AI Governance Strategy

AI Governance Strategy

(Public Commitment – September 2025)

Our Belief

AI should augment human judgement, not supersede it.

Our AI solutions exist to help teams think better, decide better, and act better – while staying in control.

We see privacy, transparency and human oversight as core design principles, not just compliance requirements. They guide how we build, how we work, and how we earn trust.

Our Core Principles

1. Privacy by Design

a. User data is a guest, not an asset – we protect it, respect it, and never keep it longer than the user chooses.

b. Encrypted and under user control – all prompts, outputs and logs are encrypted at rest; users can access, revoke or delete their data at any time. If a user requests deletion, any Guardian metadata is pseudonymized or unlinked from the user within 7 days, except where retention is required for security, fraud prevention, or legal obligations (e.g. detection of terrorism related content, human trafficking, CSAM). Such retained metadata contains no user data, is strictly access controlled, and is deleted or fully anonymized after 90 days.

c. Minimal and intentional collection – we gather only what is needed to deliver value, never for hidden or obfuscated purposes.

d. User-owned traceability – we store raw prompts, outputs and agents run logs so users can audit, replay or delete them – however these remain private to them unless they choose to share access with us.

2. Transparency and honesty

a. AI is part of Stallo core – we don't hide it, but we also don't label every sentence or document. Users know when they are interacting with an AI-powered system.

b. Clear about capabilities and limits – we explain what our product can and cannot do, including risks like bias or hallucination.

c. Plain-language model & data overview – we publish which model providers we use (e.g. Anthropic, OpenAI) and explain how we handle prompts, outputs and storage.

d. Contextual transparency – we provide signals, help text and explanations when it matters (e.g. before sending something to a client, or when a workflow is autonomous).

3. Human-in-the-loop by Default

a. Stallo recommends, drafts and summarizes – but it doesn't take irreversible action silently.

b. Users have full visibility and can review, approve or reject actions before they go live.

4. Autonomy as a User Choice

a. Autonomy is opt-in, with levels users can configure:

- i. Assistive: No actions taken, only suggestions
- ii. Semi-Autonomous: Safe, reversible actions (e.g. draft + delay-send).
- iii. Autonomous: Fully proactive ambience, with audit logging and undo where possible.

b. Guardian-Enforced Safety: every output passes through Stallo's Guardian, which grades content against a documented risk taxonomy (e.g. privacy leakage, disallowed use case, bias severity):

- i. Low-risk: Proceeds as normal
- ii. Medium-risk: Shown with a warning, workflows pause for human-in-the-loop.
- iii. High-risk: Blocked by default, but can be expanded by the user.
- iv. Critical: Blocked and flagged; Cannot be expanded or viewed.