Suhas Bhairav

Safe AI Agent Manual

A practical, beginner-friendly guide to safe AI agent use in business

A practical, ready-to-apply framework to define safe AI agent use in your company.

Suhas Bhairav

suhasbhairav.com

What is an AI Agent?

An AI agent is software that carries out tasks using AI-powered decision making. It operates under defined rules and requires human checks for sensitive actions.

Key move **Start with a single safe task and expand only after review**

PRACTICAL CHECKLIST

01	An AI agent is software that carries out tasks using AI-powered decision making.
02	It can operate across apps like email, calendars, chat, documents, and data systems.
03	It follows defined rules and requires human checks for sensitive actions.
04	It can automate routine, repetitive work to save time and reduce errors.
05	Data access and tool permissions must be explicitly granted and logged.
06	Outputs should be explainable and traceable for accountability.
07	It should be designed to fail safely and escalate when uncertain.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

What AI Agents Can Do

AI agents can handle structured, repeatable tasks by following rules. They excel at data gathering, scheduling, drafting, and monitoring processes.

Key move **Pilot one safe task first; measure results before scaling**

PRACTICAL CHECKLIST

01	Data lookup from approved sources with activity logging and source attribution.
02	Draft routine emails and messages using approved templates and tone guidelines.
03	Schedule meetings, send invites, and update calendars automatically.
04	Aggregate data into predefined reports and dashboards.
05	Provide simple decision support within defined rules and thresholds.
06	Monitor processes and raise alerts when deviations occur.
07	Run routine data quality checks and flag anomalies.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

What Needs Human Review?

Not every action should be automatic. Define where humans must approve or intervene before execution.

Key move **If unsure, pause and ask a person**

PRACTICAL CHECKLIST

01	Actions with financial impact above a threshold require human approval.
02	Actions affecting customer data must be reviewed by a person.
03	High-risk decisions require escalation to a domain expert.
04	Irreversible operations require human sign-off before execution.
05	Uncertain inputs or conflicting data should trigger human review.
06	Policy or rule changes must be validated by governance process.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Allowed Actions Template

Document the actions the agent may perform without human review. Use concrete examples and constraints.

Key move **Start with one simple allowed action and verify end-to-end**

PRACTICAL CHECKLIST

01	Action name, Purpose, Data inputs, Output type, and Preconditions must be documented.
02	Allowed data sources only; list sources and access levels.
03	Required logging: event, timestamp, actor, and outcome.
04	No irreversible actions without human sign-off.
05	Trigger rules: on-demand or automated schedule, with fail-safes.
06	Review requirement: specify if and when human review is mandatory.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Restricted Actions Template

Define actions the agent may not perform. These restrictions prevent harm and preserve control.

Key move **If in doubt, mark as restricted and review**

PRACTICAL CHECKLIST

01	Prohibited actions: actions that delete or alter critical data without approval.
02	Access to PII or sensitive data requires policy-compliant authorization.
03	Altering system configurations or security settings is forbidden without clearance.
04	Sending unverified external communications must not be automated.
05	Automated purchases or financial transactions require human consent.
06	Access to confidential plans or proprietary algorithms is prohibited.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Human Approval Steps Template

Outline how humans approve actions that require oversight. This keeps governance explicit.

Key move **Always document approvals and reasons**

PRACTICAL CHECKLIST

01	Trigger: Action requiring approval is flagged by risk, cost, or data sensitivity.
02	Approver role: Assign a named role such as manager or data steward for the decision.
03	Criteria: Approval depends on objective criteria (impact, risk, policy).
04	Method: Use a formal channel (ticketing or approval tool).
05	Timeline: Provide a max response time (e.g., 24 hours).
06	Audit: Log decision, approver, rationale, and timestamp.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Tool Access Rules Template

Define which tools the agent can access and how it uses them.

Key move **If a tool is deprecated, revoke access immediately**

PRACTICAL CHECKLIST

01	Tool access list with roles and constraints.
02	Access level per tool: read, write, admin with limits.
03	Use time-limited tokens and credential vaults.
04	Session isolation: separate sessions for each task.
05	Audit trail for tool actions and failures.
06	Immediate revocation: terminate access if abuse is detected.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Escalation Rules Template

Escalation rules specify when to involve humans to handle risk or failure.

Key move **Red flags: set an automatic escalation if delays exceed limit**

PRACTICAL CHECKLIST

01	Trigger: escalation occurs for high risk, data issues, or repeated failures.
02	Paths: route to supervisor, data steward, or risk/compliance.
03	Timing: set a max response time for escalations.
04	Communication: define who is notified and what data is shared.
05	Resolution tracking: capture status, actions taken, and outcomes.
06	Post-escalation: decision re-run, manual intervention, or halt.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.

Audit Log Requirements Template

Audit logs provide traceability and accountability for every agent action and decision.

Key move **If you can't verify logs, you can't trust automation**

PRACTICAL CHECKLIST

01	Log actions: task, data inputs, outputs, timestamp, user, tool used.
02	Record approvers and decision rationale for each approved action.
03	Store logs securely with access controls and immutability.
04	Retention policy: keep logs for compliance period.
05	Audit visibility: limit view to authorized roles.
06	Tamper-detection: sign and hash log entries.
07	Periodic reviews: monthly or quarterly audit of agent activity.

Suhas Bhairav AI Resource

A practical, ready-to-apply framework to define safe AI agent use in your company.