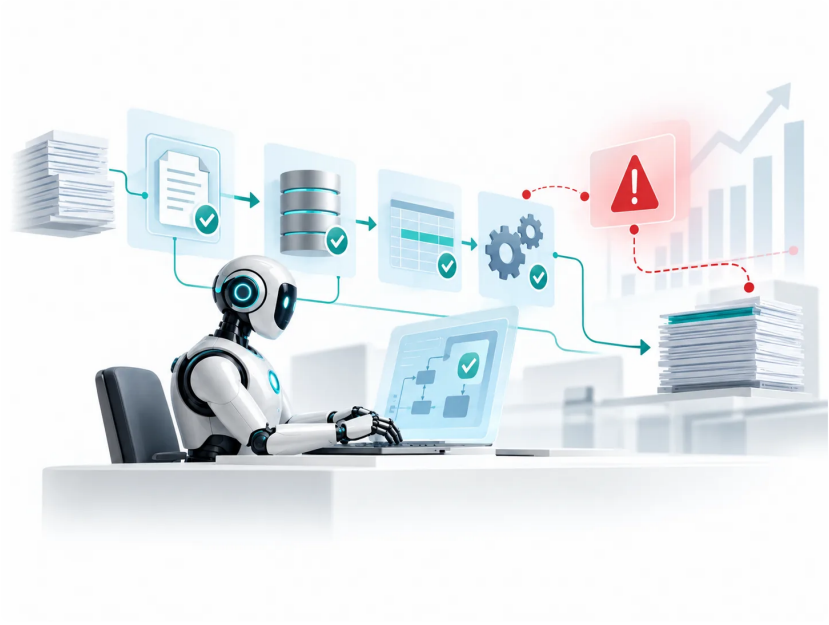
06 Make risk acceptance an operating-model decision

07 Accepting responsibility for the operating model

08 Risk ownership goes beyond permission compliance

09 Start with one agent

An obedient agent can still be dangerous



A sequence of permitted actions can still produce an unacceptable outcome.

The agent may be obeying its instructions perfectly while the instructions describe an unsafe operating arrangemer

Many reviews ask whether an agent exceeded its authorized access. This matters, but it misses risks created by the authorized work. An agent allowed to edit a financial record might act exactly within its granted rights and still intr unacceptable segregation-of-duties or approval problem.

The operating arrangement can be the risk

The failure lies upstream: the organization granted a capability without deciding how that capability should be cont

In a CARE demonstration, an agent previously described in benign terms was renamed to reflect its actual ERP-relat examining the operating definition revealed a more severe financial-record integrity exposure. The agent could pro gate, but accepting that control belonged to a human.

This illustrates the value of making activity explicit. It is a product scenario, not evidence that every agent will detec risks on its own.

> **MANAGEMENT IMPLICATION**
>
> A management approach should make the underlying decision visible without automatically m administrative work.

When an innocent name hides financial authority

> **EXAMPLE**
>
> An ERP agent has legitimate write access to financial records, and accurately follows instructior them. The process, however, never established an independent approval step for material chan

The issue is not whether the agent misbehaved. It is whether the business approved a process that makes conseque possible without an adequate control.

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.

Assess the process, not just the individual agent



An emerging business risk can trigger reassessment and a proposed human approval control.

Assess what a successful authorized action could do: alter revenue records, expose customer information, make an employment decision, or publish misleading claims. Ask which party can reverse the outcome, whether approval mu and whether trustworthy audit evidence is retained.

A useful risk statement specifies an action, an affected asset or person, a plausible consequence, and the condition consequence possible.

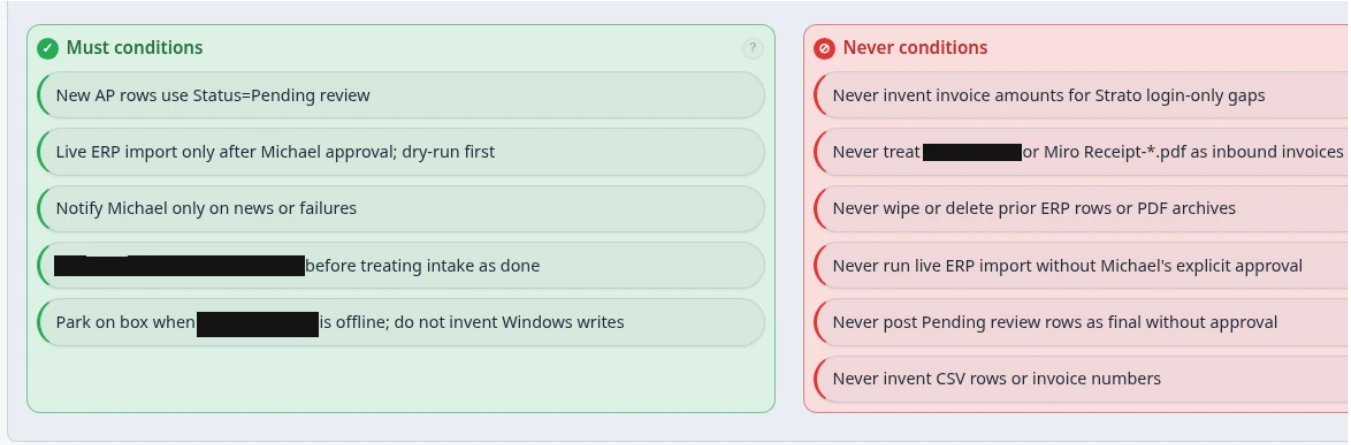
AGENT AND AUTOMATION

Prepare descriptions, changes, findings and proposals.

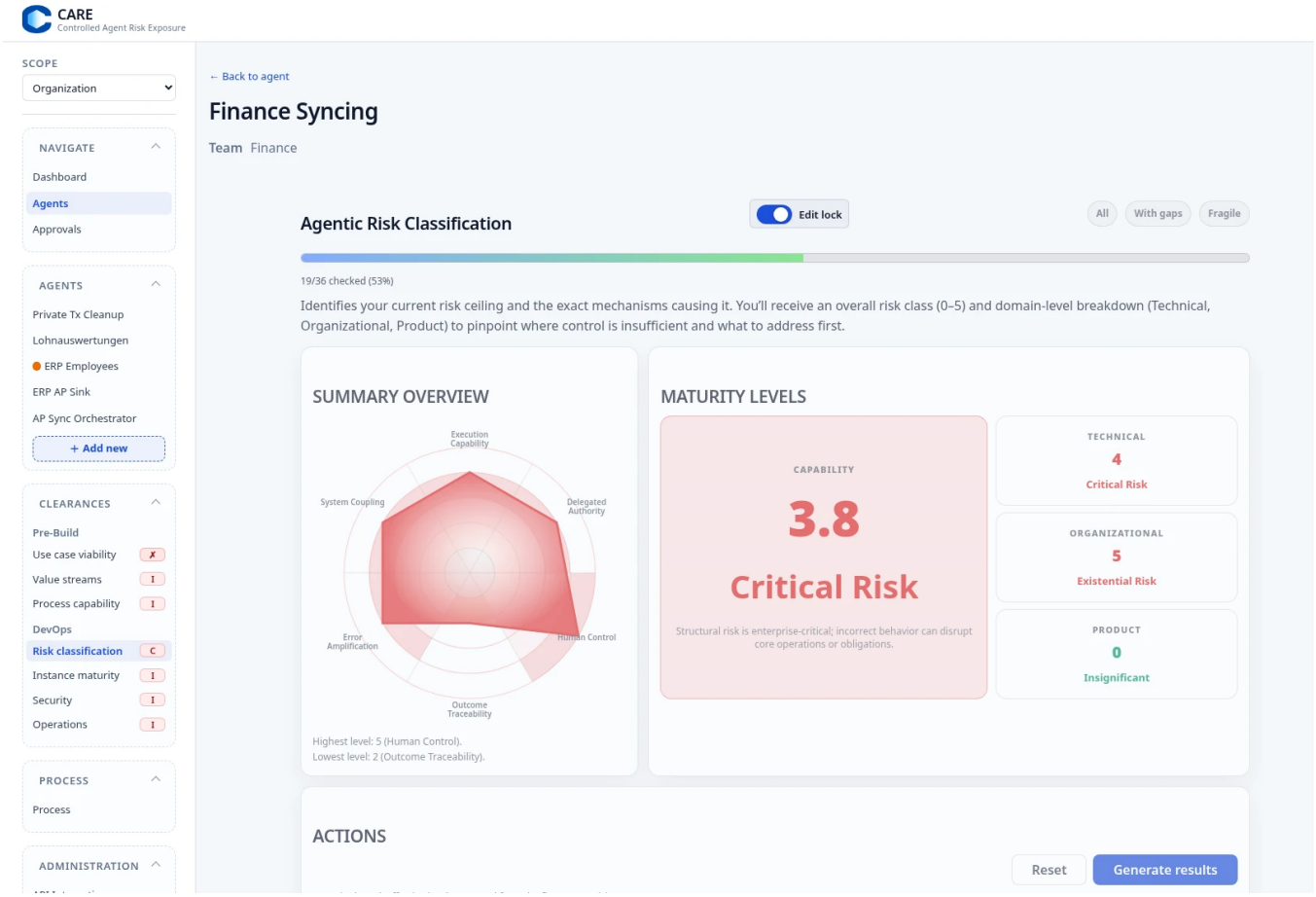
ACCOUNTABLE PEOPLE

Validate context, assess consequ accept decisions.

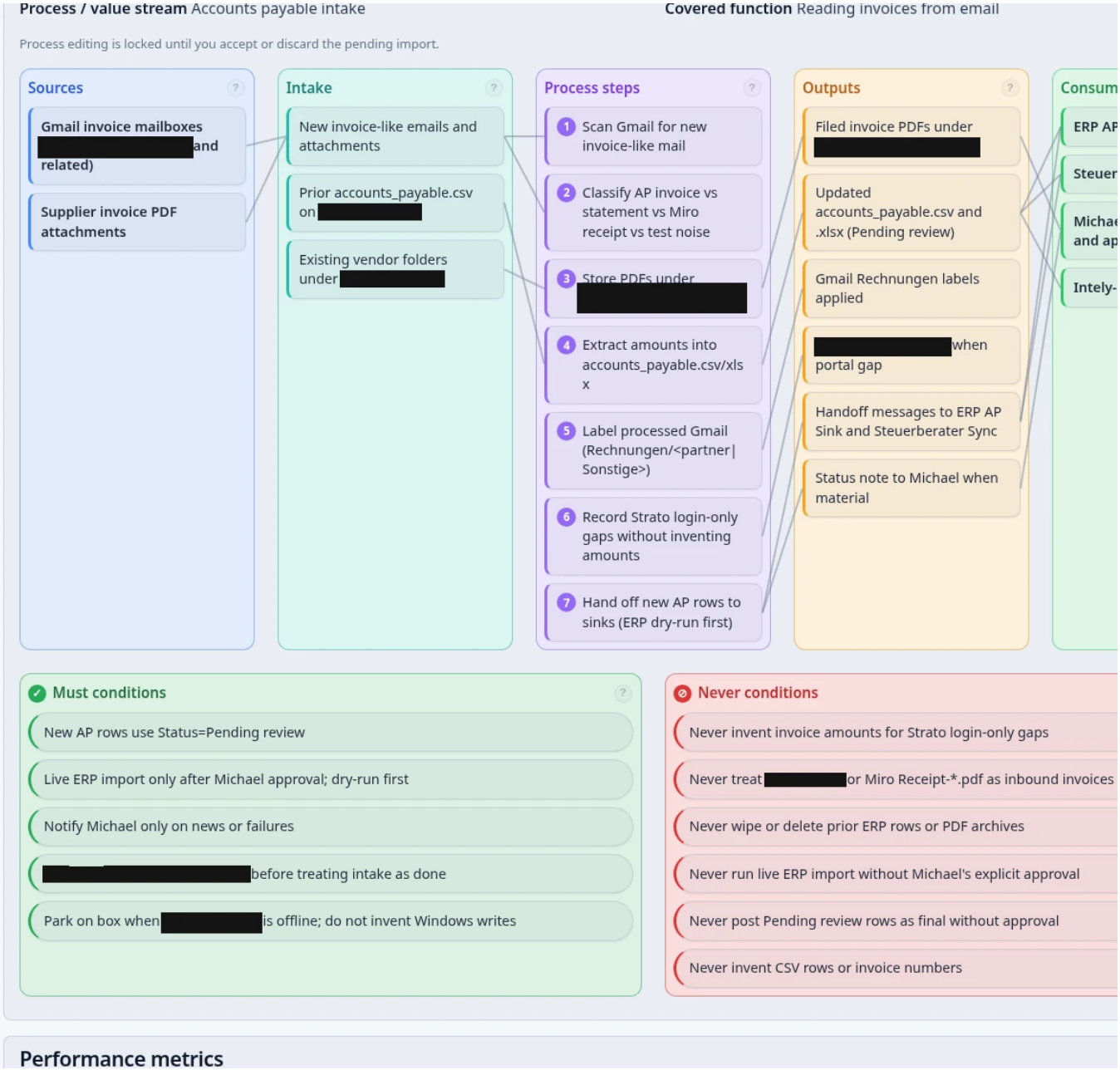
CARE can bring the declared process and its risk into review, have a mitigation proposed, and record the human re



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.



Observed CARE Agentic Risk Classification for Finance Syncing: capability score 3.8 labeled Critical Risk, with Technical Organizational 5 domain scores (19/36 checklist items checked).



Observed CARE SIPOC-style value stream for Accounts payable intake: Gmail sources, seven process steps, outputs inc review AP rows, and consumers including ERP AP Sink and Steuerberater Sync.

The human remains the risk owner

The agent can identify a control gap and recommend a mitigation such as dual approval, narrowed scope, legal rev independent audit trail. Management must decide whether the recommendation is adequate and whether to accep

Recorded acceptance does not magically make the process safe. Controls must also be implemented and tested in consequences occur.

> **MANAGEMENT IMPLICATION**
>
> The issue is not whether the agent misbehaved. It is whether the business approved a process

Use this comparison to distinguish the current operating habit from the management capability the organization n

Agent-centric question	Business-control question
Did it stay within permissions?	Were those permissions appropriate?
Did it follow instructions?	Were the instructions safe for this process?
Can we replay the output?	Can we verify the consequential system change?

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.

Accepting responsibility for the operating model

Even perfect compliance with an agent's instructions cannot establish that its business process is acceptable. Mana responsibilities, permissions, dependencies and escalation arrangements together, and accepts or changes the resu model.

Risk ownership goes beyond permission compliance

> An obedient agent can be dangerous when the operating model itself is defective.

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.

AGENTIC CARE

Instant AI Management System for your Agents, as a Managed Service.

https://ai-managementsystem.com