SECURITY BRIEF: CLAUDE CODE

August 2026

Guidance for the alternative investment industry

Claude Code is Anthropic's agentic coding assistant: given access to a codebase, it can plan, write, and edit code, run tests, and iterate until a task is complete, with the same tool-call and MCP connector access as Claude CoWork. By default it asks before touching anything outside its designated folder, but that guardrail is a setting, not a wall, and the user can turn it off. Atlas rates Claude Code R3: MODERATE, appropriate for company computers under defined guardrails, not risk-free by default.

R1 – MINIMAL	R2 – LOW	R3 – MODERATE	R4 – HIGH	R5 – CRITICAL
Governed Hosted Chat	**Governed SaaS Chat**	Constrained SaaS Agent	Privileged SaaS Agent	Unmanaged Local Gateway

The R3 rating comes down to one guardrail: Claude Code prompts for most actions by default, where CoWork instead grants scope upfront and gates only specific moves like deletion. It also still runs through Anthropic's certified cloud rather than an unmanaged local gateway, unlike OpenClaw.

Data Protection and Exfiltration Risk

Claude Code's value is how actively it reaches into a codebase, and that's also where the exposure sits. Two points matter most for capital markets:

Active Reach Across Files and Networks

Claude Code can find, read, and iterate across large directories, network folders, and websites more actively than Claude Chat, and combined with tool calling, it can locate and mount network drives on its own, if directed/permited.

Prompt Injection Through What It Reads

A file Claude Code encounters can carry malicious instructions it reads as commands, and MCP connectors raise the stakes: integrated services can be called automatically, without needing to trace back to anything the user actually asked for.

Governance, Auditability, and Defensibility

For institutional clients, the question is whether Claude Code's actions can be reconstructed after the fact. Right now, two gaps stand out:

Action Without Consultation

Claude Code can make changes on the user's behalf without prior sign-off, and tool calling lets it launch programs, including scripts it wrote itself; the permission prompts guarding this are on by default but can be switched off by the user.

Limited Auditability

Every step Claude Code takes isn't necessarily logged & monitored, even though the underlying data sits on Anthropic's ISO 27001 and SOC 2 Type 2 certified cloud by default, or in the client's own Azure tenant where that's preferred.

Atlas Guidance: Allow w/ Guardrails

Allow Claude Code on company computers now, under three guardrails: lock the folder-permission prompts on so write access outside the designated folder can't be switched off by the user; vet and whitelist MCP connectors before granting access, since an unvetted one can act on what it reads without tracing back to a request; and turn on logging wherever the deployment allows, so actions can be reconstructed after the fact.