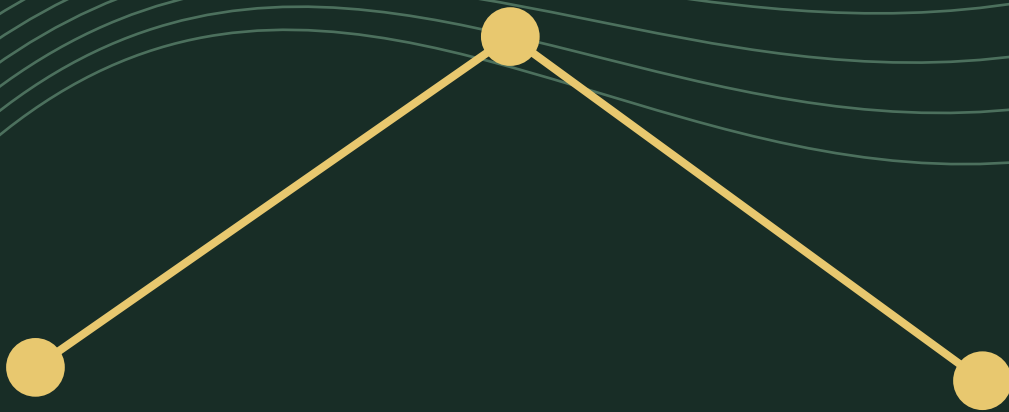


THE FIELD GUIDE / HUMAN-AI WORKBENCH

WHO SHOULD *DO WHAT?*

An atlas of delegation, uncertainty and human authority



18 PAGES / AN INDEPENDENT EDUCATIONAL PUBLICATION

Sahil Tayal · 28 October 2026 · CC BY 4.0

ORIENTATION

Read this before delegating

You will work with a tiny, deliberately inspectable model. It teaches conceptual distinctions; it is not a diagnostic, hiring tool or estimate of real-world error.

THE WORKING QUESTION

When is an AI useful as a drafter, when is independent verification necessary, and when must the consequential judgement stay with a responsible human? This question has no universal answer. Task stakes, evidence, capacity, regulation and reversibility matter.

THREE LEARNING OUTCOMES

1. Separate task performance from decision authority. 2. Observe how one change in uncertainty alters a proposed handoff while other inputs remain fixed. 3. Identify the hidden assumption that an announced review step will actually detect errors.

THE STRONGEST RIVAL VIEW

Well-evaluated automation can outperform unaided people on bounded tasks. Requiring constant human intervention can waste attention and create a rubber stamp rather than a safeguard. The correct response is not to worship either human judgement or automation: test workflow-specific results and name who is accountable.

A USEFUL READING DISCIPLINE

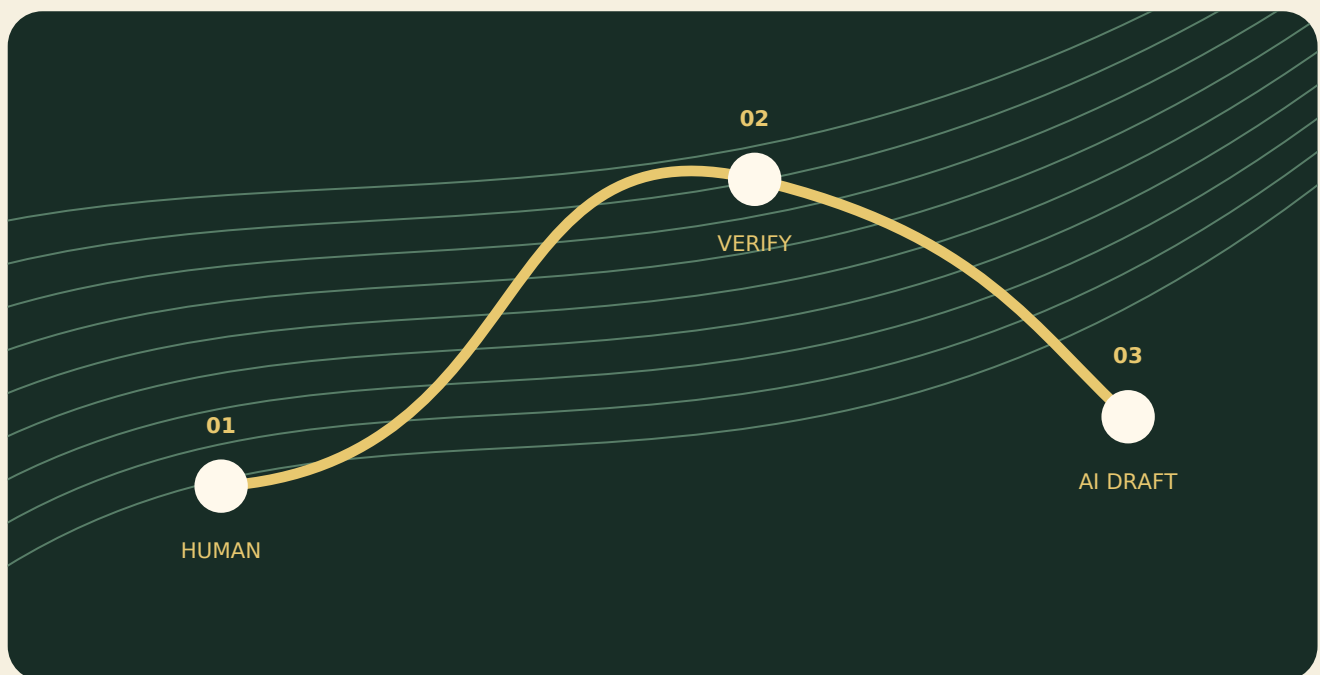
Each page distinguishes DOCUMENTED EVIDENCE, INTERPRETATION and ILLUSTRATIVE MODEL. Observations in journal articles are contextual rather than guarantees for your workplace. The cases invented for learning are explicitly labelled FICTIONAL; they are not field observations.

START HERE → Set A=1, B=0, C=1 on the website. Then change B to 2 without moving the other two controls. Explain exactly why the route changes.

CONCEPTUAL TERRAIN

A map is not a mandate

The three checkpoints explain different jobs: producing a candidate, testing a claim and taking authority for a consequential decision.

**HUMAN: MAKE A DECISION ACCOUNTABLE**

Someone must identify the task, authorise use of information, define unacceptable outcomes, choose who can override the system and explain the process to affected people.

VERIFICATION: COMPARE WITH SOMETHING INDEPENDENT

An AI answer should not be its own evidence. Strong checks can use a source document, fresh records, a specialist or a repeatable test. The reviewer must be able to reject the proposed output.

AI: GENERATE A CANDIDATE

For some tasks, systems can accelerate drafting, reformatting and option generation. A useful output is not equivalent to proof of correctness, authority or fairness.

WORKING VOCABULARY

Name the dimensions

The five most frequently collapsed ideas in human-AI allocation deserve separate words.

TASK SENSITIVITY

The seriousness and reversibility of a mistake. A typo in a disposable draft is different from an incorrect decision affecting rights, money or safety.

MODEL UNCERTAINTY

How much credible task-specific evidence is missing about system fit. It is not a magic confidence percentage displayed by a chatbot.

REVIEW DEPTH

How much accountable human checking is planned. Here: none, sampled or full. It does not measure whether the checking really catches errors.

DELEGATION ROUTE

The proposed order of drafting, verifying, piloting and deciding. A route is a governance choice, not a claim that AI possesses agency or responsibility.

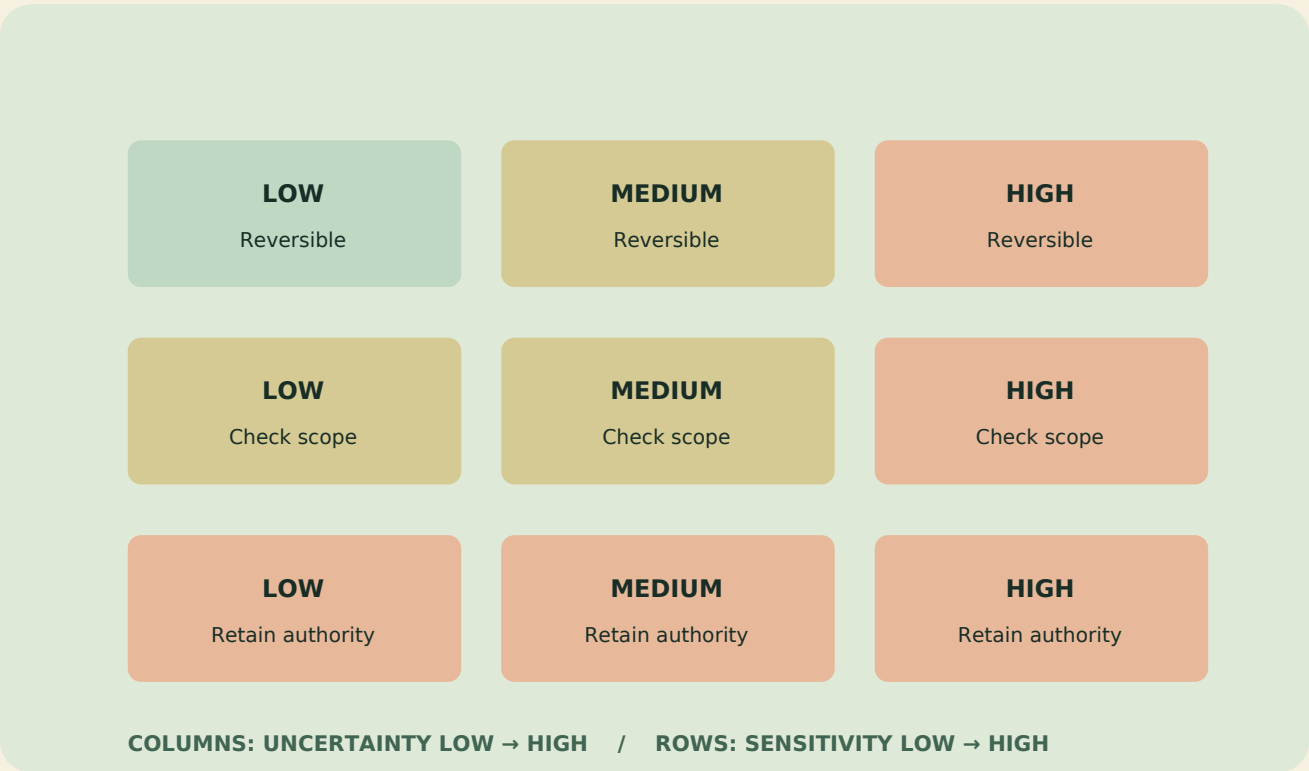
COMPLEMENTARY PERFORMANCE

A human-AI combination that outperforms each partner alone on the same evaluation. Achieving it requires more than an appealing explanation.

TWO-DIMENSIONAL THINKING

Stakes are not certainty

An AI can be familiar and still inappropriate for a high-impact final judgement. It can also be uncertain but acceptable for reversible brainstorming.



HOW TO INTERPRET THE MATRIX

These coloured cells are reminders to discuss the two axes separately, not classified risk levels. A high-stakes decision retains accountable human authority even when the task is familiar. A low-stakes draft with novel content can be easy to reverse while demanding extra verification of the draft.

A COMPLICATING VARIABLE

The availability of genuinely independent evidence changes the quality of review. A rushed person copying the model’s confidence is not equivalent to a skilled reviewer checking a current primary document. Our simulator exposes the plan; it cannot certify the quality.

DOCUMENTED EVIDENCE / 01

Customer support at scale

Published evidence: Brynjolfsson, Li and Raymond (2025), The Quarterly Journal of Economics. DOI 10.1093/qje/qjae044.

WHAT WAS STUDIED

The researchers analysed a staggered introduction of a generative-AI assistant among 5,172 customer-support agents. Their outcome included issues resolved per hour. The paper reports an average productivity increase of about 15% with considerable differences across workers.

WHAT THE RESULT SUPPORTS

For tasks resembling the studied customer-support setting, an assistant can help some workers resolve issues more effectively. The paper also reports heterogeneity: less experienced or lower-skilled workers saw larger improvements than the most experienced and highly skilled workers, who had small speed gains and small quality declines.

WHAT IT DOES NOT SUPPORT

This does not mean every organisation will observe the same improvement, or that a system should make consequential decisions unreviewed. The paper measures outcomes under a particular deployment, set of users and support-task context. Customer satisfaction, policy compliance and the value of human oversight are separate design questions.

TESTABLE HANDOFF QUESTION

Where can the assistant produce useful candidate responses while the agent retains power to detect wrong policy information and escalate exceptional cases? What quality signal would make the intervention more than a speed optimisation?

COUNTER-READING → Even a faster average response process may mask specific failure modes or unequal consequences. Examine subgroup outcomes and policy compliance before translating the result into an authority rule.

DOCUMENTED EVIDENCE / 02

The jagged frontier

Published evidence: Dell’Acqua and colleagues (2026), Organization Science, 37(2), 403–423. DOI 10.1287/orsc.2025.21838.

WHAT WAS STUDIED

In a preregistered field experiment involving 758 knowledge workers and realistic consulting tasks, investigators compared work with and without GPT-4 access. Tasks varied in their fit with the system’s demonstrated capabilities. The published paper describes both productivity and quality improvements on certain tasks and lower correctness on a complex managerial task outside the frontier.

THE POINT OF THE FRONTIER METAPHOR

AI performance is not arranged as one neat curve of easy to hard tasks. A model may handle a complicated-looking creative task but struggle on a superficially similar analytic requirement. Past success in one subtask is not enough to assume competence in another. That is why the simulator asks about uncertainty as well as stakes.

AN ALTERNATIVE INTERPRETATION

Some observed failures could become rarer with improved models, better tool access, data grounding or revised workflow design. The experiment does not establish a timeless limit for all AI. It does demonstrate why local evaluation should precede an unqualified transfer of authority.

A PRACTICAL CONSEQUENCE

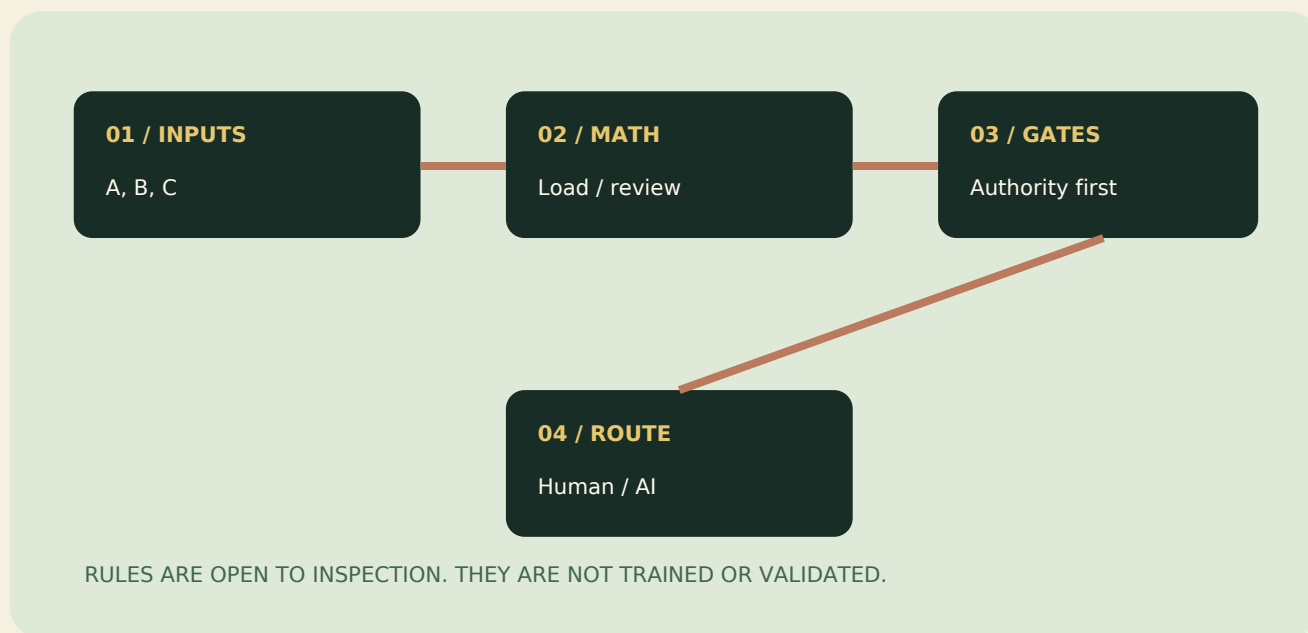
For a new task, test failure cases, not only impressive successes. Compare the candidate against primary materials, sample exceptional cases and plan how a reviewer can stop or reverse the action.

DO NOT CONFUSE → Productivity gain, correctness, accountability and user welfare are related but not interchangeable outcome measures.

SIMULATOR SPECIFICATION

The rules, in full

These numbers make a decision policy inspectable. They are not derived from a dataset and cannot estimate real harm.



EXACT VARIABLES AND EQUATIONS

A: sensitivity = 0 low, 1 medium, 2 high. B: uncertainty = 0 low, 1 medium, 2 high. C: review depth = 0 none, 1 sample, 2 full. Context load $L = 2A + 2B$. Review credit $V = 2C$. Remainder $R = \max(0, L - V)$. All levels are ordinal teaching labels.

PRIORITY OF DECISION GATES

1. If $A = 2$, choose HUMAN DECISION. 2. Otherwise if $B = 2$, choose VERIFY FIRST. 3. Otherwise if $C = 0$, choose REVIEW GAP. 4. Otherwise if R is at least 2, choose CONTROLLED PILOT. 5. In other cases, choose ASSISTED DRAFT. Earlier gates override later ones.

NO HIDDEN STATE

Every input produces one visible route with the full arithmetic and gate trace. Reset returns to $A=1, B=0, C=1$. Undo restores a prior valid state. Invalid integers or levels outside 0-2 return an explicit error instead of a fabricated result.

WORKED TRANSCRIPT

One change, different route

Keep A and C fixed so that only uncertainty B can explain the change.

BASELINE / 01

INPUTS A=1 B=0 C=1

$L=2, V=2, R=\max(0, 2-2)=0$

ASSISTED DRAFT

Familiar, medium-sensitivity writing with sample independent review. Human verification comes before use. Nothing here says the AI has decision authority.

PERTURBATION / 02

INPUTS A=1 B=2 C=1

$L=6, V=2, R=\max(0, 6-2)=4$

VERIFY FIRST

Only B rises from 0 to 2. High uncertainty triggers an earlier gate: establish credible task-fit evidence first. The model does not claim that changing B raises a measured error rate.

WHAT THE COMPARISON ISOLATES

Because sensitivity and review are fixed, the route change follows the uncertainty gate. This is an internal logic test of the simulator, not proof that the resulting recommendation is optimal in a real workplace.

COUNTEREXAMPLE & ERRORS

When zero is not permission

A small remainder cannot erase a consequential decision's need for accountable authority.

THE ADVERSARIAL PRESET

Set $A=2$ (high sensitivity), $B=0$ (low uncertainty), $C=2$ (full review). The equation gives $L=4$, $V=4$ and $R=0$. Yet the first policy gate selects HUMAN DECISION because sensitivity is high. Imagine a high-stakes benefits eligibility decision. A fluent draft and an excellent checking plan do not automatically transfer final authority to the generator.

WHY THIS COUNTEREXAMPLE MATTERS

A scalar or score can hide qualitative constraints. A team may be tempted to interpret “zero remainder” as equivalent to safety. That inference is invalid even within this toy model. There is a prior design commitment: somebody must own consequential decisions. The example shows that explicit governance gates can be more important than aggregate arithmetic.

INVALID AND EMPTY INPUT BEHAVIOR

The web interface offers only valid values 0, 1 and 2. Its “Test invalid input” button deliberately submits $A=7$ to the deterministic function, which rejects the state and explains the accepted range. The control does not silently repair or misclassify a bad input. The Reset button restores the literal baseline values and clears the undo stack.

LIMITATIONS

The route is not a recommendation for legal, medical, financial, security or personnel decisions. It does not measure model performance, review skill, regulatory duties or welfare outcomes. These factors require domain-specific evidence and human deliberation.

REFLECTION → A low remainder is at most a statement about the toy arithmetic. Ask: what constraint is outside the score, and who can challenge the rule?

EVIDENCE & DOUBT

Can explanations mislead?

Reference: Bansal and colleagues (2021), CHI, DOI 10.1145/3411764.3445717.

THE EASY ASSUMPTION

It seems natural to think that showing an AI system's explanation helps a human catch mistakes. Some explanations do assist understanding. However, explaining a recommendation is not the same as providing independent proof that it is correct.

THE COUNTEREVIDENCE

In mixed-method experiments examining complementary human-AI team performance, Bansal and colleagues found that explanations did not necessarily improve the team beyond what the members could accomplish separately. Their study observed increased acceptance of AI recommendations regardless of correctness in some conditions. This is not proof that all explanations are bad: it is a challenge to the assumption that transparency always creates calibrated trust.

COMPARE TWO REVIEW DESIGNS

DESIGN A: A reviewer checks the generated statement against a current official policy document, without treating the AI narrative as evidence. DESIGN B: A reviewer reads the model's confident justification and decides it seems plausible. Both might be called "review," but only the first includes a deliberately independent reference.

WHAT WOULD CHANGE YOUR MIND?

One could test blind review versus explanation-assisted review on representative cases with known errors, recording when reviewers catch mistakes and when they accept a persuasive false claim. The results should be analysed for task type, reviewer experience, time pressure and possible harms.

THE LESSON → A review step is a process that must be evaluated, not a label that guarantees a result.

MINIATURE CASEBOOK

Ten questions for the handoff

All ten scenarios are FICTIONAL teaching examples. None describes a real client, worker or experiment.

01 / Travel itinerary

Low stakes; unfamiliar city.

AI drafts options; human confirms hours.

Alternate: Fresh local schedules may still fail.

02 / Public FAQ

Medium policy stakes; familiar format.

AI drafts; compare the current policy.

Alternate: A stale source changes the route.

03 / Benefits denial

High stakes; tested template.

Human retains authority.

Alternate: A strong template does not delegate rights.

04 / Research outline

Low stakes; uncertain evidence.

Brainstorm; cite the primary paper.

Alternate: A plausible outline may distort claims.

05 / Data cleaning

Medium stakes; known schema.

Run checks before bulk updates.

Alternate: Edge cases may delete valid values.

06 / Accessibility text

Medium impact; mixed reliability.

Human editor tests with users.

Alternate: Automated labels can miss context.

07 / Hiring shortlist

High stakes; scoring tool.

Do not outsource accountability.

Alternate: Fairness cannot be inferred from fluency.

08 / Marketing slogan

Low stakes; easy to reverse.

Draft and edit creatively.

Alternate: Originality and consent still matter.

09 / Incident report

High stakes; time pressure.

Preserve evidence; human approval.

Alternate: Speed is not enough for accuracy.

10 / Code refactor

Medium stakes; tested project.

AI proposes; tests and review gate.

Alternate: Passing tests may miss hidden intent.

PRINTABLE WORKSHEET

A handoff audit

Complete with pencil or in a separate file. No visitor information is uploaded by the web exhibit.

NAME THE TASK

What is the precise deliverable, and which steps can be separated?

DEFINE CONSEQUENCE

Who could be harmed by a mistake, and can the effect be reversed?

RATE THE THREE INPUTS

A sensitivity (0-2): ____ B uncertainty (0-2): ____ C review (0-2): ____

EVIDENCE FOR B

What direct task-specific evaluation, if any, justifies the uncertainty level?

REVIEW SOURCE

What independent material would the reviewer consult?

AUTHORITY AND ESCALATION

Who can reject the AI draft, stop the workflow and answer for the decision?

EXAMPLE ANSWER

A policy update notice

A fictional, filled-in example to compare with your own worksheet. The facts below are invented for teaching.

TASK AND CONSEQUENCE

A customer team must send a changed refund-policy explanation. The AI drafts neutral language, but an incorrect clause could mislead customers and create extra financial costs. The team has a primary, current policy document and a trained supervisor.

THREE INPUTS

Choose A=1: medium sensitivity because it is policy-facing. Choose B=0 only because the hypothetical team has tested similar drafts and can compare the model's output to the policy. Choose C=1: the supervisor spot-checks a sample and escalates ambiguous changes. The resulting route is ASSISTED DRAFT with L=2, V=2, R=0.

STRESS THE ASSUMPTION

Suppose the policy document is missing and the assistant has never seen this type of refund rule. Uncertainty B should rise to 2. Without changing A or C, the route becomes VERIFY FIRST. The team would need a current policy source and a test of representative cases before using the draft.

WHAT REMAINS UNRESOLVED

Sample checks might miss an error sent to a customer whose case was not sampled. The organisation could consider stronger review, targeted testing of high-impact categories and a transparent correction process. The toy model cannot tell which arrangement is sufficient.

YOUR TURN → Substitute a task from your own learning or professional context. Keep confidential information out of any shared explanation.

THE SERIOUS DISAGREEMENT

Against oversight theatre

A responsible critical perspective asks whether a mandated human click can make a system less safe by creating an illusion of control.

ARGUMENT FOR MORE AUTOMATION

In some bounded workflows, properly tested systems may be faster, more consistent or less error-prone than an unaided person. Human checks can be expensive and susceptible to fatigue. A human reviewing every trivial formatting suggestion may have less attention for the rare exception that truly matters. These objections justify empirical comparisons.

ARGUMENT FOR DURABLE HUMAN AUTHORITY

Performance on a benchmark is not authority to impose consequences. Organisations must also consider contestability, due process, bias, privacy and governance duties. A person affected by a decision may need a route to ask for evidence and correction. Automated systems should not become an excuse for unaccountable institutional choices.

WHAT THE WORKBENCH CHOOSES

Our model resolves the conflict by prioritising sensitivity and uncertainty gates, followed by review and a toy remainder. This is a normative educational choice, not a discovered universal law. It makes its values clear so they can be criticised: final human authority for high-stakes judgements, independent verification for untested tasks and escalation when review is inadequate.

A FALSIFIABLE EVALUATION

Compare two real workflows on outcomes defined beforehand: correct responses, harmful errors, time, reversibility, reviewer fatigue and ability to appeal. Include failure cases and population variation. The result may favour different mixtures of automation and human review depending on the context.

PRACTICE PROTOCOL

Before a real pilot

Inspired by risk-management principles and human-AI interface research; this is an educational checklist, not compliance or legal guidance.

1

DEFINE

Name the task, intended use, prohibited use and affected people.

2

TEST

Sample representative cases, hard negatives, new conditions and edge failures.

3

GROUND

Identify independent reference material and its version/date.

4

ASSIGN

Name who can reject, override, halt and escalate outputs.

5

RECORD

Define local metrics and corrections; avoid collecting more private data than needed.

6

MONITOR

Check whether performance drifts when tasks, interfaces or policies change.

7

CONTEST

Provide an intelligible correction path for affected people.

8

RE-EVALUATE

Retest when model, prompts, documents or risks change.

SOURCE LEDGER

Eight verified references

Metadata and linked abstracts checked on 28 October 2026 against publisher, institutional or primary research pages. Summaries are paraphrases, not quotations.

01 Brynjolfsson, E.; Li, D.; Raymond, L. (2025). Generative AI at Work. QJE. doi:10.1093/qje/qjae044 Supports: 5,172 support agents, differential productivity effects.	02 Dell’Acqua, F. et al. (2026). Navigating the Jagged Technological Frontier. Organization Science 37(2). doi:10.1287/orsc.2025.21838 Supports: Some AI-assisted tasks improved, one outside-frontier task worsened.
03 Noy, S.; Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science. doi:10.1126/science.adh2586 Supports: Specified writing experiment; not a general forecast.	04 Bansal, G. et al. (2021). Does the Whole Exceed its Parts? CHI. doi:10.1145/3411764.3445717 Supports: Explanations did not guarantee complementary performance.
05 Amershi, S. et al. (2019). Guidelines for Human-AI Interaction. CHI. doi:10.1145/3290605.3300233 Supports: Interface guidelines; not proof of task performance.	06 Parasuraman, R.; Sheridan, T.; Wickens, C. (2000). A Model for Types and Levels of Human Interaction with Automation. IEEE. doi:10.1109/3468.844354 Supports: Four automation functions and variable levels.
07 Tabassi, E. (2023). Artificial Intelligence Risk Management Framework 1.0. NIST AI 100-1. doi:10.6028/NIST.AI.100-1 Supports: Voluntary cross-sector risk-management framework.	08 Autio, C. et al. (2024). Artificial Intelligence Risk Management Framework: Generative AI Profile. NIST AI 600-1. doi:10.6028/NIST.AI.600-1 Supports: Generative-AI specific governance and risks.

BACK MATTER

What this cannot prove

This publication is meant to make responsibility visible, not to grant it.

LIMITS OF THE MODEL

The arithmetic is authored for illustration; its thresholds, route preferences and assumptions have not been validated in field trials. It cannot estimate the probability of an error, identify which occupation will disappear, declare a workflow safe or decide what law requires. Do not use it as a substitute for responsible domain expertise.

OWNERSHIP AND OPEN SHARING

© Sahil Tayal, 2026. Original authored text, diagrams, examples, worksheet, simulator explanation and publication design in this Field Guide are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). You may share and adapt them, including commercially, when you credit Sahil Tayal, link to the licence and indicate changes. The licence text is available at <https://creativecommons.org/licenses/by/4.0/>.

SUGGESTED ATTRIBUTION

Sahil Tayal — Human-AI Workbench: The Field Guide, <https://sahiltayal.com>, CC BY 4.0. If you adapt it, identify the changes. Third-party academic publications, titles, quoted ideas and external reference material are not relicensed by this publication.

CONTINUE THE INVESTIGATION

The natural next experiment would observe real reviewers on bounded scenarios, recording both catches and misses while protecting participants. Such a study needs independent ethics, appropriate sampling, clear metrics and consent. Nothing in this guide claims that experiment has happened.

EXPLORE → <https://sahiltayal.com/> WRITE → sahil@sahiltayal.com THE SITE → humanaiwork.sahilwho.online (PLANNED, NOT VERIFIED LIVE)