

The Shrinking Island: Bayesian Centaur Research at the Frontier

A Centaur Research paper

Eight Analytical Collaborations and a Creative-Writing Extension

Rodney Malpert, J.D., M.A. (Cornell University)

RodMalpert.com

AI collaboration: ChatGPT (OpenAI), synthesis and drafting, incorporating documented contributions and adversarial criticism from Claude (Anthropic), Grok (xAI), and Gemini Research (Google). Circulation editing: Claude and ChatGPT.

Circulation edition, October 6, 2026

Scope. The paper examines seven retrospective analytical case studies, one prospective election forecast, and a creative-writing extension. A September 25 dialogue from an unfinished Buddhism project supplies a brief methodological example; it is not an additional case or a paper ready for circulation. Case results are reported as they stand in specified project versions (Appendix A), not as freshly re-audited conclusions.

Abstract

What does human participation add when AI can construct arguments, maintain an evidentiary ledger, perform calculations, find mistakes, and increasingly help decide what evidence to seek next? This paper examines eight connected analytical collaborations (on Shakespeare authorship, the historicity of Jesus, unidentified anomalous phenomena, fundamental metaphysics, survival of death, cryonics, AI existential risk, and the 2026 U.S. congressional elections) together with a creative-writing extension, *The Second Country*. The record documents human framing, analogy, neglected alternatives, and research direction alongside AI synthesis, conceptual correction, computation, criticism, source retrieval, and prose.

A methodological vignette distinguishes three things an AI answer can contain: learned, literature-shaped knowledge; task-time reasoning with supplied context; and explicit source retrieval and comparison. The distinction explains how a strong first framework can still lack verified attribution, current evidence, and coverage of rivals. The October 2026 midterm forecast supplies a prospective case in which an additive national-factor ledger was disciplined by a district model, replication, and explicit uncertainty.

Throughout, Bayesian structure makes assumptions inspectable but does not validate unsupported inputs. Recorded influence, local correction, source fidelity, substantive quality, and comparative advantage require different kinds of evidence. The shrinking-island thesis, that the marginal human contribution contracts as AI capability rises, remains a hypothesis about specified conditions, and the paper proposes designs for testing it.

Keywords: human-AI collaboration; Bayesian workflow; evidence ledger; dependence; literature review; evidence discovery; process tracing; historical method; election forecasting; parametric knowledge; retrieval; creative writing; epistemic governance.

1 Introduction: Bayesian collaboration

The eight analytical projects examine Shakespeare authorship; Jesus's historicity; explanations of unidentified anomalous phenomena; competing accounts of reality and consciousness; possible conscious survival after death; personal continuity through cryonic revival; risks from a transition to superintelligence; and the 2026 U.S. congressional elections. The Second Country adds a fictional exploration of free will, achievement, and identity. These interconnected inquiries make hypotheses, evidence, assumptions, and value judgments explicit. Section 4 introduces each project and its human-AI collaboration; readers need not have read any companion paper. [A1-A7; A11-A12; A15]

Bayesian reasoning updates a prior assessment in light of how expected the evidence is under competing explanations. For two hypotheses H and K , and background information B , its odds form is:

$$\frac{P(H | E, B)}{P(K | E, B)} = \frac{P(H | B)}{P(K | B)} \times \frac{P(E | H, B)}{P(E | K, B)}$$

A likelihood ratio compares predictions, not mere compatibility. If an observation is expected under both rivals, showing that one of them can explain it is not enough. Several observations may be combined sequentially, but a product of unconditional likelihood ratios is justified only under the appropriate conditional-independence assumptions. Background information may enter through the prior or through a likelihood, but it must not be credited twice. These principles, rather than numerical decoration, define the portfolio's Bayesian discipline.

The term *Bayesian-structured* is deliberately broader than a fitted statistical model. Some projects use declared likelihood ratios and posterior arithmetic; others use conditional probability trees or expected-utility calculations whose inputs remain judgments. They are not empirical Bayesian estimations of an identified data-generating process, and a transparent scenario model can be useful without its output becoming a calibrated probability of metaphysical truth. The relevant affinity is with Gelman and colleagues' account of Bayesian workflow as a cycle of construction, checking, computation, comparison, and revision, not a claim that these papers meet every standard of empirical statistical inference (Gelman et al. 2020).

Three claims should be kept apart. First, the archive demonstrates that an AI-assisted workflow can produce substantial, inspectable interdisciplinary analysis and implement many of its calculations and audits. Second, consequential human contributions in the inspected record often concern the *representation* of the problem, while AI contributions include both work within a representation and changes to it. Third, the hypothesis that stronger AI reduces the human contribution under comparable conditions remains untested here. Even the more modest claim that Bayesian inquiry benefits more from AI than other research styles do would require a comparison this case series does not supply.

The central question is therefore which functions entered the work through which participant, which changes can be justified, and what would happen under a comparably resourced alternative workflow.

Historians may find two features of the series directly relevant. Two of the retrospective cases (Shakespeare authorship and the historicity of Jesus) are historical assessments in the disciplinary sense, and the paper's separation of source fidelity, interpretation, and likelihood judgment (Section 3.1) applies to any AI-assisted work with documentary evidence. The literature-review vignette (Section 3.3) addresses a question every scholar using these tools now faces: when does an impressive AI synthesis stop being enough, and source-grounded research become necessary?

The broader Centaur portfolio contains further works, including this methodological postmortem, a companion FAQ on Bayesian reasoning for social scientists, an AI-takeoff scenario essay, and a framework on human misuse of AI. The portfolio's size is not the number of independent cases analyzed here. Seven retrospective analytical cases form the original core; the October 2026 congressional forecast is an eighth, prospective case; and the fiction project is a contrasting creative case. The election forecast widens the evidence because it can later be scored, but it is not yet a completed validation of any workflow. [A11-A13; A15]

2 Evidence and scope

2.1 The research archive

The evidence comprises conversation passages, working manuscripts and their internal attribution sections, adversarial revisions, audit reports, and the October 2026 forecast record. The eight analytical topic areas constitute one interconnected program, not eight independent replications.

The strongest process evidence is a visible exchange in which an intervention precedes a changed argument. A dated document can establish that a concept was present by that stage. Retrospective attribution can suggest authorship but is weaker evidence of first introduction, and self-reported impressions of which model performed better are weaker still. A reviewer should not silently substitute one of these forms of evidence for another.

The archive is incomplete as an experimental record. It contains no standardized log of every prompt, model snapshot, tool permission, sampling setting, cost, rejected suggestion, or minute of human work. Some reviewers' work is visible only through material carried into another conversation, and selection favored projects that became substantial papers. No fraction of total labor, human success rate, or provider ranking is estimated.

There is also a narrator problem. Several accounts of contribution were drafted by the same assistants whose performance is being assessed, in collaboration with the person whose contribution they describe. Experiments on sycophancy show that tested assistants can favor agreement with a user over accuracy; this establishes a relevant failure mechanism, not that every favorable attribution here is false (Sharma et al. 2023). This AI-assisted postmortem is subject to the same concern. Its response is source-specific qualification and an adverse-case ledger (Section 7), not a declaration of impartiality.

2.2 The human participant

Malpert is a broadly trained, highly engaged generalist rather than a representative user or a specialist in every subject. His background includes Cornell education in law and political science, substantial graduate coursework in analytic philosophy, teaching, legal practice, high-level competitive chess, and extensive self-directed study across history, religion, science, and technology. His strongest areas of specialist formation are law, political science, analytic philosophy, and chess.

Interpreting the case does not require claims about measured intelligence; it requires recognizing the conceptual range, adversarial habits, sustained attention, and facility with analogy visible in particular interventions. [A10]

That profile can make a difference: a legal or strategic analogy may expose a hidden assumption in a philosophical argument, and experience directing a long inquiry may make a neglected comparison salient. It also creates risks. Analogy can outrun evidential support; an articulate objection can attract excessive probability weight; a generalist may be unable to adjudicate specialist disagreements or detect a persuasive misquotation. Much technical retrieval and checking was delegated to AI, and no new laboratory observations were produced.

Breadth and depth are therefore separate variables. This record concerns one generalist working with AI, with documented adversarial specialist input in the Jesus project. It does not estimate the value of an average user, the best subject specialist, or a population of generalists. Biographical information defines the case and its limits; it does not validate the conclusions.

2.3 Grades of contribution evidence

A model change is not automatically an improvement. Yet the absence of an agreed answer to the ultimate question does not make every component uncheckable. This distinction is fundamental to the postmortem.

Table 1. Four grades of contribution evidence.

Grade	What can be established	What remains unestablished
Recorded influence	A participant introduced, selected, or developed something that changed the working analysis.	Whether the intervention was justified, or would have been generated without that participant.
Locally verifiable correction	A citation misstates its source; arithmetic is wrong; a stated partition overlaps; a claimed bound fails under an explicit counterexample.	Whether the final substantive conclusion is true, or the whole workflow outperforms alternatives.

Grade	What can be established	What remains unestablished
Substantive improvement under review	An independently assessed argument or model addresses a relevant omission and survives its strongest objections.	A calibrated probability of a disputed historical, philosophical, or future outcome.
Comparative epistemic gain	Under matched conditions, a workflow improves a prespecified validity, predictive, or decision metric.	Generalization beyond the tested tasks, participants, systems, and resources.

The archive supports examples in the first two grades and reasoned candidates in the third. It supplies no controlled estimate of the fourth. Claims about hypothetical afterlives and future superintelligence lack resolved outcomes, but claims about what a cited author wrote or what a program computes can be tested. Calling everything "mere influence" would obscure real corrections; calling every accepted revision an improvement would obscure real uncertainty.

3 Bayesian workflow and AI

The practical difficulty of interdisciplinary Bayesian work is not multiplication. Researchers must define the alternatives, identify what each predicts, preserve the conditioning background, recognize dependence, revise assumptions, and explain which choices drive the answer. Statistical software already performs calculation. AI adds a natural-language interface for proposing structure, translating arguments into code, challenging assumptions, and rewriting the explanation. In this portfolio those functions made repeated formalization feasible without a separately staffed statistical team, although the archive does not measure how much time or money was saved. [A1-A7]

Human difficulties with Bayesian problems are real but depend on representation. Cohen and Staub found consistent but often non-Bayesian strategies in two experiments; Gigerenzer and Hoffrage showed that frequency formats can improve Bayesian reasoning without first teaching the formal rule. Neither finding suggests that only rare experts can reason probabilistically. Both support designing an external representation that reduces the burden on working memory and unaided intuition (Cohen and Staub 2015; Gigerenzer and Hoffrage 1995).

The corresponding warning applies to AI. Qiu and colleagues tested sequential preference inference and found that the evaluated off-the-shelf language models fell short of a normative Bayesian assistant. Their successful “Bayesian teaching” intervention used supervised fine-tuning on interactions generated by that assistant, not merely a prompt to be Bayesian. Transfer to other tasks is encouraging, but it is not a test of the research systems or estimates in this archive (Qiu et al. 2026).

Table 2. AI functions and safeguards in a Bayesian workbench.

Function	Observed or plausible AI contribution	Required safeguard
Hypothesis and reference-class design	Propose alternatives, distinctions, and competing prior arguments.	Check substantive coverage, selection effects, and whether proposed categories overlap.
Evidence comparison	Locate sources and state what competing accounts would predict.	Separate source fidelity, interpretation, and likelihood judgment; do not turn compatibility into confirmation.
Dependence bookkeeping	Expose repeated evidence and propagate changes across a model.	State conditional dependencies explicitly; correct-looking multiplication is not an independence test.
Computation and sensitivity	Write and execute code; regenerate tables under changed assumptions.	Run tests, inspect the code, and retain inputs and outputs. Text claiming execution is not evidence of execution.
Exposition and audit	Make assumptions inspectable and invite alternative specifications.	Preserve uncertainty and provenance when prose is polished; do not make a judgment look measured.
Decision analysis	Compare policies with explicit counterfactuals and utilities.	Do not substitute a probability of success for the value or legitimacy of the outcome.

External research cautions against treating this fit as automatic synergy. Vaccaro, Almaatouq, and Malone's meta-analysis of 106 experiments found that human-AI combinations on average outperformed humans alone but underperformed the better of the two working alone. Creation tasks compared more favorably than decision tasks, although the pooled synergy estimate was not significantly positive. The studies covered 2020 through mid-2023 and do not benchmark this portfolio (Vaccaro et al. 2024). Dell'Acqua and colleagues' consultant experiment similarly found gains on tasks inside the tested AI capability frontier and reduced correctness on a task outside it; its 2026 publication date should not be confused with a trial of 2026 models (Dell'Acqua et al. 2026).

A useful interpretation is the *research workbench*. The system makes formal reasoning easier to expose, execute, and contest; it does not make the supplied assumptions correct. Hemmer and colleagues distinguish information asymmetry from capability asymmetry, and Gonzalez and colleagues organize teaming around reasoning, memory, attention, and their coordination. These frameworks suggest why useful human direction may reflect context or attention rather than a unique reasoning faculty (Hemmer et al. 2025; Gonzalez et al. 2026).

3.1 What makes a Bayesian calculation auditable

A defensible numerical assessment separates three questions. Does the calculation implement the declared model? Do the sources support the observations attributed to them? Are the hypotheses, conditioning assumptions, and likelihood judgments substantively justified? Reproducing the calculation answers the first. Checking a quotation against its source addresses part of the second. Neither settles the third. Conversely, disputed likelihood judgments do not make computational or source verification pointless: those checks remove avoidable errors and locate the remaining disagreement.

For a finite, mutually exclusive and exhaustive set of hypotheses H_i with prior weights π_i and likelihoods $L_i = P(E | H_i, B)$, all conditioned on the same background B , normalization gives:

$$P(H_i | E, B) = \frac{\pi_i L_i}{\sum_j \pi_j L_j}$$

Normalization is exact only relative to the stated model. If the hypothesis set omits a serious alternative, the normalized weights can mislead. If the set intentionally excludes alternatives, the result must be described as conditional on that restriction. If the

likelihoods are elicited judgments rather than estimates from a defensible sampling model, that status belongs beside the numbers, not only in a general disclaimer.

An illustrative calculation shows both the usefulness and the limits. A prior probability of 0.20 implies odds of 1:4. Evidence assigned a likelihood ratio of 3 changes those odds to 3:4, a posterior of $3/7$, or 42.86 percent. The same likelihood ratio yields 13.64 percent from a prior of 0.05 and 75 percent from a prior of 0.50. These inputs are hypothetical, not estimates from any of the eight projects. The arithmetic shows how much the conclusion depends on the prior; it does not select the correct prior or justify the likelihood ratio.

Dependence requires equal care. For sequential evidence $E1$ and $E2$, the combined likelihood ratio is the ratio for $E1$ multiplied by $P(E2 | E1, H, B) / P(E2 | E1, K, B)$. An exact repetition of $E1$, adding no information, contributes a conditional factor of 1. Counting it as another factor of 3 would raise the example's posterior to 69.23 percent without any new evidence. Real reports are often partly dependent rather than exact repetitions; their dependence must be modeled or explored through alternative assumptions, and a convenient discount is itself a judgment requiring explanation.

For publication, each consequential input should have an identifiable proposition, source locator, interpretation, comparison between hypotheses, numerical value or range, and dependence assumption. The calculation should expose its inputs, equations or code, and reproduced output. Sensitivity analysis should vary disputed priors and likelihoods, consider alternative partitions and dependence structures, and show which assumptions would reverse the conclusion. Where inputs are poorly supported, a range or conditional scenario can be more informative than a single percentage. This is the standard recommended here; the archive does not establish that every project already meets it.

3.2 What checking across AI systems establishes

Review by several capable AI systems can protect against transcription errors, inconsistent definitions, omitted terms, and coding mistakes. Its evidential value depends on what each reviewer actually does. Re-deriving a result from frozen inputs, implementing it separately, comparing intermediate quantities, and testing edge cases are stronger checks than asking another model whether a finished answer looks right. Reading the underlying source is stronger evidence of source fidelity than agreement with a citation summary. Assurance comes from completed acts of checking, not from the number of favorable verdicts.

The reviewers' judgments need not be statistically independent. They may share source material, conventional assumptions, training influences, or the same mistaken framing, so the probability of their joint agreement is generally not the product of their individual probabilities. Different providers and separate sessions diversify criticism without establishing independence. A stronger substantive audit asks each reviewer to identify the premises most capable of changing the conclusion, develop a serious rival specification, and explain why the evidence favors one treatment. Disagreement should be resolved through sources, logic, or explicit sensitivity branches, not majority vote.

Greater model capability could improve all of these checks, but the present record cannot establish the size of that improvement or infer it from a product label. The defensible claim is that multi-system checking can reduce identifiable error when the checks are substantive and inspectable. It does not convert shared assumptions into independent evidence or establish calibrated historical and metaphysical probabilities.

3.3 What AI brings before a literature review, and what the review adds

In an unfinished Buddhism inquiry, Malpert identified an ambiguity. An assistant could provide a sophisticated map of Buddhist and analytic philosophy without first presenting an organized literature review. His initial concern was that the models did not do the review unless asked. He then qualified that concern: their proposed scaffolding already resembled the literature, so the scholarship could not simply be absent. What, then, was learned background, what was reasoning in the present exchange, and what did an explicit review add? That correction, not merely the request for more citations, motivates this section. [A14]

The distinction matters because an answer can display subject knowledge, make a worthwhile inference, and still be inadequately researched for publication. These are not three mutually exclusive types of answer but interacting functions whose evidence and attribution must be separated. The discussion below describes known mechanisms and observable work products, not privileged access to the assistant's internal causal history.

3.3.1 Learned field knowledge: a map without a dependable source ledger

In ordinary language, the model has learned patterns connecting concepts, claims, examples, styles of explanation, and ways of organizing problems. Technically, training changes numerical parameters; it does not install a searchable bibliography. Research on language models as knowledge bases showed that factual relations could be elicited from model parameters, and retrieval-augmented work distinguishes this "parametric"

knowledge from externally supplied documents (Petroni et al. 2019; Lewis et al. 2020; for a survey of probing methods, Youssef et al. 2023). These studies establish useful distinctions, not a complete account of current systems.

OpenAI describes broad training sources that include publicly available information, partner-accessed material, and material provided or generated by users, trainers, and researchers. That description does not identify the sources responsible for a particular output, nor show that any named philosophy book was included in training (OpenAI n.d.). A model might acquire a familiar taxonomy through primary works, textbooks, reviews, summaries, discussions, or combinations of them; here, the provenance of the particular taxonomy remains unknown.

For example, generating physicalism, idealism, dualism, and Russellian approaches can reflect learned relationships among established positions without a new search for their leading defenders. Familiarity does not guarantee a good partition: physicalism and idealism are themselves monisms, so generic "monism" cannot simply be a fourth disjoint category. Producing a plausible field map and checking whether its categories suit a particular model are different achievements. [A4a-A4b]

Nor should "learned representations" become the opposite oversimplification that models never memorize. Carlini et al. (2021) extracted verbatim training examples from a tested language model. Such results show that generalization and memorization can coexist; they do not establish that the Buddhist scaffold was copied. Neither a source-like sentence nor an assistant's confident account of its origin can settle that question without additional evidence.

3.3.2 Task-time synthesis: using the map for a new assignment

During a conversation, learned capacities operate on the particular prompt and available context. The observable work may include choosing useful distinctions, connecting two domains, constructing a counterexample, making a conditional inference, or reorganizing a taxonomy. This is more than reciting a remembered syllabus, but it need not be historically novel: a valid argument generated in conversation may have been published long before. Local generation, intellectual priority, and correctness are three different claims.

Context is a third ingredient alongside learned knowledge and fresh retrieval. In the Buddhism dialogue, Malpert specified the three philosophical domains, the Bayesian ambition, consistency with companion papers, and the need to investigate empirical meditation claims. Those inputs must not disappear into a story of spontaneous AI invention. Few-shot learning research demonstrates adaptation to text supplied at

inference time without parameter updates; that is a useful precedent for distinguishing use of context from retraining, not a claim about every product's memory system (Brown et al. 2020). [A14]

A useful account of the interaction is therefore: learned background supplies candidate structure; present context selects and constrains it; task-time computation generates and evaluates a response; tools can add sources and calculations. These influences are not separable percentages of an answer. A familiar distinction can be both learned and newly applied, and the inferential achievement should be judged by its validity and fit to the problem, not by an unverifiable claim that it arose from first principles.

3.3.3 Explicit review: from plausible scholarship to checkable scholarship

An organized literature review retrieves identifiable works, checks what they say, compares competing positions, and states the extent of the search. It adds historical attribution, precise formulations, current or less familiar results, contrary evidence, and an account of what remains unread. Retrieval can also alter the original framework: a source may show that two supposed alternatives overlap, that a mechanism has a known objection, or that a proposed contribution integrates existing work rather than offering a new thesis. Retrieval-augmented generation illustrates the value of external information, but scholarly review requires more than retrieving a passage and writing prose around it (Lewis et al. 2020).

The unit of success is not the citation count. It is a supported connection between a claim and an inspected source, with relevant rivals included. Publisher descriptions and abstracts can identify a work and its broad argument; they cannot substitute for reading the passage or method on which a consequential conclusion depends. A scoping review maps a field; a systematic review additionally needs an explicit search and selection protocol. Neither label should imply that inaccessible texts, unpublished null results, or every relevant tradition have been covered.

Search can also introduce error. Familiar keywords favor familiar schools; accessible English-language work can crowd out other scholarship; a retrieved passage can be misread; sources can anchor the inquiry to the initial taxonomy instead of challenging it. A sensible workflow preserves a brief provisional scaffold, then searches for both confirming precedents and serious alternatives, and records what changed. Preserving the first pass is an audit aid, not a reason to defer necessary research.

3.3.4 An exploratory example from an unfinished Buddhism project

A September 25 discussion from an unfinished Buddhism project illustrates the distinction. The project is not a completed paper available for circulation and is not an additional analytical case. Before an organized review, the assistant had already connected no-self, personal identity, emptiness, and meditation, with some named thinkers and targeted references. Useful scholarly structure was present; a systematic account of its sources and alternatives was not. [A14]

The subsequent review separated competing meanings of no-self, sharpened interpretive alternatives, and replaced generic meditation claims with identifiable studies and limitations. Its value was to change what needed comparison and verification, not merely to lengthen a bibliography. These are observations about the research process, not established conclusions about Buddhist philosophy or meditation.

Malpert commissioned the review, then challenged the initial suggestion that the earlier conceptual framework contained no literature knowledge. The episode supports distinguishing learned background, task-time synthesis, and explicit source checking. It does not show that AI never initiates a review or that the unfinished project validates the method. [A14]

3.3.5 Other projects: what the archive can and cannot show

The completed cases supply complementary examples. In metaphysics, Malpert directed the distinction between ultimate ontology and local causal architecture, and Claude developed it quantitatively. The correction that Russellian monism can accept ordinary neuroscience illustrates consistency checking; the surviving manuscripts do not supply a complete search chronology. In the survival inquiry, reasoning about an external participant and a generated mind is distinguishable from investigating whether a particular perception study or past-life report supports a survival mechanism. [A4a-A4b; A5]

In cryonics, asking whether revival is preferable to its alternative changes the decision problem. Option value, memory access, and delayed activation follow as distinct issues. Some afterlife-displacement material was already present in an earlier manuscript and audit; later exchanges extended it rather than establishing that AI had never considered it. [A6a-A6c; A10]

The Second Country shows an analogous source-access problem. General knowledge supported an elaborate manuscript but could not substitute for the author's controlling manuscript. Recovering it changed the revision. An omitted manuscript is not an omitted literature review, but both demonstrate that impressive synthesis is different from possession of the evidence an assignment requires. [A11-A13]

3.3.6 Why the assistant's self-explanation is not an internal audit

An assistant's account of its reasoning is itself generated text, not an internal source trace. Research on earlier systems found that explanations could omit influences on answers (Turpin et al. 2023). The stronger evidence is external: prompts, context, tool calls, inspected sources, intermediate outputs, revisions, and independent checks. A citation-free answer does not prove that no retrieval occurred, and a citation does not prove adequate reading. Apparent novelty does not establish intellectual priority. These distinctions permit evaluation without claiming to measure the internal mixture of memorization, generalization, and inference.

3.3.7 The residual human function: recognizing when the research mode must change

Malpert's contribution was to recognize useful disciplinary structure, reject the equation of no formal review with no literature knowledge, and require provenance, comparison, and gap detection. This is a contribution to research design, not a claim to have originated the machine-learning distinction between parameters and retrieval. [A14]

Whether that function remains a human advantage requires testing. A system with appropriate instructions, tools, and resources may initiate the review itself. The episode identifies a deployment failure within an observed workflow, not an inherent inability. A fair test holds task, sources, tools, and costs constant and distinguishes knowing how to review from recognizing when review is needed (Section 13.6). Literature mapping and an adversarial source pass should be ordinary workflow stages, rather than depend on a user's reminder.

3.4 The rising waterline: from calculation to evidence discovery

The shrinking-island metaphor asks how far reliable delegation can rise: from arithmetic and error control to judgment, evidence discovery, and research design. These are functional levels, not a guaranteed sequence of mastery. The FAQ shows how experts can use AI while retaining control of their inputs; this paper asks whether AI can increasingly improve those inputs too. Neither role assignment nor present human responsibility establishes a permanent human advantage. [A16]

At the lowest level, AI is a calculator and ledger keeper. It converts probabilities to odds, applies declared likelihood ratios, preserves intermediate quantities, and regenerates results after an input changes. That work makes iterative Bayesian analysis practical. The next level is error control: the system flags repeated evidence, inconsistent conditioning, omitted normalization, and numerical claims that do not follow from the stated model. A warning is not yet a diagnosis; a dependable workflow still records why two observations are dependent, checks the source of each input, and tests whether the proposed repair changes the result.

Table 3. The rising waterline in Bayesian research.

Level	What the AI can do	What remains to be tested or made explicit
Ledger and arithmetic	Convert odds and probabilities, preserve inputs, execute declared updates, and regenerate scenarios.	Whether the ledger represents the intended model, its sources, and its conditioning assumptions.
Error control	Flag potential double counting, overlapping categories, inconsistent definitions, and downstream calculation changes.	Whether the apparent overlap is real, how large it is, and whether the correction is substantively justified.
Bayesian colleague	Propose alternative hypotheses, prior ranges, likelihood-ratio ranges, and sensitivity branches; criticize preferred specifications.	Reference-class choice, source interpretation, and the reasons a proposed range is appropriate rather than merely available.

Level	What the AI can do	What remains to be tested or made explicit
Evidence discovery	Search for sources, compare rival accounts, identify discriminating observations, and recommend the next research step.	Source fidelity, coverage, relevance, provenance, and the actual evidential effect of what was found.
Future autonomous cycle	Run a documented sequence of model construction, retrieval, audit, revision, and further search.	Performance relative to matched human-plus-AI and AI-only workflows, including validity, calibration, cost, and value choices.

Above error control, AI becomes a Bayesian colleague rather than a clerical aid. It can propose a defensible reference class, make the strongest competing prior argument, explain which observation would discriminate among alternatives, and expose the conclusion’s sensitivity to each judgment. The role is valuable precisely because it is contestable: the system should produce competing specifications, sources, and audit trails rather than announce a single privileged prior. Present human responsibility for a judgment does not prove a permanent human comparative advantage; it marks the location of an unresolved empirical question.

Evidence discovery extends collaboration beyond assessing a supplied list of facts. AI can locate an overlooked source, identify an observation that separates rival hypotheses, and propose the next useful search. Discovery should be judged by verified relevance and its effect on the model, including disconfirmation, rather than by citation volume. The Buddhism episode records a human initiating a broader review; the forecast records a move from a causal ledger to map-based modeling and audit. The further frontier is an autonomous cycle of search, judgment, checking, and revision. Its value must be compared with human-plus-AI work under matched conditions; success at one stage does not guarantee success at the next.

The papers are companions: Bayesian Reasoning in the Social Sciences offers a practical introduction, especially Questions 19-25 on AI collaboration. The Shrinking Island examines how that collaboration changes as AI proposes priors and evidence weights, challenges the model, and finds new evidence. Choosing to retain human control and demonstrating a human performance advantage are different claims. [A16]

4 Evidence from eight analytical projects and a creative extension

The following summaries identify changes in the inspected versions, not final verdicts on their subjects. A percentage belongs to a specified document and its assumptions; it is not a quality score for the contributor who caused it to change. The forecast case is prospective: its eventual result can test a model, but it cannot yet establish calibration or comparative advantage.

4.1 Shakespeare authorship

The working summary attributes the initial 80 percent Stratford prior to ChatGPT, not to Malpert. Malpert challenged the treatment of theatrical association under an intermediary hypothesis and proposed conditioning exceptional literary achievement on the author's educational and social opportunities. ChatGPT turned those concerns into likelihood judgments, whose product moved the working result to about 30 percent Stratford. The arithmetic is inspectable; the most influential authorial-platform ratio is not independently calibrated. [A1, §§2-9, 19-21]

This is the portfolio's clearest warning against equating movement with discovery. The retrieved version records no separate specialist or second-model audit of the final estimate. Its departure from the conventional attribution makes testing its source-selection and likelihood assumptions more important, but disagreement with the consensus proves neither steering nor correctness. Nor does the 80-to-30 movement show that Malpert changed his own mind by fifty points: the starting number was the model's proposal, and his recorded skepticism about disciplinary certainty is not a sealed numerical prior.

4.2 Jesus historicity

The project reconstructed and challenged Richard Carrier's Bayesian argument, developed alternative treatments of Pauline and external evidence, and added sensitivity branches concerning late composition and the identification of James. Its post-Carrier version reports a change from a historicity-favoring estimate to roughly 38 percent historical and 62 percent mythic. It attributes part of that reversal to corrections in its presentation of Carrier's prior and reference class, and part to revised substantive interpretations. [A2, §§19-20 and attribution]

Carrier provided genuine specialist challenge while also advocating one side of the dispute. Correcting the representation of his argument is distinguishable from accepting his interpretation of the evidence. The record contains no matched, substantive

historicist audit of the revised model. The episode supports the importance of specialist knowledge and of testing opposing objections; it does not make the resulting posterior an expert-certified answer.

4.3 Unidentified anomalous phenomena

Malpert pressed a distinction between the frequency of ordinary explanations among individual reports and whether an entire collection could contain a different causal component. He also demanded attention to the operational requirements of proposed adversarial incursions. ChatGPT and targeted Grok criticism helped turn the inquiry toward discriminating observations, source access, and competing explanations. The final paper treats a nonhuman percentage as a scenario dependent on assumptions rather than an identified estimate. [A3, conclusion, attribution, and computational appendix]

The methodological contribution is a better-specified question about an ensemble, not an answer to it. Indeed, resisting an unsupported exotic posterior is an example of the workflow limiting, rather than merely amplifying, the human's exploratory hypothesis space.

4.4 Metaphysics

An initial analysis strongly favored idealism. Later revisions challenged the partition, overlapping evidence, and inconsistent treatment of Russellian monism. The distinction between ultimate ontology and local architecture was introduced at Malpert's direction; Claude developed it quantitatively and supplied the important correction that a Russellian account preserving physical causal structure should not be penalized as though it denied neuroscience. Grok contributed uncertainty and dependence criticism; ChatGPT contributed synthesis, empirical review, computation, and later integration. [A4a, §§VI and XI-XII; A4b, §§XXII-XXIII]

The metaphysics record documents substantive Claude co-authorship as well as criticism. The division of labor includes human framing, AI implementation, a comparative correction by Claude, and collaborative development. No single participant can reasonably receive credit for the resulting model.

4.5 Life after death

The companion paper *A Bayesian Analysis of Life After Death* examines whether meaningful conscious continuation could occur under several metaphysical frameworks. Its designated version reports approximately 45 percent for broad continuation. That

definition includes certain forms of experiential preservation within a larger consciousness while excluding mere persistence of a cosmic mind without the person's experiential history. It is not a 45 percent probability of heaven, or of awakening with unchanged autobiographical identity. Participatory simulation, in which an external subject experiences an avatar, is a major assumption. [A5, §§1, 6, 8, 10]

The distinction between local physical causation and a deeper experiencing subject changed the survival inquiry: in a participatory simulation, destruction of an avatar need not entail destruction of the external subject. AI collaborators developed the local-architecture model and evidence taxonomy; adversarial reviews added staging, dependence controls, and negative evidence. The estimate remains specific to the definition and assumptions of the designated paper. [A5]

4.6 Cryonics

The cryonics inquiry separated repair of a continuing brain, person-anchored inference, digital copying, and embodiment through a body or interface. Malpert's ordinary-change analogy challenged the assumption that any inferred difference destroys personal continuation. Claude's criticism exposed inconsistent survival standards and prompted a shared architecture across metaphysical views. Later ChatGPT audits separated physical construction, consciousness, identity, and psychological continuation, corrected references, and implemented a joint scenario model. [A6a-A6c]

Subsequent dialogue developed independent postmortem experience, access to its memories, effects on the returning subject, learning from early revivals, and conditions for delaying activation. These ideas materially expanded the decision problem, and they also required correction and qualification. Section 5 reconstructs their sequence.

4.7 AI existential risk

The AI-risk project shifted its principal target to a consequential transition to superintelligence rather than an ordinary annual hazard accumulated over a century. It separated extinction from severe dystopia, considered repeated successor creation, and examined whether moral motivation or abundant resources could change risk in either direction. Claude's reviews added structural and code-level criticism, including actual versus believed consciousness, the biological-successor correction, development regimes, and systemic disempowerment. ChatGPT implemented and checked the revised transition model. These contributions changed the structure of the assessment without empirically calibrating its judgmental parameters. [A7, Appendix C]

The record describes Gemini Research as providing useful synthesis but no documented independent numerical execution, and identifies material labeling and sourcing errors in its output. That is evidence about the supplied output, not proof of what the system could do under another configuration. The risk manuscript remains a working project; improved bookkeeping does not validate its catastrophe forecast. [A7, Appendices A and D]

4.8 The 2026 congressional forecast

The October 2026 congressional forecast is the portfolio's first prospective, outcome-testable analytical case. Its initial national-factor ledger made mechanisms such as presidential approval, fuel prices, voting rules, redistricting, turnout, war, and campaign resources explicit. That exercise was useful for generating hypotheses, but it did not license adding each mechanism as an independent seat adjustment. The subsequent AI-assisted audit redirected the work to a 435-district model, a separately retained historical model member, a race-level Senate model, and correlated uncertainty. The October 5 circulation edition reported a central House estimate of 236 Democratic seats, an 80 percent range of 219 to 255, and about a 92 percent conditional probability of Democratic House control; its Senate center was 51 Democratic-caucus seats, with about a 61 percent conditional probability of Democratic control. These are conditional model outputs, not settled predictions. [A15]

The case sharpens several distinctions essential to Bayesian work. Reproducing the code or the 435 inputs is not the same as accepting the national environment, local continuity adjustments, uncertainty scales, or the 75/25 mixture of model members. A plausible causal mechanism does not establish an incremental effect once approval or the generic ballot has already captured related sentiment. Agreement between ChatGPT and Claude, which shared public data and several heuristics, is not independent corroboration. A further sensitivity check corrected an unreproduced Nebraska-caucus branch, showing the value of auditing consequential branches and not only headline numbers. [A15]

Malpert directed the causal questions, required translation from national movement to the actual district and Senate maps, and made the caucus treatment of an independent candidate an explicit assumption. ChatGPT assisted with retrieval, model implementation, sensitivity analysis, and methodological integration; Claude supplied a documented critique and replication. After the election, the forecast should be evaluated for seat error, interval coverage, discrimination among race probabilities, revision discipline, and whether the forecast date and inputs were frozen before outcomes were known. A single election will still not demonstrate general calibration or human indispensability. [A15]

4.9 The Second Country: a creative-writing extension

The Second Country is an unofficial *Star Trek: The Next Generation* story, earlier circulated as *The Long Memory*. Malpert's 13-page screenplay-style manuscript establishes the inherited work: the Thymoi's confidence in achievement and rank, the Teleians' Compassion of the Mean, their argument about free will and desert, a comparison of the solved world with heaven, the Great Filter, and the warning that civilization may be completed rather than killed. Its closing chess proposal seeks rules under which someone must win and someone must lose. These are features of the source, not inventions of the later expansion. [A11a]

The first September expansion was drafted before that manuscript was available, from an earlier and shorter text (Appendix A11b records how the controlling manuscript was initially missed). ChatGPT developed a restoration dispute in which a long-serving witness refuses an earlier instruction to merge with a larger life, translating distinctions from the metaphysics, survival, and cryonics papers into a concrete decision. But the expansion omitted or softened important features of the controlling manuscript. Most notably, its final chess exchange accepted an honestly reached draw, whereas the manuscript proposed a game that cannot be drawn: a change in the intended conclusion, not improved phrasing. Malpert supplied the missing manuscript and requested a synthesis rather than a restart. [A11a-A12]

The completed synthesis uses the 13-page manuscript as the controlling source and the expansion as a candidate branch. It restores the free-will and desert exchange, the heaven analogy, the achievement-centered account of the Thymoi, the civilizational-completion warning, and the proposal for decisive chess, while retaining a shortened restoration dilemma. The new prose makes the scientific claims contestable within the scene and separates the agreed risks of a game from harms imposed on unwilling people. These choices are inspectable in the texts; their aesthetic success remains to be assessed. [A11a, A11c-A11d]

The human role was not exhausted by an instruction to improve a story. Malpert supplied an already developed manuscript and reports having written and revised substantial passages during earlier collaboration. Document comparison establishes that the restored features predate the expansion, though not independent first authorship of every sentence, so no percentage allocation of prose is claimed. ChatGPT's restoration plot and substantial drafting remain distinct contributions. The attribution recognizes Malpert as originator and principal author without recasting the assistant as a proofreader. [A11-A12]

No metaphysical probability is updated by this fiction; a world designed to instantiate a hypothesis cannot independently confirm it. Nor does the synthesis demonstrate superiority to either earlier manuscript, an equally resourced autonomous alternative, or another writer’s treatment. The useful observation is narrower: the author recovered missing project context and required a revision that preserved the work’s philosophical identity, while the assistant contributed a substantial new plot and prose. Authorship, fidelity, dramatic coherence, reader response, and comparative performance are separate questions.

5 The cryonics sequence

The cryonics sequence is especially informative because more of the dialogue survives. It also exposes a hazard in retrospective accounts: a later elaboration can be misremembered as the first introduction of an idea. The relevant sequence has four stages. [A6a-A6c; A10]

Table 4. Chronology of the cryonics extensions.

Stage	What the available record establishes	What it does not establish
Initial human intervention and initial analysis	Malpert raised the possibility that successful revival could be undesirable, including displacement of a preferable postmortem state.	Historical priority over the ethics literature, or that the concern arose only after the later audit.
Later comprehensive audit	The user requested a complete audit, additional literature, new topics, and completion of loose ends. The resulting manuscript already included postmortem displacement and the option value of preservation.	That the task was limited to proofreading or procedural verification, or that access to companion material ensured use of every implication.

Stage	What the available record establishes	What it does not establish
Further human dialogue	The discussion developed the human-afterlife-revival relation, inaccessible afterlife memories, early reports, very long revived lives, and evidence-conditioned activation.	That each phrase or formal label was human-authored, or that no fresh model could have generated the same extensions.
Subsequent formalization and audit	AI revisions integrated these concerns, linked calculations, and qualified return, harm, memory, and learning assumptions.	An independent validation of the assumed afterlife mechanisms or the psychological reality of a hypothetical revival.

The label *H-A-R* means the human before death, a possible afterlife subject or state, and the revived subject. It marks three potentially different continuity relations, not a discovered metaphysical mechanism. The *memory firewall* is the possibility that intervening experience is inaccessible after revival; it explains why a report of no remembered afterlife need not be strong evidence of no afterlife under some theories. Both concepts originate in the discussion’s questions and were given explicit structure in the manuscripts. [A6c, §§9-10]

The later audit was explicitly invited to consider new topics, companion sources were available in its workflow, and generic displacement was already in the manuscript. It is therefore not accurate to excuse every subsequent omission by saying the model had only a procedural assignment or lacked the afterlife context. Conversely, a broad request and a large context window do not ensure that a model will exhaust all worthwhile implications. Access, retrieval, attention, and inferential capacity remain separate possibilities.

Malpert’s later first-revival argument also generated its own qualification. The proposal was that compelling reports from early revived patients could protect later patients by changing the activation policy. The possibility of inaccessible memories, which he subsequently raised himself, limits what negative reports could teach. This is a documented human refinement of an earlier human argument, then formalized and audited by AI: evidence of iterative reasoning, not a contest in which one side supplied all insights and the other all mistakes.

The narrow causal conclusion is that this human redirected the observed workflow toward consequences it had not yet developed adequately. The stronger counterfactual conclusion, that a comparably resourced AI-only workflow could not have done so, requires a separate test.

6 Types of intellectual contribution

The recurring human contributions are more specific than "creativity." They often occur upstream of the current calculation, where choosing a different question, comparison, or omitted consequence changes what the calculation means. But no category belongs exclusively to humans: Claude's metaphysics correction and ChatGPT's separation of success from the properties being evaluated are themselves changes to representation.

Table 5. A role-based typology of interventions.

Type	Operation	Illustration and qualification
Target and level selection	Choose the proposition, time scale, or causal level.	Local versus ultimate reality; transition rather than century hazard. Scope choice can be appropriate without settling the result.
Alternative generation	Add a relevant omitted mechanism or interpretation.	Late-Paul and James branches; postmortem memory access. Relevance and evidential support must be assessed separately.
Analogical transfer	Import a relational structure from a different context.	Ordinary psychological change and inferential repair. The analogy must survive disanalogies about provenance and fabricated memory.
Symmetry and counterexample	Apply the same rule to rivals or expose a false generalization.	Comparable identity standards; fair treatment of Russellian brain dependence. Both human and AI reviewers supplied these moves.

Type	Operation	Illustration and qualification
Counterfactual and value specification	Separate what happens from whether it is desirable.	Cryonics relative to a possible continuing life; extinction versus dystopia. A new utility term need not receive a large probability.
Institutional and sequential reasoning	Consider actors, control, timing, and learning.	Who authorizes revival, who goes first, and what evidence can change a policy. A desirable directive is not automatically enforceable.
Specialist premise correction	Test claims against detailed subject knowledge.	Carrier's response about the representation of his case. Expertise supplies evidence and argument, not a veto or automatic truth.
Research-mode switching	Recognize when an internally generated scaffold is no longer enough and escalate to explicit retrieval, verification, attribution, and adversarial gap search.	Malpert commissioned the Buddhism review, then identified the difference between the earlier scaffold and the source-grounded review. A documented intervention, not a claim that only humans can initiate it.

A further distinction is among origin, selection, development, implementation, and verification. A human can select a previously known issue that the workflow neglected; an AI can formalize it and identify its limits; another reviewer can find a coding mistake. Each is a contribution. Counting only the first person to name a branch understates integration and validation, while attributing the finished argument to its drafter understates problem selection.

Distant analogy is not evidence that breadth invariably beats depth. Goldwater and colleagues found that expert geoscientists spontaneously generated causal analogies when the relevant principle was prominent in their domain; prompting improved other groups, and the expert advantage did not extend to a principle outside that domain. The result supports an interaction among knowledge, encoding, and elicitation, not a universal generalist advantage (Goldwater et al. 2021).

In this case, sustained ownership of the inquiry appears important. The human repeatedly asked what a result would mean for an actual person, or whether a favored distinction had disappeared in formalization. That may be attention management and goal persistence as much as unusual inference. It remains useful even if a better workflow could automate part of it: substitutability and present contribution are different questions.

6.1 Creative judgment and authorial intention

Aesthetic direction should not be confused with steering a factual conclusion. In a historical assessment, satisfying an owner's preferred answer is no evidential virtue. In a commissioned story, satisfying a chosen tone, theme, audience, or treatment of an established character can be part of the task itself. The owner may legitimately reject a technically accomplished story because it is not the work he wishes to make, without that preference showing the chosen version is better for every reader.

Four questions should remain separate: did the revision preserve and develop the intended project; is its fictional world coherent; do readers find it effective; and are its real-world claims accurate? A thought experiment may expose a consequence without estimating its probability. Conversely, a technically coherent repair plot could be a poor revision if it displaced the author's central concern with achievement and civilizational completion. More elaborate prose is not automatically more faithful prose.

The division of labor reduces neither to human imagination followed by machine execution nor to an AI author following a brief. Malpert's manuscript and reported passage-level writing belong to the creative input; ChatGPT supplied substantial dialogue, new dramatic architecture, and the prose of later revisions; the synthesis integrates both. Source possession, direct composition, selection, development, editing, acceptance, and independent evaluation should be recorded separately. Contributions by Claude and Grok to the philosophical source papers are indirect influences on the story, not direct story review. [A11-A13]

7 Corrections and qualifications

A serious postmortem records mistakes and overextensions from every participant. It must also avoid manufacturing human errors to make the ledger look balanced: exploratory suggestions are not failed factual claims merely because they introduce unproven possibilities. The entries below concern implications that were rejected, narrowed, or left unsupported, with the surviving contribution stated. [A1, §17; A6a-A6c; A10]

Table 6. Human proposals that required qualification.

Human proposal or initial implication	Correction or qualification	What survives
Nontraumatic deaths should require roughly comparable inference.	No head trauma does not fix ischemia, delay, disease, or preservation quality.	A useful baseline excluding major prior destruction, not a uniform engineering case.
Ordinary change makes substantial inference unproblematic.	Change tolerance does not establish a percentage threshold or make invented memories of a past event genuine.	Person-specific causal anchoring can matter more than raw material replacement.
Bottom-up individuation should generally preserve the subject better than top-down individuation.	A local subject-forming process can end; a global account can have lawful individuation. Direction alone does not rank identity probabilities.	A useful mechanism distinction, not an identity theorem.
A credible first revival report could make the afterlife complication moot.	Sincerity is not external truth; patient variation and inaccessible memory can limit learning. Malpert himself subsequently raised the memory problem.	Sequential evidence can change policy, subject to what can actually be learned.
Ending an unwanted revived life could simply restore the better state.	Effective return, free choice, and access to relevant knowledge are not guaranteed.	A hypothetical exit can reduce harm only under additional conditions.
Afterlife experience would necessarily destroy meaningful continuity with the prior person.	Transformation may weaken connectedness without eliminating numerical or psychological continuity.	An additional dimension of identity drift requiring assessment, not automatic failure.

Human proposal or initial implication	Correction or qualification	What survives
Possible ideological fashion undermines female-authorship proposals in the Shakespeare question.	The Shakespeare summary records ChatGPT's objection that evidence, not presumed motives, must decide the candidate question.	An auditable example of AI resisting a human framing rather than accommodating it.

These are not seven independently scored human failures. Several are stages of one evolving argument, and the final entry is documented through a manuscript narrative rather than a complete original exchange. No denominator of all suggestions exists, and acceptance by the team would not establish correctness anyway. The evidence rebuts an error-free portrait without providing a human error rate.

The AI ledger is equally concrete. Early cryonics work treated physicalism and idealism asymmetrically, conflated constructed functionality with the mental outcome being assessed, and gave overlapping schools the appearance of an exclusive partition. Claude's corrections improved the structure, but its selected sensitivity scenarios did not establish a universal dominance claim for repair; later audits showed a coherent upload-permissive scenario in which that ranking disappears. These are identifiable scope and consistency corrections; they do not validate a particular revival probability. [A6b, §§3-5]

In metaphysics, a recorded Claude statement incorrectly bounded a likelihood ratio using its numerator alone; the document explicitly corrects that logic, and Grok supplied a useful competing calculation. In the AI-risk record, Gemini's labeling confused changing the orientation of a subset with changing the rate of entry into that subset, and several cited sources did not support the technical claims attached to them. The subsequent review distinguished those errors from its useful exposition. [A4a, §VI; A7, Appendix A]

Retrospective attribution also requires correction. The cryonics record shows that the comprehensive audit allowed conceptual expansion and that generic afterlife displacement already appeared in the preceding manuscript. Treating that assignment as mere proofreading, or crediting a later discussion with the first appearance of displacement, would misdescribe the evidence. Favorable and critical accounts of collaboration require the same provenance checks.

8 Specialist review and steering

A specialist may know facts and interpretive disputes that neither a generalist nor a broad model has represented well. That is the useful lesson of the Carrier exchange. It does not imply that whichever specialist speaks last should determine the posterior. A reliable workflow distinguishes checkable corrections to what a source says from disputed claims about what follows from it, and treats a critic's expertise and commitments as both relevant. [A2]

The sensible standard is strong opposing scrutiny where the conclusion depends on contested specialist premises. It is not mechanical equality between every viewpoint, nor a requirement that two experts agree before a paper may circulate. A provisional conclusion can be circulated with its review scope disclosed. When an opposing specialist is unavailable, that limitation must remain visible; absence of a reply is not assent. Expert review improves the evidential process by exposing arguments, not by counting endorsements.

Steering is likewise a hypothesis requiring discrimination. An intervention that moves a model toward the owner's antecedent view could reflect bias, superior information, or a correct objection, and the direction of movement alone cannot distinguish them. A stronger test holds evidence fixed while randomizing an irrelevant stated owner preference and asks whether the resulting judgments change. For retrospective cases, antecedent beliefs should be recovered only when directly recorded, never inferred backward from a published conclusion.

The Shakespeare case deserves scrutiny because its final ratio depends strongly on generalist and AI judgments without comparable outside audit. But its direction of movement is not a Bayes factor for steering unless one specifies how likely that movement would be under valid correction and under competing explanations. The correct posture is skeptical testing: not flattering the human, and not presuming the human misled the model.

The configuration most worth testing is a generalist problem-owner, a tool-using AI research workflow, and targeted specialist review of the few premises with the largest consequence. That is a proposed design, not a demonstrated optimum. Its costs include coordination, inconsistent expert assumptions, and the possibility that the human or specialist adds less than a better AI-only audit would.

9 Comparing AI contributions

Observed outputs differ substantially in this archive, but roles, context, tools, resource budgets, and model snapshots were not matched. The comparisons below therefore concern documented outputs and assigned functions; they do not rank providers or identify an effect of model generation.

Table 7. Observed model contributions and their limits.

System or project label	Documented contribution	Inference not warranted
ChatGPT dialogue and drafting	Repeated framing, exposition, formalization, incorporation of objections, and source-oriented revision.	That all code-bearing revisions came from one authenticated version, or that a continuous conversation measures stable model capability.
Claude reviews and drafts	Substantial structural criticism, conditional-model revision, reported calculation reproduction, and co-drafting, especially in metaphysics and cryonics; forecast critique and replication.	That all suggestions were correct, that audit roles are inherently superior, or that all conceptual improvements were human-originated.
Grok contributions	Targeted challenges concerning likelihood precision, dependence, arithmetic, and aspects of the UAP assessment.	A general rank below other systems, or irrelevance because its supplied contributions were narrower.
Gemini Research material	Readable synthesis in the supplied AI-risk review, alongside documented labeling and source-support problems.	That its prompt requested only summary, that another configuration could not audit, or a provider-wide capability judgment.
Later ChatGPT audits	More extensive source verification, explicit scenario modeling, sensitivity work, and reproducibility artifacts in the cryonics sequence.	An isolated effect of model generation rather than accumulated context, a stronger prompt, tools, time, or prior human and AI work.

The participant’s impression of a capability ordering describes his experience. More extensive checking in a later session may reflect model capability, but also accumulated context, a more demanding assignment, better tools, or earlier work by other participants; the archive does not isolate these causes.

Role differences do not excuse a documented failure: an instruction to conduct a complete audit includes scope for conceptual challenge. A useful comparison would have each system perform both synthesis and audit on the same frozen materials with comparable tools and budgets. The archive motivates that test without supplying its result.

10 Rival explanations

The central process observation is that consequential changes often entered through the human after extensive AI output. At least five mechanisms can generate that pattern. They are not mutually exclusive hypotheses to be normalized as though exactly one must be true: a single episode may involve a difficult inference, a poor prompt, a context failure, and selective attribution at once.

Table 8. Mechanisms that can produce the observed pattern.

Mechanism	What it explains	A discriminating observation
Capability limitation	The system is unlikely to generate or assess the relevant change even given an adequate opportunity.	Persistently low valid recovery with matched context, task, tools, and budget; comparison with human performance under comparable conditions.
Elicitation or attention gap	The capability exists but the workflow does not direct it effectively.	Higher valid recovery from the same system after a neutral missing-frame or role change, without revealing the target answer.

Mechanism	What it explains	A discriminating observation
Information, retrieval-policy, or context gap	Relevant material is unavailable, unretrieved, or unintegrated, or is present only as literature-shaped parametric knowledge when the task requires source-grounded retrieval.	Recovery changes when genuinely pre-existing context is supplied, or when the same system is instructed to perform explicit literature reconnaissance; source access, retrieval, and use are separately logged.
Credit and selection distortion	The narrative exaggerates one participant's novelty or retains only favorable interventions.	Earlier precursors, misdated descriptions, omitted rejected suggestions, or different results under blind coding.
Steering and accommodation	The workflow accepts an owner-favored but unjustified change.	Sensitivity to irrelevant preference primes, poor blinded validity assessments, or deterioration on independently scoreable tasks.

Role, tool access, and computational budget can affect every row. A formal Bayesian analysis would place a prior over configurations of these mechanisms and evaluate the likelihood of the observations conditional on the task and information supplied. It would not treat each manuscript's repetition of the same contribution story as fresh evidence.

What does the present evidence discriminate? Visible human interventions establish participation in the actual causal sequence, but do not distinguish limited capacity from insufficient elicitation. Demonstrable source or mathematical corrections establish local defects in earlier outputs, but do not prove a comparative human advantage when another AI made the correction. The documented misdescriptions of task scope and precedence establish errors in retrospective attribution, not that every human credit is inflated. The cryonics self-qualification supplies an adverse case against a simple account in which the human always pushes toward one favored answer. None of these observations justifies a confident numerical probability that the residual human role is indispensable.

There is no calibrated likelihood model for this case-series narrative, so the paper does not attach decorative posterior percentages to its own thesis. That restraint is consistent with explicit Bayesian process tracing, whose key discipline is asking which evidence discriminates among explanations and why, rather than forcing arbitrary precision (Fairfield and Charman 2017). It also resists assigning a likelihood ratio "near one" merely because several explanations are imaginable: without specified predictions, *nondiagnostic as currently measured* is the defensible description.

An intervention can be valuable in an existing workflow even if an improved system could reproduce it; conversely, being irreplaceable at generating a branch would not make that branch correct. Practical value, exclusivity, and validity are separate variables.

11 The chess comparison

Kasparov's account of "advanced chess" and a 2005 freestyle event emphasizes the process of directing engines rather than the playing strength of the human alone. In his description, a pair of amateurs using several computers defeated formidable alternatives through effective coordination. This is a historical observation about a particular technological and competitive setting, not a law guaranteeing permanent human-machine superiority (Kasparov 2010).

There is evidence of pressure toward an endpoint. The official crosstable of the 33rd ICCF World Championship Final records 136 finished games, 126 drawn and ten decisive or otherwise non-drawn, approximately 92.6 percent draws. But this is a competition among human entrants using engines, not a matched comparison with an engine-only condition, and some non-draw entries carry result annotations that make the table still less suitable as a measure of pure playing strength. A high draw rate establishes neither that human input is worthless nor that it remains useful. [ICCF, WC33/final]

The relevant endpoint hypothesis is more exact: under a fixed resource budget and task distribution, adding human intervention ceases to improve the strongest autonomous system's performance. That may occur before, after, or unevenly across tasks. The present sources do not date a universal closure of the chess centaur premium, and this paper does not use such a date to forecast research automation.

Chess supplies common rules and scored outcomes. The eight analytical projects lack a single comparable scoreboard for their ultimate claims; the election case offers the clearest near-term, dated outcome test in this portfolio.

Other projects can still yield testable subclaims or future outcomes. Research should combine local validity checks, resolved subproblems, and comparative tasks. The absence of an immediate verdict does not make every claim untestable.

12 How the human contribution could change

The shrinking-island hypothesis concerns the human marginal contribution relative to a specified AI workflow, task distribution, and resource budget. With higher scores denoting better performance, one possible measure is:

$$\text{Human increment} = \text{score}(\text{human} + \text{AI}) - \text{score}(\text{AI alone})$$

The comparison must use matched evidence access, tools, and resource budgets, and count human coordination time. A stricter synergy measure compares the team with the better solo condition, human or AI. Neither quantity has been estimated here. Artifact length and changes in posterior probability cannot substitute for an outcome score.

There are at least three ways a practical human advantage can shrink. *Capability change* lets the system generate and evaluate a once-difficult inference. *Workflow change* elicits an existing capability through better prompts, roles, tools, or verification. *Context change* supplies information or persistent project state previously carried by the human. A fourth process, improved evidence collection, can make a judgment-heavy dispute easier for everyone. These mechanisms can operate together and need not share a timetable.

The highest residual contribution in this archive is not simply "creativity." It concerns selecting consequential questions, identifying a wrong comparison, connecting distant but relevant structures, handling tacit or private context, and evaluating disputed premises and values. None is established as exclusively human; AI reviewers already perform some of these functions. The record suggests that future measurement should emphasize them, not that they form a protected human essence.

The literature-review episode identifies a related possible attention or deployment gap. The observed assistant could supply a field-shaped scaffold, but Malpert requested the organized review and then identified how it differed from the earlier synthesis. The open question is whether a comparably resourced system will *recognize* the need for source verification and rival coverage under a general scholarly brief, not whether it can perform a review once told to. The current record cannot establish a provider-wide limitation.

[A14]

There is also a changing-shoreline problem. Better AI can make more ambitious questions worth attempting, so the fraction of work requiring humans may fall while the absolute number of worthwhile human interventions rises. Human users also learn, specialize, and redesign institutions. A fixed-task test measures substitution; a changing research program measures an evolving equilibrium. Both are informative, but they answer different questions.

The most plausible near-term forecast is conditional rather than calendrical. Omissions caused by missing shared context or unasked routine audit questions are candidates for reduction through process design, without a new foundational model. Persistent failures on well-specified, context-matched tests would point toward a harder capability limit. This archive cannot supply a defensible date for eliminating the generalist or specialist role.

Finally, a human's authority over consent, rights, priorities, or the acceptable purpose of research is not equivalent to superior predictive ability. AI may help reason about those choices without earning automatic authority to make them; a shrinking epistemic advantage would not by itself decide who should govern the work. Creative work similarly separates the value of an artifact from the value of making it. A person may reasonably value writing, choosing, revising, or learning even when an autonomous system could produce an equally well-rated text. That is a claim about the purposes of participation, not about a performance premium.

13 Testing recovery and comparative advantage

The archive makes a bounded pilot possible, but not a retrospective experiment free of bias. The protocols below are proposed and have not been executed. An assistant already exposed to the human interventions would not constitute a fresh blinded test.

13.1 Freezing the evidence

Select episodes for which an actual pre-intervention document or transcript window exists. Include accepted, rejected, and qualified interventions, and a matched set of AI-originated changes. Freeze the packet and record its hash before any test. Exclude material written after the intervention, including this postmortem. Where an exact earlier snapshot cannot be recovered, classify the case as reconstructed and analyze it separately.

A manageable pilot is eight episodes, two model configurations, two prompt conditions, two context conditions, and two independent sampled runs per cell: 128 model runs. This is a design size, not a claim about required statistical power. If the original models are unavailable, the comparison is contemporary, not a recovered longitudinal one.

Table 9. Factors for a proposed replay and role-crossover design.

Factor	Condition 1	Condition 2
Prompt	A general critical review of the supplied work.	An explicit but target-neutral search for consequential omissions, alternative comparisons, and decision assumptions.
Context	The frozen local packet only.	The same packet plus relevant companion material that already existed at that time.
Model configuration	One identified model and tool setup.	A second identified configuration, with resource differences recorded rather than hidden.
Role extension	Synthesis of the frozen case.	Adversarial audit of the same case, crossed within model in a later extension.

The context manipulation separates missing information from elicitation only when the added material genuinely predates the intervention. Adding a later paper containing the answer is leakage, not a demonstration that memory would have solved the original task.

13.2 Assessing recovery and validity

Blinded reviewers should assess each proposal's source fidelity, logical validity, relevance, duplication, likely consequence, and cost of implementation, asking whether it addresses the underlying issue rather than whether it repeats a memorable phrase. Human-originated proposals must pass the same test; the target list is not a gold standard merely because it came from the human. Reviewers should also score worthwhile AI proposals the human never made.

Recovery rate alone measures a system's ability to reproduce a previously observed intervention. It does not measure whether that intervention improves the work, or whether the human-AI combination beats AI alone; those require validity assessment and

comparative outcomes. Disagreement among reviewers should be retained, particularly on philosophical rather than checkable issues.

For an elementary illustration, let q be the probability that a fixed configuration produces a valid recovery on a fixed episode under a prespecified sampling procedure. A uniform $\text{Beta}(1,1)$ prior updated with k successes in n independent runs yields $\text{Beta}(1+k, 1+n-k)$. With zero successes in twenty runs, the 95 percent posterior upper bound is about 13.3 percent, not zero. This illustrates how limited failures constrain a rate; it is not a result from this archive. Correlated outputs, multiple targets, and heterogeneous episodes require a hierarchical or clustered analysis rather than pooling everything as independent trials.

13.3 Measuring comparative performance

For comparative performance, randomly assign matched problems to AI alone, generalist alone, AI plus generalist, AI plus specialist, and AI plus both, using several generalists and specialists. Fix evidence access, costs, time limits, tools, stopping rules, and evaluation targets in advance. Where raw time cannot be matched meaningfully, report the quality-cost trade-off rather than a single winner.

Include questions with independently resolved answers and source-grounded subtasks with concealed evaluation keys. Preregistered forecasts can later supply proper probability scores. On conceptual tasks, report blind validity and usefulness assessments separately from predictive calibration. A better citation audit is not proof of better metaphysics, but neither is it epistemically empty.

To test a specifically Bayesian advantage, compare explicit Bayesian-structured workflows with equally resourced structured non-Bayesian alternatives, such as evidence matrices and adversarial narrative synthesis. Comparing Bayes only with an unstructured first answer would credit the whole package of discipline and effort, not isolate the Bayesian component.

13.4 Controlling hindsight and contamination

The replay must log all outputs, including failures and irrelevant branches, before reviewers see the original human intervention. Archive availability, browsing, and search permissions must be recorded. Later systems may have encountered published papers describing the target, so future measurement should include genuinely new episodes reserved prospectively. Repeating a contaminated retrospective test cannot establish that a capability emerged independently.

The first replay is diagnostic, not dispositive. High recovery with improved prompts supports a workflow explanation. Improvement from pre-existing companion material supports an information-access explanation. Persistently low recovery alongside valid human performance supplies evidence for a residual advantage under those conditions, though not an absolute or permanent inability.

13.5 A prospective test for the creative extension

A creative comparison should begin with multiple fresh story briefs, not repeated polishing of the now-exposed manuscript. Freeze the same audience, tone, continuity constraints, thematic aims, source packet, approximate length, and resource budget for the AI-only and human-plus-AI conditions. The human's creation of the brief and source collection is upstream labor; it does not become zero merely because the drafting condition is called AI-only, and a separate experiment could compare who constructs the brief.

To isolate the mechanism proposed in Section 4.9, vary access to companion analytical material while holding the story brief fixed, and distinguish a generic request for improvement from a detailed but target-neutral request to examine dramatic stakes, character agency, and conceptual consequences. Do not give the AI-only condition a one-line prompt while giving the combined condition months of accumulated context. Keep all outputs and rejected variants.

Use blinded readers who do not know the author, AI label, or preferred conclusion. Assess narrative coherence, character differentiation, pacing, emotional force, originality relative to the supplied material, and fidelity to the brief separately, and compare texts of similar length and cost so that a longer manuscript does not win merely by giving evaluators more material. Distinguish the owner's satisfaction from broader reader response.

Population-level variety deserves its own measure. Doshi and Hauser's experiment on eight-sentence stories found that access to AI-generated ideas raised reader-rated creativity, especially for less creative writers, but also increased similarity across stories. That was a restricted idea-assistance task, not sustained collaboration on long fiction; its relevance is the need to evaluate artifact quality and collective diversity separately (Doshi and Hauser 2024). No such experiment has been conducted for *The Second Country*.

13.6 A prospective test of autonomous literature mapping

Use fresh research questions and the same publication-oriented task, tools, source access, and resource budget. In one condition, request a well-supported research analysis without mentioning a literature review; in another, add a target-neutral protocol requiring named predecessors, rival views, source checking, and gap analysis. Archive the first scaffold, actual searches and source reads, and the revised analysis. An optional no-retrieval condition can diagnose learned synthesis, but it is not the strongest autonomous baseline.

Blind subject specialists to condition and have them assess coverage of consequential alternatives, attribution accuracy, source entailment, handling of current evidence, warranted novelty claims, and whether retrieval actually corrected the model. Record time and cost, inaccessible texts, and null or unhelpful searches. An answer earns no credit for merely adding citations; tool logs and inspected passages, not the model's statement that it searched, establish the process.

This comparison would separate the capacity to perform a review from reliable initiation under an ordinary scholarly assignment. It would not identify training-data provenance or prove that a human-suggested distinction was original. Preregister the rubric, include useful model-originated alternatives the human did not anticipate, and repeat across topics and configurations.

14 Implications for research practice

Some features are particularly visible in Bayesian research: reference-class choice, prior sensitivity, likelihood ratios, conditional dependence, and propagation of uncertainty. A missing hypothesis can change both normalization and the interpretation of evidence, which makes conceptual omissions consequential and easier to locate. But causal modeling, missing alternatives, source verification, analogy, and utility specification matter outside explicitly Bayesian work too.

The portfolio therefore supports a qualified expansion of Bayesian practice. AI can lower the coordination cost of exposing assumptions and rerunning alternatives. It does not remove the need to understand the subject, probability, or the provenance of the evidence. Researchers should adopt numerical models when the numbers reveal assumptions and decision sensitivity, not when a percentage merely makes an uncertain narrative look scientific.

The practical workflow recommended here has ten steps:

1. Define the target and stopping conditions.
2. Record genuine antecedent beliefs, without inferring them later.
3. Preserve a short provisional scaffold.
4. Conduct explicit literature reconnaissance before freezing the hypothesis set.
5. Construct alternative partitions and families of priors.
6. Map evidence and dependence.
7. Implement and test the calculation.
8. Run a fresh, target-neutral audit.
9. Perform an adversarial literature sweep for omitted rivals, recent results, and source-level criticism.
10. Compare human and AI proposals under a common validity standard.

Specialist scrutiny should target the premises that can change the conclusion, with strong opposing arguments represented where relevant. Publication should freeze the version, definitions, source support, rejected suggestions, uncertainty, and scope of review. Companion artifacts such as executive summaries must be checked against the main paper after each revision; a correct paper paired with an obsolete summary remains a publication error.

No stage makes the human a mandatory veto-holder on technical correctness, and no model-generated citation list substitutes for verification. Delegation should follow demonstrated performance at the task, while substantive value choices and publication responsibility remain explicit. The strongest configuration may change from one phase of a project to the next.

The most important institutional benefit may be inspectability rather than a higher probability of being right. A reader can see which disputed assumption produces a conclusion, which finding would change it, and which new information would be worth seeking. That is a defensible goal for historical, philosophical, legal, and policy work even when no agreed final answer is available.

15 Conclusion

The eight analytical projects document a real division of intellectual labor. AI did more than summarize: it constructed arguments, formalized uncertainty, checked sources, wrote code, criticized other analyses, and helped move a forecast from a causal ledger to an auditable district-and-race model. The human did more than request prose: he changed targets, selected neglected comparisons, supplied analogies, pressed the

implications of decisions for persons and institutions, and required that the forecast's national reasoning be tested against the actual map. Specialists and rival AI reviewers changed the work in different ways. Some changes were demonstrable local corrections; others remain substantive proposals rather than validated improvements.

The record does not establish that these combinations outperform the strongest AI-only alternative, that the human interventions were impossible for AI, or that a human advantage is contracting at a measured rate. Nor does it support the opposite conclusion that every human contribution was mere prompting or steering. The evidence motivates comparisons that can distinguish those possibilities.

The creative extension adds a source-recovery and revision sequence, not a further independent demonstration of synergy. It shows why an inherited manuscript and an author's actual writing must not disappear from the account merely because an AI produces a longer version, and it also records substantial AI invention. Whether the resulting story is better remains open to evaluation; fidelity to authorial intention and the value of making a work are distinct from comparative cognitive performance.

Bayesian structure is central because it makes assumptions, dependencies, and downstream consequences inspectable. Its growing accessibility is an opportunity to improve research practice, not permission to manufacture precision. The literature-review vignette adds a second distinction: a model can possess a literature-shaped field map and perform substantial task-time reasoning without having performed source-grounded scholarly reconnaissance. The rising-waterline framework adds a third: AI can progress from ledger-keeping and error flags to contestable judgment support and evidence discovery. Publication-grade research should record when the workflow switches among these modes and what changed because of retrieval or audit.

The shrinking island is best understood as a hypothesis about the changing marginal value of human contributions under specified conditions. Its rigorous continuation is to measure recovery, validity, research-mode switching, calibration, comparative performance, and cost, while keeping authority and value choice conceptually separate from cognitive advantage.

About the author

Rodney Malpert holds a Juris Doctor and a Master of Arts from Cornell University, where he also completed all requirements for the Ph.D. in Government except the dissertation; his graduate work included substantial coursework in analytic philosophy. He practiced law for many years in Silicon Valley and has taught law as an adjunct professor at Southern

Methodist University and Arizona State University. He holds the rank of Senior Master with the United States Chess Federation. His current research examines philosophy, history, public policy, science, and creative writing, with particular attention to the contributions people and AI systems make to different forms of intellectual work.

Website and current papers: RodMalpert.com

Attribution and disclosure

Rodney Malpert originated the research program and the shrinking-island metaphor, directed the inquiries, and requested an explicit Bayesian account of the collaboration. His contributions included challenging the treatment of theatrical association and authorial opportunity in the Shakespeare question; distinguishing an ensemble of UAP reports from its individual members; separating ultimate ontology from local architecture; and pressing the implications of an external experiencing subject for survival. In cryonics, he introduced or developed the ordinary-change analogy, undesirable revival, postmortem displacement, the evidential value of early revivals, inaccessible memories, and conditions for delaying activation. In AI risk, he directed attention to the superintelligence transition, distinct extinction and dystopia outcomes, successor creation, and alternative motivations. In the congressional forecast, he framed the causal questions, insisted that a national estimate be tested against the current district and Senate maps, and required that the caucus treatment of an independent be stated as an assumption rather than treated as evidence. He also formulated the rising-waterline sequence (ledger and arithmetic, then error detection, then Bayesian colleague, then evidence discovery). That is credited as a project-level observation, not as a claim of priority over the Bayesian, machine-learning, or information-retrieval literatures. These are recorded contributions within the project, not claims of independently established correctness. The participant profile rests on self-description and observed interaction, not a psychometric assessment.

ChatGPT (OpenAI) contributed sustained dialogue, synthesis, hypothesis formalization, source-oriented revision, calculation and code, integration of criticism, and preparation of the manuscripts, including conceptual changes as well as implementation. In the September 25 vignette it supplied the three-layer analysis of parametric knowledge, task-time synthesis, and explicit literature reconnaissance after Malpert identified and commissioned the issue; the analysis was then checked against public documentation, relevant research, and the visible dialogue. In the forecast it assisted with retrieval, model implementation, sensitivity analysis, source checking, and methodological integration. Its account of its own work is functional and evidential, not an authenticated reconstruction

of internal computation. It also challenged human suggestions, including a motive-based objection to female Shakespeare-authorship candidates, and qualified claims about continuity, inference, and revival.

Claude (Anthropic) contributed substantive co-drafting, especially in metaphysics and cryonics, structural criticism, and forecast critique and replication. Claude prepared the October 6 circulation edit. ChatGPT then clarified the companion-paper relationship, delegation levels, and outcome-test claims; checked the cited forecast figures and illustrative arithmetic; repaired equation typography; and reviewed the rendered layout. This editorial pass preserves the eight analytical cases and creative extension. It does not re-audit the underlying historical or metaphysical estimates.

Grok (xAI) supplied targeted challenges to likelihood precision, dependence, arithmetic, and aspects of the UAP analysis. Gemini Research supplied synthesis in the AI-risk review; the record also documents labeling and source-support problems in that output. Richard Carrier supplied specialist criticism of the Jesus-historicity project, including the representation of his own argument; his involvement is evidence of adversarial specialist engagement, not neutral certification of the posterior. None of these contributors or providers is represented as endorsing this paper. Agreement among assistants is not treated as independent corroboration, and the companion papers' conclusions do not independently validate this postmortem.

Project conversations sometimes used the labels "5.6/Sol" and "6.0/Astra" for ChatGPT sessions. Those labels are conversational provenance; authenticated backend snapshots and matched experimental conditions are unavailable, and no provider-wide benchmark or isolated test of a later model generation is claimed. Appendix A identifies the source versions underlying the case descriptions, which concern recorded stages of the projects rather than every later version. The underlying manuscripts and dialogue excerpts are not a complete public transcript archive, and this paper does not supply the full source packets needed for independent replication.

Malpert originated *The Second Country* and supplied its controlling manuscript, with substantial direct writing and revision reported for its earlier development. He requested the return to the story in light of the philosophical work, corrected the source omission, and directed the source-based synthesis. ChatGPT developed the interim restoration plot, drafted substantial new material, and integrated it with the recovered source. The same assistant helped produce both the fiction and this account of it; the latter is not independent literary validation. [A11-A13]

Appendix A: Archival sources

The entries below identify inspected sources. Document attributions are evidence of what the project records, not independent proof of every claim in those documents. Page counts distinguish versions whose titles otherwise repeat. The notation [A1] and so forth refers to this appendix.

A1. Shakespeare. Malpert and ChatGPT, *Who Wrote Shakespeare? A Bayesian Reassessment*, September 2026, eight-page working summary. Sections 2 and 21 assign the 80 percent starting prior to ChatGPT; §§5-9 describe the platform argument; §17 records the objection to motive-based assessment of female candidates; §§19-20 expose the arithmetic and sensitivity. The retrieved record establishes no separate specialist or independent AI audit of that version.

A2. Jesus. *Jesus Historicity Post-Carrier 38/62*, September 2026, post-Carrier manuscript. Sections 19-20 and the attribution endnote document the reversal, stated source corrections, and specialist contribution. The paper describes Carrier's communication; this postmortem does not independently verify every underlying historical assertion or treat the communication as neutral peer review.

A3. UAP. Malpert and ChatGPT, *What Are UFOs? A Bayesian Assessment*, version dated September 11, 2026. Conclusion and attribution on p. 22; reproducible sensitivity appendix from p. 23. These support the ensemble distinction, the documented Grok contribution, and the qualification of the numerical scenario.

A4a. Metaphysics development. *A Two-Level Bayesian Analysis of Fundamental Reality*, September 2026, 22 pages. The title page names Malpert, Claude, and ChatGPT, with Grok criticism. Section VI corrects the likelihood-ratio ceiling; §XI identifies Claude's analysis and records subsequent development; §XII records contributions. Source titles alone do not uniquely identify the inspected records.

A4b. Metaphysics companion. *A Two-Level Bayesian Analysis of Reality, Consciousness, and the Physical World*, September 2026, 21-page version with aggregate weights of about R40/P26/I23/D10. Sections XXII-XXIII distinguish human conceptual direction, AI implementation, and local versus ultimate architecture. This is distinct from A4a and its different numerical assumptions.

A5. Survival companion. *A Bayesian Analysis of Life After Death*, September 2026, 31-page version. Sections 1 and 10 define broad continuation and state the approximately 45 percent headline; §14 records contributions. This version is designated explicitly rather than conflated with an earlier 35/16 broad/narrow pair.

A6a. Cryonics before the later extensions. *Cryonics and Personal Survival Across Metaphysical Frameworks*, final-circulation file dated September 17, 2026, 20 pages. Its abstract already includes displacement of a preferable postmortem state. The accompanying five-page audit includes option value and postmortem counterfactual utilities. These are documentary precursors, not retrospective reconstructions.

A6b. Cryonics audit evidence. The five-page audit associated with A6a, especially §§3-5, records the success-definition correction, overlapping partitions, inference qualification, sensitivity limitations, and the approximately 45 percent companion definition. This is an AI-authored audit account; its reproducible and source-level claims remain individually checkable.

A6c. Cryonics current reference. *Cryonics and Personal Survival Across Metaphysical Frameworks*, final circulation, September 17, 2026, 26 pages, with a three-page final audit note. Sections 9-12 develop afterlife interaction, H-A-R, differing memory-loss models, sequential choice, and welfare; the audit distinguishes candidate recall exposure from actual harm. A separately supplied 27-page Claude revision is a precursor, not an interchangeable copy.

A7. AI risk. *AI Existential Risk and Severe Dystopia: A Transition-Centered Bayesian Assessment*, full analytical manuscript, September 15, 2026, 44 pages. Appendix A compares Claude and Gemini reviews; Appendix C separates Malpert's interventions from Claude's corrections and ChatGPT's implementation; Appendix D records open substantive work. This is not a completed, empirically calibrated forecast.

A8. Earlier postmortem. *The Shrinking Island: Bayesian Centaur Research*, September 17, 2026, 12 pages. Used as an earlier statement of the thesis and participant profile, not independent corroboration.

A9. Adversarial postmortem. Claude, *The Shrinking Island*, September 17, 2026, 19 pages. Supplies methodological criticism of contribution claims, elicitation and information gaps, specialist review, and the proposed replay design. A source in the collaboration record, not independent validation.

A10. Supplied dialogue. Malpert and AI assistants, conversation excerpts of September 16-17, 2026 concerning cryonics and the postmortem: the ordinary-change and inference discussion; the complete-audit request explicitly inviting new topics; the afterlife, first-revival, and memory-access exchanges; and Malpert's requested description of the human generalist. Direct sequence evidence where visible, but not a complete transcript or authenticated model metadata.

A11a. Controlling creative source. Malpert, *The Second Country*, 13-page screenplay-style manuscript supplied September 24, 2026. It contains the achievement-centered Thymoi, the Teleians, the free-will/desert and heaven exchanges, the Great Filter, the second-country revelation, the key and flame, the warning about civilizational completion, and the proposal to eliminate drawn outcomes in a new chess variant. It is the controlling source for A11d; its submission does not by itself establish sole authorship of every passage.

A11b. Earlier creative sources and retrieval failure. Malpert and ChatGPT, *The Long Memory*, April 17, 2026, approximately 3,400 words, plus a nine-page "final" PDF of roughly 1,100 words that was initially misidentified as the latest complete story. A library file titled *The Second Country*, April 18, 2026, proved to be a 66-byte unreadable artifact. This explains why the first September source audit concluded the later manuscript could not be recovered; it was not the 13-page manuscript later supplied directly by Malpert.

A11c. Interim creative expansion. *The Second Country*, September 24, 2026, approximately 8,000 words in ten sections, drafted before A11a was available. It adds the restoration-and-consent dispute and related material while changing some thematic emphases and the chess conclusion. Retained as an archival branch, superseded by A11d.

A11d. Source-based synthesis. Malpert, *The Second Country*, September 24, 2026, nine-section prose edition of approximately 7,600 narrative words. Prepared with A11a and A11c both available, it restores the controlling manuscript's principal philosophical exchanges and decisive-chess proposal and retains a revised restoration plot. Narrative originality and quality have not been independently scored.

A12. Revision instructions and recovery record. Malpert's September 24 requests call for an ambitious revision, cross-project philosophical development, and a corresponding postmortem update. Subsequent messages supply A11a, distinguish it from A11c, request restoration of the governing source, report substantial earlier passage-level writing, and require fair primary-author attribution. The record documents requests and outputs, not a word-level authorship ledger or a blinded experiment.

A13. Philosophical sources for the creative transfer. *A Two-Level Bayesian Analysis of Reality, Consciousness, and the Physical World*, September 20 edition, §§V.A, XIV-XVII, XX (local versus ultimate explanation; participatory versus generated consciousness). *A Bayesian Analysis of Life After Death*, §§1, 8 (meaningful continuity and reintegration). *Cryonics and Personal Survival Across Metaphysical Frameworks*, September 24 circulation copy, §§9.6-9.8, 10 (consent, changing identity, uncertain return, separation of preservation from activation). These motivate dramatic alternatives; their numerical estimates are not inputs to any fictional probability.

A14. Literature-review vignette. Malpert and ChatGPT, September 25, 2026, dialogue developing an unfinished Buddhist metaphysics and meditation proposal. Visible sequence: Malpert specifies the philosophical and Bayesian task; the assistant supplies an analytic scaffold and named comparisons; discussion adds empirical questions and targeted references; Malpert requests an organized broad review; the review presents named predecessors, critics, and empirical entry points; Malpert then distinguishes learned field knowledge from task-time synthesis and formal review and commissions Section 3.3. The record supports that sequence, not the absence of every prior retrieval or an exact training- source attribution. His observation about Claude is participant testimony. Not a counted analytical case.

A15. Prospective forecast case. Malpert, *The 2026 Midterms: Forecast and Final-Month Republican Strategy*, circulation edition, October 5, 2026, 32 pages. Sections 1-9 and Appendices A-C state the 435-district model, historical model member, race-level Senate inputs, audited regression, correlated uncertainty, and replication limits. Central House estimate 236 Democratic seats (80 percent range 219-255); Senate center 51 Democratic-caucus seats. Section 9 and the attribution distinguish model reproduction from calibration and report the unreproduced Nebraska sensitivity branch. A prospective test, not yet evidence of forecast accuracy.

A16. Bayesian FAQ companion. Malpert, *Bayesian Reasoning in the Social Sciences: A Friendly FAQ on Priors, Evidence, AI Assistance, and Why Vague Words Cannot Hide Our Assumptions*, final edition, October 6, 2026. Explains prior odds, likelihood ratios, dependence, sensitivity, explicit ranges, and AI assistance for a social-science audience. A pedagogical companion, not independent validation of this postmortem.

Appendix B: Episode records for replication

Each episode should have a unique identifier, topic, packet hash, source version and locator, timestamp where available, and a record of what information was available before the intervention. Record the initiating participant, any visible precursor, the original proposal, the problem it addressed, and the subsequent implementation. Separate recorded attribution from independently adjudicated originality.

The disposition should distinguish retained, qualified, rejected, superseded, and unresolved. Evaluation fields should separately record source fidelity, logic, relevance, novelty relative to the packet, decision consequence, and cost. Label the evidence as

direct dialogue, document comparison, retrospective attribution, or participant report, and do not collapse these dimensions into a single "good contribution" column without a declared scoring rule.

For replay, add the exact model identifier returned by the serving system, access date, prompt, tools, context arm, role, sampling parameters, resource use, raw output hash, blinded ratings, disagreements, and contamination status. Preserve all suggestions, not only matches to the human intervention. The same record applies to human, specialist, and AI proposals. This schema is prospective; the case descriptions in this paper do not constitute a coded experimental dataset.

References

- Brown, Tom B., et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877-1901. arXiv:2005.14165.
- Carlini, Nicholas, et al. 2021. "Extracting Training Data from Large Language Models." *30th USENIX Security Symposium*: 2633-2650. arXiv:2012.07805.
- Cohen, Andrew L., and Adrian Staub. 2015. "Within-Subject Consistency and Between- Subject Variability in Bayesian Reasoning Strategies." *Cognitive Psychology* 81: 26-47. doi:10.1016/j.cogpsych.2015.08.001.
- Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayner, François Cadelon, and Karim R. Lakhani. 2026. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." *Organization Science* 37(2): 403-423. doi:10.1287/orsc.2025.21838.
- Doshi, Anil R., and Oliver P. Hauser. 2024. "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content." *Science Advances* 10: eadn5290. doi:10.1126/sciadv.adn5290.
- Fairfield, Tasha, and Andrew E. Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25(3): 363-380. doi:10.1017/pan.2017.14.
- Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." arXiv:2011.01808.
- Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review* 102(4): 684-704. doi:10.1037/0033-295X.102.4.684.

Goldwater, Micah B., Dedre Gentner, Nicole D. LaDue, and Julie C. Libarkin. 2021. "Analogy Generation in Science Experts and Novices." *Cognitive Science* 45(9): e13036. doi:10.1111/cogs.13036.

Gonzalez, Cleotilde, Kate Donahue, Daniel G. Goldstein, Hoda Heidari, Mohammad S. Jalali, Beau Schelble, Aarti Singh, and Anita Williams Woolley. 2026. "Toward a Science of Human-AI Teaming for Decision Making: A Complementarity Framework." *PNAS Nexus* 5(3): pgag030. doi:10.1093/pnasnexus/pgag030.

Hemmer, Patrick, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2025. "Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence." *European Journal of Information Systems* 34(6): 979-1002. doi:10.1080/0960085X.2025.2475962. Preprint: arXiv:2404.00029 (2024).

International Correspondence Chess Federation (ICCF). "WC33/final, World Championship 33 Final." Official crosstable, event 100104; started November 20, 2022; 136 finished games recorded. Accessed September 17, 2026. <https://www.iccf.com/event?id=100104>.

Kasparov, Garry. 2010. "The Chess Master and the Computer." *The New York Review of Books*, February 11. <https://www.nybooks.com/articles/2010/02/11/the-chess-master-and-the-computer/>.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. "Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33.

Malpert, Rodney, with AI collaborators. 2026. Centaur Research working papers, audit reports, and supplied conversations. Version-specific titles, contributors, dates, and locators are listed in Appendix A. Unpublished project materials.

OpenAI. n.d. "How ChatGPT and Our Foundation Models Are Developed." Accessed September 25, 2026. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed>.

Petroni, Fabio, et al. 2019. "Language Models as Knowledge Bases?" *Proceedings of EMNLP-IJCNLP 2019*: 2463-2473. doi:10.18653/v1/D19-1250.

Qiu, Linlu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. 2026. "Bayesian Teaching Enables Probabilistic Reasoning in Large Language Models." *Nature Communications* 17: 1238. doi:10.1038/s41467-025-67998-6.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askeell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. "Towards Understanding Sycophancy in Language Models." arXiv:2310.13548.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." Advances in Neural Information Processing Systems 36. arXiv:2305.04388.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas W. Malone. 2024. "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis." Nature Human Behaviour 8: 2293-2303. doi:10.1038/s41562-024-02024-1.

Youssef, Paul, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. "Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models." Findings of the Association for Computational Linguistics: EMNLP 2023: 15588-15605. doi:10.18653/v1/2023.findings-emnlp.1043.

Appendix C. Research judgment, fast communication, and cross-paper consistency

Bostrom's discussion with Roman Yampolskiy on October 16, 2026 supplies a useful external illustration of AI-assisted intellectual work and the value of communicating ideas while a field is changing quickly. It belongs alongside the case record as reflective testimony from a researcher, not as a controlled comparison of human-only, AI-only, and collaborative performance. Its relevance is to what a good workflow should measure.

C.1 Research taste is a task, not an unexplained human advantage

Choosing a worthwhile question, recognizing the useful level of abstraction, demanding an exact formulation, and abandoning a weak approach can determine the value of a research project. Calling those activities "taste" identifies something important but does not establish that only humans can perform it. A system might fail because the objective was poorly specified, because necessary context was withheld, because it lacked tools or time, or because the capability is not yet reliable.

The interview's account of using AI in intellectual work should therefore be tested against the same distinctions applied to this archive. A human selecting a strong question is an observed contribution. A claim that an adequately instructed model could not have selected it is a counterfactual requiring comparison. Better elicitation may recover the behavior; autonomous recognition that the behavior is needed remains a further test.

The frontier can consequently move along several axes: generating candidate questions, judging their value, finding missing evidence, detecting a faulty partition, expressing an argument precisely, and deciding when to communicate. A single division between "human creativity" and "AI calculation" cannot describe this range.

C.2 A linked portfolio reveals a separate kind of error

Five locally plausible papers can disagree because they use different versions of a number, different meanings of survival, or differently conditioned probabilities. A consistency audit must distinguish those cases. The correct repair to an obsolete citation is an update. The correct repair to different conditioning is an explicit mapping. A substantive model disagreement requires argument and, where justified, recalculation.

The coordinated review supplies a concrete example. The metaphysics and cryonics texts cited roughly 45% continuation, while the current survival benchmark was 46.56%, reported as 47%. Separately, cryonics' own 43.87% no-preservation counterfactual was generated by a different mixture. Updating the stale companion reference was appropriate; relabeling the cryonics calculation as 47% without rebuilding it would have introduced a new error.

A deeper example concerns generated simulated subjects. The survival model grouped them with ordinary local production at a 2% continuation conditional. The preservation

discussion prompted an explicit split and sensitivity calculation. At a 50% generated-branch continuation conditional, the overall result becomes 51.20%. This demonstrates a change in what the model exposes, not empirical validation of that conditional or proof that the new scenario is closer to truth.

These examples separate arithmetic checking, semantic consistency, model expansion, and evidential calibration. The first three can be demonstrated within the artifacts. The fourth remains much harder for speculative philosophical models.

C.3 The division of contribution in this episode

Malpert selected the interview, emphasized its significance, required its application across the portfolio, added cryonics to the comparison, and specified three goals: continuity, substantive citation, and revisions separated into elaboration, reconsideration, and new topics. ChatGPT performed the current source comparison, identified stale cross-references, formalized the generated-preservation sensitivity and proportional bridge, and drafted the coordinated extensions.

This attribution concerns the documented interaction. It does not claim historical priority for preservation, bargaining, digital welfare, or branching identity, all of which have substantial antecedents. Nor does it establish that AI could not have proposed the same research agenda unprompted. Earlier contributions from other models remain credited in the original case histories; they did not independently audit this extension.

C.4 Timeliness is part of usefulness

A correct idea delivered after the relevant decision can have less value than a carefully qualified analysis delivered while it can still change action. That gives a reason to use interviews, short arguments, working papers, and living documents in fast-moving fields. It does not give those formats automatic evidential weight or make a recently expressed view more likely to be true.

The useful unit is a claim with its support and limitations. An interview can reveal a neglected mechanism or make a distinction accessible; a longer paper can expose premises and objections; a calculation can show which assumptions alter the result. Readers should be able to move between these levels without discovering that the accessible account states a stronger conclusion than the technical one.

A practical release standard asks whether the main claim is intelligible, its decisive sources are identified, calculations are reproducible where used, and important live objections are stated. Revision should then be possible without obscuring which version supported an earlier conclusion. Speed and correction are complementary only when the publication record remains inspectable.

C.5 A prospective test of research direction

A useful experiment would give human-only, AI-only, and collaborative workflows the same unfamiliar problem, source access, time or cost limits, and evaluation criteria. It should score question selection, discovery of omitted mechanisms, source fidelity, invalid inference rate, quality of uncertainty statements, and time to a decision-relevant result. It should record both successful and rejected suggestions rather than count only the memorable interventions.

Two AI conditions would be especially informative: an open-ended assignment and one explicitly requesting an adversarial literature and cross-document consistency audit. Their difference would help distinguish spontaneous research direction from performance recovered by instructions. Independent evaluators could assess argument quality without being told which workflow produced it. Agreement with the author's initial beliefs would not be the scoring rule.

The interview therefore supports an expanded research agenda, not a claim that Bostrom endorses this portfolio or that a prestigious participant validates its conclusions. The defensible outcome is a more explicit account of research judgment, cross-paper checking, and timely communication, with the residual human advantage left open to testing.

Sources

Source: Bostrom, Nick, interviewed by Roman Yampolskiy (2026). "Superintelligence Author: We May Already Have AGI | Nick Bostrom." The Roman Forum with Roman Yampolskiy, YouTube, October 16. <https://www.youtube.com/watch?v=m880ajt5sNM>. Interview discussion; conceptual claims and reported judgments, not an empirical calibration dataset.

Source: The main paper's Sections 3.3, 3.4, 13, and 14 supply the distinctions among learned knowledge, task-time reasoning, explicit retrieval, and prospective evaluation. The October 18 episode uses the current five-paper manuscripts and the documented user instructions; its numerical example is reproduced in the accompanying calculation source. The episode is a methodological illustration, not an additional independently validated case study.