Corresponding essay: **E11** “The Sentence That Owns the Agent”

RUNTIME DEFENSE FOR ONE AGENT LOOP THAT READS OUTSIDE CONTENT

Four-Corner Runtime Defense Checklist

Take one agent workflow that reads outside content, map its input streams and the lethal trifecta, then work the four corners. A missing corner is a corner a sentence can walk through.

BRING One agent workflow that reads email, web, tickets, issues, documents or tool responses

Map the loop and find the trifecta

List the workflow's input streams, find the lowest-trust one a stranger can write into, then ask whether the same session also holds private data and an outbound channel.

Question	Your answer
Workflow and its input streams (email, web, tickets, issues, documents, tool responses, logs)	
Lowest-trust stream - the one a stranger can write into	
Does the same session hold private data?	
Does the same session hold an outbound channel?	
Trifecta present? If yes, which corner do you build first?	

Corner 1 - Source labeling

Who said this token? Enforced at the context builder. Treat the model's instruction hierarchy as a useful prior, never as the trust boundary.

- [ ] The context builder labels each span by source: system, user, approved tool, retrieved web, customer content, public issue, private file.
- [ ] Untrusted spans are visibly marked as content to process, not instruction to obey.
- [ ] The harness owner is named as the accountable owner for this corner.

Corner 2 - Trifecta break

Does one session hold private data, untrusted content and an outbound channel? Enforced at the session planner.

- [ ] No single session holds all three of private data, untrusted content and an outbound channel.
- [ ] Where all three are genuinely needed, the workflow is split with a human or deterministic checkpoint between the halves.
- [ ] The security architect is named as the accountable owner for this corner.

Corner 3 - Capability flow control

Can untrusted content choose the next tool, URL, command, recipient or file? Enforced at the tool gateway.

- [ ] Control flow stays in trusted code; the model transforms data inside a constrained lane only.
- [ ] The tool gateway rejects any action whose selector came from untrusted text.
- [ ] Untrusted content can be summarized but cannot choose which private file is fetched or where the result is published.
- [ ] The runtime owner is named as the accountable owner for this corner.

Corner 4 - Output gating

What leaves the boundary, and under whose policy? Enforced at a final egress checkpoint.

- [ ] A final outbound checkpoint verifies data provenance, applicable policy, destination and approval before any side effect.
- [ ] Sensitive fields are blocked or redacted, and high-risk sends route to human review with a trace identifier.
- [ ] Sandboxing and constrained network access limit what a successful injection can reach.
- [ ] The platform or security owner is named as the accountable owner for this corner.

Notes - which corner you build first and why

Record the corner you will build first for this workflow and the decision that follows.

Companion worksheet to **Essay 11 · The Sentence That Owns the Agent**, in the series **Architecting the AI Coworker**. · Dr Peter McCann Strain · Fill this in against one real agent, action class or vendor.

Series Companion + all 22 worksheets: **Release_v12/Series_Companion.pdf**