

The Jonesi Consultancy | Houston, TX

EXECUTIVE AI ADOPTION · AGENT HANDOUT

AI GUARDRAILS FOR AGENTS



Freedom to Act. Boundaries to Keep It Safe.

Guardrails are not friction. They are what makes it responsible to let an agent move fast.

1 Why Guardrails Matter Five Outcomes They Protect

An AI agent does not just answer. It takes action. Guardrails are the controls that let an agent move quickly inside a boundary you set deliberately, so speed never outruns accountability.

PROTECT PEOPLE

Prevent harm, bias, and unethical outcomes

STAY COMPLIANT

Meet legal, regulatory, and policy requirements

BUILD TRUST

Raise confidence in agent decisions

REDUCE RISK

Minimize errors, misuse, and surprises

ENABLE SCALE

Keep behavior consistent as agents multiply

2 The Six Control Areas Design Every Agent Against All Six

1 BOUNDARIES AND SCOPE

Define what the agent can and cannot do.

- › Restrict actions by role, data set, and business context.
- › Set topic and output limits in the system instructions.
- › Apply privacy and data handling rules at the tool level.

EXAMPLE: The agent can read customer records but cannot access payroll data or make hiring decisions.

2 APPROVAL REQUIREMENTS

Decide where a human has to sign off.

- › Require approval for high impact or irreversible actions.
- › Set dollar limits and risk thresholds in writing.
- › Escalate anything the agent flags as uncertain.

EXAMPLE: The agent drafts the contract, but legal review is required before it sends.

3 MONITORING AND LOGGING

You cannot govern what you cannot see.

- › Log every action, input, and decision the agent takes.
- › Track error rates, anomalies, and tool usage over time.
- › Set real time alerts on thresholds that matter.

EXAMPLE: Alert the owner if the agent hits repeated policy blocks or unusual call volume.

4 HUMAN OVERSIGHT

Accountability does not transfer to the agent.

- › Name an owner for every agent in production.
- › Publish clear escalation paths for exceptions.
- › Preserve the right to override or stop at any moment.

EXAMPLE: A manager reviews weekly agent summaries and can reverse any action taken.

5 TRANSPARENCY

Decisions people cannot explain do not get trusted.

- › Show the reasoning, the data used, and the confidence level.
- › Keep audit trails easy for a non technical reviewer to read.
- › Disclose to end users when an agent is involved.

EXAMPLE: The agent explains its recommendation with the key factors and sources behind it.

6 SECURITY AND ACCESS

Give the agent the least access that still works.

- › Enforce least privilege on every system it touches.
- › Protect keys, secrets, and credentials from the prompt layer.
- › Scope third party tool permissions and review them regularly.

EXAMPLE: The agent can read customer data but cannot delete or export it.

3 Guardrail Checklist

Before An Agent Goes Live

DEFINE

- ☐ Boundaries documented and approved
- ☐ Owner named and accountable
- ☐ Permissions configured and tested
- ☐ Approval routes established

OBSERVE

- ☐ Audit logging enabled
- ☐ Monitoring and alerts active
- ☐ Performance thresholds set
- ☐ Review cadence scheduled

RESPOND

- ☐ Human review process in place
- ☐ Override and stop procedure tested
- ☐ Incident response plan ready
- ☐ Update and retraining path defined

SAFE AGENTS. STRONGER OUTCOMES.

Guardrails are not friction. They are what makes it responsible to let an agent move fast. Empower your agents. Protect your people. Build with confidence.

© 2026 The Jonesi Consultancy. All rights reserved.

Freedom to Act. Boundaries to Keep It Safe.

We help organizations set the lines, train to them, and put governance behind them.

