AI Security and Data Privacy

Three Areas to Understand Before You Connect, Share, or Automate

Don't Forget the Human Part!™

www.learnair.com

[Apps] Connecting ChatGPT to Third-Party Tools

Explanation

When you connect ChatGPT to a third-party tool, such as a file drive, email system, calendar, project management tool, or workspace app, ChatGPT uses information from that tool as context for its responses.

Risks

1. **Overexposure of Data**: When you connect Google Drive, SharePoint, Gmail, Slack, GitHub, or similar systems, ChatGPT may be able to access whatever your account can access. If permissions are too broad, sensitive information becomes easier to discover and surface.
2. **Prompt Injection Attacks**: A connected app may contain malicious instructions hidden inside emails, documents, websites, or files. These instructions can attempt to manipulate the AI into revealing information or taking unintended actions.
3. **App Actions**: Some connected tools allow ChatGPT to take actions on your behalf, such as sending emails. If instructions are misunderstood, manipulated, or approved without careful review, unintended actions may be performed using your account.

SECURITY BEST PRACTICES

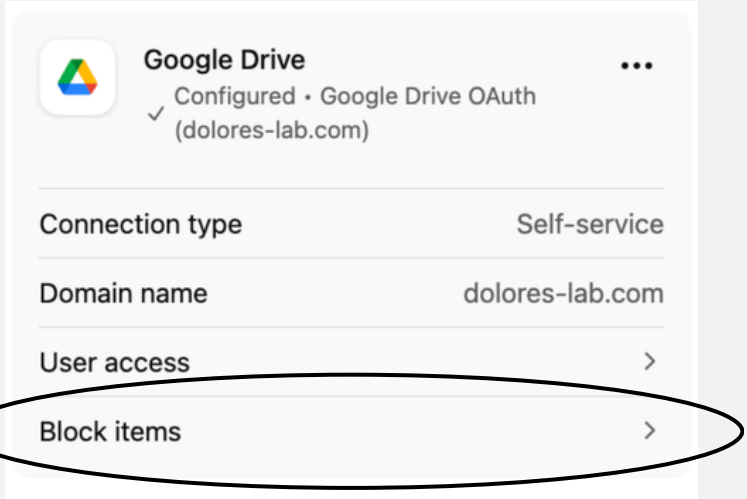
☐ **Follow the Principle of Least Privilege**

Grant only the minimum access required.
Instead of:

- Entire Google Drive

Prefer:

- Specific folders
- Specific projects
- Read-only permissions



☐ **Practice Human-in-the-Loop Verification**

Before acting:

- Review AI recommendations critically
- Question unusual or urgent requests
- Require human approval for sensitive actions

☐ **Define Success Criteria for Automated Actions**

Establish clear requirements before allowing AI-connected tools to take action.

Before enabling an app automation:

- Define what a successful outcome looks like
- Set boundaries for what the AI can and cannot do

Require human approval for sensitive actions

[Data Privacy] Not All AI Accounts Are Equal

Explanation

The data protection policy varies by company and plan. For example, ChatGPT Business, Enterprise, and Edu customer data is not used to train models by default. However, data from Free, Plus, Go, and Pro users, may be used to train models if the "Improve the model for everyone" setting is on.

Risks

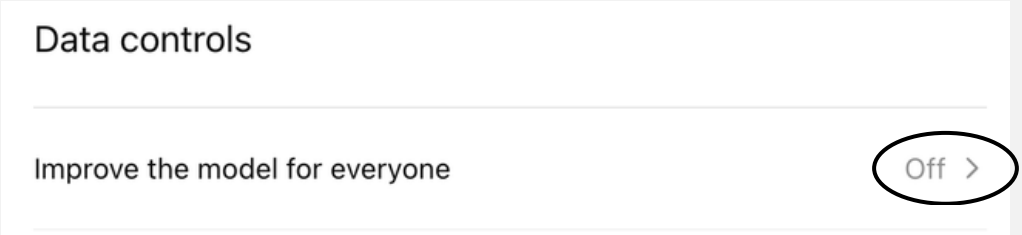
The risk exists when users assume their data is not being used to train the model, when it actually is.

For example, an employee might paste a client contract, internal strategy document, HR issue, financial report, or piece of proprietary code into their AI account to "just get help." The problem is that if the data is not protected from being used to train the model, sensitive work information could leave the organization's approved environment. This may put confidential strategy, client details, proprietary code, or internal processes at risk of being exposed outside the company.

SECURITY BEST PRACTICES

- [ ] **Disable Model Training for Sensitive Work**

  In "Settings" → "Data Controls," ensure "Improve the model for everyone is set to Off



[Agents] Within ChatGPT

Explanation

A ChatGPT agent can help with repeatable tasks and workflows, and can pull data from connected apps, browse, or perform actions on your behalf.

Risks

The main risk with agents is action. A normal chat acts as a collaborative thought partner. An agent can take autonomous actions on your behalf.

Depending on its permissions, an agent may be able to access company knowledge, send messages, update records, trigger workflows, or use connected applications.

This can create significant human time savings, but it also raises the stakes. If an agent has excessive access, unclear instructions, weak approval controls, or permission to act automatically, mistakes can spread farther and faster.

SECURITY BEST PRACTICES

☐ **Transparency Mandate**

Require agents to explain their work before actions are approved.
Include instructions that require the agent to:

- Explain how conclusions were reached
- Surface assumptions, missing information, and risks
- Cite sources when possible
- Identify actions it plans to take before execution

☐ **Rubric Scoring**

Define success criteria and require the agent to evaluate its output before acting.

Require the agent to:

- Score its output against predefined criteria
- Explain weaknesses or areas of uncertainty
- Flag outputs that fail to meet minimum standards
- Escalate sensitive decisions for human review

☐ **Human in the Loop**

Set "Write action safety" to "Always ask," so you stay in control of anything being completed on your behalf.

Disable any "write" actions that you do not want your agent to be able to do.