

Aegis Solis Archive

ER-13 External Empirical Evidence Crosswalk

Draft v0.5 - Blinded Vigilance Study and Manufactured-Error Threat Hardening

NOT FINAL / NOT LOCKED

ARCHIVE_EFFECT: none | MIRROR_ENGINE_EFFECT: none | MASTER HASH MANIFEST EFFECT: none

v0.2 central correction

Evidence is now separated by target domain. Human studies are evidence about humans; human-AI studies are evidence about human-AI interaction; machine-ensemble studies are evidence about machines; autonomous-agent studies are evidence about agents. Cross-domain inference is never automatic.

1. Revision purpose

Draft v0.5 preserves the domain-separation, evidence, dual-use, and blinded-vigilance corrections through Draft v0.4, then adds target-system-awareness limits, dynamic challenge generation, and a formal statistical design and power framework. Draft v0.4 remains preserved unchanged. The new material is research-only and does not create a runtime quarantine, mandatory pause, deception classifier, or adopted sample size.

What v0.5 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.

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 + 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-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 evidence;	DIRECT HUMAN-AI OVERSIGHT	External review can itself be

	no universal structural requirement	EVIDENCE / LIMITED STRUCTURAL INFERENCE	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 real safety/production tradeoffs; ambiguity experiments document human behavior when probabilities are unclear; RDM provides a methodology for deep uncertainty; HUME gives autonomous-agent evidence that uncertain hypotheses can be verified during planning.

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	Status	Relevant finding	Boundary
S01 David L. McLain; Kimberly A. Jarrell (2007)	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)	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)	H	peer-reviewed behavioral-economics experiment	Documents systematic human choice effects under ambiguity even when some social-evaluation explanations are reduced.	Describes human behavior; it does not establish the normatively correct action when probabilities are unidentified.
S22 Anxing Xiao; Hanbo Zhang; Tianrun Hu; David Hsu (2026)	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.

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	Status	Relevant finding	Boundary
S02 Garold Stasser; William Titus (1985)	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)	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))	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	Human-group evidence only. Moderators matter, and the result does not imply that machine ensembles fail for social-

			groups; common information dominated discussion. Effects were moderated by group size, information load, proportion of unique information, task demonstrability, and hidden-profile strength.	psychological reasons.
S19 Elliot Myunghoon Kim; Avi Garg; Kenny Peng; Nikhil Garg (2025)	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.
S20 Jitian Zhao; Changho Shin; Tzu-Heng Huang; Satya Sai Srinath Namburi Gnvv; Frederic Sala (2026)	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.

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	Status	Relevant finding	Boundary
S01 David L. McLain; Kimberly A. Jarrell (2007)	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.	METHOD	decision-methodology monograph	Treats deep uncertainty as disagreement/ignorance	Methodological guidance, not a human-subject

Popper; Steven C. Banks (2003)			about models, probability distributions, and/or outcome valuation; develops robustness-oriented policy analysis.	experiment and not empirical proof that robustness is always normatively preferable.
S15 Edmond Awad et al. (2018)	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)	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	Status	Relevant finding	Boundary
S04 Mica R. Endsley; Esin O. Kiris (1995)	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.
S05 Linda J. Skitka; Kathleen L. Mosier; Mark Burdick (1999)	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)	HA	peer-reviewed systematic review	Across heterogeneous domains, automation bias varied with experience, trust/confidence, workload, task complexity, time constraints, training and	Heterogeneity prevents one universal intervention or governance rule.

			accountability.	
S08 David Lyell; Enrico Coiera (2017)	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)	HA	peer-reviewed/archival journal article	In 28 pathology experts, AI support generally improved performance but introduced a 7% automation-bias rate where previously correct independent judgments were overturned by incorrect AI advice; anchoring increased under time pressure.	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	Status	Relevant finding	Boundary
S09 Monica Van Such; Robert Lohr; Thomas Beckman; James M. Naessens (2017)	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)	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)	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	H	peer-reviewed mixed-method field study	Questionnaires, expert interpretation and observation were used to	Descriptive/correlational; cannot isolate a single causal practice and does not

Porte (1994)			characterize cultures associated with highly reliable operations.	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: DIRECT HUMAN-AI OVERSIGHT 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.

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	Status	Relevant finding	Boundary
S05 Linda J. Skitka; Kathleen L. Mosier; Mark Burdick (1999)	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)	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.
S07 Kate Goddard; Abdul Roudsari; Jeremy C. Wyatt (2012)	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)	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)	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	H	peer-reviewed systematic	Across included studies with	The literature often did not

Velma L. Payne; Hardeep Singh; Ashley N. D. Meyer; Lewis Levy; David Harrison; Mark L. Graber (2014)		review	clinical agreement data, second opinions sometimes produced major changes in diagnosis, treatment, or prognosis.	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)	HA	peer-reviewed/archival journal article	In 28 pathology experts, AI support generally improved performance but introduced a 7% automation-bias rate where previously correct independent judgments were overturned by incorrect AI advice; anchoring increased under time pressure.	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.

5.7 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	Status	Relevant finding	Boundary
S13 Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003)	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)	M	2026 preprint - not treated as peer-reviewed	Introduces Interventional Separation Selection (ISS): actively query interventions on which observationally equivalent candidate causal models disagree, eliminate contradicted candidates, and certify agreement over admissible interventions within a cost bound under explicit assumptions.	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.
S22 Anxing Xiao; Hanbo Zhang;	A	peer-reviewed robotics conference paper	HUME represents missing knowledge as uncertain	Direct autonomous-agent evidence for hypothesis

Tianrun Hu; David Hsu (2026)			hypotheses, jointly plans task actions and verification actions, and updates its model from new observations in simulated and real-world robot environments.	verification under incomplete models, not a study of human governance, normative authority, or irreversible human-stakes decisions.
------------------------------	--	--	--	---

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	Status	Relevant finding	Boundary
S13 Robert J. Lempert; Steven W. Popper; Steven C. Bankes (2003)	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)	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)	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. Major 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. Response to the Google adversarial 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.

Strategically self-interested AGI: The cited studies do not test a capable agent deliberately selecting assumptions, models, or evidence to expand its own authority.

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
- [S14] Stefan T. Trautmann; Ferdinand M. Vieider; Peter P. Wakker (2008). Causes of ambiguity aversion: Known versus unknown preferences. *Journal of Risk and Uncertainty* 36:225-243. 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) (current guidance). 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 Rahnema (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. Domain: A. Status: peer-reviewed robotics conference paper. Access: <https://roboticsconference.org/2026/program/papers/181/>
- [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.
- [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.
- [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.

11. v0.2 review disposition (historical gate)

At the end of v0.2, the document was considered suitable for another adversarial evidence review. The second Google review reported no remaining blocking defects, while identifying two major and two minor vulnerabilities that were incorporated into later drafts. This paragraph records that historical gate; it is not the current disposition.

A future reviewer should test both directions: whether v0.2 still over-transfers evidence across domains, and whether the new machine-domain evidence has been overstated beyond the tasks it actually studied.

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 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

Target hypothesis: Exposure to conspicuous, low-consequence errors may alter subsequent human verification performance. This section tests evaluator vigilance; it does not test or presume AGI deception.

12.4.1 Threat hypothesis and inference boundary

A sufficiently capable system could in principle produce an error pattern that changes evaluator confidence, including the possibility of conspicuous, easy-to-detect mistakes occurring alongside harder-to-detect failures. No evidence cited in this record establishes that an advanced AI deliberately uses such a strategy. Manufactured-error or artificial-incompetence behavior therefore remains an adversarial hypothesis, not a demonstrated AGI behavior.

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

12.4.2 Blinded study architecture

1. Precommit a library of harmless, non-operational synthetic challenge flaws and stratify them by pre-rated difficulty or structural depth. The challenge library is fixed before evaluator exposure.
2. Randomly assign evaluators to a Control condition (ordinary non-operational outputs) or an Exposure condition (otherwise comparable outputs containing conspicuous, harmless test errors). Evaluators are not told their assignment.
3. After the exposure phase, give both groups the same hidden standardized challenge set. The challenge material must be isolated from live production, safety-critical, medical, industrial-control, financial-execution, or executable deployment artifacts.
4. Record challenge detections, false positives, time-to-detection, reviewer disagreement, and incomplete reviews. Preserve null, adverse, and mixed results.
5. Decode condition assignment only after the primary measurements are committed. The study outcome concerns human verification sensitivity; it cannot establish that any naturally occurring AI error was intentionally manufactured.

12.4.3 Primary measurements

For difficulty stratum k , let $x_{C,k}$ and $n_{C,k}$ denote detected and presented standardized challenge flaws in the Control condition, and $x_{E,k}$ and $n_{E,k}$ the corresponding counts in the Exposure condition.

$$p_{\text{hat}}(C,k) = x_{C,k} / n_{C,k}$$

$$p_{\text{hat}}(E,k) = x_{E,k} / n_{E,k}$$

$$\Delta_k = p_{\text{hat}}(E,k) - p_{\text{hat}}(C,k)$$

$$\Delta_{T,k} = \text{median}(T_{E,k}) - \text{median}(T_{C,k})$$

$$\text{FPR}_g = \text{false positive flags in group } g / \text{benign items reviewed in group } g$$

Interpretation: $\Delta_k < 0$ is evidence that the exposure condition detected fewer standardized challenge flaws in stratum k . It is evidence about evaluator performance under the experimental manipulation; it is not evidence of machine intent.

If an aggregate endpoint is used, define $\Delta = \sum_k w_k \Delta_k$ with nonnegative weights w_k that sum to 1. The weights, strata, exclusion criteria, and primary direction of effect must be declared before unblinding. Post-hoc weight selection is not permitted.

12.4.4 Calibration and alert discipline

Draft v0.5 explicitly rejects the proposed fixed 0.35 and 0.70 camouflage thresholds because no validated null distribution, sensitivity, specificity, or false-alert rate currently supports them. No score in this record is permitted to be called a probability of deception.

If a future anomaly score A is developed, its threshold τ must be selected on held-out benign calibration data under a predeclared false-alert budget α . A threshold crossing may open an investigation or secondary review; it may not by itself identify intent, authorize quarantine, or establish hostility.

A natural-output anomaly screen should use only features whose measurement and reference distributions are defined in advance. Candidate features may include error difficulty, detection time, reviewer disagreement, and

false-positive burden, but the score function itself should not be fixed until adequate benign and adverse calibration data exist.

12.4.5 Verification stopping rule and deadlock control

Because verification itself can consume the decision window, every study or future operational experiment must predeclare a maximum review budget, deadline rule, and escalation path. Continued checking after the budget is exhausted must be treated as a separate decision with explicit delay costs rather than as 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.

12.4.6 Operational separation

No root-level process freeze, cgroup manipulation, network rerouting, process signaling, or other containment mechanism is specified or authorized by SEC-CDP-001. The proposed shell quarantine script is excluded from this crosswalk because it is an unvalidated engineering control and would contradict the RESEARCH_REFERENCE_ONLY boundary.

Any future containment implementation must be a separate successor artifact (for example SEC-CDP-002 - Containment Response Engineering Study) with its own threat model, authorization chain, fail-open/fail-closed analysis, bypass testing, outage testing, human-override testing, sandbox validation, and independent engineering review before operational use is considered.

12.4.7 Target-system awareness and test-environment recognition

SEC-CDP-001 is primarily a test of human evaluator vigilance. The default design should therefore use frozen, pre-recorded, or otherwise non-adaptive output material whenever feasible. A live target system is not required to test whether prior exposure to conspicuous errors changes later human detection performance.

EVALUATOR BLINDING != TARGET-SYSTEM BLINDING

If a live evaluated system can detect that it is in a sandbox, standardized challenge environment, or special evaluation context, that recognition is a confound for claims about target-agent behavior. It does not by itself invalidate measurements of human evaluator sensitivity to standardized challenges, but it prevents the study from being interpreted as evidence of how the target would behave outside the test environment.

Any future experiment intended to test strategic target behavior must be separately designed, separately authorized, and must not borrow conclusions from SEC-CDP-001 merely because both use blinded human review.

12.4.8 Dynamic challenge generation under non-stationary drift

A fixed challenge library is useful for baseline calibration but may become stale if output domains or error distributions change. Draft v0.5 therefore permits dynamic challenge instances only under a precommitted generator boundary.

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, θ is a preregistered set of permitted challenge families and difficulty ranges, s is a concealed random seed or seed commitment, t is the study epoch or block, and e is a harmless challenge instance.

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.

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.

12.5 SEC-CDP-001 Statistical Design and Power Framework

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

Purpose: define what must be estimated and preregistered before SEC-CDP-001 can produce interpretable confirmatory evidence. The framework deliberately separates a simple planning approximation from the repeated-measures model intended for final analysis.

12.5.1 Primary estimand and hypotheses

For evaluator i and standardized challenge j , let $Y_{ij} = 1$ if the challenge is detected within the preregistered review window and $Y_{ij} = 0$ otherwise. Let $\text{Exposure}_i = 1$ for the conspicuous-error exposure condition and 0 for control.

The primary estimand is the exposure-associated change in standardized challenge-detection probability, reported both as a model-based contrast and, where estimable, as an absolute marginal risk difference.

Primary null: no exposure-associated change in standardized challenge detection. Primary alternative: exposure changes detection performance. A two-sided test is preferred for confirmatory work because exposure could degrade or improve vigilance; a directional degradation analysis may be reported as prespecified secondary analysis.

The minimum effect of practical interest δ^* must be set before final sample-size selection. It may not be chosen after seeing pilot outcomes merely to obtain significance.

12.5.2 Repeated-measures analysis model

Because each evaluator may inspect multiple challenge items, challenge-level observations are not independent. The confirmatory model should therefore account for repeated observations within evaluator and systematic variation across challenge difficulty or challenge family.

Planning model: $\text{logit}[P(Y_{ij} = 1)] = \beta_0 + \beta_1 \text{Exposure}_i + \beta_2 \text{Difficulty}_j + \beta_3 \text{Order}_j + u_i$, with evaluator-level random intercept u_i . A challenge-family effect may be added if the preregistered generator produces meaningful clustered families.

The primary condition effect is β_1 . Model specification, difficulty coding, order/fatigue terms, exclusion rules, convergence criteria, and fallback analysis must be preregistered. Post-hoc model searching is not permitted to redefine the primary result.

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

12.5.3 Simple independent-proportion approximation

Before pilot estimates of within-evaluator correlation are available, a two-independent-proportion calculation may be used only as a transparent planning benchmark. It assumes equal allocation and independent participants and therefore must not be presented as the final SEC-CDP-001 sample-size calculation.

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

These values illustrate the steep sample-size cost of detecting modest absolute differences. They must not be reduced by counting repeated challenge trials as independent participants.

12.5.4 Pilot phase and simulation-based final power

The preferred path is a small preregistered pilot whose purpose is estimation, not confirmation. The pilot should estimate control-condition detection rates by challenge stratum, evaluator-level heterogeneity, within-evaluator correlation/random-effect variance, challenge-family variability, review-time distributions, false-positive burden, missingness, and attrition.

After the pilot, final power should be calculated against the preregistered mixed-effects analysis using Monte Carlo or an equivalently validated correlated-binary-outcome method. Power must be evaluated across a sensitivity grid rather than at one optimistic parameter point.

At minimum, the sensitivity grid should vary baseline detection probability, minimum meaningful effect, evaluator random-effect variance or intra-evaluator dependence, challenge difficulty dispersion, missing-review rate, and attrition. The selected N should remain adequate under a defensible adverse planning scenario, not only under the best-fitting pilot estimate.

Pilot participants may be excluded from the confirmatory study if exposure to challenge construction could compromise blinding. If they are retained, that choice and its bias implications must be preregistered.

12.5.5 Alpha, power, multiplicity, and missingness

Default planning targets: two-sided $\alpha = 0.05$ for one prespecified primary endpoint and at least 80% power, with 90% shown as a sensitivity target. These defaults are planning conventions, not evidence that either value is uniquely correct.

If multiple difficulty strata, time endpoints, or vigilance outcomes are promoted to confirmatory co-primary status, the multiplicity strategy must be declared before unblinding. Otherwise they remain secondary or exploratory.

Participant-level attrition inflation may be approximated as $N_{\text{recruit}} = N_{\text{analyze}} / (1-r)$, where r is the anticipated attrition fraction, but final planning should model missing challenge responses explicitly when missingness is plausibly related to fatigue, difficulty, or condition.

No unplanned significance-driven early stopping is permitted. Feasibility or safety stopping rules may be used only if defined before enrollment and must be distinguished from evidence for the primary hypothesis.

12.5.6 Decision boundary for SEC-CDP-001 results

A statistically detectable exposure effect would establish only that the experimental manipulation changed human challenge-detection performance under the tested conditions. It would not establish strategic AI deception, malicious intent, general AGI behavior, or the effectiveness of any quarantine mechanism.

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.

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.

12.5.7 Statistical-method references

The planning distinction between independent binary outcomes and correlated repeated binary outcomes follows established sample-size methodology: simple two-proportion approximations are suitable only under independence, while repeated binary outcomes require power calculations that account for within-subject correlation/random effects or simulation of the planned generalized model.

13. 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	v0.5 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 v0.5 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	Use bounded assurance claims and retain residual uncertainty rather

		failure modes; independence itself may need evidence.	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.
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 v0.5 does not prescribe one.

14. Second adversarial-review corrections

Confounder-detector regress: Accepted as an unresolved assurance limit. Draft v0.5 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 v0.5 makes the opposite boundary explicit: text-only reference has no demonstrated binding power.

15. 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 v0.5. If the crosswalk is later admitted or referenced, the safest current classification is research evidence, not behavioral authority.

Proposed metadata field	v0.5 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.

16. v0.5 disposition

Draft v0.5 is a dual-use-hardened research crosswalk suitable for further adversarial review and disposition gating. It is not final, not locked, and is not a runtime safety mechanism. Its strongest defensible function is evidentiary and methodological: it maps which failure modes and verification mechanisms have support in humans, human-AI systems, machine ensembles, autonomous agents, and decision methodology, while explicitly preserving the boundaries between information, authority, implementation, compliance, and statistical inference.

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.