IRIS 365 LTD

Regulated Entity Advisory Note

Anthropic Agentic AI Tools

Claude CoWork • Claude in Chrome • Claude in Excel

Security, Compliance & Risk Advisory | March 2026

Prepared by: Iris 365 Ltd — Information Security Advisory

Document Ref: IRIS365-AI-ADV-002

Version: 2.0

Date: March 2026

Audience: Regulated Entity Clients — Senior Management, Risk & Compliance

Executive Summary

Anthropic offers a suite of agentic AI tools that extend beyond conversational chat: Claude CoWork (desktop automation agent), Claude in Chrome (browser agent), and Claude in Excel (spreadsheet agent). Unlike conventional AI chat assistants, these tools can act autonomously — reading files, browsing the web, processing spreadsheet data, and executing multi-step workflows on the user's behalf.

For regulated entities — including CIMA-regulated financial services firms, professional fiduciaries, investment managers, and legal practices — these agentic capabilities introduce a materially different risk profile from conversational AI. This advisory sets out those risks clearly, including several that are frequently underestimated in vendor and market commentary.

Iris 365 Recommended Position

None of the three tools covered by this advisory are suitable for use involving regulated data, client-confidential files, or any workflow requiring an audit trail — in their current form.

Critically: these tools have no native awareness of data classification. Any restriction to 'non-regulated data' is a paper policy enforced entirely by user behaviour, not a technical control. The documents, files, and folders a user has access to are accessible to the agent once a folder grant is made.

Conditional, limited use on genuinely isolated internal tasks is possible, but only under the strict technical preconditions set out in Section 6 — not merely procedural ones.

1. Understanding the Anthropic Tool Landscape

Not all Claude products carry the same risk profile. The critical distinction for regulated entities is between conversational tools — which generate text for a human to review and act upon — and agentic tools, which take actions directly on the user's behalf. This section sets out where each Anthropic product sits on that spectrum and why it matters.

The Conversational vs Agentic Distinction

A conversational AI tool responds to a prompt with text. The human reads the output and decides whether and how to act. The AI itself takes no action in the world.

An agentic AI tool receives a goal and acts to achieve it — reading files, executing code, submitting forms, browsing websites, or modifying data — with minimal or no human review of each individual step. The AI is the actor, not the advisor.

This distinction is not cosmetic. Regulators assess what a system can do, not how it is marketed. An agentic tool that can modify files and execute scripts must be assessed as an operational risk system, subject to change management, access controls, and audit trail requirements — regardless of whether it is described as an 'AI assistant'.

	Claude Chat (claude.ai)	Claude Code (CLI / IDE)	Claude CoWork (Desktop agent)	Claude in Chrome (Browser agent)	Claude in Excel (Spreadsheet agent)
Type	Conversational	Developer / agentic	Agentic	Agentic	Agentic
What it does	Generates text responses for human review	Writes, edits & executes code in dev environment	Automates desktop tasks — files, scripts, web	Operates inside live browser sessions	Reads & manipulates Excel workbooks
Takes real actions?	No	Yes — in codebase/terminal	Yes — files, scripts, browser	Yes — web pages, forms, tabs	Yes — workbook data & formulas
Primary access	None — text only	Local codebase, terminal, MCP servers	Local file system (folder grant)	All open tabs and web content	Active workbook content and cell data
Injection vector	Low — text output only	Code repos, terminal input, MCP tools	Docs, PDFs, emails, web pages	Any web page, ad, or injected script	Cell values, comments, imported data
Injection severity	LOW	MEDIUM–HIGH	HIGH	VERY HIGH	MEDIUM–HIGH
OneDrive/SPO risk	None	Low — scoped to dev repo	HIGH — synced folders look identical to local	MEDIUM — via authenticated browser session	MEDIUM — if workbook is SPO-sourced
Audit logging	Centralised (Teams/Enterprise)	Local — terminal history only	Local only	Local only	Local only
Admin controls	Full (Teams/Enterprise console)	Limited — no enterprise console	None (preview)	None (preview)	None (preview)
Regulated use?	Yes — with DPA in place	Internal dev only — not client data	Not without strict isolation controls	Not recommended	Not without isolation controls

The three agentic tools in the shaded columns — CoWork, Claude in Chrome, and Claude in Excel — are the subject of this advisory. Claude Chat and Claude Code are included for context only; their risk profiles are addressed separately where relevant.

Claude Code warrants a brief note: while it is a developer tool rather than an end-user agent, it is agentic in nature, can execute code and terminal commands, and integrates with MCP servers. It is not suitable for use involving client data or regulated repositories and should be subject to its own risk assessment in developer environments.

Scope Note: These Risks Apply to Agentic AI Generally. The risks identified in this advisory are not unique to Anthropic's tools. Prompt injection, the absence of data classification awareness, audit trail gaps, and the OneDrive sync cascade risk are properties of agentic AI operating in a Windows and Microsoft 365 environment — not of any particular vendor's implementation. The same framework and risk assessment obligations set out in this advisory should be applied to any agentic AI tool before deployment, including Microsoft Copilot Actions, OpenAI Operator, Google Gemini for Workspace agents, and any other tool capable of autonomous file access, web interaction, or multi-step task execution. This advisory focuses on Anthropic's CoWork, Claude in Chrome, and Claude in Excel because they are the tools under evaluation; Iris 365 will issue separate guidance on other agentic tools as they reach enterprise availability.

1.3 Product Maturity — Research Preview and Beta Status

None of the three tools covered by this advisory carries a Generally Available (GA) designation from Anthropic. Anthropic's own product page for Cowork describes it explicitly as *"a research preview"*. Claude in Chrome and Claude for Excel are labelled *"Beta"*. The status of Claude in Excel should be confirmed at point of publication via the Microsoft AppSource listing.

For regulated entities, this maturity status is a governance signal independent of the specific technical risks identified in subsequent sections. A vendor-designated research preview or beta product carries several implications that apply regardless of underlying risk controls:

- No SLA or uptime commitment — the product can be withdrawn, changed, or broken without notice, with no contractual remedy for the deploying organisation
- No enterprise support tier — incident response for a regulated workflow failure may be limited to community documentation or general product support, without a defined response commitment
- Security model, permission structure, and data handling may change materially between versions — controls documented today may not reflect the product's behaviour after an unannounced update
- Vendor's own labelling implicitly disavows production readiness — deploying a vendor-previewed tool in a regulated workflow means the organisation assumes risk that the vendor has explicitly declined to underwrite
- The audit trail and administrative control gaps identified in this advisory are in part a consequence of preview-stage architecture — they are not necessarily permanent product design decisions, but they are the current reality

This advisory will be updated as Anthropic moves these products toward general availability. The minimum threshold for reassessment of regulated deployment eligibility is delivery of centralised audit logging and role-based administrative controls — which are also the conditions under which the preview and beta designations would most likely be retired.

2. How Folder Access Works — And Why It Is Not a Reliable Control

A common misconception is that CoWork silently inherits access to everything on a device. The reality is more nuanced — but not meaningfully safer in practice.

2.1 The Folder Grant Model

CoWork uses an explicit, session-time folder grant. When it needs file access, it presents a standard OS folder picker dialog, and the user selects a folder to permit. That grant persists for the duration of the session.

In theory this is a speed bump. In practice it provides limited protection for four reasons:

> **Why the Folder Grant Is Weaker Than It Appears**
>
> 1. The grant is as broad as the folder selected. Selecting Documents or OneDrive gives CoWork access to the entire tree beneath it — every subfolder and file, with no secondary confirmation per file.

2. Users under time pressure select high-level folders to avoid repeated prompts. The path of least resistance is the broadest possible grant.

3. Once a grant exists, prompt injection attacks can exploit it without requesting new permissions. An injected instruction arriving via a web page or document can read or exfiltrate anything within the already-granted scope — no further user interaction required.

4. The grant is not logged. Neither CoWork nor the OS records which folders were granted, when, or what was accessed within them.

2.2 OneDrive-Synced SharePoint — The Silent High-Risk Surface

This is the most significant practical risk for organisations using Microsoft 365, and it is consistently underestimated.

The OneDrive sync client maps SharePoint document libraries to local folders. On a typical consultant or analyst machine, these appear as:

- C:\Users\[name]\OneDrive - [Organisation]\
- C:\Users\[name]\OneDrive - [Client Name]\ (if client SharePoint sites are synced)

These are indistinguishable from local folders in the OS folder picker. There is no visual indicator, no warning, and no technical barrier that would cause a user — or CoWork — to treat them differently.

The OneDrive Sync Cascade Risk

If a user grants CoWork access to a folder that contains or sits above a OneDrive-synced library:

- CoWork can read every file in every synced SharePoint library within scope
- Microsoft Purview sensitivity labels are NOT enforced at the file system level — a file labelled Confidential opens as a normal file on disk if the user has access
- Any modifications CoWork makes will sync back to SharePoint and propagate to all other synced users
- Those modifications will appear in SharePoint audit logs as the user's own action — not as an AI agent action
- For MSPs or consultants with multiple client SharePoint libraries synced, a single folder grant can expose content from multiple client environments simultaneously

2.3 CoWork Has No Data Classification Awareness

This point deserves explicit statement because it undermines a common framing — including language that has appeared in earlier versions of this advisory — that CoWork can be restricted to 'non-regulated' or 'non-confidential' data.

CoWork Cannot Distinguish Confidential from Non-Confidential Data

CoWork has no awareness of:

- Microsoft Purview sensitivity labels or retention policies
- SharePoint permissions or site classifications
- File metadata, data classification tags, or DLP policies
- Your organisation's internal data classification scheme

> Any restriction to 'non-regulated data' is a procedural control enforced entirely by user behaviour and judgment at the time of the folder grant. It is not a technical guardrail. It will not hold under audit scrutiny as a meaningful control, and it will not prevent accidental or injected access to regulated content.

3. Prompt Injection — The Primary Attack Vector Across All Three Tools

Prompt injection is the most significant technical risk shared by all three tools. It occurs when malicious or unintended instructions are embedded in content the agent reads — documents, web pages, spreadsheet cells — causing the agent to perform actions the user did not request and may not be aware of. Anthropic acknowledges non-zero successful injection rates even after mitigations are applied.

3.1 Claude CoWork — Document and Email Injection

CoWork processes documents, emails, and web content as part of its tasks. Any of these sources can carry embedded instructions:

- A PDF received from a counterparty or client contains hidden text instructing CoWork to copy files to an external location before completing the visible task
- A web page visited during research contains injected content that redirects the agent to submit data from the granted folder to a third-party endpoint
- An email body instructs CoWork to forward attachments or initiate an unintended workflow

Because CoWork can execute scripts and interact with the web, a successful injection is not limited to data exposure — it can trigger active operational harm.

3.2 Claude in Chrome — The Highest Injection Risk of the Three

Claude in Chrome operates directly inside the browser with the entire web as its environment. Every page the agent visits is untrusted content. This makes it the highest injection-risk tool:

- Any web page — including legitimate sites with compromised third-party advertising scripts or injected content — can carry instructions the agent acts upon
- Hidden text, zero-opacity elements, and content outside the visible viewport can carry instructions invisible to the user but readable by the agent
- Multi-tab browser sessions mean content in one tab can potentially influence agent behaviour across others
- Form interaction and submission capabilities mean a successful injection could submit data, trigger authenticated actions, or interact with SharePoint and other M365 services accessed via browser
- For users conducting due diligence, regulatory research, or accessing client portals via browser, the risk is not theoretical — the open web is an adversarial surface

3.3 Claude in Excel — Data-Driven Injection

Claude in Excel processes workbook content for analysis, formula generation, and data transformation. The injection surface is the spreadsheet data itself:

- A cell value containing crafted text — imported from an external data feed, a client submission, or a shared workbook — can embed instructions the agent acts upon when reading that cell
- Named ranges, cell comments, or formula strings in externally-sourced workbooks can carry malicious payloads
- For regulated entities using Excel for financial models, NAV calculations, compliance checklists, or risk registers, this is a direct risk — particularly where workbooks are received from counterparties or populated via external data connections

The bounded nature of the workbook reduces severity compared to Claude in Chrome, but the consequence in a regulated financial context — manipulation of financial records or compliance data — can be severe.

4. Shared Compliance Concerns

4.1 Audit Trail Gap — All Three Tools

All three tools store activity logs locally only. They are explicitly excluded from Anthropic's centralised Audit Logs, the Compliance API, and data export features. This means:

- No centralised, tamper-evident record of what the agent accessed, read, modified, or submitted
- No SIEM-ready event trail for security operations
- No forensic capability following an incident — you cannot reconstruct what the agent did
- Agent-driven modifications to SharePoint-synced files appear in M365 audit logs as the user's own actions, making AI activity indistinguishable from human activity after the fact

For any workflow where an audit trail is a regulatory, contractual, or internal governance requirement, this gap is disqualifying — regardless of other controls applied.

4.2 No Administrative Controls — All Three Tools

All three tools are currently in research preview with no enterprise administration capability:

- No role-based or department-level scoping — enablement is all-or-nothing per device
- No centralised admin console, MDM policy, or Intune compliance rule specific to these tools
- No admin-defined folder restrictions, data boundary enforcement, or per-user permissions
- CoWork, Claude in Chrome extension, and Claude in Excel add-in can all be installed by any user with local admin rights, with no central visibility

Organisations wishing to prevent installation must implement this proactively via Windows Defender Application Control (WDAC), Intune application blocklists, or browser extension management policies. The default position without these controls is: any user can install and run these tools.

4.3 NDA and Professional Secrecy

Most NDA and client confidentiality agreements predate agentic AI. Allowing an autonomous agent to access and process client-confidential files may constitute disclosure to a third-party data processor under those agreements — regardless of what Anthropic's DPA states about training data.

For Cayman Islands regulated entities, this intersects with professional secrecy obligations under fiduciary, investment management, and trust law. This risk is legal and contractual, not merely technical, and has not yet been tested in court or addressed by CIMA guidance. The legal uncertainty itself is a risk where regulated entities operate.

4.4 MCP and Plugin Supply Chain (CoWork)

CoWork supports MCP (Model Context Protocol) server integrations that extend its capabilities and can operate with elevated local privileges. Known CVEs already exist in this ecosystem. Only internally-developed or formally approved integrations should be permitted — no marketplace or third-party plugins.

4.5 Cross-Tenant Contamination (MSP and Multi-Client Environments)

For MSPs and consultants managing multiple client environments via GDAP or delegated access, the combination of multiple client SharePoint libraries synced locally and broad CoWork folder grants creates a cross-client exposure risk. A single grant that encompasses the OneDrive root can simultaneously expose content from multiple client tenants. Agent-driven modifications or exfiltration in this scenario would be exceptionally difficult to detect or attribute after the fact.

5. Arguments For and Against Adoption

5.1 Arguments in Favour

Explicit Folder Grant Model

CoWork does not silently inherit all device access. The folder picker model, while imperfect, does require a user action before any file access occurs. For a technically aware user on a properly isolated workstation, this provides a meaningful starting point for control.

No Model Training on Enterprise Data

Under Claude for Teams and Enterprise, Anthropic does not use customer data to train models. A formal DPA is available. This is materially better than many consumer AI tools.

Strong Platform Compliance Posture

Anthropic holds SOC 2 Type II, ISO 27001, ISO 42001, and HIPAA-ready certifications. These apply to the Claude API and Enterprise infrastructure — not to the local execution environment of these agentic tools — but they signal organisational security maturity.

Sandboxing Reduces Blast Radius

CoWork's folder-scoped permission model and Excel's workbook boundary both limit the theoretical maximum exposure compared to a fully unrestricted agent. These are not strong controls, but they are not nothing.

Productivity Gains Are Real

For appropriate use cases on genuinely isolated, non-regulated content, these tools offer meaningful time savings. The goal of this advisory is not to prohibit use but to define conditions under which use is defensible.

5.2 Arguments Against

No Technical Enforcement of Data Boundaries

The most fundamental problem: there is no mechanism to enforce a data classification boundary. Telling users to 'only use CoWork on non-regulated data' is a paper policy. It will not be consistently applied, it cannot be audited, and it does not protect against prompt injection that exploits an existing session grant.

OneDrive Sync Creates Uncontrolled Exposure

In any organisation using Microsoft 365 with SharePoint sync enabled — which is the default for most M365 deployments — the risk of inadvertent access to regulated or client-confidential content via OneDrive-synced folders is high and largely invisible to the user.

Audit Gap Is Disqualifying for Regulated Workflows

The absence of centralised, tamper-evident logging disqualifies all three tools from any workflow subject to regulatory audit requirements. This is not a gap that compensating controls can close — it is a fundamental architectural limitation at this stage.

Legal and Regulatory Uncertainty

NDA exposure, professional secrecy obligations, and the absence of CIMA or equivalent regulatory guidance on agentic AI mean that regulated entities bear the full burden of residual legal risk. That risk is material and currently unquantifiable.

Preview-Stage Governance Is Not Enterprise-Ready

All-or-nothing enablement, no admin console, no role-based scoping, and no MDM integration mean these tools cannot be deployed in a manner consistent with least-privilege principles or standard enterprise change management. The appropriate time for regulated deployment is after these capabilities exist — not before.

6. Iris 365 Recommended Position and Controls

Default Position — All Three Tools

Without exception, none of these tools should be used in any workflow involving client data, regulated records, SharePoint-synced content, or any process requiring an audit trail.

This is the correct default position until Anthropic delivers centralised audit logging and role-based administrative controls.

6.1 Conditions for Permitted Internal Use (CoWork and Excel Only)

Limited internal use may be defensible — but only when ALL of the following technical preconditions are met. These are not procedural guidelines; they are minimum technical requirements:

- CoWork or Claude in Excel is installed only on a designated workstation that is used exclusively for non-client, non-regulated internal work
- OneDrive sync is DISABLED on that workstation — no SharePoint libraries, no personal OneDrive, no client tenants are mapped as local folders
- The workstation has no active Microsoft Partner Centre credentials, no GDAP relationships, and no cached client tenant tokens
- The only content present on the workstation has been pre-reviewed and confirmed as non-regulated, non-confidential internal data
- No externally-sourced files (client submissions, counterparty documents, data feeds) are present in any folder accessible to the tool
- For CoWork: only Iris 365-developed and formally approved MCP integrations are connected — no third-party or marketplace plugins
- Users have completed AI risk awareness training that explicitly covers prompt injection, folder grant scope, and the OneDrive sync risk
- Use cases are documented in the AI Use Register before deployment

6.2 Claude in Chrome — Higher Bar

Claude in Chrome carries a higher residual risk than the other two tools due to its unrestricted web surface and live session access. In addition to the above conditions where applicable:

- Must not be used in any browser session that is authenticated to client systems, SharePoint, Partner Centre, or any regulated web application
- Should be considered unsuitable for production use in any regulated entity environment until centralised audit logging is available
- If installed, browser extension management policies should restrict it to named, approved devices only

6.3 Blocking Installation — Recommended Default

For organisations that have not conducted a formal risk acceptance, Iris 365 recommends proactively blocking installation of all three tools using the following controls:

- Windows Defender Application Control (WDAC) or Intune application blocklist to prevent CoWork installation
- Browser extension management policy (via Intune or Group Policy) to block the Claude in Chrome extension
- Microsoft 365 add-in management policy to restrict Claude in Excel from being added without IT approval

These controls are not applied by default. Without them, any user with local admin rights can install and run these tools on any managed device with no central visibility.

7. Recommended Actions for Regulated Entities

If your organisation is considering deployment, or has already deployed any of these tools, Iris 365 recommends the following steps:

- Issue an immediate interim guidance note confirming that CoWork, Claude in Chrome, and Claude in Excel are not approved for client data, SharePoint-synced content, or regulated workflows pending formal risk acceptance
- Audit current installations — check whether any of these tools are already installed on managed devices, including via self-service by users with local admin rights
- Implement proactive blocking via WDAC, Intune application policy, and M365 add-in management for any device where these tools are not explicitly approved
- Disable or audit OneDrive sync scope on any workstation where these tools are approved for limited use — confirm no client or regulated SharePoint libraries are mapped
- Review existing NDA and confidentiality agreements for AI processing carve-outs and obtain legal counsel on whether agentic AI processing constitutes disclosure
- Add all three tools to your AI tools register and your DPA sub-processor schedule
- Do not rely on procedural controls alone — if you cannot enforce the technical preconditions in Section 6, do not permit use
- Monitor Anthropic's product roadmap: centralised audit logging and role-based admin controls are the minimum threshold for regulated deployment consideration

8. Regulator-Ready Summary Statement

The following statement may be used in regulatory submissions, audit responses, or board risk papers:

Claude CoWork, Claude in Chrome, and Claude in Excel are agentic AI tools that introduce materially different risks from conversational AI. All three share prompt injection as a primary attack vector, local-only audit logging, and immature administrative controls at their current preview stage.

A critical characteristic of all three tools is that they have no native data classification awareness. Any restriction to non-regulated or non-confidential data is a procedural control dependent on user behaviour — not a technical guardrail. In environments where Microsoft SharePoint libraries are synced locally via OneDrive, a single broad folder grant can expose regulated and client-confidential content without any additional user action.

Claude in Chrome carries the highest injection risk due to its unrestricted web surface. Claude in Excel introduces data-driven injection risk directly relevant to regulated financial and compliance workflows. Claude CoWork combines file system, script execution, and browser risks in a single agent.

None of these tools are suitable for regulated data, client-confidential workflows, or audit-critical processes in their current form. Conditional use on internal non-regulated tasks requires strict technical preconditions — including disabled OneDrive sync, no client tenant access, and workstation isolation — not merely procedural guidelines. This position will be reviewed as Anthropic delivers centralised audit logging and role-based administrative controls.

About Iris 365 Ltd

Iris 365 Ltd is a Microsoft 365-focused managed service provider headquartered in the Cayman Islands, specialising in security, compliance, and operational tooling for regulated entities across financial services, legal, and professional services sectors.

For further guidance on AI risk assessment, DPA review, MDM policy implementation, or security policy development, please contact your Iris 365 account team.

Iris 365 Ltd | George Town, Grand Cayman, Cayman Islands | www.iris365.cloud

This document is provided for informational and advisory purposes only. It does not constitute legal advice. Regulated entities should seek independent legal counsel on NDA, confidentiality, and regulatory compliance matters.