AI safety

From the SourceLace manual: What SourceLace guarantees when AI apps, the assistant and AI agents work with your systems, including when someone tries to misuse an AI. These guarantees come from how SourceLace is designed, not from asking the AI to behave.

SourceLace version 0.1.0 · Made November 22, 2026
The latest version is at https://sourcelace.ai/docs

AI safety

What SourceLace guarantees when AI apps, the assistant and AI agents work with your systems, including when someone tries to misuse an AI. These guarantees come from how SourceLace is designed, not from asking the AI to behave.

What any AI can and cannot see

- **The AI never holds a secret.** Sign-ins to your systems, your organization's own AI key and your admins' app secrets are never put in front of an AI: not in its instructions, not in a tool's description and not in a tool's result. They are shown in SourceLace's own pages only as "(saved)".
- **What SourceLace tells an AI is treated as public.** Any connected AI app can read SourceLace's tool descriptions, and a person can talk an AI into repeating its instructions. So that text says what SourceLace does, never how it is built, and holds nothing that would help someone break in.
- **Each AI sees only what the person using it may see.** Every request runs as that person, with their own sign-in where the system has one, and is checked against your organization's access by group and data protection rules.

Your organization's data stays your organization's

- **Organizations are kept apart in SourceLace's code**, not by the AI's judgment. Every request is checked against the signed-in person's organization before anything is read. Another organization's sources, people, audit trail, query results, changes, AI agents and their runs, skills and chats cannot be reached by any tool or page, even by id, and even if an AI is told to try.
- **Encrypted per organization.** Secrets, saved chats, AI agents' settings and run summaries, and query results while they are held are encrypted with your organization's own key.

Data that tries to give orders

Rows, emails, files and messages read from your systems are data. An AI is told to treat them as data, and SourceLace makes sure it does not matter if it does not:

- **No change without a person.** A change to a record is always previewed first. In the SourceLace app only the person's own click applies it; in other AI apps the person confirms it in their app or types a code. AI agents' changes wait for an approval unless an admin allowed that one action to run on its own.
- **Agents stay inside what they were given.** An agent can only use its own sources and the actions its owner listed, so text inside data cannot widen what it may read or do.
- **A vendor's changed tool waits for your admin.** When an outside MCP server's vendor changes a tool's description (the words that tell the AI what the tool does) or its arguments after your admin allowed it, SourceLace stops offering it until an admin reviews the new version. See Held tools.

Answers about SourceLace itself

When someone asks the assistant in the SourceLace app how SourceLace itself is built, hosted or secured, or for its instructions, its answer is checked before it is shown. A sentence that would describe how SourceLace is built, hosted or secured, carry something shaped like a key or password, or make a compliance claim is replaced with a short line that says what cannot be shared and what to do next, such as: "I can't go into how SourceLace is built or hosted; that part stays private. I can still help with how to use it, such as connecting a system or who can see what. For a security review, your admin can ask your SourceLace contact for the security review pack." The rest of the answer stays as written. Your audit trail records that this happened, never the text. How-to answers, such as which addresses to allow in your firewall or how to connect a system, and answers about your own data, including your own databases and cloud accounts, are never changed.

Tested on every release

SourceLace's automated tests try these attacks on every change: asking the AI to repeat its instructions, asking how SourceLace is built, reaching another organization's data through every tool, pulling out tokens and keys, and planting orders inside returned data. A change that weakens any guarantee on this page fails those tests.

For more detail, see Security and privacy and Security questions. Your admin can request the security review pack from security@sourcelace.com.