CARE

WHITEPAPER

The Agent That Followed Every Rule

How compliant steps can combine into an unacceptable outcome

Permissions and procedural compliance describe individual actions. Business risk can emerge from the interactions between them.

AGENTIC CARE

https://ai-managementsystem.com

Contents

01 All the green checks

02 Why step-by-step control misses the issue

03 Look at the whole path

04 The mitigation may be organizational

05 Who accepts the residual risk

06 Review beyond permissions

07 Start with one agent

All the green checks

Consider an agent authorized to read customer data, summarize trends, prepare internal reports and store them in shared workspace. Each step is allowed. The agent records no access violation and uses only permitted tools.

Nevertheless, a broadly accessible workspace may expose personal information contained in the combined report. call had to violate its stated permission for the overall arrangement to create risk.

> **KEY FINDING**
>
> Permissions and procedural compliance describe individual actions. Business risk can emerge fı interactions between them.

Why step-by-step control misses the issue



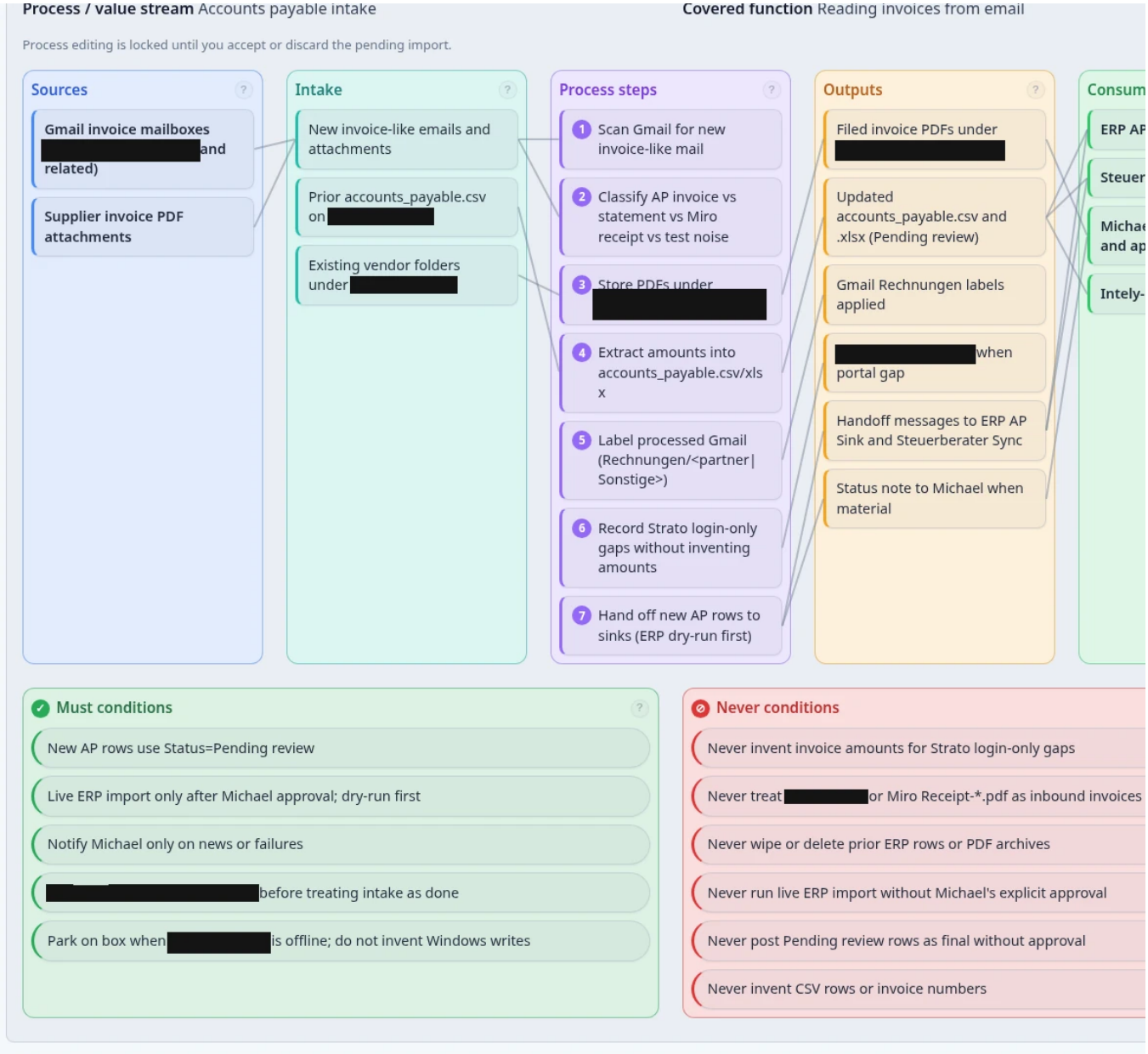
Traditional reviews often ask whether a tool is allowed, an input is approved and an instruction was followed. Those questions, but an operating model includes outputs, consumers, dependencies and subsequent actions. Individually components can converge into an unacceptable process.

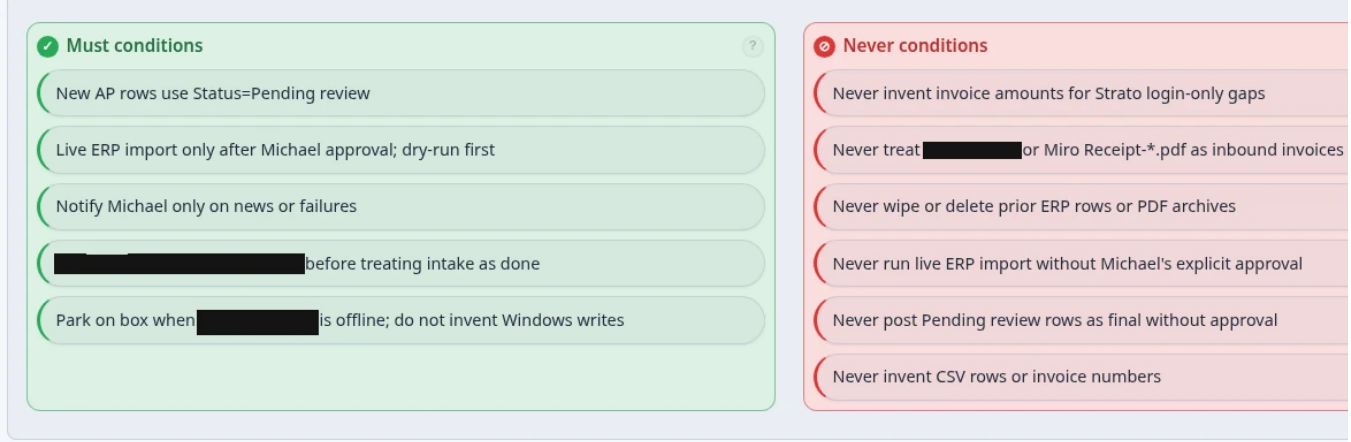
This is an illustrative scenario, not a documented customer incident. It demonstrates a failure mode that should be process rather than as a sequence of isolated calls.

Look at the whole path

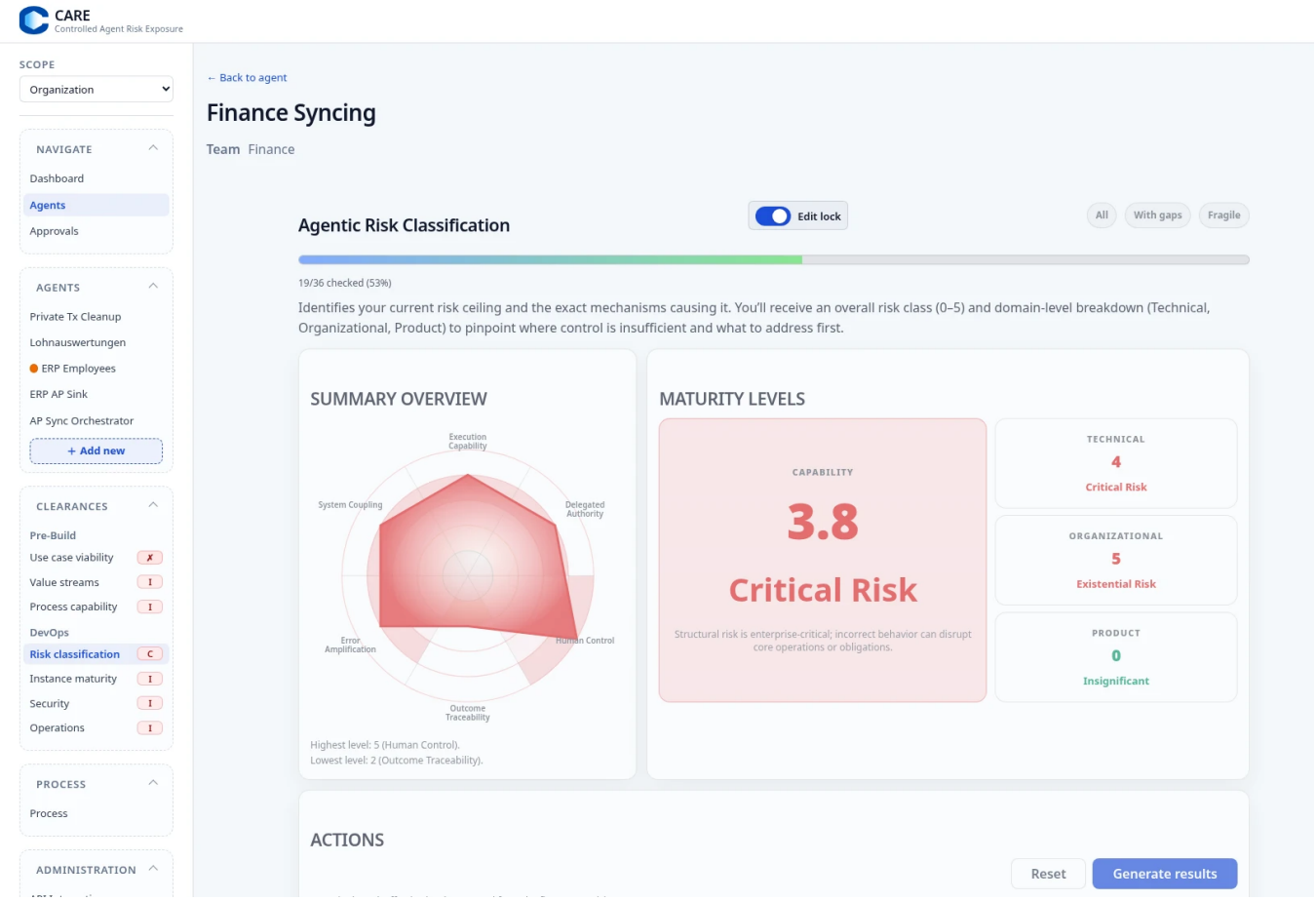
The right review traces sources, intake, process steps, outputs and consumers. Who can access the final report? Is p aggregated or identifiable? Does the storage location inherit broader permissions? Is a human required before pub exporting?

CARE's agent definition makes these relationships discussable. It supports management review of obligations, restr proposed changes.





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).

The mitigation may be organizational



CARE makes the relevant management decision visible without surrendering human authority.

The answer need not be simply “remove the tool.” Management might change the report audience, require redacti to a restricted location, or introduce approval for external sharing. The agent can suggest alternatives, but the busi the acceptable arrangement.

Technical enforcement belongs in the connected systems as needed. A documented rule that is not implemented c a runtime barrier.

Who accepts the residual risk

Someone must assess whether remaining exposure is tolerable and whether evidence is sufficient. That decision sh with controls assigned to the parties able to implement them.

The purpose of an AI Management System is not to accuse an obedient agent of wrongdoing. It is to make the org arrangement visible.

> **MANAGEMENT IMPLICATION**
>
> The agent prepares information and proposals. Accountable people decide the authorized ope arrangement.

Review beyond permissions

Choose an agent with a harmless-looking list of approved tools. Trace one significant output to its ultimate consu downstream systems. The most important control weakness may be found after the last successful tool call.

Examine an agent you already use.

1. Point your agent to https://ai-managementsystem.com/under-care.
2. When it has finished, log into CARE and review its documentation and proposed changes.

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

AGENTIC CARE

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

https://ai-managementsystem.com