

The Human Approval Gate

A practical guide to keeping a person at the decision point in AI-assisted work

A human approval gate is a controlled decision point where a named, qualified person checks defined evidence against written criteria before an AI-assisted output may affect real work.

Keith Staggars, MS, RN-BC
Nurse Leader and AI Trainer | Keith Staggars Studio
keithstaggars.com | August 2026

Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
<https://creativecommons.org/licenses/by/4.0/>



Direct answer

The gate is not a glance at a screen. It is a hold on consequence. It ends with one of four visible decisions: accept, revise, reject, or escalate.

This guide turns a vague promise of human review into an observable control with a named owner, evidence, criteria, decision paths, and a recovery record.

How to use this guide

Start with the consequence the workflow could create. Apply the CLEAR test, complete the companion worksheet, then run the synthetic cases that resemble the failure modes of your workflow. Keep consequential actions disabled until every required control has evidence.

Why ordinary "human review" is too vague

Many workflow descriptions say that a person stays in the loop. That promise leaves important questions unanswered:

- Which person is responsible?
- What must that person be qualified to judge?
- Which evidence must be inspected?
- What counts as acceptable?
- Can the reviewer reject the output without workarounds or pressure?
- What happens if the tool fails or the evidence is incomplete?
- Is the approved version the same version that is used?

If those questions do not have concrete answers, human review may be ceremonial rather than protective.

When to require a gate

Use a human approval gate before an AI-assisted workflow can:

- send or publish content outside a private draft space
- alter a record, schedule, permission, account, or system state
- make or recommend a decision that affects a person's access, opportunity, care, employment, money, or reputation
- create a financial commitment or charge

- expose, transform, or transfer sensitive information
- give professional, legal, financial, clinical, safety, or compliance guidance
- trigger another automated action that is hard to reverse
- proceed when the source evidence is incomplete, conflicting, or uncertain

Not every low-risk draft needs the same level of control. The strength of the gate should match the possible consequence, reversibility, sensitivity of the information, and difficulty of detecting an error.

The CLEAR test

C	Consequence held No real-world action occurs before approval.
L	Listed reviewer A named, qualified person owns the decision.
E	Evidence checked The reviewer inspects defined evidence, not polish alone.
A	Allowed outcomes Accept, revise, reject, and escalate stay available.
R	Record and recovery The decision, version, reason, and safe response are recorded.

C: Consequence held

The workflow must stop before the real-world action. Drafting and analysis may occur, but sending, publishing, changing, charging, or authorizing remains disabled until approval.

Evidence that this control exists can include a draft state, disabled send action, pending status, separate approval queue, permission boundary, or tested manual hold.

L: Listed reviewer

Name one person or role that owns the decision. Match the reviewer to the consequence. A fluent writer may judge clarity but may not be qualified to approve a clinical, legal, financial, security, or personnel decision.

Avoid shared ownership such as "the team will review." A backup reviewer may be named, but the decision record should identify the actual reviewer.

E: Evidence checked

List what the reviewer must inspect. The AI output by itself is rarely enough. Depending on the task, evidence may include:

- the original source and its date
- the complete input or approved redacted version
- calculations performed independently
- policy or regulatory references
- known limitations and confidence signals
- a version comparison
- test results and exception logs
- confirmation that restricted data was not used

The reviewer should be able to reach the decision from the evidence, not from trust in the tool.

A: Allowed outcomes

Define four outcomes before the workflow runs:

- 1 Accept: the evidence satisfies every required criterion.
- 2 Revise: the output can be corrected, then returned through the same gate.
- 3 Reject: the output must not be used.
- 4 Escalate: the named reviewer lacks authority, expertise, evidence, or certainty to decide.

Do not treat silence, delay, partial review, or a missing reviewer as approval. Do not let a deadline quietly convert an escalation into acceptance.

R: Record and recovery

Record the workflow, output version, reviewer, decision, date, evidence location, reason, limits, and next action. Keep the record proportionate to the risk and retention rules that apply.

Define recovery before a failure. The recovery path should say how to keep the work manual, prevent duplicate actions, revoke or correct an output, restore the prior state, and notify the right owner when needed.

Seven-step implementation

1. Name the consequence

Write the exact action that the gate prevents until approval. "Review the answer" is vague. "Hold the customer-facing email in draft and keep sending disabled" is observable.

2. Name the decision owner

Identify the person or role with the expertise and authority to approve the action. Record a backup or escalation route.

3. List prohibited inputs

State what must not enter the tool or workflow. Use only approved systems, accounts, purposes, and the minimum information needed. Synthetic or properly approved test data should be used for testing.

4. Define the evidence packet

List each source, calculation, test, version, disclosure, and policy check the reviewer must see. Make missing evidence a stop condition.

5. Write the decision criteria

Use criteria that can be answered consistently. Examples include factual support, completeness, privacy, authority, tone, accessibility, reversibility, and alignment with the approved purpose.

6. Test the gate

Run normal, missing-information, privacy, unsupported-claim, high-impact, duplicate-action, rejection, outage, and post-approval-change cases. A gate that works only when everything goes right is not ready.

7. Record the decision and monitor drift

Save the final decision and the exact version approved. Retest when the model, prompt, connected tool, policy, data source, user population, or consequence changes.

Worked synthetic example

Scenario

A fictional community organization uses an AI tool to draft a public workshop announcement from an approved event brief. The draft includes the title, date, location, accessibility statement, and registration link.

Consequence held

The announcement remains in a private draft. Publishing permissions are not available to the drafting workflow.

Listed reviewer

The communications coordinator approves public accuracy and tone. The event owner confirms the date, location, capacity, and accessibility details.

Evidence checked

- approved event brief
- public registration page in preview
- link check
- accessibility checklist

- difference between the generated draft and approved final copy

Decision criteria

- every logistical fact matches the event brief
- no invented sponsor, credential, quotation, or attendance claim appears
- the registration link resolves to the correct event
- the accessibility statement is present and accurate
- no private attendee or staff information appears

Allowed outcome

The reviewer may accept, request a revision, reject the draft, or escalate conflicting event details to the event owner. Publishing occurs only after acceptance.

Record and recovery

The record includes the approved copy version, reviewer, date, source brief, and link-check result. If an error is discovered after publication, the coordinator corrects or removes the announcement and records the change.

Gate worksheet

Complete these fields before the workflow affects real work:

- 1 Workflow and purpose
- 2 Consequence being held
- 3 Decision owner and backup
- 4 Prohibited inputs
- 5 Evidence the reviewer must inspect
- 6 Acceptance criteria
- 7 Revision conditions
- 8 Rejection conditions
- 9 Escalation route
- 10 Decision record and retention location
- 11 Manual fallback and recovery steps
- 12 Change conditions that require retesting

The companion one-page worksheet provides space for each field.

Failure signals

Stop and redesign the gate when:

- the workflow can act before approval
- nobody can name the decision owner
- the reviewer sees only the generated output
- acceptance depends on a feeling rather than written criteria
- rejection is discouraged or automatically overridden
- the approved version can change before use
- the same click both approves and performs an irreversible action without confirmation
- missing evidence is treated as permission to continue
- a failure can create duplicate messages, charges, records, or changes
- nobody owns correction after an error

Measurement without false certainty

Useful implementation measures include the percentage of cases with a named owner, complete evidence packet, recorded decision, tested stop, tested rejection path, and tested recovery path. These are control-completeness measures. They do not by themselves prove safety, compliance, accuracy, or improved outcomes.

Track false approvals, false rejections, escalations, reversals, duplicate actions, recovery time, and reasons for revision when those measures are appropriate and can be collected without exposing protected information.

Limits and responsible use

This guide is educational. Adapt it to the laws, contracts, privacy rules, security controls, professional duties, records requirements, accessibility needs, and organizational policies that apply. A worksheet cannot make a prohibited use acceptable or make an unqualified reviewer qualified.

Use public or authorized sources. Use synthetic information for demonstrations. Do not enter patient information, employee records, employer material, client data, passwords, credentials, or confidential information into an unapproved tool.

The Human Approval Gate and CLEAR test are proposed practical methods. They have not been validated as research instruments and should not be represented as a certification, standard, guarantee, or substitute for expert review.

Public source foundation

- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). <https://doi.org/10.6028/NIST.AI.100-1>
- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. <https://doi.org/10.6028/NIST.AI.600-1>
- U.S. Department of Health and Human Services. Minimum Necessary Requirement. <https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/>
- U.S. Department of Health and Human Services. HIPAA and Cloud Computing. <https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/>

These sources inform the risk, governance, privacy, and oversight principles in the guide. They do not endorse this kit or Keith Stagers.

Suggested citation

Stagers K. The Human Approval Gate: A Practical Guide to Keeping a Person at the Decision Point in AI-Assisted Work. Version 1.0.2. Keith Stagers Studio; 2026.

<https://www.keithstagers.com/human-approval-gate/>

License

This Version 1.0.2 release is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>. See LICENSE.txt for the attribution notice. Third-party source material remains subject to its original terms.