The Right to Stop

Evidence paper · Free complete edition

Human authority is a working control.

An AI workflow can produce many artifacts without producing an accepted result. The right to interrupt matters when another correction repeats a failed method, damages a preserved requirement, or consumes review effort without observable progress. This paper separates that authority from claims about what a form, a vendor or an AI agent can enforce.

Recognition of qualitative fit can differ from manual production expertise. A person may calibrate a vision without knowing Photoshop. Our position is that common sense is innate: a philosophy, not proof of fully formed judgment in everyone. Exercises apply human sense, not manufacture it. AI is trained on data; technical expertise remains valuable.

For HR and internal proposal authors

Use this paper to define decision rights, identify a qualified reviewer and plan a small, observable trial. Do not use it as a validated selection test, a financial forecast or a legal compliance certificate. The proposed stop/reframe protocol is an editorial synthesis; it has not been validated by these studies.

Reading the evidence

Research findings [1–5], voluntary governance [6], director-reported observations and constructed illustrations are distinct. No case-level savings, causal diagnosis or measured cognitive benefit is established here. Edition 2.0 · November 13, 2026.

Who can judge the target?

Research and bounded interpretation

Separate the operator from the domain arbiter

The operator runs the production system and diagnoses execution constraints. The domain arbiter judges whether the result fits its purpose: a chef, editor, art director, clinician or other qualified specialist. One person may hold both roles, but software fluency alone does not establish the relevant judgment. The accountable owner authorizes delegation and accepts consequences.

What the studies actually support

Jacobi et al. [1] surveyed 3,092 biology and physics scientists across four countries and found associations between personality and aesthetic engagement. This is evidence of variation in that population, not evidence that aesthetic judgment is fixed or that a software-free reviewer will outperform an operator.

Sackett et al. [2] revisit predictor validity and its variability in personnel selection. The implication for HR is caution about unvalidated proxies and generalization. A proposed near-miss/overcorrection audition would need job analysis, structured scoring, independent qualified evaluators and piloting for reliability and fairness. Hiring must remain human-accountable.

AI assistance is not excluded

Doshi and Hauser [3] found AI story ideas improved rated individual creativity while increasing similarity between stories. Kranzle and Sharratt [4] report moderate expert alignment under particular frameworks and few-shot prompting. Maier et al. [5] found collaboration mode affected idea quality, diversity and perceived ownership in experiments. None validates this paper’s stop threshold or proves AI review is impossible.

Name the failure before another fix.

Illustration versus reported observation

Constructed illustration: the steak

A customer says a steak is undercooked; it returns burnt. Nuance was missed, not solved. This illustrates overcorrection: a response overshoots the desired middle. It is invented, not a restaurant case or a measured loss. A chef would specify the calibrated target and verify it, rather than merely reverse the first complaint.

Four mechanisms, not one diagnosis

Overcorrection overshoots a target. Regression breaks something that previously worked. Drift changes scope or an approved direction. Repetition repeats equivalent failed steps with no observable progress. They may overlap, but should not be treated as one proven technical fault.

Separate director-reported chronologies

November 11, 2026: a website director challenged unrequested design drift. November 12: a separate comparison hero image appeared to have two left hands, noticed after another hero was selected. These are distinct observations, not demonstrated corrections.

November 12: a roughly 60-second animated review cut was rejected and self-rated 5.2/10 by its director, not an audience panel. Mouth movement, acting, pacing and audio were criticized. No final accepted MP4 or verified fix is established; some clips may be reusable.

November 12: a separate connector dialog remained unresolved through five user-reported failed attempts. The human directed a change of reasoning. Root cause, successful connection, time and cost are unknown. These accounts are not independently verified; no private case identities are disclosed.

The economics of another pass.

Proposed accounting · No measured savings

Count burdens, not only output

A loop can incur provider usage, tool or production charges, operator time, qualified review, rework and supervision. Keep separate currencies and billing categories; one service allowance is not another service's credit balance. Preserve reusable artifacts rather than treating every rejected deliverable as wholly wasted.

A proposed internal business case

Record the attempt, method, preserved baseline, failed criterion and observable change. Attribute each human hour once, even if that hour includes supervision and review. Leave unknown costs and hours unknown rather than substituting zero. A proposed incremental-cost ledger can add attributable charges to non-overlapping human time valued at explicitly stated internal rates. No rates or totals are supplied here.

Compare a repeated method with a genuinely different method in a bounded pilot. Measure acceptance against a preset criterion, reviewer agreement, regressions and total attributable effort. Do not infer avoided costs from a hypothetical counterfactual or claim a causal return on investment without a defensible comparison.

More agents are not automatic independence

Several agents can compound cost if they repeat the same assumptions and lack independent acceptance criteria. Independent AI checks can sometimes help, especially when criteria and evidence are inspectable; human review can also be biased. Choose the review design for the risk, not the number of agents. The cited studies do not quantify this family's failure-loop economics.

Pause. Reframe. Verify.

Proposed human-controlled protocol

1 / Set authority before delegation

Name the accountable owner, operator and qualified domain arbiter. Define preserved requirements, an acceptance criterion, permitted AI actions, an approval record and a stop mechanism. Confirm who can actually pause the vendor, queue or tool; a charter alone cannot grant that ability.

2 / Interrupt unresolved repetition

After two materially equivalent failures without observable progress, pause further equivalent attempts. This two-attempt trigger is a proposed local rule, not a research-validated optimum or NIST requirement. Stop immediately for a serious safety, privacy or irreversible-action risk. Preserve the evidence and accepted baseline.

3 / Reassign judgment and change method

Separate observed state from hypotheses. The human may override the agent, choose a qualified specialist, redefine a calibrated criterion or escalate to the party with real authority. Choose a substantively different method; a paraphrased instruction is not enough. Credit the person who identified and interrupted the failure.

4 / Resume only after observable acceptance

Test one bounded proof, record the result and reviewer decision, and check preserved requirements for regression. A film proof could test one speaking shot for lip sync, audible audio and consistent acting; this is a proposal, not a completed fix. Record approver, artifact version, criteria and date before broader spending resumes.

Governance is not certification

NIST AI 600-1 [6] addresses documented responsibilities (GOVERN 2.1), differentiated human-AI oversight (GOVERN 3.2), and deactivation/escalation procedures (MANAGE 2.4). This protocol interprets that guidance; it is not endorsed or certified by NIST. The Authority Map documents intentions, not external enforcement, informed consent or legal adequacy.

Sources and evidence limits.

References · Clickable primary source links

[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 survey; not a hiring validation.
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. Selection-validity synthesis; not an AI-arbiter audition study.
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. Online short-story experiment; context-specific.
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. Framework and prompt choices affected alignment; not universal expert equivalence.
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. Interface experiments; not proof of enterprise savings.
https://doi.org/10.1145/3772318.3791185

[6] NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). Voluntary governance guidance, companion to AI RMF 1.0; not certification or an effectiveness trial.
https://doi.org/10.6028/NIST.AI.600-1

Source scope: publisher text for [1, 2, 5], publisher abstracts/metadata for [3, 4], and the NIST profile for [6] were checked. Private canonical records could not be accessed; reconciliation is pending. Edition 2.0, November 13, 2026. No measured case costs, accepted fixes or validated hiring procedure are asserted.