Human Oversight Reliance Review

A practical method for deciding what an organisation can rely on human oversight to do

Maja Kurek, AIGP, CIPM, CIPP/E · RuleBridge Advisory

Abstract

Human Oversight Reliance Review is a platform-neutral method for assessing what an organisation can rely on human oversight to accomplish under specified operating conditions. It addresses the gap between assigning a human role and establishing that relevant cases reach qualified people, intervention can take effect in time, and shared capabilities can meet competing commitments.

Six steps — Define, Trace, Examine, Reconcile, Decide and Revisit — take a specified oversight function from its requirements through evidence assessment to a separately recorded operating decision. Bounded claims are examined against supporting and challenging evidence. The review covers selection and information preparation before the human decision, and execution and consequences after it. It examines whether commitments can be fulfilled together across shared expertise, information, intervention routes and external providers. Replacement arrangements and proposed remedies face the same scrutiny, including their effects on other work and people affected by action or delay.

Findings remain distinct from operating authorisation: permission to operate is not evidence that oversight performs as intended, and a planned improvement is not a capability available now. Material changes prompt reassessment of affected claims and decisions. Handovers carry unresolved concerns, relevant evidence, responsibility and the remaining time to act.

The output is an evidence-linked account, rather than an aggregate score, of what oversight can support, what remains unresolved, what is permitted now and which conditions require a renewed decision. The guide, editable records, two synthetic worked examples and practice checks support application through the supplied workbook or existing organisational systems. Across design, deployment and continuing operation, including agentic workflows, the method supports decisions to retain, restrict or redesign AI use.

1. Start with the decision the organisation needs

A short example

An AI service drafts offers. Staff approve every offer before it reaches a customer. Management treats that approval as protection against inaccurate content, unauthorised commitments and inappropriate handling of confidential information.

Those are three functions, not one. The employee may be qualified for standard content but require a specialist for unusual commitments. The service may already have sent information to an external processor before the employee sees the offer. The specialist may also be responsible for urgent complaints elsewhere.

The review therefore asks which function the employee can perform, what must happen earlier, what requires another capability, and whether all commitments can be met together. A signature at the end cannot answer those questions. A restriction may be justified for one part of the work while another continues under an established arrangement.

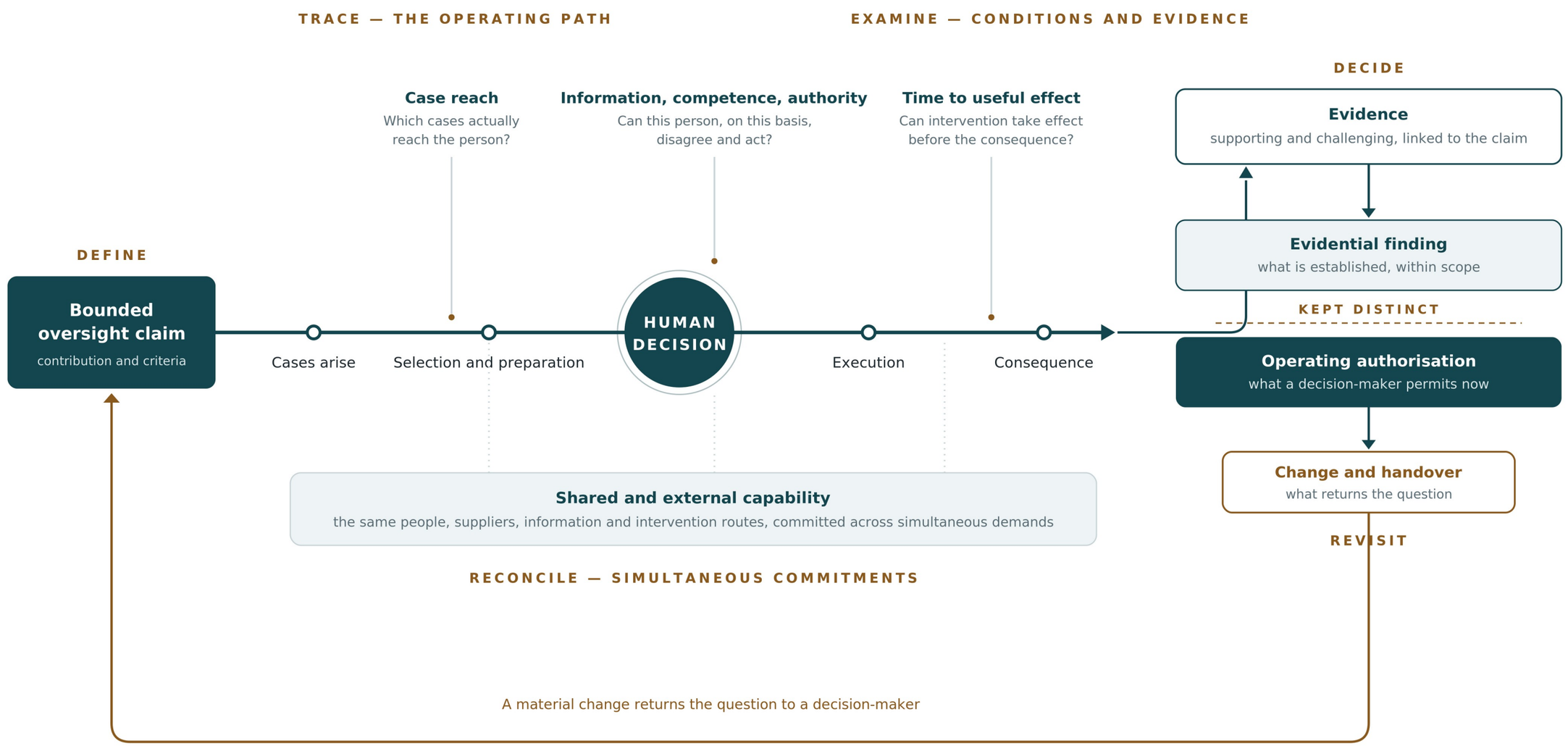
Six steps, one connected record

Step	Question	Working record
1. Define	What must oversight accomplish, for whom and against which criteria?	Scope; Functions
2. Trace	What happens before, at and after the relevant human decision?	Functions; Claims
3. Examine	What capabilities and evidence support the expected contribution?	Claims; Evidence; Evidence Links
4. Reconcile	Can the linked commitments be met together, including under relevant changes?	Dependencies; Joint Review
5. Decide	What is established, and what does an authorised person decide may happen?	Claims; Decisions
6. Revisit	Who detects a change to a material condition, and what happens meanwhile?	Change

Start with one process or one shared oversight capability and a real decision: launch, extension, redesign, continued operation or transfer. Keep the scope small enough to examine, but follow an important dependency beyond the initial boundary when ignoring it could change the conclusion. Record the remaining boundary rather than treating unseen work as absent.

First use. Read this page and sections 2–4. Open the blank workbook at Start. Create a Review ID, describe the operating context, and complete one function before creating its claims. Use Example A to see the full record; use Example B for selection and handover. The last columns in Claims and Evidence Links check record structure only. They do not assess oversight or permit operation.

Figure 1. The Reliance Review Map



2. Purpose, people and proportionate use

When to use the method

Use it when a decision depends on people being able to scrutinise, authorise, intervene, investigate a pattern, resolve an exception or consider a challenge to AI-assisted work. It is especially useful where several processes share expertise, where a supplier holds information or execution capability, or where changes alter what reviewers see and can do.

A simple approval point may need only a small set of records. Connected processes may require joint scenarios. High-consequence or contested questions require appropriate specialist input and stronger evidence, not merely more rows. There is no minimum number of claims, reviewers or meetings.

Begin with existing records, not a replacement inventory. A usable service record, quality log, change process or decision register may remain the system of record. Reference it and map the necessary fields. Completing this workbook is optional; preserving the method's distinctions is not.

Roles

The **review owner** keeps scope, evidence and open questions coherent. **Process owners and oversight personnel** explain the work and examine the proposed arrangement. **Capability owners** account for shared people, information, supplier services or execution routes. **Qualified specialists** resolve technical, domain, legal, privacy, security or human-factors questions within their competence. The **decision authority** has the mandate to decide the relevant operating scope and a workable route to implement that decision.

One person may hold several roles. Record material conflicts and arrange a credible opportunity to challenge consequential assumptions. A facilitator does not gain authority to permit use by completing the record. A capability owner does not become accountable for every aspect of every dependent AI system.

Include people who can describe the effects of both action and delay, and the route by which an affected person can seek review. Direct participation, representatives, existing complaints and appropriate research can all contribute; record what was used and whose perspective is missing. Do not expose confidential complainants in a general workshop.

What reliance means

In this method, reliance is a bounded organisational judgement that a specified human-oversight function can support a stated control or decision objective under recorded operating conditions, based on evidence about case reach, relevant capability, executable intervention and material dependencies. Reliance is claim-specific, conditional and revisable; it is not a general property of a person, role or workflow.

Boundaries

The method supports organisational design and decisions. It is not a legal opinion, technical model evaluation, independent assurance, certification, incident-response procedure or determination that AI is safe, fair or compliant. It supplies no universal acceptable error rate, staffing ratio or autonomy threshold.

Urgent safety, incident, security, privacy, employment and remedy processes take precedence. Do not wait for a completed workbook before acting through existing authority. Capture the decision and evidence afterwards where necessary. A decision to accept uncertainty cannot waive a mandatory obligation.

3. Keep the record intelligible

Three things that must remain separate

A **requirement** states what the oversight function needs to achieve and why. A **claim** is a specific proposition being assessed about that function in a defined state and context. A **decision** states what an authorised person permits or requires. Evidence supports or challenges the claim; it does not make the decision automatically.

Example: a service rule requires nonstandard commitments to be checked before dispatch. The proposition that all such offers were held during a defined period is a claim. An authorised restriction on dispatch is a decision. The rule alone does not establish that the hold worked, and the restriction does not prove that the reviewer makes good judgments.

One bounded claim, several evidence links

Give a claim one ID and one assessable proposition. State the population or cases, configuration, period, mode and relevant conditions. Split a sentence when its parts could have different findings or different decisions: timely review and competent review are usually separate claims.

Link every relevant supporting and challenging item to that same claim. Each link states what the item shows, its limitation and why it matters. Do not create separate copies of the same claim merely because one source is a policy and another is an observed record. Reconcile their relationship in a single assessment. An unexamined conflict remains visible.

Create a new claim or a preserved revision when the proposition, relevant state or scope changes. A proposed design and a current arrangement are not the same claim. Do not overwrite the history used for an earlier decision.

Three independent descriptors

Target state: Current, Proposed or Historical. This describes the arrangement being assessed, not the source used to assess it.

Evidence type: Design or rule; Account; Controlled check; Operational record or observation; Analytical derivation. These describe how an item originated. They are not a ranking and do not confer credibility automatically. A log is an operational record; whether it is complete or authentic remains a separate question. An arithmetic result is a derivation dependent on its inputs.

Assessment state: Not assessed; Assessed; Reassessment due; Superseded. A previous conclusion may remain recorded while its current applicability needs review. Never display it as fresh simply because its text has not changed.

A design document can accurately describe an installed control. A proposed control can have a sound design without existing in operation. The field called Target state prevents those facts from being merged.

4. Rules for evidence and conclusions

Evidence conclusions

Conclusion	Meaning
Supported within scope	The relevant evidence and reasoning support the specified proposition under the recorded conditions. Material challenges have been considered. This is not a universal guarantee.
Contradicted within scope	Relevant evidence establishes that the proposition does not hold in its stated form, or that the stated criterion was not met. Preserve what is established and the boundary of the contradiction.
Not established	The evidence is absent, inaccessible, too limited, conflicted or otherwise insufficient to settle the proposition at the required scope. State what is missing and its decision relevance.

Leave the conclusion empty only while Assessment state is Not assessed. After assessment, use one of the three conclusions. A record that is awaiting assessment is not the same as an assessed finding of insufficient evidence. Reassessment due preserves the dated prior conclusion without renewing it.

Describe the problem separately. A missing capability, missed referral, late intervention, disputed criterion or poor judgment is an issue description, not an alternative evidential status. A missing capability can itself be counterevidence. Several issues can coexist, and none establishes its cause automatically.

How to reach a conclusion

Record the proposition and criterion before selecting evidence where practicable. Where a review starts from a known event, state that fact. Identify both supporting information and plausible challenges, then explain why the evidence bears on the proposition. Treat copies, shared sources and mutually dependent accounts as such; do not decide by counting sources.

For sampled evidence, record the eligible population, period and selection route. A targeted set is useful for exposing a weakness but does not estimate prevalence. An observed miss can contradict a universal claim even when its frequency is unknown. A single successful review cannot establish universal performance. If the available result concerns a smaller scope, narrow the claim explicitly and preserve the original question as unresolved where it still matters.

Distinguish an observed outcome from a causal explanation. An action sent before review could reflect a timeout, a configuration defect or an unauthorised route rather than overload. More staffing is not a demonstrated remedy until it addresses the relevant problem.

Protect the evidence and the people

Keep sensitive source material in an appropriate controlled system. Record a usable locator, custodian, access basis, date and limitations rather than copying entire conversations. Evidence links must still allow an authorised reviewer to examine the basis of the conclusion. Where that is not possible, state the limit; a supplier's confidence is not an inspected result.

Do not use the workbook to rank individual workers, expose complainants or aggregate personal performance scores. Retention and access must fit the purpose and applicable obligations. Preserve dissent about the arrangement without turning it into a dossier on the person raising it.

5. Define the expected contribution

Step 1 — Purpose, criteria and scope

Write each oversight function in terms of an action or judgment and its intended contribution. Identify the people or interests affected, relevant obligations and who is qualified to interpret them. Useful sources include domain standards, contractual commitments, authorised service rules, impact assessments and substantiated stakeholder needs. None is automatically interchangeable with another.

The method does not allow management to set a trivial target merely to make the assessment pass. Record why the selected function and criterion address the relevant consequence. If that choice is disputed or a necessary specialist conclusion is missing, create an explicit claim or open question and identify who can resolve it.

Separate preventive approval, pattern detection, exception handling and a person's opportunity to challenge a decision. Retrospective sampling can detect a pattern without preventing each individual action. A right or appropriate opportunity to obtain human consideration cannot be removed simply because automated decisions appear more accurate on average. Due-diligence guidance provides a broader context for considering effects on people and remedies rather than internal throughput alone. [5]

Criterion design. State what would count as meeting the requirement, what would count against it, and how ambiguous cases are adjudicated. Use qualified judgment where a numeric threshold would be artificial. For empirical estimates, obtain suitable statistical advice where needed; do not borrow a threshold from a worked example. A change to the criterion receives its own reason, authorisation where relevant and effective date, rather than retroactively rescuing a failed result.

What to write in Functions

Record the expected contribution; its source and rationale; the eligible cases and exclusions; the relevant timing or intervention window; and the route from input to action. Include the responsible function and affected people. State the actual operating configuration: task, model or version where known, interface, prompts or relevant instructions, data routes, tools, permissions, operating hours and applicable constraints. A stable product label alone is not sufficient context.

Create only decision-relevant claims. A function usually needs a small number covering its reach, capability or quality, and important joint dependencies. Reuse the same shared-capability record rather than inventing separate availability promises in each process file.

Exit question: could a second competent reader tell what oversight is meant to do, for which cases, and what evidence could change the proposed conclusion?

6. Trace the actual work

Step 2 — Before the reviewer, through the decision, to its effect

Start where the need for oversight arises, not at the queue the organisation already calls reviewed. Follow selection, summarisation, prioritisation, translation, attachment processing, merging and closure. Determine whether another concern or required decision can disappear when one part of the work finishes.

Compare the source population and its destinations where records permit. Completed cases, pending work, redirected work, processing failures and bypasses should not silently drop out. When the population cannot be reconstructed, describe that limitation and any separately established recordkeeping deficiency. A review of escalated cases establishes nothing by itself about cases the selection mechanism omitted.

For each relevant action, identify who can decide and how the decision becomes effective. An escalation may send a message, hold a particular item, stop future execution or disable an entire service. Those are different interventions. Trace the particular control or agreed execution route rather than infer its scope from a label.

Time to useful effect

Consider the whole path: detection, understanding, decision, communication and execution. Compare it with the last point at which the intended consequence can still be changed. Keep a service response target separate from a deadline that changes an affected person's options. Preserve the original clock across handover. Record time zones and uncertainty when relevant; do not invent an exact deadline.

A route through a qualified administrator can be adequate; the reviewer need not personally hold every technical permission. Conversely, a stop control may leave queued, delegated or completed actions outside its effect. Stopping new messages cannot retract information already received elsewhere.

For agentic systems, describe what can happen without a new human decision: tool calls, onward delegation, spending, data access, persistent activity and external communication. Autonomy is context, not an automatic recommendation for a particular form of oversight. The OECD and IMDA provide related conceptual and organisational guidance. [13,14]

Evidence to seek

Use configuration records, actual transaction traces, relevant observations and safely conducted checks. Determine what happens when approval is unavailable, a request times out or a component fails. A requirement for positive approval must not become a timeout permission merely because a queue is busy. A designed test in a non-live environment must be described as such, with its configuration and limits.

Exit question: what exactly could the human decision alter, for which cases, and before or after which consequence?

7. Examine capability and contribution

Step 3 — More than a named person

Examine authority, information, competence, capacity, consequence-safety and traceability as diagnostic prompts. Their relevance depends on the oversight function and the way capabilities are distributed across people, technology and the organisation. They are not six conditions that each individual must personally possess or a substitute for evidence of the expected contribution. [1,3,11,16]

Information. Is the available representation sufficient and usable for the particular judgment? Sometimes a well-designed summary and access to clarification suffice; sometimes the source and attachments are necessary. Check fidelity and omissions where they matter. Universal access to all raw material is neither required by this method nor necessarily appropriate.

Competence and judgment. Match knowledge to the case class, not merely a role title or course completion. Examine justified acceptance as well as correction, escalation and recognition of insufficient information. Include correct, defective and ambiguous material where controlled checks are appropriate. Use relevant quality criteria and disclose the adjudication process. Neither a high nor a low override rate is intrinsically good.

Authority and execution. Establish mandate, practical leverage and the route to act. A technical permission does not establish permission to change every affected use. A managerial mandate without an executable route is likewise incomplete. Role concentration may warrant challenge or compensating arrangements; it does not alone prove a bad decision.

Capacity and incentives. Account for relevant expertise, availability windows, other work, preparation, handover, interruptions and escalation. Ask whether throughput targets, fees or blame make appropriate questioning costly. Seek concrete evidence and preserve workers' ability to disclose limits without penalty for raising them.

Traceability. Establish what can be reconstructed, and why reconstruction is needed. A complete approval log does not show the quality of each judgment; a missing log does not necessarily prove no review occurred. A retention failure, where established, is a distinct finding.

Evidence of the contribution itself

Research does not support assuming that combining a person and AI necessarily improves task performance. [2] Identify evidence appropriate to the expected contribution: a working hold for prevention, adequate case judgments for quality, referrals against an appropriate reference point for coverage, or an accessible and meaningful response for challenge.

Use existing evidence where suitable. When a new check is needed, define cases, criteria and boundaries before running it. A workshop response is not routine-use performance. Do not turn a limited success into a general claim about fairness, safety or error reduction.

Exit question: which conclusions are supported by evidence of the relevant work, and which still describe an intended arrangement?

8. Reconcile linked commitments

Step 4 — A dependency must account for everything material it serves

Record an important capability once: a specialist team, source, supplier service, tool needed for intervention or authority capable of resolving a conflict. Identify its owner, operating limits, associated functions and known other commitments. Unknown external commitments are a limitation, not unused capacity. Private supplier information may remain private, but the customer's assessment must then state what it can and cannot establish.

In Joint Review, describe what each linked function requires, when it requires it, which resources are interchangeable and which are not. Analyse the whole relevant set. Pairwise feasibility is not enough. Three separate 60-minute tasks require 180 minutes even when each pair fits a 150-minute window. Conversely, one legitimate review serving two purposes should not be counted twice. These numbers illustrate arithmetic, not acceptable workload or quality thresholds.

A daily sum cannot settle whether two urgent demands need the same person at the same time. A different skill, language, permission or location may prevent substitution. Include pending work and the cost of the proposed response, not just new arrivals. Shared capability can be beneficial; the aim is a workable arrangement, not the elimination of sharing. Related work describes portfolio capacity and the pace, incentives and resource constraints of monitoring in use. [6,12]

Scenarios that can change the decision

Select plausible conditions material to the scope: ordinary operation; concentrated or more complex demand; a missing specialist or source; a supplier limitation; model or interface change; incomplete handover; or a failed intervention. Record the reason for excluding an apparently relevant condition. Use observed conditions where available, and distinguish them from assumptions and constructed tests.

Look for common causes, not only resource shortage. Two reviewers using the same defective summary are not two independent information safeguards. A backup supplier may depend on the same unavailable expert or service. Automated triage can reduce workload but introduces a selection dependency; examine the cases it excludes by a route not determined solely by its own uncertainty flag.

Test the proposed response against the same constraints

Additional staff must bring the required capability and time. A manual fallback using the same overloaded people provides no extra availability. Reassigning a duty may improve one process while depriving another. Holding work may prevent one consequence while harming timely access to a service. Record who is affected, possible alternatives and a route for urgent circumstances.

A scenario can support a conditional conclusion without proving how often that condition will occur. Keep its assumptions visible and create an Analytical derivation evidence item when reasoning or arithmetic is used. No scenario result becomes a measured population outcome.

Exit question: can the relevant commitments be met together, and does the proposed answer merely move the problem elsewhere?

9. Make the finding useful for a decision

Step 5 — Assessment is not authorisation

For each material claim, record an evidence conclusion, its reason and its boundary. Consider contradictory information explicitly. Do not average away a known departure, count source types as votes, or treat a design record as stronger than a reliable observation solely because it is formally approved. Where conflict cannot be resolved, use Not established and explain the conflict.

Then prepare the organisational decision separately. The decision authority identifies what may operate now, what is excluded or held, the conditions, alternatives considered, material uncertainties retained, and the actions or specialist determinations needed. The method supplies no authority to waive a binding obligation and no default permission when the record is incomplete.

An existing approval may be recorded even if its evidential basis has not been assessed. Mark that gap explicitly; do not manufacture a supporting finding retrospectively. Likewise, an assessor may finish a finding while the authority has not decided. Use Decision disposition Pending rather than an empty row that looks like a completed review.

Choose the response by the diagnosed problem

An unreviewed action that remained held and an action released without required review need different responses. Poor judgment despite adequate timing differs from an inaccessible evidence source. A record gap differs from proof that no review took place. Describe cause only to the extent established; do not allocate blame by inference.

Possible responses include continuing without change in the assessed scope, narrowing cases or permissions, changing when oversight occurs, improving the information provided, protecting qualified time, arranging an executable escalation or replacing the design. Removing unnecessary review may be reasonable where its purpose is no longer justified, provided relevant obligations and affected interests are addressed. There is no requirement to create a new committee or corrective task when the evidence supports the existing arrangement.

Current operation, planned correction and past effects

A sound future design is not an installed safeguard. Keep its claim marked Proposed. To use it as a basis for current operation, establish the appropriate implementation and capability evidence and record the transition. An authorised conditional release can exist before installation, but the permission must not activate until its stated conditions have been met.

Address the affected period separately when a gap may have mattered: which people, actions or records need examination, correction, communication or no further action, and who determines that? A repaired control or empty queue does not settle the past. Record relevant effects on other processes and the practical owner of implementing a restriction, not just the name of the person who decided it.

Exit question: can a reader distinguish what is established, what is only recommended, what was actually decided and what is in effect now?

10. Keep the decision connected to change

Step 6 — Detect, route and revisit

For each material condition, identify a signal or information source, the person who will notice or receive it, the route back to decision authority and what is permitted while a new decision is pending. A review date without a responsible person and a defined response is not a monitoring mechanism.

Reassess proportionately when a change can affect the basis of a conclusion: different cases or languages, additional tools or external reach, a changed supplier, altered source material, new review incentives, loss of expertise, changed intervention route, different model behaviour or shared demand. Start either from the affected process or from the changed capability and follow the links in both directions.

A model update does not automatically make a person's earlier knowledge wrong, but it also does not prove that the earlier evidence remains applicable. Research on human–AI teams demonstrates that an accuracy improvement can alter compatibility with users' learned expectations. [7] State what carries forward, why, and what needs another check.

Handover preserves unfinished work

Separate assignment, acceptance and capability. The incoming person may accept routine work but lack the access or mandate to resolve an open concern. Record those states separately, the interim owner and what the service does while transfer remains incomplete. Receiving a queue or a signature is not sufficient evidence that an oversight function can be performed.

Carry the concern, relevant evidence, current configuration, decision route and time still available. A closed individual case must not silently close a wider concern. Conversely, an unsupported or obsolete concern needs a reasoned resolution rather than becoming a permanent warning. Responsible Handover supplies an established foundation for treating transfer as an exchange of requirements and capabilities, not simply delivery. [8]

The successor must be able to question the inherited interpretation. Protect relevant information and retain only what the purpose requires. Preparing, reading and testing a handover take time; include that work in the commitment, without turning the record into an archive of every conversation.

Use the organisation's existing follow-up process

The Change sheet identifies standing conditions and response routes. Link actions to the organisation’s existing incident, change or corrective-action record, with an owner and completion criteria. RuleBridge’s AI Change Accountability & Reauthorisation is an optional companion for tracing a specific change from signal through decision to completion; HORR can be used independently.

Preserve the record on which earlier decisions relied. Mark a claim Reassessment due or Superseded rather than silently refreshing its date. A current view should point to the applicable revision and retain access to the earlier basis.

Exit question: when a condition changes, who learns it, who decides, and what happens before that decision?

11. Working with the workbook

The blank workbook and completed example workbook use the same structure. Each table contains preformatted entry space, filters and frozen identifiers. Grey check columns provide limited structural warnings. They never assign a conclusion or an operating decision. Read the field notes on Dictionary before extending or importing records.

Sheet	Complete it when	What belongs there
Scope	At the start of every review	Review purpose, configuration, boundaries, roles, people affected and parallel processes.
Functions	For each relevant contribution	Requirement, criteria, cases, actual path, timing and responsibility.
Claims	For each decision-relevant proposition	Target state, bounded claim, assessment state, conclusion, integrated reasoning and unknowns.
Evidence	Once per distinct evidence item	Type, locator, scope, collection or selection method, source relationships and access limits.
Evidence Links	Whenever an item bears on a claim	Supports, Challenges or Context; the precise inference and its limits.
Dependencies	For a material capability	Owner, capacity or coverage limits, associated functions, other commitments and substitutes.
Joint Review	For relevant shared or changed conditions	Joint requirements, assumptions, analysis, linked claims and consequences of the response.
Decisions	For a needed or recorded decision	Authority, current permission and exclusions, reasons, conditions, implementation basis and pending issues.
Change	For material continuing conditions or a handover	Source of change, detection and response owners, interim rule, clock and transfer states.
Readout	To read one claim or decision clearly	Select an ID to retrieve its recorded text; a view, not a new assessment.

Record conventions

Use stable identifiers within the workbook: R for reviews, F for functions, C for claims, E for evidence, L for evidence links, P for dependencies, S for joint scenarios, D for decisions and H for change conditions. Prefix cases where several reviews share a file. The examples use A and B. Separate several linked IDs with a semicolon. Do not put an external document title where an internal ID is required; use a locator field.

One claim can depend on several functions. One evidence item can bear on several claims through separate links. One decision can refer to several claims or state No assessment recorded. Do not copy a claim merely to attach another item. In a revised review, preserve the earlier rows or a dated read-only export and record the previous ID or revision reference. The workbook does not enforce an append-only audit trail.

Blank templates contain no specimen evidence or preselected findings. The example workbook contains only synthetic material. Evidence type inside those examples describes a fictional source in the story, not real organisational evidence collected for this method.

12. A usable claim and a usable decision

Claim card

A claim can be expressed in ordinary language. For example: “In the twelve scripted dispatch attempts on configuration Q7, a nonstandard offer could not be sent before specialist approval.” That is more assessable than “human oversight works”.

The claim record states the target state, exact cases and conditions, criterion, evidence links, assessment state and conclusion. Its reasoning should explain why the evidence supports or challenges the proposition and what remains outside it. If operational data later show an unapproved dispatch, that does not negate the result of those twelve tests. Record the operational-performance claim separately and revisit any broader decision that relied on the tests. Evidence must address the same proposition and scope before it is treated as counterevidence.

Use the conclusion to say no more than is established. A controlled exercise can demonstrate the tested path under its conditions. It does not establish ordinary-use frequency, the quality of every specialist decision or a lower rate of harm. Where a broader operating decision relies on extrapolation, record that assumption in the decision and identify how it is monitored.

Decision card

Required content	Completion question
Authority and time	Who is entitled to make this decision, for which scope, and when was it made?
Present operating position	What is permitted now, what is held or excluded, and what evidence shows required conditions are in place?
Reasons and limits	Which findings, unresolved matters and alternatives were considered? Where is no assessment recorded?
Effects and actions	Who else is affected, including people who wait? Who implements the response and considers the affected period?
Continuing conditions	Which change or deadline returns the matter, who notices, and what applies if no new decision arrives?

An organisation may keep this content in its existing approval or service record. The example decision text is a teaching device, not a suggested legal delegation. Record disagreement and unresolved authority. Where no competent authority can be identified, the method cannot substitute for that absence; escalate through the organisation's existing governance and do not represent the review as approval.

Stop adding detail when it no longer changes the decision

The minimum useful review includes a defined scope, an appropriate function and criterion, at least one material claim with its evidence or explicit uncertainty, a check of relevant dependencies, and a decision or visible pending decision with a response route. Explain why joint review is unnecessary in a genuinely isolated case. Do not fill optional sheets with fictional complexity to make a small arrangement resemble a large one.

13. Methodological foundation and relationship to other work

HORR brings established human-oversight research, evidence-based argument and organisational control analysis into a connected review procedure. Its focus is the contribution attributed to an oversight function under actual operating conditions, especially where several uses depend on shared or external capabilities.

Structured assurance arguments connect scoped propositions with evidence, assumptions and reasons open to challenge. PRAISE applies this discipline to ethical acceptability. [9] HORR uses the distinction between a claim, its evidence and its assumptions to support organisational review; it is not a formal assurance case. STPA provides a related foundation for examining problematic control interactions even where individual components have not failed. [10]

Gaube and colleagues connect oversight architecture to monitoring, decisions and intervention. [3] Vanynh combines control-pattern selection with meaningful-review readiness. [11] Human Approval Gate Kit provides a practical approval-gate approach. [17] HORR concentrates on reconciling the contributions attributed to oversight across actual work, shared capabilities and change. These approaches may be complementary; their relative usefulness depends on the task and has not been comparatively evaluated here.

Context crosswalk — not a compliance checklist

Source context	Where this method provides a practical record	What the record does not establish
NIST Playbook MAP 3.4–3.5 [4]	Functions, Claims and Evidence connect required proficiency and oversight assessment to the task.	NIST approval, certification or a prescribed competence test.
OECD responsible-AI due diligence [5]	Purpose, affected people, business dependencies, response effects and review of results.	Completion of all due diligence or fulfilment of every stakeholder obligation.
EU AI Act, Articles 14 and 26(2)–(3), where applicable [15]	Role, support, intervention and information records can contribute to examining the relevant arrangement.	Legal classification, applicability, conformity or discharge of all provider/deployer duties.
Agentic-AI guidance [13,14]	The actual action path, available controls, permissions and changes are made explicit.	A universal autonomy threshold or proof that an agent is adequately contained.

The legal crosswalk is a contextual reference to Articles 14 and 26, not a statement of the obligations applicable to a particular deployment. The Commission article pages cited reproduce the original 2024 text. Current consolidated law, transitional provisions, roles, jurisdiction and dates require qualified assessment. This method contains no universal legal implementation timetable.

14. Application, authorship and use

Application and evidence

The worked examples and practice checks are constructed demonstrations of the method. Organisational effectiveness has not been evaluated. Conclusions in an actual review depend on its criteria, evidence, operating conditions and the relevance of that evidence to the decision at hand.

Use the review with qualified contributors, usable evidence and a decision authority able to act. Identify unseen work and uncertain dependencies explicitly. Keep sensitive evidence and decision history in the appropriate organisational systems; the workbook supplies linked records and structural checks, not access controls or automatic enforcement.

Authorship

Maja Kurek, AIGP, CIPM, CIPP/E, is Founder and Principal of RuleBridge Advisory. Her work connects organisational AI governance, human oversight and decision accountability across external and agentic systems. Her judgment draws on nearly a decade of regulation-driven operational change across Google, Photomath and Uber. This perspective informs the method's treatment of shared capability, practical authority, evidence and continuity.

Licence and citation

Original text, templates and synthetic examples are licensed under CC BY 4.0. Credit the author and source, link to the licence and indicate changes. Referenced third-party works retain their own rights. https://creativecommons.org/licenses/by/4.0/

Suggested citation: Kurek, M. (2026). Human Oversight Reliance Review: A practical method for deciding what an organisation can rely on human oversight to do. RuleBridge Advisory. https://rulebridge.eu/human-oversight-reliance-review.html

Resource: https://rulebridge.eu/human-oversight-reliance-review.html
Contact: office@rulebridge.eu

References

Foundational research

[1] Sterz, S., et al. (2024). **On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives**. ACM FAccT. DOI 10.1145/3630106.3659051. Sections on causal power, epistemic access, self-control and fitting intentions.

https://arxiv.org/html/2404.04059v1

[2] Vaccaro, M., Almaatouq, A., and Malone, T. W. (2024). **When combinations of humans and AI are useful: A systematic review and meta-analysis**. Nature Human Behaviour. DOI 10.1038/s41562-024-02024-1. Study scope and overall findings.

https://arxiv.org/abs/2405.06087

[7] Bansal, G., et al. (2019). **Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff**. Proceedings of AAAI 33(1), 2429–2437. DOI 10.1609/aaai.v33i01.33012429.

https://ojs.aaai.org/index.php/AAAI/article/view/4087

Governance and risk frameworks

[4] National Institute of Standards and Technology. **AI Risk Management Framework Playbook — MAP**. MAP 3.4 and MAP 3.5: proficiency and defining, assessing and documenting oversight processes.

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

[5] OECD (2026). **OECD Due Diligence Guidance for Responsible AI**. DOI 10.1787/41671712-en. Due-diligence steps on impacts, business relationships, prevention, tracking implementation and results, and remediation.

https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/oecd-due-diligence-guidance-for-responsible-ai_7831bb49/41671712-en.pdf

[6] Rao, A. K., et al.; NIST (2026). **Challenges to the Monitoring of Deployed AI Systems**. NIST AI 800-4. DOI 10.6028/NIST.AI.800-4. Sections 3.1.3–3.1.5 on pace, incentives and resources.

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-4.pdf

[13] OECD (2026). **The agentic AI landscape and its conceptual foundations**. OECD Artificial Intelligence Papers No. 56. DOI 10.1787/396cf758-en. Descriptive concepts and dimensions of agentic systems.

https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[14] Ministry of Digital Development and Information / IMDA, Singapore (2026). **Model AI Governance Framework for Agentic AI — official launch account**. Official launch account, 22 January 2026; organisational responsibility, action boundaries and human oversight.

https://www.mddi.gov.sg/newsroom/singapore-launches-new-model-ai-governance-framework-for-agentic-ai--/

[15] European Parliament and Council (2024). **Regulation (EU) 2024/1689 (Artificial Intelligence Act)**. Articles 14 and 26(2)–(3), as displayed in the European Commission AI Act Service Desk (original 2024 text). Context only: check current consolidated law, legal role, classification and applicable dates for the actual deployment.

https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-14

Article 26: https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-26

Adjacent methodological work

[3] Gaube, S., et al. (2026). **Keeping an Eye on AI: A Framework for Effective Human Oversight of AI Systems**. Preprint, arXiv:2605.16278v1. Sections 3–4; documentation template, printed pp. 10–11.
https://arxiv.org/pdf/2605.16278v1

[9] Porter, Z., Habli, I., McDermid, J., and Kaas, M. (2023; inspected preprint). **A principles-based ethics assurance argument pattern for AI and autonomous systems**. AI and Ethics. DOI 10.1007/s43681-023-00297-2; preprint arXiv:2203.15370. PRAISE and structured, defeasible argument.
https://arxiv.org/pdf/2203.15370

[10] Leveson, N. G., and Thomas, J. P. (2018). **STPA Handbook**. MIT. Introduction and unsafe-control-action analysis.
https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

[16] Zhu, L., Lu, Q., Ding, M., Lee, S. U., and Wang, C. (2026). **Designing meaningful human oversight in AI**. AI and Ethics. DOI 10.1007/s43681-026-01147-7. Operational and evaluative agency and oversight patterns.
https://link.springer.com/article/10.1007/s43681-026-01147-7

Practice references

[8] Sense about Science (2024). **Responsible Handover: How to share and implement data analytics and AI tools**. Implementation guide; sections on operating requirements, receiving-party requirements, transfer of context and acceptance.
https://senseaboutscience.org/wp-content/uploads/2024/11/Responsible-Handover-implementation-guide.pdf

[11] Vanynh, V. (2026). **Designing Human Control for AI-Assisted Work**. Self-published practitioner report; especially printed pp. 15–17.
https://vanynh.com/wp-content/uploads/2026/06/Designing_Human_Control_for_AI_Assisted_Work.pdf

[12] Policani, M. (2026). **Oversight Capacity, Not Model Capability, Is the Ceiling on AI Scale**. Self-published white paper; especially printed p. 5.
https://policani.net/governance/whitepapers/marco-policani-oversight-capacity-ai-scale.pdf

[17] Staggers, K. (2026). **Human Approval Gate Kit**. Public repository, practitioner guide and worksheet.
https://github.com/Kcstaggers/human-approval-gate