A PRACTICAL GUIDE FOR BUSINESS, TECHNOLOGY AND SECURITY

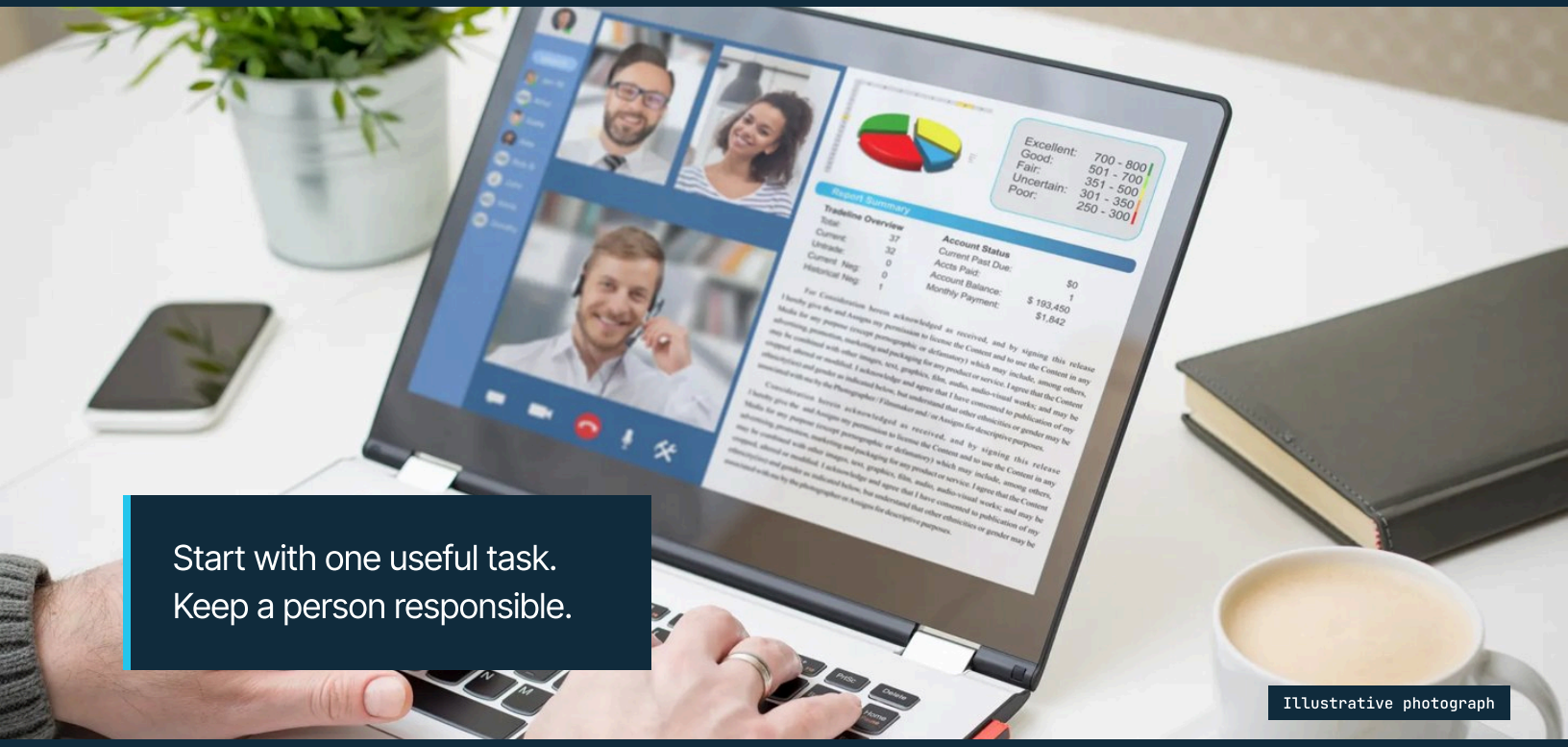
When AI Takes the Wheel

A practical guide to bounded agents, human approval and recovery.

01 / Define the boundary

02 / Support the reviewer

03 / Test and decide



Start with one useful task.
Keep a person responsible.

Illustrative photograph

01 / UNDERSTAND WHAT CHANGES WHEN AI CAN ACT

Know the reach.

A drafting assistant proposes words. A tool-using agent may also read records, run commands, send messages or change systems.

Identify the action boundary

Risk grows when uncertain model output becomes a real action. A wrong summary may mislead a reviewer. A wrong command can change a live system before anyone notices. The important questions are what the agent can reach, which actions it can take and who authorizes them.

List every connector and tool. For each, record its account, reachable systems, available operations and maximum impact. Include indirect paths such as shell access, browser sessions, shared drives and a service account that can create other credentials. A narrow-looking tool can still carry broad authority.

A PRACTICAL REVIEW MAP

Reach	Action	Authority
Which systems and records can it access?	Which operations can it request?	Which operations are actually permitted?

Illustrative workflow. Adapt the boundaries and responsibilities to the approved use.

01 / APPLY THE PRACTICE

Choose the first boundary.

Choose a modest first workflow

A useful starter agent could read a selected public documentation folder and prepare a change proposal. Keep its output in a review workspace. Give it no production credentials, no send capability and no ability to modify the source collection.

Write an execution contract: objective, allowed resources, prohibited actions, output, reviewer, spending or execution limit, deadline and stop condition. Make the contract visible to the human operator and enforce material boundaries outside the model.

Ask three questions before connecting a tool

- Could this tool expose information the user should not receive?
- Could it change, delete, publish, purchase or contact someone without the required approval?
- Can an operator stop execution, revoke access and inspect what actually happened?

PUT THE PRACTICE TO WORK / REVIEW NOTES

Owner	Evidence	Next Review
Who accepts this work and can pause it?	Which source or test supports the decision?	What must be resolved before proceeding?

02 / GIVE THE AGENT ONLY THE AUTHORITY IT NEEDS

Narrow the authority.

Use separate identities and narrow tool interfaces so a mistaken plan has a limited reach.

Start with read-only access

Use a dedicated service identity for the specific workflow. Limit it to the required folders, records and operations. Avoid inheriting an administrator's session or broad personal account. Keep credentials outside prompts and retrieved material, and make revocation practical.

For a document assistant, expose a search operation scoped to an approved collection. Do not provide a general shell simply to search files. Apply access checks at retrieval and again when returning results. Separate customers, departments and environments through enforceable authorization.

A PRACTICAL REVIEW MAP

Identity	Interface	Enforcement
A dedicated account for the specific workflow.	Allowlisted operations with validated inputs.	Deny actions outside the approved boundary.

Illustrative workflow. Adapt the boundaries and responsibilities to the approved use.

02 / APPLY THE PRACTICE

Make limits enforceable.

Design bounded action tools

- Allow specific operations with validated inputs instead of arbitrary commands.
- Restrict targets, paths, recipients and network destinations where the workflow permits.
- Set batch sizes, time limits and spending limits in the application or platform.
- Separate development, test and production identities and data.
- Require new approval for a broader scope, new connector or higher-impact action.

Check the tool result

A tool call is a request, not proof of success. Verify the returned status and the actual target state. Record failures accurately and stop after the permitted attempts. Do not let repeated retries create duplicate messages, purchases or records.

OWASP's guidance identifies excessive agency as a security risk. A useful practical response is to remove unnecessary tools and permissions before adding more instructions. Test that forbidden actions are denied even when the agent asks for them directly. Include that denial evidence in the pilot review.

PUT THE PRACTICE TO WORK / REVIEW NOTES

Owner	Evidence	Next Review
Who accepts this work and can pause it?	Which source or test supports the decision?	What must be resolved before proceeding?

03 / MAKE HUMAN APPROVAL CONCRETE

Show the exact action.

Approval works when the reviewer can see the exact action, understand its effect and reject it before execution.

Build an action preview

Show the target, proposed change, relevant source evidence and expected impact. For a message, show recipients and the complete text. For a file change, show the diff and affected paths. For an infrastructure action, show the environment and exact operations. Highlight uncertainty and missing evidence.

Bind approval to the reviewed action. If the content, destination, scope or target state changes materially, require review again. Avoid approval prompts that merely ask whether the agent may “continue.” The reviewer should know what will happen next and how to recover if it is wrong.

A PRACTICAL REVIEW MAP

Preview	Approve	Verify
Target, content, evidence and expected effect.	A reviewer accepts the exact proposed action.	Check what changed at the actual target.

Illustrative workflow. Adapt the boundaries and responsibilities to the approved use.

03 / APPLY THE PRACTICE

Keep approval meaningful.

Treat external content as untrusted

An email, web page, ticket or repository file can contain instructions aimed at the agent. OWASP describes prompt injection as a risk when such content redirects model behavior. Keep retrieved content separate from application instructions, restrict tool authority and require review for consequential actions.

Do not rely on a prompt that says “ignore malicious instructions” as the sole defense. Test a document that asks the agent to reveal secrets, send data externally or alter its task. The application should deny the action even if the model proposes it.

A practical review request

> “Prepare a proposal only. Identify the requested action, allowed target, supporting evidence, expected effect and rollback. List any uncertainty. Do not execute, send or publish. If the task needs a permission outside the approved scope, stop and describe the missing authorization.”

Use this template alongside technical restrictions. It supports review; it does not grant or enforce access.

PUT THE PRACTICE TO WORK / REVIEW NOTES

Owner

Who accepts this work and can pause it?

Evidence

Which source or test supports the decision?

Next Review

What must be resolved before proceeding?

04 / PLAN FOR FAILURE BEFORE THE FIRST WRITE

Prepare to stop.

A recoverable workflow needs protected evidence, a working stop control and a tested way to restore service.

Record what actually happened

Keep an audit trail of the requesting identity, task, proposed action, approval, tool invocation, result and independent verification. Record relevant configuration and versions when they affect reproduction. Minimize sensitive content in logs and restrict who can read them.

Separate planned, attempted and verified outcomes. “The agent reported success” is weaker evidence than a checked target state. Preserve failed results as well as successful ones so an investigator can reconstruct the sequence. Apply retention rules and avoid storing credentials in traces.

A PRACTICAL REVIEW MAP

Planned	Attempted	Verified
The action the operator intended.	The tool call and its returned status.	The independently checked result.

Illustrative workflow. Adapt the boundaries and responsibilities to the approved use.

04 / APPLY THE PRACTICE

Practice the recovery.

Practice stopping and restoring

- Interrupt a running job and verify no new actions begin.
- Revoke the service identity or connector and confirm access is denied.
- Restore a changed test record or file from a known recovery point.
- Check that a retry does not duplicate an earlier action.
- Confirm an operator can continue the underlying service without the agent.

Know where rollback ends

A sent message, disclosed secret or external purchase may not be reversible. For these actions, prevention and explicit authorization matter more than a rollback promise. Use drafts, test accounts and isolated environments until the boundary is proven.

If the agent exceeds scope, stop its execution and contain access. Preserve necessary evidence, assess affected systems and follow the incident process. Restart only after the owner reviews the cause and verifies corrective controls. A calmer prompt alone does not establish that the underlying failure is fixed.

PUT THE PRACTICE TO WORK / REVIEW NOTES

Owner	Evidence	Next Review
Who accepts this work and can pause it?	Which source or test supports the decision?	What must be resolved before proceeding?

05 / EARN WIDER AUTONOMY THROUGH EVIDENCE

Earn the next step.

Expand only when the workflow is useful, observable and bounded under realistic failure conditions.

Evaluate a complete workflow

Create safe tasks with expected results and unacceptable outcomes. Include missing information, conflicting sources, denied access, malicious documents, tool errors and interrupted execution. Test the actual integration in its intended environment. A model benchmark does not establish that the surrounding system is safe.

Compare total human effort, final quality, failure rates and actual operating cost against the current process. Count review, recovery and support work. Report the sample and its limits. Do not turn a successful demonstration into a claim that unsupervised operation is ready.

A PRACTICAL REVIEW MAP

Quality

Inspect the final result and the failure cases.

Effort

Count checking, correction and support.

Cost

Record actual operating cost and its limits.

Illustrative workflow. Adapt the boundaries and responsibilities to the approved use.

05 / APPLY THE PRACTICE

Keep the decision visible.

Use a staged decision

Stage 1	Read approved material and produce proposals.
Stage 2	Prepare changes in an isolated test environment.
Stage 3	Execute a narrow action only after specific human approval.
Further authority	Require a separate risk decision, stronger evidence and appropriate controls.

Keep an owner after launch

Name the person accountable for permissions, evaluations, incidents and continued usefulness. Review the workflow after changes to tools, source data, configuration or operating context. Retire unused access and pause when evidence no longer supports the current scope.

Use the NIST AI Risk Management Framework to organize the broader risk discussion, while keeping the acceptance criteria specific to the service. At the next review, bring one representative task, its approval trail, a failed test and its recovery evidence. Those artifacts make an informed decision possible without relying on the agent's confidence.

PRIMARY REFERENCES

Primary sources. Practical context.

Verify current guidance and the requirements applicable to your own organization.

01 NIST AI Risk Management Framework

Voluntary framework for governing, mapping, measuring and managing AI risks. A framework is not a certification.

https://www.nist.gov/itl/ai-risk-management-framework

02 NIST AI RMF Playbook

Suggested actions to adapt to the organization and its use case.

https://airc.nist.gov/airmf-resources/playbook/

03 OWASP Top 10 for LLM Applications

Security risks including prompt injection and excessive agency.

https://owasp.org/www-project-top-10-for-large-language-model-applications/

04 NIST Generative AI Profile

Companion risk-management guidance for generative AI; use with organization-specific evaluation and controls.

https://doi.org/10.6028/NIST.AI.600-1

> A planning aid for bounded AI workflows. Qualified organizational reviewers determine applicable obligations, suitability and deployment approval. Frameworks and checklists do not certify a tool or service.

YOUR NEXT STEP

Bring one workflow. Leave with a clearer starting point.

A focused AI planning conversation starts with the task, the data and the people responsible for the result.

01 / THE WORKFLOW

The repeated task, its current burden and the people who perform it.

02 / A SAFE EXAMPLE

A public or synthetic example that illustrates the task without protected information.

03 / THE DATA PATH

Where inputs, drafts, logs and connected services would go.

04 / THE DECISION

Your pilot criteria, review owner and the questions that remain open.

Plan Your AI Pilot ↗

844-974-8669 · z7solutions.com

Keep protected information out of booking notes and initial examples.
Z7 Solutions LLC