FREE EVIDENCE BRIEFING · HR & INTERNAL PROPOSAL AUTHORS

Who Gets to Say It's Good?

Recognising qualitative fit is not the same job as making the artefact.

A person may know what to preserve, what to change and when a result has gone too far—even without knowing Photoshop. AI can assist production. Neither software fluency nor an impressive output, by itself, settles who is qualified to accept it.

Editorial argument. Treat production expertise and evaluative judgment as potentially distinct, overlapping capabilities. Establish the relevant domain knowledge, the criteria and the authority to decide; do not infer competence from a job title or tool list alone.

WHAT THIS PAPER ESTABLISHES—AND DOES NOT

Project perspective. We regard common sense as innate human sense, not something an exercise manufactures. This is a philosophical position, not proof that everyone is born with fully formed judgment. The studies inform our argument; they do not establish innateness or validate a hiring test for ‘taste’. [1–5]

Proposed practice. Make qualified acceptance a named responsibility. Pilot structured review exercises, retain human accountability and evaluate fairness before using results in hiring. The governance matrix on page 5 is a discussion aid, not an automated decision rule.

Reading path: distinguish roles → weigh evidence → calibrate review → assign authority. Source URLs and limitations appear on page 6. No measured business benefit, cost reduction or universal claim about AI incapacity is made.

01 / EVALUATING VERSUS PRODUCING

Different work. Shared responsibility.

Producing means executing a design, analysis, film, lesson or other artefact. **Evaluating** means judging its fit for a purpose, people and constraints. **Accepting** adds accountable permission to use it. One person can perform all three, but the roles should not be silently conflated.

Interpretation. An editor can recognise broken pacing without operating the edit suite; a teacher can judge an explanation without animating it; a clinician can assess a clinical claim without writing the code. Exercises apply human sense to decisions; they do not manufacture it. These role illustrations do not prove anyone is qualified. Domain experience, source knowledge and demonstrated judgment still matter.

PUBLISHED RESEARCH / SCOPE MATTERS

[1] Jacobi et al. (2024): observational study. In a four-country survey of 3,092 biologists and physicists, personality traits were associated with aesthetic engagement and sensitivity in scientific work. This supports taking aesthetic experience seriously outside the arts. Association is not causation; self-reported engagement is not a hiring test, and the study does not establish innate, fixed reviewer ability.

[2] Sackett et al. (2023): research synthesis and applied implications. Revised selection-validity estimates challenge earlier overcorrections and emphasise variability and validity–diversity trade-offs. The authors' discussion supports careful, job-relevant selection design—not dismissing experience or selecting on an unvalidated 'eye for quality' exercise.

[3] Doshi & Hauser (2024): experiment. Access to AI story ideas improved evaluated short-story creativity and quality, especially among less creative writers, while AI-assisted stories became more similar to one another. Evidence for benefit and a diversity trade-off in that task is not evidence that every discipline needs the same reviewer or workflow.

Proposal for an internal business case. Separate the production deliverable from the acceptance obligation. Specify why this domain needs a reviewer, what they must demonstrate and how disagreement will be handled. Measure local outcomes before promising savings or improved selection.

02 / A CORRECTION CAN MISS IN THE OTHER DIRECTION

Not ‘the opposite’. The right amount.

A customer says the steak is undercooked. It returns burnt.

Constructed illustration. This is not a report of a restaurant incident. It shows why reversing a qualitative criticism is not the same as finding the intended middle. The desired correction may be small; the original strengths and other constraints must survive.

Overcorrection overshoots the requested adjustment. **Regression** breaks something previously working. **Drift** changes work outside the authorised scope. **Repetition without progress** leaves the failed state unchanged despite equivalent attempts. These mechanisms can overlap, but none proves the others or identifies a technical root cause.

Interpretation. Recognising ‘too far the other way’ requires criteria and a qualified sense of the target. The right evaluator might be an editor, clinician, sound director, designer or teacher—not necessarily the engineer operating the agents. A qualified reviewer also needs authority to stop or withhold approval.

PROPOSED CORRECTIVE PRACTICE / NOT VALIDATED

Before another attempt, name **what must remain**, **the specific change** and **the evidence that would justify acceptance**. If direction is unresolved, HOLD and reframe. If the change is clear, request a bounded revision. Do not equate producing a new version with successful correction.

Use independent criteria for a second review, not merely a second agent repeating the same instruction. Compare observed results and unresolved constraints. Repeated attempts warrant a human no-progress review; an attempt count alone should not impose a verdict.

Extra AI calls, production rework, specialist review and human cognitive effort are separate cost categories. None is measured here. Keep unknown quantities unknown; do not treat time spent supervising an unsuccessful loop as successful AI execution.

03 / AI ASSISTANCE IS NOT INDEPENDENT PROOF

A second opinion is only as directed as its criteria.

[4] Kranzle & Sharratt (2025): creative-evaluation study. The institutional abstract reports moderate alignment with human experts when GPT-4 used a base framework and spectrum-based few-shot prompting. The practical question is how evaluation is designed, not whether machines can ever evaluate. Moderate task-specific alignment is not universal reliability or qualification for clinical or hiring decisions.

[5] Maier et al. (CHI 2026): controlled experiments. The conference abstract reports a study of 486 participants and a separate validation context of 640. Model-led rewriting improved idea quality but reduced diversity and perceived ownership; reflective human-led modes improved quality while preserving them. This concerns co-creation interface design, not proof that a second AI judge can certify an output.

Interpretation. AI review can help locate contradictions, compare criteria or propose alternatives. Shared assumptions, biased examples and the same missing evidence can also survive a second agent. Treat its conclusions as review inputs. Test its task-specific usefulness against qualified assessments and retained source evidence.

PROPOSED SELECTION EXERCISE / PILOT BEFORE USE

Show candidates a relevant near-miss and an overcorrection. Ask what to preserve, what to change, what evidence is missing and when they would hold. Use a consistent rubric, more than one qualified evaluator and recorded reasons. This is a proposed audition, not a validated predictor of job performance. [2]

Offer accessible formats, reasonable accommodations and clear instructions; avoid judging familiarity with a particular production tool unless that skill is genuinely required. Pilot evaluator agreement, job relevance and group outcomes. Check applicable law, invite challenge and keep hiring decisions human-accountable.

Do not replace résumé review with unexplained ‘taste’ rankings. Recognising evaluative talent may broaden who is considered, but fairness and predictive usefulness must be demonstrated rather than asserted.

04 / PROPOSED GOVERNANCE DECISION MATRIX

Authority follows the stakes.

Use this discussion aid to assign review responsibility. It is neither a validated selection instrument nor legal or clinical advice. All actions below are proposed practice, informed by the NIST Generative AI Profile's risk-management orientation. [6]

Situation	Review & evidence	Accountable action
Low-stakes internal draft	Purpose-aware editor checks relevance and stated constraints.	Revise or accept for the limited use; identify the owner.
Public creative output	Qualified domain reviewer checks fit, originality, rights and audience impact.	Accept only against named criteria; hold unresolved rights or material concerns.
Hiring or employment decision	HR and relevant specialists assess job relevance, evidence, fairness and recourse.	Human decision owner retains accountability; pilot before using a new exercise.
Clinical or other high-stakes claim	Appropriately qualified specialist checks critical sources and intended use.	Hold unverified claims; source attestation is not independent proof.
No progress or an overshoot	Reviewer compares changed and unchanged conditions; distinguishes failure mechanisms.	Hold/reframe or name a bounded revision; never infer a veto solely from attempt count.

Acceptance record. Name the purpose, stakeholders, criteria, source checks, unresolved risks, reviewer and decision owner. Retain the reasons for cleared concerns. Provide a route for challenge and define what evidence would reopen the decision.

The site's Decision Gate records a person's reasoning. Its source checks and clearances are self-reported; its editable history lasts only in the unsaved session. It does not independently verify evidence, detect overshoots or decide who is qualified.

SOURCES / EVIDENCE CUTOFF 13 NOVEMBER 2026

Trace the claim to its source.

[1] Jacobi, C., Varga, P. J., Jessani, Z., & Vaidyanathan, B. (2024). Individual differences in scientists' aesthetic disposition, aesthetic experiences, and aesthetic sensitivity in scientific work. Frontiers in Psychology, 14. Observational research.
https://doi.org/10.3389/fpsyg.2023.1197870

[2] Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. Industrial and Organizational Psychology, 16, 283–300. Research synthesis / applied implications.
https://doi.org/10.1017/iop.2023.24

[3] Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10, eadn5290. Experimental research.
https://doi.org/10.1126/sciadv.adn5290

[4] Kranzle, T., & Sharratt, K. (2025). Evaluating Creative Output With Generative Artificial Intelligence: Comparing GPT Models and Human Experts in Idea Evaluation. Creativity and Innovation Management, 34, 991–1012. Creative-evaluation research.
https://doi.org/10.1111/caim.70007

[5] Maier, S., Schneider, M., & Feuerriegel, S. (2026). Partnering with Generative AI: Experimental Evaluation of Model-Led and Human-Led Interaction in Human-AI Co-Creation. CHI '26. Controlled experimental research; conference abstract checked.
https://doi.org/10.1145/3772318.3791185

[6] NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Voluntary guidance, not a causal study. This is the Generative AI Profile, not the base AI RMF 1.0.
https://doi.org/10.6028/NIST.AI.600-1

Review limits. This is an editorial synthesis, not a systematic review. Publisher/institutional abstracts and metadata were checked; the Cambridge article and NIST profile were also consulted. The six sources do not directly validate the proposed role distinction, audition or matrix. No independent clinical, legal, hiring-outcome or cost evaluation was conducted.

Evidence key. Research = a scoped published finding. Interpretation = this paper's reasoning. Illustration = a constructed example. Proposal = practice requiring local validation. No private case identifiers or partner identities are disclosed.