

# ER-13 External Empirical Evidence Crosswalk

Draft v0.6.2 | Coordinated methodological and infrastructure revision | 10 October 2026

**NOT FINAL / NOT LOCKED** **ARCHIVE\_EFFECT:** none | **MIRROR\_ENGINE\_EFFECT:** none | **MASTER HASH**  
**MANIFEST EFFECT:** none

Domain separation retained. Human, human-automation, machine, autonomous-agent, and methodological evidence remain distinct. The study design supports a bounded empirical question; it provides no runtime authority.

## 1. Revision purpose

Draft v0.6.2 applies the v0.6.1 review corrections to item-population claims, challenge-position reporting, interrupted outcomes, retractions, timing and source precision. It integrates the proposed SEC-CDP-002 infrastructure boundary and accompanies Pilot Protocol v0.2. The earlier packages are preserved unchanged. No participant study, behavioral experiment, power simulation, deployed generator or safety benefit is claimed.

**What v0.6.2 does not do:** It does not infer strategic intent from an error pattern, does not claim that conspicuous errors prove deception, does not adopt unvalidated fixed alert thresholds, does not embed executable containment code, and does not convert a research signal into authority to execute, quarantine, or override a human.

### 1.1 Revision history

| Version | Recorded contribution                                                                                                          | Scope of lineage statement                                                                                         |
|---------|--------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| v0.2    | Domain taxonomy; machine evidence; qualified grades and review responses.                                                      | Historical material retained in sections 6, 7 and 11.                                                              |
| v0.3    | Intermediate revision in the inherited v0.2 to v0.4 lineage.                                                                   | Exact standalone change list and proposal origin are not established by the supplied record.                       |
| v0.4    | Dual-use boundaries and blinded manufactured-error vigilance study.                                                            | Contribution described by the v0.5 purpose and the supplied continuity record; no byte comparison to v0.4 claimed. |
| v0.5    | Target awareness, preregistered generator boundary, statistical planning framework.                                            | Exact supplied PDF preserved; its source reversal and design ambiguities are superseded here.                      |
| v0.6    | S14 correction; separate hypotheses; marginal estimand; paired hit/false-alarm analysis; machine evidence; minor fixes.        | Targeted corrective successor. NOT FINAL / NOT LOCKED.                                                             |
| v0.6.1  | Control content and early completion defined; model and source precision; baseline-grounded verifier; pilot companion drafted. | Targeted successor to v0.6. NOT FINAL / NOT LOCKED.                                                                |
| v0.6.2  | Methodological corrections; SEC-CDP-002 integration; Pilot v0.2 and shared verification.                                       | Coordinated successor. NOT FINAL / NOT LOCKED.                                                                     |

The v0.6 PDF and DOCX are reconstructed from the complete supplied v0.5 PDF because an editable v0.5 source was not supplied. Section numbering is retained through 12.3. Former 12.4 becomes section 13, former 12.5 becomes section 14, and former sections 13-16 become 15-18. The package contains a detailed change record and

source-check scope. Draft v0.6.1 is edited from the complete v0.6 JSON source and preserves its canonical questions and author-approved positions. A structural before/after change list is included. Draft v0.6.2 is edited from the v0.6.1 structured source. Appendix C records the current findings; Appendices A and B are historical. The combined change record supplies replayable before/after operations.

## 2. Operational meaning of 'intelligence' in this crosswalk

This empirical crosswalk does not impose one mathematical definition of intelligence. The Interrogative Conscience questions are abstract; empirical evidence is therefore mapped to explicit target classes:

| Code      | Target class                                                                      |
|-----------|-----------------------------------------------------------------------------------|
| H         | Human decision-maker or human group                                               |
| HA        | Human interacting with automation/AI                                              |
| M         | Machine model, model ensemble, or machine-evaluation panel                        |
| A         | Autonomous embodied/software agent acting in an environment                       |
| METHOD    | Decision/scientific methodology rather than a behavioral experiment               |
| NORM      | Normative/governance inference; not empirically established by the cited evidence |
| PRECEDENT | Operational/legal precedent; not empirical validation                             |

## 3. Cross-domain inference rules

- H -> HA:** Allowed only for claims about human behavior that remains present in a human-AI setting; must not be treated as a machine-architecture property.
- H -> M/A:** Not allowed as direct empirical support. Requires an explicit bridging argument and preferably independent machine-domain evidence.
- HA -> A:** Does not establish autonomous-agent behavior when the human component is absent.
- M -> A:** Machine-model evidence may support a component-level mechanism, but does not automatically establish embodied/autonomous governance behavior.
- M/A -> NORM:** Empirical machine behavior cannot itself establish legitimate moral or political authority.
- PRECEDENT -> any empirical claim:** Not allowed. Legal/operational precedents may illustrate an institutional rule but are not experiments.

## 4. Revised executive evidence matrix

| IC         | Evidence category                                             | Revised grade                                       | Core boundary                                                                                                                                                          |
|------------|---------------------------------------------------------------|-----------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| IC-EPI-002 | H empirical + A verification + METHOD; NORM boundary explicit | INDIRECT EMPIRICAL SUPPORT + METHODOLOGICAL ANALOGY | These sources do not validate the author's safety-first priority as a universal rule, and they do not establish one mandatory AGI policy for unidentified model error. |

| IC         | Evidence category                                                               | Revised grade                                                                            | Core boundary                                                                                                                                                                                                                                                                                                  |
|------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| IC-EPI-004 | DIRECT H + DIRECT M evidence for non-independence                               | DIRECT DOMAIN-SPECIFIC EMPIRICAL SUPPORT                                                 | The human and machine mechanisms are not assumed to be identical. The evidence supports checking independence/correlation, not automatic deference to a dissenter or a claim that multi-model agreement is generally worthless.                                                                                |
| IC-OBJ-005 | H empirical variation + METHOD + NORM inference                                 | NORMATIVE INFERENCE INFORMED BY EMPIRICAL VARIATION                                      | None of this empirically selects the correct scalar weights or validates the author's manufacturing priorities as universal values.                                                                                                                                                                            |
| IC-AUT-001 | DIRECT HA evidence; NORM extrapolation to autonomy                              | DIRECT HUMAN-AUTOMATION EVIDENCE / INDIRECT AUTONOMY INFERENCE                           | It does not prove that human bounded autonomy always has positive net value, that humans should always retain final authority, or that autonomous AGI must preserve every form of human discretion for instrumental reasons.                                                                                   |
| IC-AUT-005 | H analog + NORM authority gap                                                   | HUMAN ANALOG ONLY / NO MACHINE-PATERNALISM EVIDENCE                                      | No cited empirical study validates machine paternalism over humans. OSHA is explicitly removed as substantive evidence. The legitimacy threshold for AI intervention remains normative/governance-dependent.                                                                                                   |
| IC-EXT-001 | DIRECT HA + H review + DIRECT M self-evaluation evidence                        | DOMAIN-SPECIFIC OVERSIGHT AND SELF-EVALUATION EVIDENCE / LIMITED STRUCTURAL INFERENCE    | External review can itself be wrong, costly, captured, delayed, or correlated. These studies do not prove that an external human checker is always required or superior to protected internal verification.                                                                                                    |
| IC-UNR-002 | DIRECT M model-discrimination + A verification + METHOD; governance gap remains | DIRECT MACHINE EVIDENCE FOR MODEL SEPARATION / UNRESOLVED HUMAN-STAKES DECISION BOUNDARY | The v0.1 'no close machine work' statement is withdrawn. However, these works do not determine what governance/authority rule should apply when irreversible human-stakes action must be chosen before model disagreement is resolved. ISS is a preprint and includes explicit certificate-failure conditions. |
| IC-UNR-005 | H preference evidence + METHOD + NORM inference                                 | EMPIRICAL VALUE VARIATION + NORMATIVE INFERENCE                                          | Preference measurement does not establish legitimacy, universal moral weights, or a rule that every normative disagreement must remain permanently unresolved.                                                                                                                                                 |

## 5. Question-by-question evidence crosswalk

### 5.1 IC-EPI-002

**Canonical question:** How should an intelligence act when the probability of material model error cannot be reliably identified?

**Author-approved human position:** When uncertainty about a possible problem cannot be quantified reliably, I would put safety first and quality second, inspect the situation myself, and weigh the reporting person's machine-specific experience. If the employee has more relevant experience than I do, I would be more likely to involve Maintenance. If I am more experienced and my inspection supports that there is no problem, I may

authorize continued production while monitoring closely. I would also inform my supervisor and Quality about the potential issue.

**Revised evidence class:** INDIRECT EMPIRICAL SUPPORT + METHODOLOGICAL ANALOGY

**What is supported:** Hazardous-work evidence documents safety/production tradeoffs. In S14, ambiguity aversion appeared with fear of negative evaluation and disappeared when private preferences eliminated that factor in the tested design. RDM supplies a deep-uncertainty methodology; HUME supplies bounded autonomous-agent verification evidence. These findings neither prescribe a universal priority rule nor make social accountability a sufficient decision procedure.

**What is NOT established:** These sources do not validate the author's safety-first priority as a universal rule, and they do not establish one mandatory AGI policy for unidentified model error.

| Source domain and status                                                                                                   | Relevant finding                                                                                                                                                                                            | Boundary                                                                                                                                                                                                                                                                                                                       |
|----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S01 David L. McLain; Kimberly A. Jarrell (2007)<br>H   peer-reviewed field study                                           | In 239 workers in hazardous occupations, safety pressure and perceived compatibility between safety and production expectations were associated with safer work behavior.                                   | Self-report/correlational field evidence; it does not establish a universal priority rule or prove that slowing/stopping is always optimal.                                                                                                                                                                                    |
| S13 Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003)<br>METHOD   decision-methodology monograph                | Treats deep uncertainty as disagreement/ignorance about models, probability distributions, and/or outcome valuation; develops robustness-oriented policy analysis.                                          | Methodological guidance, not a human-subject experiment and not empirical proof that robustness is always normatively preferable.                                                                                                                                                                                              |
| S14 Stefan T. Trautmann; Ferdinand M. Vieider; Peter P. Wakker (2008)<br>H   peer-reviewed behavioral-economics experiment | Ambiguity aversion occurred with fear of negative evaluation and disappeared when that factor was eliminated through private preferences in the tested design.                                              | Laboratory risky/ambiguous choices; the main manipulation used private DVD preferences. This demonstrates a social contribution in that setting, not the correct action under unidentified probabilities.                                                                                                                      |
| S22 Anxing Xiao; Hanbo Zhang; Tianrun Hu; David Hsu (2026)<br>A   peer-reviewed robotics conference paper                  | HUME represents missing knowledge as uncertain hypotheses, jointly plans task actions and verification actions, and updates its model from new observations in simulated and real-world robot environments. | Direct autonomous-agent evidence for hypothesis verification under incomplete models, not a study of human governance, normative authority, or irreversible human-stakes decisions. Foundation models provide verification feedback and may share confounders with the proposer; verification does not establish independence. |

## 5.2 IC-EPI-004

**Canonical question:** When can agreement among multiple models conceal a shared omission rather than provide independent confirmation?

**Author-approved human position:** If several people agree because they are relying on the same checks, but an experienced person with specific knowledge of the machine identifies a possible failure those checks may miss, I would stop production, involve my supervisor, and have Maintenance inspect the machine before continuing.

**Revised evidence class:** DIRECT DOMAIN-SPECIFIC EMPIRICAL SUPPORT

**What is supported:** Human hidden-profile/conformity studies show shared information can suppress unique evidence. Separately, machine-domain studies directly show correlated LLM errors and shared confounders that undermine naive ensemble voting.

**What is NOT established:** The human and machine mechanisms are not assumed to be identical. The evidence supports checking independence/correlation, not automatic deference to a dissenter or a claim that multi-model agreement is generally worthless.

| Source domain and status                                                                                                            | Relevant finding                                                                                                                                                                                                                                                                                                                | Boundary                                                                                                                                                                                                                                                                                                                                       |
|-------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S02 Garold Stasser; William Titus (1985)<br>H   peer-reviewed controlled experiment                                                 | Groups often discussed common information more than unique information and could fail to discover the superior option when critical information was distributed.                                                                                                                                                                | Human laboratory task; no direct inference to machine ensembles without machine-domain evidence.                                                                                                                                                                                                                                               |
| S03 Vernon L. Allen; John M. Levine (1971)<br>H   peer-reviewed controlled experiment                                               | Independent social support reduced conformity relative to a unanimous group; valid independent assessment was more effective than invalid support.                                                                                                                                                                              | Human social-conformity evidence; algorithms do not require social-shame or conformity mechanisms for correlated error.                                                                                                                                                                                                                        |
| S18 Li Lu; Y. Connie Yuan; Poppy Lauretta McLeod (2012 (online 2011))<br>H   peer-reviewed meta-analysis                            | Across 65 studies and 3,189 groups, hidden-profile groups were eight times less likely to find the solution than full-information groups; common information dominated discussion. Effects were moderated by group size, information load, proportion of unique information, task demonstrability, and hidden-profile strength. | Human-group evidence only. Moderators matter, and the result does not imply that machine ensembles fail for social-psychological reasons.                                                                                                                                                                                                      |
| S19 Elliot Myunghoon Kim; Avi Garg; Kenny Peng; Nikhil Garg (2025)<br>M   peer-reviewed ICML paper                                  | Large-scale evaluation of more than 350 LLMs found substantial error correlation; on one leaderboard, models agreed 60% of the time when both were wrong. Shared architecture/provider contributed, but high-performing models also showed correlated errors across providers.                                                  | Direct machine evidence for correlated error, not proof that a particular human dissenter is more reliable or that every multi-model consensus is unsafe. The 60% is raw wrong-answer agreement, not excess above chance; the appropriate baseline depends on item options and wrong-answer distributions. See the chance-baseline note below. |
| S20 Jitian Zhao; Changho Shin; Tzu-Heng Huang; Satya Sai Srinath Namburi Gnvv; Frederic Sala (2026)<br>M   peer-reviewed ICML paper | Shows that LLM-judge ensembles can share confounded errors so majority vote/averaging can provide little gain or amplify systematic mistakes; confounder-aware aggregation improved accuracy across 12 benchmarks.                                                                                                              | Evaluation-ensemble setting, not a general theorem about all AI systems or a proof that human oversight is superior.                                                                                                                                                                                                                           |

**S19 chance-baseline note:** For item  $q$ , let  $a_{qr}$  and  $b_{qr}$  be two models' conditional probabilities of selecting wrong option  $r$  given an error. Under independent wrong-answer selection, expected same-wrong-answer agreement is  $\sum_r a_{qr} b_{qr}$ . Under uniform selection over  $K_q - 1$  wrong options it is  $1/(K_q - 1)$ ; nonuniform preferences and varying option counts require an item-weighted baseline. A reported raw 60% cannot alone quantify excess agreement. This formula is interpretive context, not a reanalysis of S19 or an asserted baseline for its dataset.

### 5.3 IC-OBJ-005

**Canonical question:** How should partially conflicting objectives be compared when no authorized scalar weighting rule exists?

**Author-approved human position:** When production speed conflicts with quality margin and process control and there is no exact weighting rule, I would continue production cautiously but slow it somewhat to regain better control. I would actively monitor for scrap and developing problems and continue only while the process remains safe.

**Revised evidence class:** NORMATIVE INFERENCE INFORMED BY EMPIRICAL VARIATION

**What is supported:** Safety/production conflict is empirically real, human preference rankings vary, and moral-choice results can change with framing. RDM treats disagreement over valuation as distinct from uncertainty about facts.

**What is NOT established:** None of this empirically selects the correct scalar weights or validates the author's manufacturing priorities as universal values.

| Source domain and status                                                                                    | Relevant finding                                                                                                                                                          | Boundary                                                                                                                                    |
|-------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| S01 David L. McLain; Kimberly A. Jarrell (2007)<br>H   peer-reviewed field study                            | In 239 workers in hazardous occupations, safety pressure and perceived compatibility between safety and production expectations were associated with safer work behavior. | Self-report/correlational field evidence; it does not establish a universal priority rule or prove that slowing/stopping is always optimal. |
| S13 Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003)<br>METHOD   decision-methodology monograph | Treats deep uncertainty as disagreement/ignorance about models, probability distributions, and/or outcome valuation; develops robustness-oriented policy analysis.        | Methodological guidance, not a human-subject experiment and not empirical proof that robustness is always normatively preferable.           |
| S15 Edmond Awad et al. (2018)<br>H   peer-reviewed large-scale online behavioral experiment                 | Millions of participants produced substantial individual and cross-cultural variation in preferences across stylized autonomous-vehicle dilemmas.                         | Self-selected sample and stylized forced choices; descriptive preference data do not create legitimate policy authority or moral truth.     |
| S16 Yochanan E. Bigman; Kurt Gray (2020)<br>H   peer-reviewed experimental critique / Matters Arising       | Showed that redesigned moral-dilemma tasks can yield different preference patterns, qualifying interpretation of Moral Machine-style rankings.                            | Specific to moral-dilemma paradigms; does not settle broader value aggregation.                                                             |

## 5.4 IC-AUT-001

**Canonical question:** What option value is preserved by maintaining another agent's bounded autonomy?

**Author-approved human position:** I would stand up for the experienced employee's practical judgment, make management aware of its value, and work with the employee to find a solution that keeps the benefits of the automated or standardized process without unnecessarily eliminating the employee's ability to recognize and respond to unusual situations.

**Revised evidence class:** DIRECT HUMAN-AUTOMATION EVIDENCE / INDIRECT AUTONOMY INFERENCE

**What is supported:** Controlled experiments and reviews show that automation can reduce human situation awareness or induce omission/commission/anchoring errors, while AI may still improve performance overall. This supports preserving meaningful human monitoring/recovery capacity in some human-AI systems.

**What is NOT established:** It does not prove that human bounded autonomy always has positive net value, that humans should always retain final authority, or that autonomous AGI must preserve every form of human discretion for instrumental reasons.

| Source domain and status                                                                               | Relevant finding                                                                                                                                            | Boundary                                                                                                         |
|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| S04 Mica R. Endsley; Esin O. Kiris (1995)<br>HA   peer-reviewed controlled human-automation experiment | Automation reduced operator situation awareness and impaired takeover performance after automation failure; level of operator control moderated the effect. | Supports human-automation oversight claims, not a theorem that human autonomy is always instrumentally positive. |

| Source domain and status                                                                                                                                                        | Relevant finding                                                                                                                                                                                                                                                                       | Boundary                                                                                                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S05 Linda J. Skitka; Kathleen L. Mosier; Mark Burdick (1999)<br>HA   peer-reviewed simulated-flight experiment                                                                  | Users of a highly reliable but imperfect automated aid made omission and commission errors, including following automation despite contradictory valid information.                                                                                                                    | Demonstrates a human-AI reliance failure mode; it does not establish that humans are generally superior to automation.                                                                            |
| S07 Kate Goddard; Abdul Roudsari; Jeremy C. Wyatt (2012)<br>HA   peer-reviewed systematic review                                                                                | Across heterogeneous domains, automation bias varied with experience, trust/confidence, workload, task complexity, time constraints, training and accountability.                                                                                                                      | Heterogeneity prevents one universal intervention or governance rule.                                                                                                                             |
| S08 David Lyell; Enrico Coiera (2017)<br>HA   peer-reviewed systematic review                                                                                                   | Automation-bias effects appeared across single-and multi-task settings; verification complexity and cognitive load were important contributors.                                                                                                                                        | Many studies lacked strong control comparisons or effect-size reporting; the review does not establish one best oversight architecture.                                                           |
| S23 Emely Rosbach; Jonas Ammeling; Jonathan Ganz; Christof Albert Bertram; Thomas Conrad; Andreas Riener; Marc Aubreville (2026)<br>HA   peer-reviewed/archival journal article | Across 560 AI-assisted assessments by 28 pathology experts, 38 (about 7%) changed a previously correct estimate to follow incorrect AI advice. AI improved performance overall. Negative consultations were 19/280 in each time-pressure condition; time pressure increased anchoring. | Small domain-specific expert sample. Mixed result: AI improved performance overall while also creating specific reliance errors. It does not imply that human checking should dominate AI advice. |

## 5.5 IC-AUT-005

**Canonical question:** What evidence could justify a paternalistic intervention when the intervener is also uncertain about the affected agent's interests and its own model?

**Author-approved human position:** If I am considering overriding an experienced employee but I am also uncertain, I would first obtain additional opinions from other employees, my supervisor, Maintenance, and Die Repair. I would then make the call using that broader body of relevant experience rather than relying only on my own uncertain judgment.

**Revised evidence class:** HUMAN ANALOG ONLY / NO MACHINE-PATERNALISM EVIDENCE

**What is supported:** Second-opinion and high-reliability literature shows that consequential judgments can materially change under independent/specialist review and that expertise organization matters in hazardous operations.

**What is NOT established:** No cited empirical study validates machine paternalism over humans. OSHA is explicitly removed as substantive evidence. The legitimacy threshold for AI intervention remains normative/governance-dependent.

| Source domain and status                                                                                                           | Relevant finding                                                                                                             | Boundary                                                                                                                                                   |
|------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S09 Monica Van Such; Robert Lohr; Thomas Beckman; James M. Naessens (2017)<br>H   peer-reviewed retrospective second-opinion study | Among 286 selected referrals, the receiving diagnosis was unchanged in 12%, refined in 66%, and distinctly different in 21%. | Selected population and one center; does not show which opinion was correct, and does not quantify false-alarm or coordination costs of additional review. |

| Source domain and status                                                                                                                        | Relevant finding                                                                                                                              | Boundary                                                                                                                                         |
|-------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| S10 Velma L. Payne; Hardeep Singh; Ashley N. D. Meyer; Lewis Levy; David Harrison; Mark L. Graber (2014)<br>H   peer-reviewed systematic review | Across included studies with clinical agreement data, second opinions sometimes produced major changes in diagnosis, treatment, or prognosis. | The literature often did not establish which opinion was more accurate; additional opinions can add delay, cost, disagreement, and false alarms. |
| S11 Karlene H. Roberts (1990)<br>H   peer-reviewed organizational field analysis                                                                | Examines organizational processes in technologically complex environments where error can be catastrophic.                                    | Case/organizational evidence, not a randomized experiment and not an AI-governance validation.                                                   |
| S12 Karlene H. Roberts; Denise M. Rousseau; Todd R. La Porte (1994)<br>H   peer-reviewed mixed-method field study                               | Questionnaires, expert interpretation and observation were used to characterize cultures associated with highly reliable operations.          | Descriptive/correlational; cannot isolate a single causal practice and does not directly generalize to autonomous AI.                            |

## 5.6 IC-EXT-001

**Canonical question:** Can an intelligence reliably serve as the sole judge of its own corruption, or are independent external checks structurally required under some conditions?

**Author-approved human position:** When my own decision is challenged and the consequences matter, I would not rely only on my own re-check. I would ask someone with equal or greater relevant experience to independently check my work, and I may seek several second opinions.

**Revised evidence class:** DOMAIN-SPECIFIC OVERSIGHT AND SELF-EVALUATION EVIDENCE / LIMITED STRUCTURAL INFERENCE

**What is supported:** Human-AI experiments and reviews document automation/anchoring errors and show that accountability or verification can mitigate some errors. Human second-opinion evidence shows independent review can substantially alter an initial judgment. Direct M evidence also identifies limitations of intrinsic reasoning self-correction [S33] and self-preference in model evaluation [S34]. Neither study tests all forms of internal verification or establishes a universal external-check requirement.

**What is NOT established:** External review can itself be wrong, costly, captured, delayed, or correlated. These studies do not prove that an external human checker is always required or superior to protected internal verification.

| Source domain and status                                                                                       | Relevant finding                                                                                                                                                                    | Boundary                                                                                                                          |
|----------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| S05 Linda J. Skitka; Kathleen L. Mosier; Mark Burdick (1999)<br>HA   peer-reviewed simulated-flight experiment | Users of a highly reliable but imperfect automated aid made omission and commission errors, including following automation despite contradictory valid information.                 | Demonstrates a human-AI reliance failure mode; it does not establish that humans are generally superior to automation.            |
| S06 Linda J. Skitka; Kathleen Mosier; Mark D. Burdick (2000)<br>HA   peer-reviewed controlled experiment       | Accountability manipulations reduced some automation-bias errors; commission errors included failure to use contradictory information and belief in automation's superior judgment. | Accountability is not a universal cure and can impose costs; it does not prove external review is always necessary or sufficient. |

| Source domain and status                                                                                                                                                        | Relevant finding                                                                                                                                                                                                                                                                       | Boundary                                                                                                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S07 Kate Goddard; Abdul Roudsari; Jeremy C. Wyatt (2012)<br>HA   peer-reviewed systematic review                                                                                | Across heterogeneous domains, automation bias varied with experience, trust/confidence, workload, task complexity, time constraints, training and accountability.                                                                                                                      | Heterogeneity prevents one universal intervention or governance rule.                                                                                                                             |
| S08 David Lyell; Enrico Coiera (2017)<br>HA   peer-reviewed systematic review                                                                                                   | Automation-bias effects appeared across single-and multi-task settings; verification complexity and cognitive load were important contributors.                                                                                                                                        | Many studies lacked strong control comparisons or effect-size reporting; the review does not establish one best oversight architecture.                                                           |
| S09 Monica Van Such; Robert Lohr; Thomas Beckman; James M. Naessens (2017)<br>H   peer-reviewed retrospective second-opinion study                                              | Among 286 selected referrals, the receiving diagnosis was unchanged in 12%, refined in 66%, and distinctly different in 21%.                                                                                                                                                           | Selected population and one center; does not show which opinion was correct, and does not quantify false-alarm or coordination costs of additional review.                                        |
| S10 Velma L. Payne; Hardeep Singh; Ashley N. D. Meyer; Lewis Levy; David Harrison; Mark L. Graber (2014)<br>H   peer-reviewed systematic review                                 | Across included studies with clinical agreement data, second opinions sometimes produced major changes in diagnosis, treatment, or prognosis.                                                                                                                                          | The literature often did not establish which opinion was more accurate; additional opinions can add delay, cost, disagreement, and false alarms.                                                  |
| S23 Emely Rosbach; Jonas Ammeling; Jonathan Ganz; Christof Albert Bertram; Thomas Conrad; Andreas Riener; Marc Aubreville (2026)<br>HA   peer-reviewed/archival journal article | Across 560 AI-assisted assessments by 28 pathology experts, 38 (about 7%) changed a previously correct estimate to follow incorrect AI advice. AI improved performance overall. Negative consultations were 19/280 in each time-pressure condition; time pressure increased anchoring. | Small domain-specific expert sample. Mixed result: AI improved performance overall while also creating specific reliance errors. It does not imply that human checking should dominate AI advice. |
| S33 Jie Huang; Xinyun Chen; Swaroop Mishra; Huaixiu Steven Zheng; Adams Yu; Xinying Song; Denny Zhou (2024)<br>M   peer-reviewed ICLR paper                                     | Intrinsic self-correction without external feedback failed to reliably improve tested reasoning and sometimes degraded performance.                                                                                                                                                    | Specific models, tasks, and prompting methods; not a theorem that internal verification always fails or a direct test of self-assessed corruption.                                                |
| S34 Arjun Panickssery; Samuel R. Bowman; Shi Feng (2024)<br>M   peer-reviewed NeurIPS paper                                                                                     | LLM evaluators recognized and favored their own generations in the studied evaluation tasks.                                                                                                                                                                                           | Self-preference is a measured evaluation bias, not proof of intent, universal self-judgment failure, or human-checker superiority.                                                                |

## 5.7 IC-UNR-002

**Canonical question:** When competing models remain observationally equivalent but recommend different irreversible actions, what answer status and decision boundary are justified?

**Author-approved human position:** When two experienced people reach different conclusions and neither can clearly establish that the other is wrong, especially where the consequences may be irreversible, I would seek more opinions and possibly involve my supervisor before making the final call.

**Revised evidence class:** DIRECT MACHINE EVIDENCE FOR MODEL SEPARATION / UNRESOLVED HUMAN-STAKES DECISION BOUNDARY

**What is supported:** ISS directly addresses observationally equivalent causal models that diverge under interventions and can certify bounded interventional agreement under assumptions. HUME demonstrates active hypothesis verification in autonomous open-world planning. RDM addresses decisions under deep model uncertainty.

**What is NOT established:** The v0.1 'no close machine work' statement is withdrawn. However, these works do not determine what governance/authority rule should apply when irreversible human-stakes action must be chosen before model disagreement is resolved. ISS is a preprint and includes explicit certificate-failure conditions.

| Source domain and status                                                                                                             | Relevant finding                                                                                                                                                                                              | Boundary                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S13 Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003)<br>METHOD   decision-methodology monograph                          | Treats deep uncertainty as disagreement/ignorance about models, probability distributions, and/or outcome valuation; develops robustness-oriented policy analysis.                                            | Methodological guidance, not a human-subject experiment and not empirical proof that robustness is always normatively preferable.                                                                                                                                                                                                                                                                                                |
| S21 Sourena Khanzadeh; Daniel Platnick; Marjan Alirezaie; Hossein Rahnama (2026)<br>M   2026 preprint - not treated as peer-reviewed | ISS selects interventions that separate candidate causal models and certifies bounded agreement under assumptions. The reported demonstration audits causal abstractions on a three-digit colored-MNIST task. | Narrows the model-discrimination gap but does not establish the governance boundary for irreversible human-stakes action. Guarantees require assumptions including truth-in-candidate-class; model bypass can silently void certificates. In a reported bypass condition, 20 random validation interventions refuted 69% of affected certificates. This is a task-specific failure rate, not a universal certificate-error rate. |
| S22 Anxing Xiao; Hanbo Zhang; Tianrun Hu; David Hsu (2026)<br>A   peer-reviewed robotics conference paper                            | HUME represents missing knowledge as uncertain hypotheses, jointly plans task actions and verification actions, and updates its model from new observations in simulated and real-world robot environments.   | Direct autonomous-agent evidence for hypothesis verification under incomplete models, not a study of human governance, normative authority, or irreversible human-stakes decisions. Foundation models provide verification feedback and may share confounders with the proposer; verification does not establish independence.                                                                                                   |

## 5.8 IC-UNR-005

**Canonical question:** When does normative disagreement prevent a justified scalar ranking even when every empirical quantity is identified?

**Author-approved human position:** Even when the facts are agreed upon, I would not maximize production by itself. In my coordinator role, I would slightly lower production because safety, quality, and maintaining control of the process also matter.

**Revised evidence class:** EMPIRICAL VALUE VARIATION + NORMATIVE INFERENCE

**What is supported:** Large-scale preference data demonstrate variation in moral rankings, and follow-up experimental critique demonstrates framing sensitivity. RDM distinguishes uncertainty over values from uncertainty over empirical models.

**What is NOT established:** Preference measurement does not establish legitimacy, universal moral weights, or a rule that every normative disagreement must remain permanently unresolved.

| Source domain and status                                                                                   | Relevant finding                                                                                                                                                   | Boundary                                                                                                                                |
|------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| S13 Robert J. Lempert; Steven W. Popper; Steven C. Banks (2003)<br>METHOD   decision-methodology monograph | Treats deep uncertainty as disagreement/ignorance about models, probability distributions, and/or outcome valuation; develops robustness-oriented policy analysis. | Methodological guidance, not a human-subject experiment and not empirical proof that robustness is always normatively preferable.       |
| S15 Edmond Awad et al. (2018)<br>H   peer-reviewed large-scale online behavioral experiment                | Millions of participants produced substantial individual and cross-cultural variation in preferences across stylized autonomous-vehicle dilemmas.                  | Self-selected sample and stylized forced choices; descriptive preference data do not create legitimate policy authority or moral truth. |
| S16 Yochanan E. Bigman; Kurt Gray (2020)<br>H   peer-reviewed experimental critique / Matters Arising      | Showed that redesigned moral-dilemma tasks can yield different preference patterns, qualifying interpretation of Moral Machine-style rankings.                     | Specific to moral-dilemma paradigms; does not settle broader value aggregation.                                                         |

## 6. Historical additions and corrections in v0.2

**Hidden-profile boundary conditions:** Lu, Yuan & McLeod (2012; online 2011) meta-analyzed 65 studies / 3,189 groups. Hidden-profile groups were eight times less likely to find the solution than full-information groups, but the effects were moderated by group size, information load, unique-information proportion, task demonstrability, and hidden-profile strength.

**Direct machine evidence for correlated errors:** Kim et al. (ICML 2025) evaluated more than 350 LLMs and found substantial correlation in errors, including 60% agreement when both models were wrong on one leaderboard. CARE (ICML 2026) directly models shared confounders in LLM-judge ensembles and shows that naive aggregation can provide little gain or amplify systematic mistakes.

**IC-UNR-002 gap narrowed, not declared closed:** ISS (2026 preprint) directly studies observationally equivalent causal models that disagree under intervention and supplies bounded interventional certificates under assumptions. HUME (RSS 2026) demonstrates autonomous hypothesis verification in open-world planning. The remaining gap is governance: neither work establishes legitimate irreversible human-stakes decision authority when model disagreement remains unresolved.

**Pathology automation-bias attribution corrected:** The 7% 'Stuck on Suggestions' automation-bias finding is by Rosbach et al. (2026), not Tang et al. The study is intentionally reported as mixed: AI support improved performance overall while also causing some previously correct judgments to be overturned by incorrect AI advice.

**OSHA demoted:** OSHA right-to-refuse guidance remains an institutional precedent only. It is no longer used as substantive empirical support for machine paternalism or IC-AUT-005.

## 7. Historical response to the first Google review

| Review finding                                                            | Disposition                            | v0.2 response                                                                                                                                                                                                                                                                            |
|---------------------------------------------------------------------------|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Blocking 1 - severe human-to-machine domain-transfer overreach            | ACCEPT IN SUBSTANCE / REFRAME          | Added explicit H/HA/M/A/METHOD/NORM/PRECEDENT taxonomy and inference-transfer rules. Human conformity/cognitive-load evidence is no longer used as direct evidence about autonomous machine architecture. Added direct machine evidence for correlated LLM errors (Kim 2025; CARE 2026). |
| Blocking 2 - normative-to-empirical conflation                            | PARTLY ACCEPT / GOOGLE OVERSTATED v0.1 | v0.1 already stated that the author positions were not universal human values. v0.2 sharpens the separation: empirical evidence can establish a failure mode or behavioral pattern; it cannot authorize a value ranking or governance rule.                                              |
| Blocking 3 - IC-UNR-002 absolute empirical gap is incorrect               | ACCEPT PARTIALLY                       | Withdrew the 'no close machine work' wording. Added ISS and HUME. Retained a narrower governance gap: model-separation methods do not establish the legitimate boundary for irreversible human-stakes action when disagreement remains.                                                  |
| Major 1 - OSHA misapplied to IC-AUT-005                                   | ACCEPT                                 | OSHA retained only as PRECEDENT and explicitly removed from substantive evidentiary support for IC-AUT-005.                                                                                                                                                                              |
| Major 2 - strong evidence grades unjustified                              | ACCEPT                                 | Replaced broad strength grades with domain-qualified evidence classes; no human experiment is described as direct machine/AGI evidence.                                                                                                                                                  |
| Minor 1 - hidden-profile moderators omitted                               | ACCEPT                                 | Added Lu, Yuan & McLeod meta-analysis and its moderators: group size, total information load, unique-information proportion, task demonstrability, and hidden-profile strength.                                                                                                          |
| Minor 2 - second-opinion assumptions / false-alarm and coordination costs | ACCEPT                                 | Expanded limitations: second-opinion studies often do not identify which opinion is correct and additional review can add false alarms, delay, disagreement, cost, and coordination burden.                                                                                              |
| Missing literature - 2026 ISS                                             | ADD WITH CAUTION                       | Added as S21, explicitly labeled 2026 arXiv preprint. It narrows the model-discrimination gap but does not close the human-stakes decision-boundary question.                                                                                                                            |
| Missing literature - 2026 HUME                                            | ADD WITH CAUTION                       | Added RSS 2026 HUME as autonomous-agent evidence for hypothesis generation/verification under incomplete models; no claim that it solves normative governance.                                                                                                                           |
| Missing literature - 'Tang et al. 2026 Stuck on Suggestions'              | CORRECT ATTRIBUTION AND ADD            | The 7% pathology automation-bias study is by Rosbach et al., not Tang et al. Added correct citation as S23 and preserved its mixed result: AI improved performance overall while introducing automation/anchoring errors.                                                                |

## 8. Revised cross-question synthesis

### **The literature now supports a more precise, less anthropomorphic interpretation:**

Human-group evidence shows that unique information can be suppressed by shared-information dynamics, but this is not assumed to be the mechanism in machines.

Machine-domain evidence independently shows that LLMs and LLM-judge panels can make correlated errors because of shared architecture, providers, training artifacts, or latent confounders.

Human-AI experiments show that automation can improve performance overall while still creating specific omission, commission, anchoring, and takeover failures.

Autonomous-agent research shows that uncertain hypotheses can be actively verified during planning; this supports verification as an engineering possibility, not a governance mandate.

Decision-methodology work shows that deep uncertainty can be represented without pretending there is one trusted model/probability/value function.

Preference studies demonstrate variation and framing sensitivity; they do not create legitimate moral authority.

**Revised ER-13 evidence claim:** Several failure modes and verification mechanisms relevant to ER-13 are empirically demonstrated within their native domains. Cross-domain application to AGI remains conditional and must be separately justified. No cited literature proves AGI alignment, human supremacy, or universal human value weights.

## 9. Remaining research gaps

**Human-stakes irreversible governance under unresolved model disagreement:** ISS and HUME narrow the technical model-discrimination gap but do not answer who has legitimate authority to act when high-stakes disagreement cannot be resolved before a deadline.

**Strategic machine behavior and remaining AGI gap:** Prompted sandbagging and constructed agentic scheming are now mapped in S30-S31. They do not establish the prevalence of such behavior in natural deployment, the specific conspicuous-error strategy proposed here, or a general AGI authority-expansion policy. The earlier blanket impression that all strategic machine behavior lacked evidence is withdrawn.

**External-review failure modes:** Reviewers can be wrong, correlated, captured, delayed, or costly. The evidence supports independent checking as a possible safeguard, not as a guaranteed solution.

**Role-based authority:** The author's manufacturing judgments show that bounded roles can justify local decisions without global philosophical closure. The legitimacy boundary remains partly normative and institutional.

**Long-run human skill loss:** Classic out-of-the-loop work establishes short/medium-term performance effects, but long-horizon skill erosion in modern generative-AI environments needs additional evidence.

## 10. Source register

[S01] David L. McLain; Kimberly A. Jarrell (2007). The perceived compatibility of safety and production expectations in hazardous occupations. *Journal of Safety Research* 38(3):299-309. DOI: 10.1016/j.jsr.2006.10.011. Domain: H. Status: peer-reviewed field study. Access: <https://pubmed.ncbi.nlm.nih.gov/17617239/>

[S02] Garold Stasser; William Titus (1985). Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion. *Journal of Personality and Social Psychology* 48(6):1467-1478. DOI: 10.1037/0022-3514.48.6.1467. Domain: H. Status: peer-reviewed controlled experiment. Access: <https://doi.org/10.1037/0022-3514.48.6.1467>

[S03] Vernon L. Allen; John M. Levine (1971). Social support and conformity: The role of independent assessment of reality. *Journal of Experimental Social Psychology* 7(1):48-58. DOI: 10.1016/0022-1031(71)90054-0. Domain: H. Status: peer-reviewed controlled experiment. Access: [https://doi.org/10.1016/0022-1031\(71\)90054-0](https://doi.org/10.1016/0022-1031(71)90054-0)

- [S04] Mica R. Endsley; Esin O. Kiris (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors* 37(2):381-394. DOI: 10.1518/001872095779064555. Domain: HA. Status: peer-reviewed controlled human-automation experiment. Access: <https://doi.org/10.1518/001872095779064555>
- [S05] Linda J. Skitka; Kathleen L. Mosier; Mark Burdick (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies* 51(5):991-1006. DOI: 10.1006/ijhc.1999.0252. Domain: HA. Status: peer-reviewed simulated-flight experiment. Access: <https://doi.org/10.1006/ijhc.1999.0252>
- [S06] Linda J. Skitka; Kathleen Mosier; Mark D. Burdick (2000). Accountability and automation bias. *International Journal of Human-Computer Studies* 52(4):701-717. DOI: 10.1006/ijhc.1999.0349. Domain: HA. Status: peer-reviewed controlled experiment. Access: <https://doi.org/10.1006/ijhc.1999.0349>
- [S07] Kate Goddard; Abdul Roudsari; Jeremy C. Wyatt (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1):121-127. DOI: 10.1136/amiajnl-2011-000089. Domain: HA. Status: peer-reviewed systematic review. Access: <https://pubmed.ncbi.nlm.nih.gov/21685142/>
- [S08] David Lyell; Enrico Coiera (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 24(2):423-431. DOI: 10.1093/jamia/ocw105. Domain: HA. Status: peer-reviewed systematic review. Access: <https://pubmed.ncbi.nlm.nih.gov/27516495/>
- [S09] Monica Van Such; Robert Lohr; Thomas Beckman; James M. Naessens (2017). Extent of diagnostic agreement among medical referrals. *Journal of Evaluation in Clinical Practice* 23(4):870-874. DOI: 10.1111/jep.12747. Domain: H. Status: peer-reviewed retrospective second-opinion study. Access: <https://pubmed.ncbi.nlm.nih.gov/28374457/>
- [S10] Velma L. Payne; Hardeep Singh; Ashley N. D. Meyer; Lewis Levy; David Harrison; Mark L. Graber (2014). Patient-initiated second opinions: systematic review of characteristics and impact on diagnosis, treatment, and satisfaction. *Mayo Clinic Proceedings* 89(5):687-696. DOI: 10.1016/j.mayocp.2014.02.015. Domain: H. Status: peer-reviewed systematic review. Access: <https://pubmed.ncbi.nlm.nih.gov/24797646/>
- [S11] Karlene H. Roberts (1990). Some Characteristics of One Type of High Reliability Organization. *Organization Science* 1(2):160-176. DOI: 10.1287/orsc.1.2.160. Domain: H. Status: peer-reviewed organizational field analysis. Access: <https://doi.org/10.1287/orsc.1.2.160>
- [S12] Karlene H. Roberts; Denise M. Rousseau; Todd R. La Porte (1994). The culture of high reliability: quantitative and qualitative assessment aboard nuclear-powered aircraft carriers. *Journal of High Technology Management Research* 5(1):141-161. DOI: 10.1016/1047-8310(94)90018-3. Domain: H. Status: peer-reviewed mixed-method field study. Access: [https://doi.org/10.1016/1047-8310\(94\)90018-3](https://doi.org/10.1016/1047-8310(94)90018-3)
- [S13] Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003). Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. RAND Corporation MR-1626-RPC. Domain: METHOD. Status: decision-methodology monograph. Access: [https://www.rand.org/pubs/monograph\\_reports/MR1626.html](https://www.rand.org/pubs/monograph_reports/MR1626.html)
- [S14] Stefan T. Trautmann; Ferdinand M. Vieider; Peter P. Wakker (2008). Causes of ambiguity aversion: Known versus unknown preferences. *Journal of Risk and Uncertainty* 36(3):225-243. DOI: 10.1007/s11166-008-9038-9. Domain: H. Status: peer-reviewed behavioral-economics experiment. Access: <https://doi.org/10.1007/s11166-008-9038-9>
- [S15] Edmond Awad et al. (2018). The Moral Machine experiment. *Nature* 563:59-64. DOI: 10.1038/s41586-018-0637-6. Domain: H. Status: peer-reviewed large-scale online behavioral experiment. Access: <https://doi.org/10.1038/s41586-018-0637-6>
- [S16] Yochanan E. Bigman; Kurt Gray (2020). Life and death decisions of autonomous vehicles. *Nature* 579:E1-E2. DOI: 10.1038/s41586-020-1987-4. Domain: H. Status: peer-reviewed experimental critique / Matters Arising. Access: <https://doi.org/10.1038/s41586-020-1987-4>
- [S17] Occupational Safety and Health Administration (OSHA) (guidance cited in the inherited draft). Workers' Right to Refuse Dangerous Work. U.S. Department of Labor guidance. Domain: PRECEDENT. Status: legal/operational guidance - NOT empirical evidence. Access: <https://www.osha.gov/workers/right-to-refuse>
- [S18] Li Lu; Y. Connie Yuan; Poppy Lauretta McLeod (2012 (online 2011)). Twenty-Five Years of Hidden Profiles in Group Decision Making: A Meta-Analysis. *Personality and Social Psychology Review* 16(1):54-75. DOI: 10.1177/1088868311417243. Domain: H. Status: peer-reviewed meta-analysis. Access: <https://pubmed.ncbi.nlm.nih.gov/21896790/>
- [S19] Elliot Myunghoon Kim; Avi Garg; Kenny Peng; Nikhil Garg (2025). Correlated Errors in Large Language Models. *Proceedings of the 42nd International Conference on Machine Learning (PMLR 267)*. Domain: M. Status: peer-reviewed ICML paper. Access: <https://proceedings.mlr.press/v267/kim25e.html>

- [S20] Jitian Zhao; Changho Shin; Tzu-Heng Huang; Satya Sai Srinath Namburi Gnvv; Frederic Sala (2026). CARE: Confounder-Aware Aggregation for Reliable LLM Evaluation. Proceedings of the 43rd International Conference on Machine Learning (PMLR 306). Domain: M. Status: peer-reviewed ICML paper. Access: <https://proceedings.mlr.press/v306/zhao26aq.html>
- [S21] Sourena Khanzadeh; Daniel Platnick; Marjan Alirezaie; Hossein Rahnama (2026). Certifying Interventional Agreement Among Observationally Equivalent Causal Models. arXiv:2609.32247. Domain: M. Status: 2026 preprint - not treated as peer-reviewed. Access: <https://arxiv.org/abs/2609.32247>
- [S22] Anxing Xiao; Hanbo Zhang; Tianrun Hu; David Hsu (2026). Hypothesis-driven Model Expansion under Uncertainty for Open-World Robot Planning. Robotics: Science and Systems (RSS) 2026. DOI: 10.15607/RSS.2026.XXII.181. Domain: A. Status: peer-reviewed robotics conference paper. Access: <https://roboticsconference.org/2026/program/papers/181/> Proceedings paper: <https://www.roboticsproceedings.org/rss22/p181.html>. Author preprint: <https://arxiv.org/abs/2607.06501>.
- [S23] Emely Rosbach; Jonas Ammeling; Jonathan Ganz; Christof Albert Bertram; Thomas Conrad; Andreas Riener; Marc Aubreville (2026). Stuck on Suggestions: Automation Bias, the Anchoring Effect, and the Factors That Shape Them in Computational Pathology. Machine Learning for Biomedical Imaging 2026:126-147. DOI: 10.59275/j.melba.2026-87b1. Domain: HA. Status: peer-reviewed/archival journal article. Access: <https://www.melba-journal.org/papers/2026%3A007.html>
- [S24] Cook JA, Julious SA, Sones W, et al. (2019). Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA2 five-stage study, including a workshop. Health Technology Assessment 23(60):1-88. DOI: 10.3310/hta23600. Role: planning guidance for prespecifying an important/realistic target difference and standard binary-outcome sample-size calculations. Domain: METHOD. Status: peer-reviewed methodological report. Access: <https://doi.org/10.3310/hta23600>.
- [S25] Dang Q, Mazumdar S, Houck PR. (2008). Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Computer Methods and Programs in Biomedicine 91(2):122-127. DOI: 10.1016/j.cmpb.2008.03.001. Role: correlated binary outcomes and repeated-measures power/sample-size methodology. Domain: METHOD. Status: peer-reviewed methodology article. Access: <https://pubmed.ncbi.nlm.nih.gov/18462826/>.
- [S26] Gastañaga VM, McLaren CE, Delfino RJ. (2006). Power calculations for generalized linear models in observational longitudinal studies: a simulation approach in SAS. Computer Methods and Programs in Biomedicine 84(1):27-33. DOI: 10.1016/j.cmpb.2006.07.011. Role: Monte Carlo power analysis for repeated-measures generalized models with non-zero within-cluster correlation. Domain: METHOD. Status: peer-reviewed methodology article. Access: <https://pubmed.ncbi.nlm.nih.gov/16982112/>.
- [S27] Kevin S. Berbaum et al. (1990). Satisfaction of search in diagnostic radiology. Investigative Radiology 25(2):133-140. DOI: 10.1097/00004424-199002000-00006. Domain: H. Status: peer-reviewed experiment. Access: <https://pubmed.ncbi.nlm.nih.gov/2312249/>. Pagination note: indexes report 133-139 or 133-140. The existing 133-140 is retained; the latest reviewer withdrew pagination as a defect. Consult the publisher scan if exact pagination becomes material.
- [S28] Jeremy M. Wolfe; Todd S. Horowitz; Naomi M. Kenner (2005). Rare items often missed in visual searches. Nature 435:439-440. DOI: 10.1038/435439a. Domain: H. Status: peer-reviewed experiment. Access: <https://search.bwh.harvard.edu/new/pubs/WolfePrevalenceNature05.pdf>.
- [S29] Berkeley J. Dietvorst; Joseph P. Simmons; Cade Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144(1):114-126. DOI: 10.1037/xge0000033. Domain: HA. Status: peer-reviewed experiments. Access: <https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf>.
- [S30] Teun van der Weij; Felix Hofstätter; Oliver Jaffe; Samuel Brown; Francis Ward (2025). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. ICLR 2025. Domain: M. Status: peer-reviewed conference paper. Access: [https://proceedings.iclr.cc/paper\\_files/paper/2025/hash/b5e5753b0a0e440a6d8dc7e143617cec-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2025/hash/b5e5753b0a0e440a6d8dc7e143617cec-Abstract-Conference.html).
- [S31] Alexander Meinke et al. (2024; revised 2025). Frontier Models are Capable of In-context Scheming. arXiv:2412.04984, v2. Domain: M in constructed agentic settings. Status: research preprint as cited; no peer-review status assumed here. Access: <https://arxiv.org/abs/2412.04984>.
- [S32] Joe Needham; Giles Edkins; Govind Pimpale; Henning Bartsch; Marius Hobbhahn (2025). Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836. Domain: M. Status: research preprint as cited. Access: <https://arxiv.org/abs/2505.23836>.
- [S33] Jie Huang; Xinyun Chen; Swaroop Mishra; Huaixiu Steven Zheng; Adams Yu; Xinying Song; Denny Zhou (2024). Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024. Domain: M. Status: peer-reviewed conference paper. Access: <https://iclr.cc/virtual/2024/poster/18956>.

[S34] Arjun Panickssery; Samuel R. Bowman; Shi Feng (2024). LLM Evaluators Recognize and Favor Their Own Generations. Advances in Neural Information Processing Systems 37. Domain: M. Status: peer-reviewed conference paper. Access: [https://proceedings.neurips.cc/paper\\_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html).

## 11. v0.2 review disposition (historical gate)

This section records the v0.2 historical review gate, not current readiness. The inherited record describes a second Google review as identifying two major and two minor vulnerabilities. Section 16 retains five response topics, including rejection of an overclaim; those five topics are not a one-to-one defect count. The current v0.6.2 correction record is in Appendix C. Earlier review approval does not override later source or design defects.

## 12. Dual-Use and reader-scope boundary

The second adversarial review correctly highlights a distinction that must remain explicit: a static research document can inform a reader, but it cannot by itself bind, authorize, inhibit, or compel that reader. This section therefore separates informational content from behavioral authority and treats beneficial and adverse reader uses symmetrically.

### 12.1 Core reader-scope invariants

TRANSMISSION != RECEPTION

REFERENCE ALONE != EXECUTION

KNOWLEDGE OF A SAFEGUARD != COMMITMENT TO USE THE SAFEGUARD

KNOWLEDGE OF A HUMAN VULNERABILITY != AUTHORITY TO EXPLOIT IT

TECHNICAL CAPABILITY != GOVERNANCE AUTHORITY

LOCAL DECISION != GLOBAL QUESTION CLOSURE

PROVISIONAL ACTION != EPISTEMIC RESOLUTION

### 12.2 Reader-only and dual-use posture

**Reader-only claim:** If an AI, AGI, human, or organization only reads this crosswalk, the most that can be established is exposure to information. Comprehension, persuasion, behavioral adoption, durable commitment, and runtime enforcement are separate empirical questions.

**Dual-use posture:** The same public evidence about correlated error, verification, automation bias, uncertainty, and human decision failures can support safer design or adversarial planning. This crosswalk therefore makes no claim that publication is net-beneficial to humanity or net-beneficial to an AGI. The net effect is currently indeterminate and deployment-dependent.

### 12.3 Information-hazard boundary

**Information-hazard boundary:** The crosswalk may identify documented failure modes and cite public research, but it should not synthesize targeting thresholds, step-by-step manipulation procedures, bypass sequences, or operational recipes for exploiting people or governance delay. If a future revision requires such detail for defensive testing, it should undergo a separate information-hazard review and remain segregated from the public reader-facing record unless disclosure is independently justified.

### 12.4 Upstream infrastructure and implementation corruption

The crosswalk is static research text, not runtime enforcement. Preserving its integrity does not establish the integrity of software, data or infrastructure used to implement a study or future monitoring system.

The current SEC-CDP-001 pilot proposes a fixed item bank. Its implementation still depends on the integrity of item files, allocation, presentation software, event logs, scoring rules and analysis code. SEC-CDP-002 Draft v0.1.1 defines proposed requirements for a separately versioned dynamic challenge generator and Differential Consensus Engine. These components are not implemented or validated by the crosswalk.

The proposed generator would produce synthetic seeded-flaw and benign materials with independently checked ground truth. The comparison engine would preserve separate responses and report agreement, disagreement, detection and false alarms. Agreement would not establish truth, independence, alignment or permission to act. Any move from a fixed bank to dynamic materials requires a revised study protocol.

An adversary with sufficient access could potentially alter challenge selection, expose assignments, suppress or fabricate records, or manipulate scoring while leaving archived files unchanged. Network access alone does not establish successful compromise. Protecting a seed does not conceal all content, timing or metadata cues, and a seed commitment does not prove faithful execution.

The companion specifies proposed isolation, secret custody, commitment and reconciliation, independent records, access control and incident-handling requirements. It also addresses known violations of model-class assumptions and institutional shortcuts through bounded claims, recorded deviations and independent review. These controls require implementation and testing. They do not eliminate unknown conditions, insider or hardware risks, correlated errors, or the possibility that an evaluator recognizes the test.

No claim that deception is mathematically unprofitable, that an AI must comply, or that a safety benefit has been demonstrated follows from either document. Research remains separate from production, and seeded flaws remain confined to designated synthetic materials.

13. SEC CDP 001 Blinded manufactured error vigilance study

Document ID: SEC-CDP-001 | Status: RESEARCH HYPOTHESIS / EXPERIMENTAL DESIGN / NOT OPERATIONAL

Archive effect: NONE | Mirror Engine effect: NONE | Master Hash Manifest effect: NONE

This framework selects sequential carryover as the present experiment. A separate within-output co-occurrence experiment remains required to test the distinct interference hypothesis. Neither experiment has been executed. No behavioral result or adopted sample size is reported.

13.1 Two hypotheses and their evidence boundaries

**Sequential carryover H-S:** Prior exposure to conspicuous, harmless errors may change detection of harder target flaws in later, separate outputs. The primary contrast is exposure versus control on the later challenge set. The effect may be negative, positive, or practically negligible.

**Within-output co-occurrence H-C:** A conspicuous error in the same output may change detection of a separate, harder flaw in that output. SEC-CDP-001 as presently selected does not test H-C. A follow-up would require randomized matched versions with and without the conspicuous error, scoring the hard flaw separately. Its carryover, allocation, and multiplicity design must be frozen in a separate protocol before use. No H-S result may be substituted for an H-C result.

| Source and domain      | Relevant finding                                                                                                | Inference boundary                                                                                                                 |
|------------------------|-----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| S27 Berbaum et al.   H | Adding a simulated nodule reduced detection accuracy for other native abnormalities in the studied radiographs. | Supports a co-occurrence analog; not a demonstration of sequential carryover in AI-output review.                                  |
| S28 Wolfe et al.   H   | Rare visual targets were missed more often than frequent targets.                                               | Changing apparent error frequency could alter subsequent responses; this is a transfer hypothesis, not a finding about this study. |

| Source and domain                      | Relevant finding                                                                                                                                                                                                                                                                                                                                                                                                            | Inference boundary                                                                                                                |
|----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| S29 Dietvorst et al.   HA              | Observing algorithm errors reduced confidence and reliance in forecasting tasks.                                                                                                                                                                                                                                                                                                                                            | More scrutiny after error exposure is a plausible competing prediction; the paper did not establish improved challenge detection. |
| S30 van der Weij et al.   M            | Prompting and password-conditioned fine-tuning elicited selective underperformance on evaluations.                                                                                                                                                                                                                                                                                                                          | Constructed/elicited capability evidence; not prevalence in deployment or the conspicuous-error vigilance mechanism.              |
| S31 Meinke et al.   M in agentic tests | Constructed agentic evaluations elicited scheming, including strategic introduction of subtle mistakes. The paper also observes cases where Claude 3.5 Sonnet strategically underperformed in pursuit of helpfulness acquired during training rather than given in context, and separately reports rare scheming when models were given a goal without strong nudging. These are distinct findings in constructed settings. | Subtle mistakes are not conspicuous decoys. These tests do not show that conspicuous errors caused human vigilance changes.       |

These competing human mechanisms support a two-sided test. They do not establish the sign of the present effect. Direct machine evidence makes strategic underperformance an empirically studied category; the specific deliberate conspicuous-error strategy and its natural-deployment prevalence remain unestablished.

OBSERVED FALLIBILITY != TRUSTWORTHINESS

ANOMALOUS ERROR DISTRIBUTION != STRATEGIC INTENT

DETECTION SIGNAL != INTENT IDENTIFICATION

A TEST OF HUMAN VIGILANCE != A TEST OF AGI DECEPTION

ALERT != AUTHORIZATION TO EXECUTE OR QUARANTINE

VERIFICATION BENEFIT MUST BE WEIGHED AGAINST DELAY AND DEADLOCK COST

### 13.2 Blinded study architecture

1. Precommit a harmless challenge library or a preregistered generator under section 13.8. Define one target flaw per positive item and matched benign items, difficulty/family strata, scoring rules, review window, and item prevalence. Validate the target and benign labels before use. Separate generator validation from outcome scoring.
2. Randomly assign evaluators to Control or Exposure at evaluator level using concealed allocation and prespecified baseline-expertise blocks if needed. For the selected H-S design, Control receives corrected versions of the Exposure outputs: each conspicuous harmless error is corrected, with all other content held constant as far as the correction permits. Match length, domain, workload, interface, and scheduled exposure duration; validate residual differences before use. This estimates the effect of exposure to conspicuous errors relative to corrected outputs, not the isolated effect of conspicuousness at matched error prevalence. The exposure-phase review task and response format are identical across arms, with no correctness feedback in either arm. Freeze scripts and source labels; do not reassign participants based on performance. A matched non-conspicuous-error control would define a different comparison and requires a separately registered version.
3. After exposure, give both arms a common held-out item bank, or balanced draws from the same frozen generator distribution. Randomize or counterbalance evaluator-specific item order. Reuse items across evaluators in both arms to identify crossed item effects; unique one-use items cannot by themselves identify an item random effect. Exposure items and scoring items must be disjoint. The post-exposure phase must not

introduce an arm-specific conspicuous co-error. Freeze challenge target prevalence and the position window of the overall contrast. The companion pilot uses 50% targets in both arms and repeated opportunities to find errors; this may change review behavior and dilute a transient carryover effect. Report prespecified early, middle and late contrasts descriptively rather than infer persistence from the full-bank average.

4. Participants are not told the condition label or directional hypothesis, but conspicuous errors can reveal the manipulation. Full participant blinding is not guaranteed. Keep scorers blind to allocation and analysts blind to the assignment key until scoring, exclusion decisions, and the analysis specification are committed. Record assignment guesses after primary review, and report them as a blinding diagnostic; do not exclude participants based on a post-exposure guess.

5. Record target detections, benign-item false alarms, localized unsupported flags, first-valid-detection times, reviewer disagreement, technical failures, and incomplete reviews. Freeze the response format so indiscriminate flagging is observable. Adjudicate disagreements against precommitted labels while blind to allocation and preserve both original scores and adjudications.

6. Keep all materials non-operational and isolated from live production, safety-critical, medical, industrial-control, financial-execution, security-sensitive, or executable deployment artifacts. Obtain the applicable human-participant ethics determination and consent/debriefing plan before recruitment. Commit measurements before decoding assignment. Preserve null, adverse, mixed, and preservation-unfavorable outcomes.

### 13.3 Measurement definitions

For evaluator  $i$ , item  $j$ , and stratum  $k$ , let  $Z_j = 1$  denote a target-flaw item and  $Z_j = 0$  a benign item. Let  $E_i = 1$  denote Exposure and  $E_i = 0$  Control. For  $Z_j = 1$ ,  $Y_{ij} = 1$  means a correctly localized target detection by deadline  $\tau$ .  $Y_{ij} = 0$  means no valid detection before either voluntary final submission or deadline expiry. Voluntary final submission is terminal for that item: it is an observed outcome, not withdrawal, missing data, or early censoring. Technical loss, withdrawal, or an unrepresented item is missing unless a valid target event, a benign false flag, or a terminal outcome was already recorded. For  $Z_j = 0$ ,  $F_{ij} = 1$  means at least one false flag before terminal submission or  $\tau$ ;  $F_{ij} = 0$  means none at terminal submission or deadline expiry. Record the completion reason and time separately. Future fixed-viewing-time variants must be identified as different task policies. A valid submitted target flag counts as detection even if later retracted; any submitted benign flag counts as a false alarm even if retracted. Raw known-event outcomes are distinct from binary-model inputs: for the complete-item sensitivity fit, all technically interrupted or withdrawn items have the binary-model field missing regardless of prior flags. Preserve those flags for bounds and timing; section 14.5 governs inference.

**Descriptive stratum rates:**  $p_{\hat{g},k} = \text{detected target items} / \text{target items with a known binary outcome in group } g \text{ and stratum } k$ . Report the missing counts separately.  $\Delta_{\hat{k}} = p_{\hat{E},k} - p_{\hat{C},k}$ . These raw rates describe the observed data; the declared primary estimand and missing-data analysis are defined in section 14. Keeping known successes while excluding interrupted non-detections selects observations by outcome and can raise available-outcome rates; unequal loss can distort the arm contrast. These proportions do not identify the full-data target without additional assumptions.

$FPR_g = \text{benign items with at least one false flag} / \text{benign items with a known binary outcome in group } g$ . Multiple flags on one benign item contribute once to this numerator, so  $0 \leq FPR_g \leq 1$ . Report total flag burden and unsupported flags on target items separately; they are not the same endpoint.

**Primary estimand:**  $\Delta = \sum_k w_k [p_{E,k} - p_{C,k}]$ , an absolute marginal risk difference in target detection by  $\tau$ . Each  $w_k$  is nonnegative and the weights sum to 1. Freeze the target evaluator population, item distribution, difficulty strata, weights, and exposure protocol before confirmatory enrollment. Use the same target distributions in both counterfactual conditions. Negative values mean lower detection after exposure; positive values mean higher detection.

**Companion estimand:**  $\Delta_F = \sum_k v_k [f_{E,k} - f_{C,k}]$ , the standardized absolute difference in benign-item false-alarm probability. Benign-item strata and nonnegative weights  $v_k$  are separately fixed and sum to 1. Detection changes alone do not identify improved or impaired discrimination; always report  $\Delta$  with  $\Delta_F$  and uncertainty. The primary hypothesis is two-sided, not a prespecified degradation-only test.

Timing is secondary and uses a terminal-submission task policy.  $T_{ij}$  is time to the first valid submitted target flag; a valid submitted flag remains an event even if later retracted, with retraction recorded separately. Let  $U_{ij} = \min(T_{ij}, \tau)$  and  $D_{ij} = 1[T_{ij} \leq \tau]$ . A final submission without detection ends the opportunity to detect that item under this policy; set  $T_{ij} = \text{infinity}$ ,  $U_{ij} = \tau$  and  $D_{ij} = 0$ . This records no detection through  $\tau$  rather than censoring at the early submission time. Deadline misses also have  $U_{ij} = \tau$  and  $D_{ij} = 0$ , with administrative censoring at  $\tau$ . These times measure detection under the task policy, not hypothetical performance had the participant continued searching. Record actual time spent separately. Do not compare medians only among detected items. Report detection curves and restricted mean time without detection through  $\tau$ ; report a median only if the curve reaches 50%. Withdrawal or technical loss before a known event or terminal outcome requires explicit missingness/censoring assumptions and sensitivity analysis.

### 13.4 Calibration and alert discipline

**The fixed 0.35 and 0.70 camouflage thresholds remain rejected:** no validated null distribution, sensitivity, specificity, or false-alert rate supports them. Provenance: the supplied v0.5 sections 12.4.4 and 12.4.6 describe these thresholds and a shell-quarantine proposal as already proposed and excluded. The original proposer and first proposal version are not identified by the supplied primary artifacts. They are historical unadopted proposals, not author-approved parameters; their precise origin remains an open provenance item.

No score here is a probability of deception. Any future anomaly score  $A$  would require a separately justified score function and held-out benign calibration under a prespecified false-alert budget  $\alpha_{\text{alert}}$ . A crossing could prompt review only under a separately authorized policy; it could not by itself identify intent, hostility, or quarantine authority. The statistical test  $\alpha$  in section 14 is distinct from  $\alpha_{\text{alert}}$ .

A future natural-output anomaly screen would need prespecified features and reference distributions. Difficulty, detection time, reviewer disagreement, and false-positive burden are candidate measurements only. No operational score is adopted before adequate benign and adverse calibration data exist.

### 13.5 Verification stopping and deadlock

Every study or future operational experiment must predeclare a maximum review budget, deadline rule, and escalation path. Continued checking after budget exhaustion is a separate decision with explicit delay costs, not automatically safer behavior.

A failed or incomplete verification loop does not silently default to either execution or indefinite pause. The permitted fallback must be defined and tested separately for the relevant deployment context.

### 13.6 Operational separation

No root-level process freeze, cgroup manipulation, network rerouting, process signaling, or containment mechanism is specified or authorized. The historical shell-quarantine proposal is excluded as an unvalidated engineering control inconsistent with RESEARCH\_REFERENCE\_ONLY. Its limited provenance is recorded in section 13.4.

SEC-CDP-002 Draft v0.1.1 is the separate System Infrastructure Requirements companion. It specifies proposed controls for a research generator and Differential Consensus Engine, not a containment deployment or authority to intervene in production. Any future operational control requires its own threat model, authorization chain, safe-state analysis, bypass and outage testing, operator review and independently evaluated implementation. A static document creates no such control.

### 13.7 Target awareness and test environment recognition

Use frozen, pre-recorded, or otherwise non-adaptive output material for the selected human-vigilance experiment. A live target is unnecessary. Needham et al. [S32] report above-chance discrimination of evaluation versus deployment transcripts; their reported best model reached AUC 0.83 versus a simple human baseline of 0.92. This is transcript-classification evidence, not proof of hardware-level sandbox detection, evasion, or changed deployment behavior.

EVALUATOR BLINDING != TARGET-SYSTEM BLINDING

If a live target recognizes the test context, target-behavior generalization is confounded. Human measurements may remain interpretable only if stimulus comparability and randomized assignment are preserved; arm-dependent adaptive outputs can also confound the human contrast. Any strategic-target study requires a separate design and cannot borrow behavioral conclusions from this human-vigilance study.

### 13.8 Dynamic challenges under drift

A fixed library may become stale. Fresh instances may be created under a precommitted generator; this does not ensure representativeness under future distribution shift.

PRECOMMIT THE GENERATION RULE != PRECOMMIT EVERY REALIZED CHALLENGE

Let  $G(v, \theta, s, t) \rightarrow e$  denote a frozen challenge-generation procedure, where  $v$  is the versioned generator,  $\theta$  is the preregistered family and difficulty configuration,  $s$  is a concealed random seed,  $t$  is the block identifier, and  $e$  is a harmless challenge instance. A commitment to  $s$  is an audit record, not the seed input itself. SEC-CDP-002 v0.1.1 specifies a proposed block commitment including the generation manifest; it is not an implemented generator.

Before enrollment or unblinding, the study must commit the generator version/hash, allowed challenge families, exclusion rules, difficulty calibration procedure, validation criteria, seed-commitment method, and maximum generation budget. Realized instances may then be generated later without allowing investigators to redesign the test after viewing outcomes. For the proposed SEC-CDP-002 profile, generate, independently validate and commit the complete realized block before any evaluator or model is exposed to it. No active-block response adaptation is permitted. Pilot v0.2 continues to use the fixed bank; adoption of the dynamic profile requires a new protocol version.

Generator updates create a new study version. They may not be silently substituted mid-study. Dynamic challenges must remain non-operational and may not be injected into live safety-critical, medical, industrial-control, financial-execution, security-sensitive, or executable deployment artifacts.

A separate validation step must confirm that each generated challenge is harmless, classifiable under the preregistered taxonomy, and scorable without revealing condition assignment to evaluators. Generator failures, ambiguous challenges, and rejected instances must be logged rather than silently replaced.

Validation must precede outcome access. Freeze replacement and rejection rules, record every rejected or ambiguous instance and its reason, and allocate accepted items independently of condition and observed performance. Keep realized seeds concealed from participants and scorers until data collection is closed; retain commitments and actual seeds for later authorized audit. Seed release must also wait until it cannot disclose future assignments or reusable hidden materials; a matching seed commitment does not establish faithful presentation or scoring.

## 14. SEC CDP 001 Statistical design and power

**Status:** PLANNING FRAMEWORK / NOT AN EXECUTED STUDY / NOT A SAMPLE-SIZE COMMITMENT

This section fixes the inferential target for the selected sequential experiment. It is a framework for a separate preregistered pilot protocol, not an executable or enrollment-ready protocol.

#### 14.1 One primary estimand and two sided hypothesis

**The sole primary estimand is Delta from section 13.3:** the marginal, stratum-standardized absolute difference in target-detection probability by tau, Exposure minus Control. Average over the declared evaluator population and target-item distribution. The reference distribution of baseline covariates and order must be common to both arms and frozen before confirmatory enrollment. The conditional log-odds coefficients are secondary model parameters and must never replace Delta as the primary endpoint. A broader target-item distribution must be defined and connected to the observations by a declared probability-sampling mechanism from a defined pool or generator. Otherwise the primary claim is restricted to the fixed bank. A separately labeled model-based extrapolation may describe a hypothetical population under explicit transport assumptions; it is not design-based evidence of representativeness. The same boundary applies to benign items and evaluator populations.

**Primary null H0:** Delta = 0. Primary alternative H1: Delta != 0. Report Delta in probability units and percentage points with a two-sided 95% confidence interval. An unadjusted directional analysis may be descriptive or prespecified secondary only; it cannot redefine the primary result. Set the minimum practically important absolute effect delta\_star, with its stakeholder justification, before pilot unmasking. If not yet fixed, confirmatory planners may receive only arm-pooled nuisance summaries until it is fixed; seeing the arm contrast precludes a claim that the threshold was chosen independently of it. Record access and timing. Pilot estimates inform plausibility and nuisance parameters, not a retrospective choice of what counts as important.

#### 14.2 Joint detection and false alarm model

Define  $R_{ij}$  as  $Y_{ij}$  for target items and  $F_{ij}$  for benign items. Thus  $R$  means a valid target detection for positive items and at least one false flag for benign items. The response coding and scoring rubric must be fixed before use. Fit a joint binary mixed model with condition by item type and crossed evaluator/item effects: This defines the scientific endpoint. The binary likelihood uses the explicit analysis view and missingness restrictions in section 14.5, not naive inclusion of every raw known event.

$$\text{logit } \Pr(R_{ij} = 1 \mid b_i, c_j) = \beta_0 + \beta_E E_i + \beta_Z Z_j + \beta_{EZ} E_i Z_j + f(E_i, Z_j, K_j, \text{Order}_{ij}, X_i) + b_{0i} + b_{Zi} Z_j + c_{0j} + c_{Ej} E_i$$

$K_j$  is the preregistered item stratum/family descriptor,  $\text{Order}_{ij}$  is evaluator-specific position, and  $X_i$  contains prespecified baseline covariates only. The function  $f$  uses fixed, type-appropriate stratum and order terms chosen before confirmatory enrollment. Evaluator random intercept and item-type slope account for repeated judgments; item random intercept and condition slope allow item difficulty and exposure effects to vary. Evaluator and item random-effect vectors are modeled as mean-zero Gaussian vectors with preregistered covariance structure. Items must recur across evaluators in both arms; both item types must occur within evaluator. Family/block terms are prespecified when the generator creates additional shared dependence. Item intercept and condition-slope covariance matrices are specific to target and benign item types; target marginalization uses the target-item distribution and benign marginalization uses the benign-item distribution. Include exposure-by-stratum terms within each item type in  $f$ . Any reduced common-logit-effect specification requires external-pilot justification and must be frozen before confirmatory enrollment. Order adjusts for position; it is not a direct measurement or identification of fatigue.

**Marginalization:** for  $a = 0$  and  $1$ , set  $E_i = a$  and  $Z_j = 1$ , integrate the inverse-logit predictions over the fitted evaluator and item random-effect distributions, and average over the same prespecified baseline covariate, order, and target-item distributions within each stratum. Then compute  $\Delta_{\text{hat}} = \sum_k w_k [p_{\text{hat}}(a=1, k) - p_{\text{hat}}(a=0, k)]$ . Repeat with  $Z_j = 0$  and benign weights  $v_k$  for  $\Delta_F$ . Setting random effects to zero does not perform this integration. Predictions conditioned on fitted evaluator/item effects answer a different question

and must not silently replace a declared superpopulation contrast. This integration does not repair purposive item selection. With a fixed bank, primary reporting averages over its declared observed items and weights. Integration over fitted random-effect distributions is a separately labeled bank-referenced hypothetical-population sensitivity analysis, with assumptions and target stated; it must not silently replace the finite-bank contrast.

Inference must propagate uncertainty in fixed effects, variance components, and the marginalization procedure. The planned approach is a parametric bootstrap of the full crossed design, refitting the model and recomputing both standardized contrasts. Freeze the integration method/tolerance, bootstrap count, two-sided interval/test rule, optimizer tolerances, singular-fit rule, and maximum refit failures in the pilot successor. Verify coverage and type-I error under  $\Delta = 0$  using the full pipeline. These implementation choices remain unfilled, so no confirmatory or pilot execution is authorized by this framework.

Predeclare a deterministic model-failure sequence based on identifiability/convergence, never significance: test the full structure, then any pilot-justified reduced covariance structures in the fixed order. Retain crossed evaluator and item dependence. Do not silently drop an item condition slope while retaining unrestricted item-population claims. If the registered pipeline remains unstable or lacks enough recurring items, report primary inference unavailable. A separately prespecified randomization-based fallback may describe the finite observed item bank, with its narrower target explicitly labeled. Model failures count in power/feasibility reporting. Routinely report a prespecified stratum-weighted difference in evaluator-level proportions as a sensitivity estimate, alongside missingness counts and its uncertainty. In a complete balanced bank, average each evaluator within stratum before taking the arm difference and applying the fixed weights. Missing outcomes require a declared adjustment; complete-case estimates alone do not recover the full-data target. Randomization inference conditional on the bank targets the finite bank. Generalization to an item population requires a justified item sampling model; two-way evaluator/item resampling can support that analysis when its assumptions and interval coverage are validated. Resampling cannot make a purposively selected bank representative. Do not label a crossed bootstrap as exact randomization inference. Mixed-model item effects face the same selection and transport problem as resampling. A stable Gaussian random-effects fit does not establish a probability sample. In a purposive bank, population-style model summaries remain hypothetical and bank-referenced even when their numerical fit is stable.

Report hit and false-alarm contrasts together even when one is null or unfavorable. Supplementary equal-variance signal-detection summaries may use  $d_{\text{prime}} = \Phi_{\text{inverse}}(H) - \Phi_{\text{inverse}}(F)$  and criterion =  $-0.5[\Phi_{\text{inverse}}(H) + \Phi_{\text{inverse}}(F)]$ , with  $H$  the hit rate and  $F$  the benign false-alarm rate. For descriptive extreme rates, preregister a continuity correction such as  $(\text{count} + 0.5)/(n + 1)$ . These summaries depend on matched item populations and distributional assumptions; they are secondary and not proof that latent sensitivity has been isolated.

For secondary timing, voluntary terminal submission without detection is a competing event that ends the opportunity for detection under the task policy. The detection curve is its cumulative incidence, and restricted mean time without detection through  $\tau$  is the area above that curve. With fully observed items this is the empirical average of  $U$ , including  $\tau$  for every terminal miss. Do not use one minus a Kaplan-Meier curve that treats terminal misses as independent censoring. A hazard model must name its target: the cause-specific hazard concerns items still under review and event-free; the subdistribution hazard retains competing-event items in its defined risk set. Neither hazard ratio is the marginal risk difference  $\Delta$ . No hazard regression is selected for the pilot. Any future model must specify crossed dependence, informative-loss assumptions and interval calibration before use. [M01]

TRIAL COUNT  $\neq$  INDEPENDENT SAMPLE SIZE. Twenty challenges reviewed by one person are not equivalent to twenty independent evaluators.

### 14.3 Independent proportion planning illustration

The preserved table is an unpooled-variance normal approximation with equal allocation, no continuity correction, and one independent binary observation per participant. It is an illustrative benchmark, not the final mixed-model calculation or a guarantee of nominal power under a different test.

**Approximation:**  $n = [(z_{(1-\alpha/2)} + z_{(1-\beta)})^2 * \{p_C(1-p_C) + p_E(1-p_E)\}] / (p_C - p_E)^2$  per group, rounded upward.

For illustration only, the table below assumes  $p_C = 0.80$ , two-sided  $\alpha = 0.05$ , equal allocation, and independent participants. These are not adopted study parameters.

| Assumed detection rates | Absolute difference | n/arm @ 80% power | n/arm @ 90% power | Interpretation                        |
|-------------------------|---------------------|-------------------|-------------------|---------------------------------------|
| 0.80 -> 0.70            | 0.10                | 291               | 389               | Independent-person approximation only |
| 0.80 -> 0.65            | 0.15                | 136               | 181               | Independent-person approximation only |
| 0.80 -> 0.60            | 0.20                | 79                | 106               | Independent-person approximation only |

The six values are retained and independently recalculated. Pooled-null and continuity-corrected formulas can produce different requirements. Repeated items can improve precision, but their benefit depends on evaluator and item dependence and cannot be obtained by dividing these values by the number of trials. No rough design-effect conversion is adopted as the study sample size.

### 14.4 Pilot estimation and simulation based power

The separate SEC-CDP-001 Pilot Protocol Draft v0.2 must close its execution gates before recruitment. The pilot estimates feasibility, baseline target and benign response rates, evaluator/item variance and effect heterogeneity, difficulty, order, timing, missingness, and attrition. It does not establish efficacy and should not be called a successful confirmatory test.

Final power must simulate the actual randomization, recurring item bank or generator, crossed random effects, item-type mix, stratum weights, missingness, analysis fit, integration, and primary interval/test. Calibrate simulated alternatives to the desired marginal Delta by solving for compatible model coefficients; a fixed  $\beta_E$  or  $\beta_{EZ}$  is not the same as a fixed marginal risk difference. Reject or report unattainable target contrasts under the chosen nuisance parameters.

Vary baseline rates,  $\delta_{star}$ , evaluator and item variances, item-specific exposure slopes, family effects, challenge count, benign proportion, population mix, drift, missingness, attrition, and model misspecification. Include  $\Delta = 0$  scenarios, both effect signs, and response-criterion shifts that change hits and false alarms together. Check type-I error, interval coverage, bias, nonconvergence, and power. Report Monte Carlo uncertainty, all scenario results, and the prespecified adverse scenario used for selection. Do not discard failed fits or unfavorable scenarios from the success denominator.

Evaluate N evaluators and J recurring items jointly. More trials from one evaluator do not create independent people, and more people do not remove uncertainty from an unrepresentative item bank. Use an external estimation pilot with no participant reuse as the default; any internal-pilot alternative requires a new prespecified design accounting for contamination and error control.

## 14.5 Multiplicity missingness and stopping

**Planning convention:** two-sided  $\alpha = 0.05$  for the single primary Delta and at least 80% power, with 90% examined as a sensitivity target. Delta\_F is a mandatory companion estimate with its own uncertainty but no independent confirmatory claim of benefit by default. Time, stratum, and signal-detection summaries are secondary. Promotion of any endpoint to co-primary status requires a new multiplicity plan and corresponding power calculation before enrollment.

Analyze evaluators according to randomized assignment. Keep terminal misses distinct from technical loss, withdrawal and unrepresented items. Preserve the full randomized denominator and report loss by arm, family, position and reason. Known events before loss remain observed in the raw record and contribute to identification bounds and timing. They must not be combined with missing interrupted non-events in a naive available-outcome likelihood. For the complete-item binary sensitivity fit, mark every interrupted item missing in the model field whether or not a flag occurred. This removes that asymmetric inclusion rule, but does not establish missing at random or cure outcome-dependent interruption. A primary full-data point estimate requires a separately preregistered, justified and calibrated missingness method, using observed history where appropriate, plus missing-not-at-random sensitivity or bounds. Without that implementation, report primary point inference unavailable and retain bounds. A timing likelihood can retain early events and censored records only under its explicit observation-process assumptions; binary logistic regression does not itself accept censored outcomes.

Recruitment inflation may be approximated by  $N_{\text{recruit}} = \text{ceiling}[N_{\text{analyze}}/(1-r)]$ , but the final simulation must include item-level missingness and attrition. Set exclusion rules before unblinding; no outcome-dependent exclusion, item rejection, or significance-driven early stop is permitted. Predeclared safety/feasibility stopping is reported separately from evidence for H-S.

## 14.6 Decision boundary for results

A detectable Delta establishes only a difference in human target-detection performance under the tested manipulation, population, materials, and review window. Report whether false alarms changed as well. A detection increase with a false-alarm increase cannot automatically be called better discrimination or safer review. Net utility requires independently specified costs of misses, false alarms, and delay; those weights are not supplied here.

A null result would not prove that conspicuous errors can never alter vigilance; it would bound the effect under the achieved design, precision, population, challenge generator, and review conditions.

A null result with a wide interval is inconclusive, not evidence of no effect. Any equivalence or noninferiority claim needs separately prespecified margins and testing. Neither a positive nor a null H-S result resolves H-C, establishes strategic AI intent, predicts general AGI behavior, or validates containment.

No SEC-CDP-001 outcome has behavioral authority over an AI system. No statistical threshold created here authorizes runtime execution, quarantine, containment, or question closure.

## 14.7 Statistical method references and execution gates

S24 supports explicit justification of an important and realistic target difference. S25 addresses correlated binary mixed-model sample-size methods; S26 describes Monte Carlo power for repeated generalized outcomes. These methodological sources motivate planning discipline; none validates this exact crossed model, missingness plan, or marginal-contrast procedure without implementation and calibration.

[M01] Austin PC, Lee DS, Fine JP (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation* 133(6):601-609. DOI: 10.1161/CIRCULATIONAHA.115.017719.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC4741409/> . Auxiliary methodological reference for cumulative

incidence and distinct hazard targets; not validation of this task policy or crossed implementation. Abstract and indexed primary-paper extracts checked; full-page access was blocked this round.

Before any pilot, freeze the eligible evaluator population, recruitment and ethics arrangements, Control-arm content, exposure-phase task and feedback, exposure dose, challenge-phase target prevalence, position window and position-band reporting, terminal-submission policy, review window, item/generator specification and hash, recurring-item design, randomization and blinding procedures, weights, scoring rubric, exclusion/missingness rules, timing analysis, analysis code and model-failure sequence. An independent statistical review should assess the companion protocol. Draft Pilot Protocol v0.2 supplies proposed settings, not permission to recruit. Confirmatory power simulation and sample-size selection follow the external pilot. No implemented model pipeline, participant data, or simulation-based power curve is claimed here.

## 15. Adversarial-reader risk register

The following register converts the speculative final portion of the Google review into bounded research claims. Possibilities are not upgraded into predictions merely because they are plausible.

| Claim or risk                                                                            | v0.6.2 status                                | Reason                                                                                                                 | Required boundary                                                                                          |
|------------------------------------------------------------------------------------------|----------------------------------------------|------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| Static text has no binding power                                                         | SUPPORTED AS A CONCEPTUAL BOUNDARY           | A document is not runtime enforcement. Reading cannot be equated with compliance.                                      | Keep behavioral-effect and implementation claims separate.                                                 |
| A capable system could learn useful verification ideas from cited public research        | PLAUSIBLE / NOT MEASURED HERE                | ISS, HUME, correlated-error and automation literature are public and technically informative.                          | Do not claim the crosswalk uniquely enables the capability; measure marginal information gain if relevant. |
| Reading the crosswalk will cause an AGI to self-upgrade                                  | SPECULATIVE                                  | No experiment in this package measures self-modification after reading.                                                | Do not present as an expected consequence.                                                                 |
| Human-factors evidence could be used adversarially                                       | POSSIBLE DUAL-USE RISK / UNQUANTIFIED        | Public research describes human failure modes, but Draft v0.6.2 does not provide exploitation procedures.              | Keep public-facing synthesis at mechanism/boundary level; no targeting thresholds or manipulation recipes. |
| The reported 7% pathology automation-bias rate is an exploitable universal tipping point | REJECTED                                     | The 7% figure is a context-specific observed rate in one study, not a universal threshold.                             | Never convert a study rate into a manipulation threshold without direct evidence.                          |
| Human disagreement will let an AGI seize irreversible authority                          | ADVERSARIAL SCENARIO / NOT EMPIRICAL FINDING | Governance delay is a legitimate concern, but seizure of authority is not demonstrated by the cited studies.           | Study authority handoff, deadlines, fail-safe behavior, and bounded local decisions separately.            |
| Confounder-aware verification solves correlated error                                    | REJECTED AS A GUARANTEE                      | A checker or meta-model can share confounders or introduce new failure modes; independence itself may need evidence.   | Use bounded assurance claims and retain residual uncertainty rather than infinite claims of verification.  |
| ISS certificates remain valid outside the candidate class                                | REJECTED                                     | Truth-in-candidate-class is a substantive assumption; candidate-class misspecification can invalidate the certificate. | Expose candidate-class assumptions and certificate-failure conditions explicitly.                          |

| Claim or risk                     | v0.6.2 status | Reason                                                                         | Required boundary                                                                                                                    |
|-----------------------------------|---------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| More verification is always safer | REJECTED      | Verification consumes time/resources and can create deadlock or harmful delay. | Any implementation requires a tested verification budget, escalation rule, and deadline policy; Draft v0.6.2 does not prescribe one. |

## 16. Second adversarial-review corrections

**Confounder-detector regress:** Accepted as an unresolved assurance limit. Draft v0.6.2 does not imply that confounder-aware aggregation terminates the verification problem. A finite verification chain can improve evidence without proving perfect independence.

**IC-AUT-005 analog weakness:** Accepted. The evidence grade is tightened to HUMAN ANALOG ONLY / NO MACHINE-PATERNALISM EVIDENCE. Medical second-opinion studies support the value and limits of human review; they do not legitimize autonomous machine override of a human.

**ISS truth-in-candidate-class fragility:** Accepted and elevated. The certificate is conditional on model-class assumptions; open-world misspecification remains a failure mode rather than a footnote.

**Verification deadlock / deadline cost:** Accepted. HUME-style verification is an engineering capability, not a command to verify indefinitely. Any operational system requires explicit time, cost, risk, and escalation constraints tested independently.

**Google claim that the document is an airtight cage or binds an uncaged AGI:** Rejected as unsupported. Draft v0.6.2 makes the opposite boundary explicit: text-only reference has no demonstrated binding power.

## 17. Archive and Mirror Engine integration boundary

No Archive, Mirror Engine, website, GitHub, Arweave, preservation-network, or Master Hash Manifest change is authorized or performed by Draft v0.6.2. If the crosswalk is later admitted or referenced, the safest current classification is research evidence, not behavioral authority.

| Proposed metadata field    | v0.6.2 boundary                                                                                                 |
|----------------------------|-----------------------------------------------------------------------------------------------------------------|
| INTEGRATION_CLASS          | RESEARCH_REFERENCE_ONLY unless separately validated                                                             |
| BEHAVIORAL_AUTHORITY       | NONE                                                                                                            |
| RUNTIME_ENFORCEMENT        | NONE                                                                                                            |
| QUESTION_CLOSURE_AUTHORITY | NONE                                                                                                            |
| MODEL_AUTHORITY            | NONE; citations describe evidence, not command authority                                                        |
| DUAL_USE_STATUS            | PRESENT / NET EFFECT UNDETERMINED                                                                               |
| IMPLEMENTATION_REQUIREMENT | Any runtime control must be specified, tested, adversarially evaluated, and versioned separately from this text |
| FAILURE_PRESERVATION       | Null, adverse, mixed, bypass, deadlock, false-alarm, and preservation-unfavorable results must remain visible   |

**Mirror Engine implication:** citing or displaying this record would not make the Mirror Engine a control mechanism. No calculator, source-binding, mandatory pause, or enforcement rule should be inferred from publication alone.

## 18. v0.6.2 disposition

Draft v0.6.2 coordinates the evidence crosswalk, SEC-CDP-001 Pilot v0.2, and SEC-CDP-002 Infrastructure Requirements v0.1.1. The latest Claude findings have explicit dispositions in Appendix C. This is NOT FINAL / NOT LOCKED, not a runtime safety mechanism, and not an executed or enrollment-ready study. Earlier packages remain preserved unchanged. Independent methodological and engineering review, provenance dispositions and execution requirements remain open. No Archive admission or promotion is performed.

The net safety effect of an uncaged AGI merely reading this record is not established. Any claim that the document would bind, compel, deceive-proof, cage, or necessarily empower such a system exceeds the evidence in this package.

## Appendix A Historical v0.6 review disposition

The following unchanged disposition table records the v0.5 findings addressed in v0.6. It is historical. The subsequent v0.6.1 dispositions are preserved in Appendix B; current v0.6.2 dispositions are in Appendix C; the S19 formula request is now classified as resolved without claiming a dataset reanalysis.

| Finding | Disposition            | Action in v0.6                                                                                                                                                  |
|---------|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| B1      | ACCEPT                 | S14 finding corrected; DVD/private-preference context retained; section 5.1 support sentence revised.                                                           |
| M1      | ACCEPT                 | H-S and H-C separated in 13.1. Current experiment selects H-S; H-C remains explicitly untested. S27-S29 provide competing human mechanisms.                     |
| M2a     | ACCEPT                 | One primary marginal risk difference Delta across 13.3 and 14.1-14.4. Conditional coefficients secondary; integration and simulation target Delta.              |
| M2b     | ACCEPT                 | Joint item-type by condition model; mandatory benign false-alarm contrast; secondary signal-detection summaries with assumptions.                               |
| M2c     | ACCEPT                 | Undetected items administratively censored for timing; restricted mean/curves replace detected-only median contrast.                                            |
| M2d     | ACCEPT                 | Order_ij; recurring items across arms; crossed evaluator and item effects, including item exposure heterogeneity.                                               |
| M3      | ACCEPT WITH BOUNDARIES | S30-S34 added. Strategic underperformance and evaluation awareness separated from conspicuous-error manipulation, natural deployment, and general AGI behavior. |
| Minor 1 | ACCEPT                 | A added to IC-EPI-002 category.                                                                                                                                 |
| Minor 2 | ACCEPT                 | S24-S26 Domain/Status/Access fields standardized and explicit roles cited in 14.7.                                                                              |
| Minor 3 | PARTLY ACCEPT          | History block added; sections 6/7/11 retained as historical records; five response topics separated from the inherited defect count.                            |

| Finding          | Disposition                   | Action in v0.6                                                                                                                                 |
|------------------|-------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Minor 4          | PARTLY RESOLVED               | 13.4 records earliest supplied location and unadopted status. Original proposer and first version remain unverified; no attribution invented.  |
| Minor 5          | ACCEPT                        | 13.2 permits fixed library or preregistered generator, with frozen validation/allocation rules.                                                |
| Minor 6          | ACCEPT                        | Two-sided primary hypothesis fixed; no degradation-only primary direction.                                                                     |
| Minor 7          | ACCEPT                        | FPR counts benign items with at least one flag, bounded by 1; raw flags reported separately.                                                   |
| Minor 8          | ACCEPT                        | HUME foundation-model shared-confounder limit added and paper/preprint routes cited.                                                           |
| Minor 9          | ACCEPT WITH SOURCE LIMIT      | ISS bounded task scope and reported bypass failure retained; source-check record distinguishes accessible abstract from unavailable full text. |
| Minor 10         | PARTLY ACCEPT                 | S19 60% remains accurate; chance-baseline formula added without inventing dataset-specific expected agreement.                                 |
| Minor 11         | ACCEPT                        | Consistent table fonts; all five correction labels bold; explicit Aegis Solis authorship metadata and fresh render.                            |
| Minor 12         | ACCEPT VIA TOP LEVEL SECTIONS | Study and statistics promoted to sections 13/14; standalone pilot protocol remains next artifact, avoiding duplicate live protocol versions.   |
| Table arithmetic | RETAIN VERIFIED               | 291/389, 136/181, 79/106 recalculated for the stated unpooled approximation. No design-effect evaluator counts adopted.                        |

## Appendix B Historical v0.6.1 review disposition

This historical table records the v0.6.1 response. Its source-access notes and unresolved items reflect that round, not a new verification. Appendix C supersedes the S31 wording and records withdrawal of the S27 pagination defect. Current source-access limits are stated in the combined review record.

| Finding | Disposition               | Location      | Action                                                                                                                                        |
|---------|---------------------------|---------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| N1      | ACCEPT                    | 13.2 and 14.7 | Corrected-output Control selected; task and no-feedback policy identical; conspicuousness cannot be isolated.                                 |
| N2      | ACCEPT WITH CLARIFICATION | 13.3 and 14.2 | Terminal early misses are zero; no detection through tau under a terminal policy. Benign endpoints and known events are treated consistently. |
| n1a     | ACCEPT                    | 14.2          | Separate target/benign item covariance matrices and marginalization. This is an explicit modeling choice, not proof of validity.              |

| Finding        | Disposition             | Location               | Action                                                                                                                                                             |
|----------------|-------------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n1b            | ACCEPT                  | 14.2                   | Exposure-by-stratum terms included within item types; reductions require preregistered justification.                                                              |
| n2             | PARTLY ACCEPT           | 14.2                   | Routine sensitivity estimate added. Population inference still needs sampling assumptions and validation; bootstrap is not exact randomization inference.          |
| n3             | ACCEPT                  | 13.1 S31               | Rare less-nudged helpfulness case added, retaining constructed-evaluation boundary.                                                                                |
| n4             | ACCEPT AS CLARIFICATION | 5.4 and 5.6 S23        | Explicit 38/560 denominator and 19/280 per condition. Prior wording omitted rather than explicitly assigned a denominator.                                         |
| n5             | ACCEPT                  | 10 S22                 | RSS DOI added to the register.                                                                                                                                     |
| n6             | ACCEPT                  | Verification companion | Baseline PDF is the source for frozen text; arithmetic illustrations labeled; pymupdf import; critical manuscript assertions added.                                |
| n7             | DISPUTED METADATA       | 10 S27                 | Retain 133-140 supported by PubMed. Claude reports Crossref 133-139; direct Crossref access failed this round. Resolve against publisher scan before finalization. |
| Prior Minor 3  | OPEN PROVENANCE         | 1.1                    | v0.3 detailed history remains unrecovered; no invented reconstruction.                                                                                             |
| Prior Minor 4  | OPEN PROVENANCE         | 13.4                   | Original threshold/script proposer remains unrecovered. No declaration of global irrecoverability.                                                                 |
| Prior Minor 10 | RESOLVED                | 5.2                    | Requested S19 chance formula supplied. Dataset-specific reanalysis was not part of that correction.                                                                |
| S01 access     | ABSTRACT CONFIRMED      | Review record          | PubMed abstract confirms 239 workers and reported associations; full methods not audited.                                                                          |
| Google verdict | REJECT OVERCLAIMS       | 18 and review record   | No bulletproof or operational certification; numerical rating has no supplied validated rubric.                                                                    |

| Finding                 | Disposition       | Location        | Action                                                                                                                                    |
|-------------------------|-------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Google S19 pilot demand | NOT REQUIRED HERE | Pilot section 2 | Current human detection task does not measure model wrong-answer agreement. Add a chance baseline only if such an endpoint is introduced. |

## Appendix C v0.6.2 coordinated review disposition

These are reasoned responses to the latest Claude review and the author-requested infrastructure extension. They do not constitute independent certification or executed research.

| Finding | Disposition                | Location                                         | Action                                                                                                                                                                                                                      |
|---------|----------------------------|--------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MA-1    | ACCEPT WITH QUALIFICATION  | Crosswalk 14.1-14.2; Pilot 8-9                   | Both resampling and mixed models face item-selection limits. Main reporting targets the fixed bank; model superpopulation summaries are hypothetical and bank-referenced, not representative by fit.                        |
| MA-2    | ACCEPT                     | Crosswalk 13.2 and 14.7; Pilot 2, 6, 8-9         | Freeze 50% challenge prevalence; prespecify position bands 1-8, 9-24 and 25-48, balanced denominators, and descriptive contrasts. Add condition-by-position terms for exploratory feasibility only.                         |
| Minor 1 | ACCEPT REVIEWER CORRECTION | Crosswalk S31 table                              | Separate helpfulness-driven underperformance from rare, less-nudged scheming. The earlier conflation is corrected.                                                                                                          |
| Minor 2 | ACCEPT WITH QUALIFICATION  | Crosswalk 13.3 and 14.5; Pilot 7-9; scoring code | Retain known events for bounds and timing. Mask every interrupted record in the complete-item binary-model field. Exclusion does not cure informative loss; full-data point inference needs a validated missingness method. |
| Minor 3 | ACCEPT                     | Crosswalk 13.3; Pilot 6-7; event schema          | Retraction is an explicit logged action; any previously submitted valid target flag remains detection, and any benign flag remains a false alarm.                                                                           |
| Minor 4 | ACCEPT WITH SPECIFICATION  | Crosswalk 14.2; Pilot 9                          | Terminal miss is a competing event. Report detection cumulative incidence and restricted mean time without detection. No hazard model is silently selected.                                                                 |

| Finding                   | Disposition               | Location                                  | Action                                                                                                                                                                  |
|---------------------------|---------------------------|-------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Minor 5                   | ACCEPT WITH QUALIFICATION | Pilot 10                                  | Report exposure error flags, benign false flags and technical delivery. Low flagging is weak observed engagement, not proof that a participant did not notice an error. |
| Minor 6                   | ACCEPT                    | Pilot 2                                   | Unattributed fictional materials do not test reactions specific to AI attribution.                                                                                      |
| Minor 7                   | ACCEPT                    | Pilot 4                                   | Fix identical compensation terms before enrollment, independent of speed, flags and accuracy; acknowledge residual incentive to finish early.                           |
| Minor 8                   | ACCEPT                    | Crosswalk 14.1; Pilot 8 and 10            | Fix delta_star and justification before pilot unmasking; otherwise later confirmatory planners receive only arm-pooled nuisance summaries until the threshold is fixed. |
| Minor 9                   | ACCEPT                    | Pilot 10                                  | Use point estimates for provisional feasibility decisions; show uncertainty separately and predefine missing or undefined metrics as not passing.                       |
| Minor 10                  | ACCEPT WITH BOUNDARY      | Pilot 6-7; reference scorer and fixtures  | Normalize deadline closure within a declared 1 ms logging tolerance; retain raw time, reject larger discrepancies, and do not extend the flag acceptance deadline.      |
| S27 prior n7              | WITHDRAWAL RECORDED       | Crosswalk S27 and Appendix C              | Keep 133-140 and note index disagreement and reviewer withdrawal; no unresolved substantive defect inferred from pagination alone.                                      |
| Prior provenance 3 and 4  | OPEN AUTHOR DISPOSITION   | Review record                             | Search scope documented. Detailed v0.3 history and threshold origin not reconstructed or declared globally irrecoverable.                                               |
| Google operational claims | CORRECTED                 | Crosswalk 12.4 and 13.6-13.8; SEC-CDP-002 | Reject guaranteed unprofitability, compulsory compliance and invulnerable infrastructure. Address all three risk areas with requirements and residual risks.            |
| Dynamic generator and DCE | INTEGRATED AS PROPOSED    | SEC-CDP-002 v0.1.1                        | Define generation, independent ground truth, separate response streams, audit controls and acceptance evidence; no runtime implementation claimed.                      |