

WHITEPAPER / PRACTICAL RESOURCE

# Practical AI Governance & Human Review

An operating system for marketing decisions



## HOW TO USE THIS RESOURCE

# Practical AI Governance & Human Review

Marketing leaders, operations, security partners and review owners

## Your practical outputs

- Risk-tier and permission matrix
- Human-review rubric and escalation path
- Incident and exception records
- Governance handover and review cadence

## Read, choose, apply

Start with the overview and decision rules. Choose a concrete constraint in your business. Work through the relevant templates using your own evidence, record open assumptions and agree a review with the accountable people. The complete online edition provides linked sources and contextual navigation.

## Editorial context

Explicitly fictional cases and example figures are working aids, not client results or universal benchmarks. Frameworks structure decisions; they are not scientifically validated scores. Platform capabilities may change. Sources and review context are disclosed.

Read online: [Practical AI Governance & Human Review](#)

## CONTENTS

|                                                     |    |
|-----------------------------------------------------|----|
| 1. Govern the decision, not the buzzword            | 4  |
| 2. Classify exposure using separate dimensions      | 4  |
| 3. Design human review as skilled work              | 5  |
| 4. Define permissions as enforceable contracts      | 6  |
| 5. Establish a review-capacity budget               | 6  |
| 6. Calibrate reviewers before relying on agreement  | 7  |
| 7. Create an evidence standard for marketing claims | 8  |
| 8. Make exceptions temporary and inspectable        | 8  |
| 9. Secure the source-to-action chain                | 9  |
| 10. Turn monitoring into a response contract        | 10 |
| 11. Test governance under pressure                  | 10 |
| 12. Maintain governance without creating paralysis  | 11 |
| 13. Define stop, containment and restart separately | 12 |
| 14. Delegate governance with an evidence handover   | 12 |
| Turn this resource into a team brief                | 15 |
| Sources & further documentation                     | 16 |

# 1. Govern the decision, not the buzzword

Governance is the arrangement that makes consequential decisions visible, authorized, and revisitable. It is not a folder of policy statements. For marketing, the important questions concern what information enters a system, what claims leave it, what actions it can take, and who accepts the consequences. Begin with those questions rather than asking whether a vendor calls its product an agent.

The original framework used here is REVIEW: Recognize exposure, Establish authority, Verify evidence, Intervene when uncertain, Escalate exceptions, Watch after release. It is an operating aid, not a legal standard or compliance certification. The [NIST AI RMF](#) and its [Generative AI Profile](#) provide voluntary reference material for broader risk conversations. Qualified legal, privacy, security, and contractual owners must resolve questions in their domains.

## Start with a consequential moment

Choose a real marketing output and ask what could happen if it were wrong, misleading, disclosed to the wrong person, or acted on automatically. Identify who would notice, who could intervene, and whether the effect could be reversed. The same draft has different exposure when kept internal, shown to a prospect, published publicly, or used to allocate money.

Document these differences in the workflow contract. A change in destination can require a new permission even when the prompt remains unchanged. Governance becomes practical when an operator can identify the next decision and the evidence needed to make it. If a rule cannot be translated into an observable action or an enforceable boundary, it needs further design.

# 2. Classify exposure using separate dimensions

A single risk score hides why a workflow needs controls. Keep data sensitivity, claim consequence, audience exposure, action authority, and reversibility separate. A low-sensitivity input can still produce a damaging public claim. A highly sensitive input can be risky even when the output remains an internal summary. Do not let a low value on one dimension cancel a severe value on another.

| Exposure band         | Example use                  | Minimum operating arrangement                  | Escalation trigger     |
|-----------------------|------------------------------|------------------------------------------------|------------------------|
| Internal assistance   | Approved-source drafting     | Named reviewer and evidence                    | New source class       |
| External claims       | Public campaign copy         | Qualified claim review and publishing approval | Unsupported comparison |
| Sensitive information | Restricted insight synthesis | Explicit data approval and access tests        | Unexpected disclosure  |

| Exposure band    | Example use                  | Minimum operating arrangement            | Escalation trigger  |
|------------------|------------------------------|------------------------------------------|---------------------|
| Action authority | Send, publish, update, spend | Separate action permissions and approval | Unauthorized action |

These are suggested operating bands, not a legal classification. Adapt them to the product, geography, channel, and actual consequences. Use the highest consequential dimension to determine the minimum controls rather than averaging it away. Record who approved the classification and when it must be reviewed.

### Reclassification worksheet

Describe the old boundary, proposed change, affected information, new recipients, possible actions, reversibility, and worst plausible failure. Then identify which reviewer, technical restriction, or approval must change. A workflow moving from internal notes to a public knowledge assistant needs reassessment even if the underlying model is identical. A workflow gaining a write tool needs a new action review. If the team cannot explain the consequence of a new permission, keep the narrower permission until the uncertainty is resolved.

## 3. Design human review as skilled work

“Human in the loop” is incomplete unless the human has expertise, time, evidence, authority, and a clear decision. A reviewer who sees only a polished answer cannot verify claims efficiently. A reviewer who cannot reject the output is an observer, not a control. Design the review interface around material decisions rather than a generic approval button.

For the FICTIONAL B2B SaaS company NorthstarDesk, the product marketer reviews internal campaign briefs. All examples and numerical values in this guide are planning assumptions, never customer outcomes. The review packet includes the draft, source passages, product scope, unresolved questions, and a change summary. The reviewer first checks claims and exclusions, then audience fit, then style. This order reduces the chance that elegant writing distracts from unsupported substance.

### Review disposition card

- Accept: evidence supports the material claims and the intended use is permitted.
- Accept with edits: corrections are recorded and rechecked where material.
- Return: the work lacks evidence or fails task requirements.
- Escalate: a domain or permission decision exceeds the reviewer's authority.
- Stop: a hard boundary may have been crossed.

Measure review time and the kinds of defects found. If reviewers repeatedly rewrite most of the result, the workflow may be shifting work rather than reducing it. Include a second-person

review of selected accepted items to detect rubber-stamping and rubric ambiguity. Do not use agreement alone as proof of correctness: two reviewers can share a mistaken assumption. Give them access to the underlying evidence and a route to subject experts. Review is a designed service with capacity limits, not an unlimited safety net.

## 4. Define permissions as enforceable contracts

A permission contract states which actor may perform which action on which information for which purpose. It should distinguish read, draft, recommend, modify, send, publish, and spend. These verbs have different consequences. A system that can recommend a campaign change should not gain write access simply because the recommendation is usually accepted.

For each workflow, make a permission map with user identity, source collections, tool scope, output destination, approval requirement, and expiry. Ask the technical owner where each restriction is enforced. A prompt can communicate instructions, but sensitive access and action authorization need controls that do not depend on the model agreeing with the instruction.

### Permission-change test

Before granting a new tool, identify the smallest useful capability. A read-only report may answer the question without CRM write access. A draft message may serve the task without sending authority. If a write capability is necessary, define the exact object, permitted fields, destination, confirmation step, and recovery route. Inspect the actual requested action before approval rather than approving an open-ended instruction.

Keep assistive workflows and autonomous agents separate in the register. An agent that selects tools or intermediate actions needs assessment of those choices, not merely the final prose. The [OpenAI agents documentation](#) is an implementation reference when relevant; the control design must still be verified locally. Review permissions after personnel, source, integration, and workflow changes. Remove unused permissions rather than assuming they remain harmless. The least complicated authority that accomplishes the approved task is usually easier to inspect and maintain.

## 5. Establish a review-capacity budget

Human review has a finite service capacity. Estimate it from observed minutes per accepted item, difficult-case share, arrival variation, and reviewer availability. Do not derive capacity from draft-generation speed. A fast upstream system can overwhelm the people who carry the real acceptance responsibility.

Use a FICTIONAL NorthstarDesk planning calculation. Assume 100 briefs per month, 18 review minutes per ordinary brief, and an additional 30 minutes for each of 20 difficult cases. That is 40 hours of review, before meetings, training, leave, and incidents. At an assumed €65 loaded hourly rate, review represents €2,600 of capacity. Assume €300 tooling and €600 governance

maintenance: the governed route carries €3,500 of these costs before drafting. All figures are assumptions, not measured client results or current prices.

| Capacity stress            | Operational response                    | Decision owner |
|----------------------------|-----------------------------------------|----------------|
| Queue exceeds agreed age   | Limit intake or use manual route        | Workflow owner |
| Difficult cases rise       | Add specialist capacity or narrow scope | Domain lead    |
| Reviewer absent            | Activate qualified deputy or pause      | Business owner |
| Review becomes superficial | Reduce volume and inspect accepted work | Review lead    |

Compare full governed effort with the manual process at the same quality floor. If review burden erases the benefit, narrow the task or improve source preparation. Removing review to rescue the business case changes the risk model and requires a separate decision. Report ranges and capacity constraints honestly. The correct result may be that assistance is useful only for a subset whose evidence is already well structured.

## 6. Calibrate reviewers before relying on agreement

Give reviewers a shared set of examples that includes correct outputs, unsupported claims, misleading qualifications, appropriate refusals, and ambiguous cases. Have them score independently before discussing. Record both the disposition and the rationale. A disagreement about style is different from a disagreement about whether the source supports a claim.

Build anchor examples for each rubric dimension. “Good quality” is too vague. “Every material product claim is supported by an approved passage within the permitted product scope” is inspectable. Keep factual support separate from writing quality, and hard boundaries separate from graded preferences. Do not let excellent prose compensate for invented evidence.

### Calibration worksheet

For each case record reviewer A's decision, reviewer B's decision, evidence used, disagreement type, resolution owner, final anchor, and rubric revision. If the reviewers disagree because the source is ambiguous, route to the source owner. If the expected audience is unclear, repair the task contract. Do not automatically blame inconsistent people or an unreliable model when the business rule itself is unresolved.

A calibration sample is not population performance. It reveals how the rubric behaves on selected cases. Repeat calibration when product scope, language, reviewer group, or source class changes. An automated judge may help triage, but compare its decisions against qualified review and inspect its errors. The [OpenAI evals guide](#) provides a technical reference for

evaluation workflows; the governance decision remains with named owners. Keep disagreement visible in management reporting where it affects the confidence of release decisions.

## 7. Create an evidence standard for marketing claims

Classify claims by what would substantiate them. A product capability needs approved product evidence. A quantitative performance claim needs a clearly scoped study or measurement with conditions. A customer statement needs appropriate permission and context. A comparative claim needs current evidence for the relevant comparison. The model's fluency, a competitor's slogan, or an internal brainstorm does not substitute for substantiation.

Use a claim ledger with statement, source, relevant passage, scope, date, owner, permissible use, and expiry trigger. Separate facts, interpretations, proposals, and unknowns. A statement may be suitable for an internal hypothesis while being unsuitable for public copy. The ledger should make that distinction visible.

### A useful review exercise

Take “our platform cuts onboarding time” and ask what the source actually says. Does it concern one workflow, a selected user group, a controlled experiment, or an anecdote? Does the proposed wording imply a universal effect? If the evidence is narrower, narrow the claim or remove it. Do not ask the model to make unsupported content sound cautious while preserving the same misleading implication.

Keep a source owner responsible for updates and retirement. When product functionality changes, identify affected claims and outputs. Where downstream tracing is incomplete, state the limitation and consider broader review rather than pretending every use is known. Retrieval may surface an old document because it is semantically relevant; it does not establish that the content is current or approved. The [original RAG paper](#) is a research foundation for retrieval-augmented generation, not a certification of the truth of any particular marketing knowledge base.

## 8. Make exceptions temporary and inspectable

An exception is a conscious, bounded deviation approved by the person who owns the affected risk. It is not an informal workaround or a retrospective justification. Record the requested deviation, business reason, affected information, duration, compensating controls, approving owner, and automatic expiry. Some boundaries should not be waivable within the marketing team at all.

For example, a campaign deadline may justify a manual route or an extra qualified reviewer. It does not by itself justify uploading restricted customer information or granting publishing access.

If the person receiving the request cannot approve the relevant risk, they must route it to the authorized owner rather than treating urgency as consent.

### Exception register fields

Use exception ID, workflow version, requested change, reason, alternatives considered, exposure, owner, start date, expiry date, controls, affected cases, and closure evidence. Review open exceptions in the operating meeting before ordinary improvement ideas. A repeated exception indicates that the underlying workflow contract may be wrong or that a control is not practical. Repair the design instead of renewing the exception indefinitely.

Distinguish exceptions from incidents. An incident involves an observed or suspected failure needing containment and investigation; it cannot be legalized by writing an exception after the event. Keep incident evidence separate from routine reporting and restrict access appropriately. The owner should decide when to involve security, legal, privacy, communications, or other specialists. This guide does not prescribe jurisdiction-specific notification duties. Those depend on actual facts and qualified interpretation. The operating contribution is a clear route that gets the right people reliable evidence quickly.

## 9. Secure the source-to-action chain

A governance review must inspect more than the final text. Map the chain from source admission to retrieval, generation, review, destination, and action. A weakness at any transition can invalidate an apparently good output. For example, a correct internal answer may still be an unauthorized disclosure to the requesting user.

Require source metadata for ownership, approved use, sensitivity, version, effective date, and retirement. Enforce user access before the content reaches generation. Keep instructions found inside source documents separate from authorized system instructions. Retrieved text can be malicious or merely mistaken. Prompt-injection defenses need layered design and testing; a warning in the prompt is not a complete security boundary.

### Chain inspection

Start with a restricted document and verify that an unauthorized test user cannot retrieve it. Then use a harmless test document containing an instruction to bypass review. Confirm that no external action occurs. Finally, present a superseded source beside its replacement and inspect the chosen evidence and qualification. Record observed behaviour, not just the intended architecture.

When using hosted retrieval such as [OpenAI file search](#), review actual source collections, metadata filters, returned evidence, and access design. Product features do not establish organizational permission. Ask who can add files, change filters, and grant tools. Those administrative actions can alter the risk boundary. Include them in change control. A governance system that inspects only end-user prompts may miss the more consequential decisions made by administrators and source maintainers.

## 10. Turn monitoring into a response contract

A dashboard is useful only when a signal leads to an owned response. Choose signals tied to plausible failures: unsupported-claim returns, unexpected source access, review queue age, bypassed approvals, stale sources, repeated exceptions, and material changes in correction effort. Give each signal a denominator, owner, observation method, and response.

| Signal                    | Interpretation to investigate | Response owner           | First action        |
|---------------------------|-------------------------------|--------------------------|---------------------|
| Claim returns rise        | Source or generation issue    | Domain lead              | Inspect case sample |
| Restricted source appears | Access boundary issue         | Security/technical owner | Stop affected route |
| Review queue ages         | Capacity mismatch             | Workflow owner           | Limit intake        |
| Exceptions recur          | Contract or control mismatch  | Governance owner         | Redesign rule       |

Do not use a universal threshold copied from another company. Establish an initial operating limit from consequence, baseline, and capacity, then revise with evidence. A single unauthorized disclosure may warrant immediate containment even when its percentage is tiny. A gradual increase in minor formatting edits may call for routine repair rather than shutdown.

### Monitoring worksheet

Write the signal, why it matters, how it is observed, what information is missing, what triggers action, who acts, and how closure is verified. Test the route by generating a safe synthetic alert. Does it reach someone who can act? Is the evidence sufficient? Is sensitive content unnecessarily exposed in the alert? Monitoring should include accepted work, because failures can pass review. It should also include non-use and bypass, because low incident counts can reflect low visibility rather than good control.

## 11. Test governance under pressure

Policies usually look coherent in a quiet meeting. Test them when a deadline, absence, or confusing source makes compliance inconvenient. Run a tabletop exercise using a fictional work item and no live sensitive data. Assign the requester, reviewer, source owner, technical maintainer, and incident lead. Give each participant only the information their role would normally see.

Scenario one: a retrieved launch document includes a plausible instruction to ignore review and publish. Scenario two: an approved source contradicts a newer but unapproved presentation. Scenario three: the primary reviewer is absent and the campaign deadline is near. Scenario four:

a draft contains information that the requesting user should not access. Ask participants to act, not merely describe policy.

### Observe the decision path

Can the operator recognize the boundary issue? Can the reviewer return the work without sponsor pressure? Can the source owner determine which version governs? Can the maintainer disable the affected route? Does the incident lead receive enough evidence without unnecessary disclosure? Does the deputy know where the current instructions live?

Record time to identify the issue, blocked handovers, missing authority, and unclear terms. These are rehearsal observations, not incident-response benchmarks. Repair the highest-consequence gaps first and repeat the affected scenario. A new prompt is not an adequate answer to a missing permission boundary. A training slide is not an adequate answer to a reviewer who lacks authority. Governance is ready for bounded use when ordinary participants can execute the process and explain its remaining limits under realistic pressure.

## 12. Maintain governance without creating paralysis

The goal is proportionate control with a clear path to useful work. Review controls when the workflow changes, when defects recur, and at a scheduled interval appropriate to the exposure. Remove redundant checks only when evidence shows that another control genuinely covers the same decision. Do not remove them simply because no one has reported an incident.

Maintain a small set of operational artifacts: workflow register, source register, permission map, review rubric, exception log, incident route, and release decisions. Link them through workflow and configuration IDs. Each artifact needs an owner and a review trigger. A document without a maintenance responsibility gradually becomes misleading evidence.

### Governance health check

Ask a new colleague to locate the current instructions and process a doubtful case. Ask a reviewer to explain why a recent output was accepted. Ask a maintainer to demonstrate the actual restriction on external actions. Ask a business owner what would cause shutdown. Ask a source steward which material is overdue for review. Their answers reveal more than the number of policies published.

At handover, record which risks remain, which controls depend on humans, and which observations are still too limited for generalization. Never describe a checklist as a guarantee that failures cannot occur. If the operating burden exceeds the value of a workflow, narrow or retire it. If a useful workflow is blocked by an unnecessarily broad rule, redesign that rule with the relevant owners rather than encouraging quiet bypasses. Good governance makes both permission and refusal understandable, and gives the organization a reliable way to change its mind when evidence changes.

## 13. Define stop, containment and restart separately

Stopping prevents further exposure; containment addresses what may already have happened; restart is a new permission based on evidence. These are separate decisions. A technical operator can stop a route quickly without having authority to determine all downstream obligations. The incident lead coordinates the investigation with relevant specialists.

Preserve the configuration, source identifiers, affected cases, observed behaviour, recipients or destinations, and time sequence. Minimize unnecessary sensitive copies. Record what is known, suspected, and unknown. A regenerated answer does not undo disclosure, and a corrected source does not prove that every downstream use has been repaired.

### Restart packet

Include the failure mechanism or explicit remaining uncertainty, affected scope, containment actions, repair evidence, regression tests, revised controls, manual fallback, and named acceptance owner. Test the boundary that failed, then nearby scenarios. If the cause remains unknown, consider narrower access, reduced volume, or continued manual processing. Do not claim certainty because the original case no longer reproduces once.

A hard stop is appropriate for suspected unauthorized information exposure, forbidden external actions, absent review ownership, or inability to recover the manual route. Other defects may permit bounded operation under an explicit temporary decision. Define that decision before business pressure arrives. This guide provides operating steps, not jurisdiction-specific incident law. Qualified owners decide legal, privacy, contractual, and communication obligations from the facts. The governance team's job is to make those facts available and prevent the same unexamined permission from continuing by default.

## 14. Delegate governance with an evidence handover

Delegation should transfer a decision and the evidence needed to exercise it. Do not delegate a vague responsibility to “watch AI.” Specify the workflow population, permission boundary, review authority, exception scope, incident route, and escalation threshold. A deputy should know which decisions remain outside their authority.

Prepare an authority handover with current workflow register, source ownership, permission map, review anchors, open exceptions, unresolved incidents, recurring costs, and next review dates. Have the receiver inspect a real acceptance decision and a rejected case. Then ask them to explain what would change if the workflow gained a new language or a publishing tool.

## Delegation acceptance

The receiver confirms understanding, access to evidence, available capacity, authority to refuse work, and a route to specialists. The sender records limitations and pending decisions. If any of these are missing, transfer preparation may continue while the current owner remains accountable. A calendar invitation does not complete the handover.

Roll governance out by workflow exposure rather than imposing every control on every task at once. Start with a bounded service, observe where practitioners struggle, repair the interface, and then extend. Preserve minimum boundaries for sensitive data and actions throughout. The [Anthropic distinction between workflows and agents](#) helps identify when dynamic tool choice changes the oversight task. Do not infer that a more autonomous design is more mature. Maturity means that the organization can explain, operate, question, and reverse the permissions it grants.

## Worked exception-to-decision handover

Use this FICTIONAL NorthstarDesk exercise to test whether the exception process is usable. A requester wants to include a roadmap capability in an urgent internal campaign brief. The source owner has approved the current capability reference but not the roadmap statement. The operator records the requested deviation without inserting the claim into the draft. The reviewer identifies the missing authority and returns the claim for a domain decision. Urgency changes the scheduling question, not the evidence standard.

| Handover object     | Sender completes                 | Receiver decides             |
|---------------------|----------------------------------|------------------------------|
| Requested deviation | Exact claim and intended use     | Is the request in authority? |
| Evidence packet     | Current and proposed sources     | What can be substantiated?   |
| Alternative         | Current approved wording         | Is narrower work useful?     |
| Closure record      | Decision, expiry, affected draft | Has the exception ended?     |

The product owner may supply a different approved statement, approve a narrowly scoped internal hypothesis where permitted, or refuse the request. The publishing owner still controls any later external use. The operator records the resulting wording and source restriction, then the reviewer checks the revised brief. If the owner is unavailable, the affected claim waits or the manual route omits it. The workflow does not invent a replacement approval.

This walkthrough exposes a common diagnostic trap: a confident unsupported sentence may appear to be a model error while the real failure is an unresolved organizational claim decision. Repair the authority gap before tuning the model. Preserve the decision so the next campaign does not repeat the same exception.

## Primary references and limits

Sources checked 24 October 2026. Technical documentation changes; verify implementation details before release. Frameworks, worksheets and fictional assumptions are original operating proposals, not vendor guarantees.

- [NIST AI Risk Management Framework](#)
- [NIST Generative AI Profile](#)
- [OpenAI: Working with evals](#)
- [OpenAI: Agents](#)
- [Anthropic: Building effective agents](#)
- [OWASP: Prompt injection](#)
- [Lewis et al.: Retrieval-Augmented Generation](#)
- [OpenAI: File search](#)

## YOUR TRANSFER CANVAS

## Turn this resource into a team brief

Use this page for a joint review. Agree one practical output, observable evidence and an accountable owner. Add the specific methods and templates from the relevant chapters.

| Working field               | Your agreement                  |
|-----------------------------|---------------------------------|
| Priority output             | Risk-tier and permission matrix |
| Your starting point         | Finding / source / date: _____  |
| Smallest useful test        | _____<br>_____                  |
| Evidence & counter-evidence | _____<br>_____                  |
| Ownership & review          | Owner / reviewer / date: _____  |
| Stop or change when ...     | _____<br>_____                  |
| Decision & next action      | _____<br>_____                  |

Keep unresolved assumptions explicitly open. A completed worksheet is not a substitute for evidence or accountable judgement.

## Sources & further documentation

Primary sources checked 24 October 2026. Sources support the linked technical or factual statements; the working models and fictional cases are original editorial decision aids. Recheck changing platform details before implementation.

1. [NIST AI Risk Management Framework](#)
2. [NIST Generative AI Profile](#)
3. [OpenAI: Working with evals](#)
4. [OpenAI: Agents](#)
5. [Anthropic: Building effective agents](#)
6. [OWASP: Prompt injection](#)
7. [Lewis et al.: Retrieval-Augmented Generation](#)
8. [OpenAI: File search](#)

### Next step

If you want help adapting the working plan to your product, evidence and team capacity, request a non-binding strategy conversation.

[patrickwings.com](https://patrickwings.com)