

Cognitive Mediation of Relational Evidence: An Investigation of Epistemic Judgment and Reliance

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

Human observers can encounter representations of the same relational contribution history yet form different beliefs about its significance and trustworthiness. This third paper in a coordinated research programme investigates cognitive processes between the structural availability of evidence and an observer's assessment of a person, claim, or contribution. Its proposed analytic objects are selective attention, evidence encoding, reconstructive memory, attribution, interpretive schemas, belief revision, epistemic vigilance, and behavioral reliance. The inquiry distinguishes evidentiary accessibility from cognitive encounter, subjective confidence from justified assessment, and expressed trust from risk-bearing cooperation. Prior work in attention, working memory, judgment under uncertainty, epistemic vigilance, and trust in automation motivates a provisional sequential-kernel architecture; that architecture remains a research proposal rather than a validated general cognitive model. AI-mediated information environments supply a historically relevant application, without defining the paper's full scope. This working draft contains all ten planned sections, covering cognitive interfaces, selective attention, memory-based trajectory reconstruction, contribution attribution, proposition-specific belief revision, testimony evaluation, behavioral reliance, empirical designs with explicit identification boundaries, and a synthesis of conditional theoretical findings. It reports no original empirical results.

Keywords: relational evidence; cognitive mediation; epistemic judgment; behavioral reliance; selective attention; reconstructive memory; epistemic vigilance

Discussion Paper Note

This is the third investigation in a coordinated five-paper research programme. Paper II developed conditional structural models of relational-history observability and evidence access. Paper III investigates bounded human attention, encoding, memory, interpretation, inferential assessment, and epistemic vigilance when representations of such histories become cognitively available. The provisional processing architecture offered here is a candidate formal research language, not an empirically established cognitive mechanism or a universal ontology of memory and belief.

The document contains an initial complete draft of Sections 1–10, with all ten research sections reviewed through the completed section-by-section revision cycle. It remains an exploratory scholarly DRAFT pending integrated mathematical and interpretive audits, source-interpretation re-auditing, author approval, empirical testing, and independent scholarly review. Hypothetical examples are non-identifiable. A source-verification ledger accompanies this document.

Responsible Use and Rights Reservation

This page provides a request for responsible scholarly use separately from the legal permissions of the licence. It does not amend the licence or statutory exceptions. The author welcomes scholarly critique, verification, teaching, and responsible adaptation. Users should preserve distinctions among hypothetical mechanisms, verified empirical findings, normative proposals, and demonstrated institutional effects. Claims concerning persons or institutions require evidence proportionate to their consequences.

The copyright notice, Creative Commons licence, and applicable permissions appear in the Notices that follow. This request for responsible use introduces no additional legal or technical restriction on uses permitted by that licence or by applicable law.

DRAFT

Notices

Publication status. Sections 1–10 are drafted and have undergone individual revision. This is an exploratory scholarly DRAFT pending an integrated manuscript audit, scholarly critique, empirical testing, and author approval. No original empirical or causal estimates are reported.

Licence. Except where otherwise indicated, copyright © 2026 Wanhong Huang. Original material is made available under Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0). The licence and applicable law govern permitted use. Full licence: <https://creativecommons.org/licenses/by-nc/4.0/>.

AI-assisted preparation. OpenAI’s ChatGPT assisted with conceptual organization, bibliographic discovery, provisional scholarly drafting, and LaTeX preparation. The human author establishes the research object, determines the scientific claims and any eventual publication, and retains responsibility for evidence, interpretation and scholarly ethics. ChatGPT is not identified as an author.

Research independence. Hypothetical research scenarios are anonymous and not representations of specific institutions. Reuse of a typesetting template does not imply endorsement by an earlier project.

Contents

Discussion Paper Note	2
Responsible Use and Rights Reservation	3
Notices	4
1 Introduction	6
2 Cognitive Interfaces to Relational Evidence	13
3 Selective Attention and Evidentiary Encounter	23
4 Memory Compression and Trajectory Reconstruction	33
5 Interpretive Schemas and Attribution of Contribution	43
6 Belief Revision Under Surprisal and Diagnostic Evidence	54
7 Epistemic Vigilance and Testimony Evaluation	64
8 Epistemic Trust and Behavioral Reliance	75
9 Research Designs and Identification Limits	85
10 Conclusion	95
Research Ethics and Source Provenance	107

1 Introduction

The cognitive recognition of a relational history depends on more than the availability of a record describing it. An observer may have access to an earlier contribution and yet fail to encounter the record, may encounter the record without retaining its provenance, or may accurately recall the record while drawing an inference that its evidential content does not support. Conversely, a relatively minor change in the format, contextual explanation, or order of presentation of an unchanged record may enable an observer to identify a relevant claim or revise an earlier judgment. The object of this paper is the sequence of observer-level processes through which available relational evidence becomes attended to, represented, interpreted, and used in assessments of knowledge, competence, and trust. Its method combines distinctions from cognitive psychology and social epistemology with a provisional formal representation of context-conditioned information processing. The argument specifies separate cognitive mechanisms, observable outcomes, and research limitations before proposing experiments capable of comparing them.

The preceding investigation in this research programme examined relational conditions under which a historical contribution becomes observable to an additional observer. It distinguished documentary existence, permission or technical access, evidential provenance, actual examination, and the possibility of warranted assessment. This paper begins at the interface where accessible evidence enters a human cognitive task. Two observers may have equivalent access to the same evidence-bearing representations and nevertheless reach different assessments. The difference may reflect attention allocation, prior experience, limited memory, representations of social relationships, uncertainty about the source, competing commitments, or the requirements of the decision being made. Some differences are justified by relevant information unavailable to one observer. Others represent failures of retrieval, inference, or calibration. The present inquiry aims to distinguish those cases rather than treating disagreement itself as evidence of bias.

The distinction has a historically specific importance where scientific, administrative, and public information is increasingly mediated by searchable repositories, automated recommendation, summarization, and artificial-intelligence systems. A document may be retrieved through a platform, condensed into a generated summary, presented with social or institutional cues, and subsequently repeated by another person. The human recipient can encounter an apparently coherent account while possessing only partial access to its provenance and supporting evidence. The frequency of encounter can grow while understanding of evidential independence remains limited. This description identifies a plausible configuration of contemporary information work, not a claim that AI-mediated communication is uniformly deceptive or epistemically inferior. The related literature on trust in automation describes how context, imperfect system performance, and the cognitive mechanisms of human operators shape appropriate reliance [51]. A cognitive theory of relational recognition must accommodate both the benefits of mediation and the errors that particular forms of mediation can introduce.

The available empirical research illustrates why interface design, performance, and reliance require separate variables. In a study involving 1,526 participants drawn from a U.S. sample, Kaiser and colleagues compared practical information-search tasks using ChatGPT or Google. They reported faster task completion and a higher frequency of correct answers in the ChatGPT condition, alongside less reliance on primary web sources; the publication is a conference extended abstract, and the results are specific to the tasks and conditions examined [41]. In an incentivized interactive experiment, Klingbeil and colleagues found that source labeling and advice could influence participants' reliance on AI, sometimes despite conflicting contextual evidence, with consequences for their own and others' experimental payoffs [45]. These studies do not estimate the proposed relational-recognition kernels or establish that AI mediation has a single directional effect on judgment. Together they motivate an empirically discriminating question: which changes arise from what evidence is consulted, which from the way a source is interpreted, and which from the decision to act on a judgment? A reduction in independent verification can coexist with greater performance on a bounded task; an increase in declared confidence can coexist with an avoidable decision error. The appropriate comparison consequently depends on the proposition to be assessed, the opportunities to consult independent materials, and the stakes of reliance.

A historical-materialist orientation further directs attention to the material organization of cognitive effort. Increases in the capacity to produce and preserve information do not provide human evaluators with unlimited time, expertise, working memory, or institutional opportunity to examine each contribution. Tasks of indexing, translation, reviewing, teaching, summarizing, and making expert judgments require labor and are embedded in historically changing organizations. Scarcity of an evaluator's attention does not by itself establish exploitation or injustice; it does create constraints under which decisions about examination become consequential. The relevant research question is how those constraints interact with cognitive selection and interpretation after records have become materially accessible. Attention and memory are properties of human processing under particular conditions, whereas the distribution of evaluative work and the design of information interfaces are social and technical arrangements. A useful account specifies their interactions while leaving the legitimacy and endogenous control of institutional arrangements for the separate investigation of power.

The historically situated mechanism proposed here is an uneven adjustment between productive and evaluative capacities. Tools may lower the time required to generate, reproduce, or translate claims while the time needed to examine provenance, substantive quality, or evidential independence remains costly and unequally distributed. This potential divergence creates several empirically distinguishable responses: some evaluators may inspect fewer claims, others may rely more heavily on summaries or referrals, and others may demand more provenance work from contributors before examination begins. The direction is conditional on improvements in searchability, the diagnostic quality of summaries, the allocation of review labor, and the availability of methods that genuinely reduce verification costs. A simple count of published items cannot establish the amount of trustworthy evidence produced or the burden of evaluating it. Historical comparison should instead reconstruct the changing organization of production and verification, recording who produces the primary materials, who performs inspection, who controls access to documentation, and who bears the opportunity cost of an incorrect judgment. Those arrangements may create unequal occasions for cognition even among observers with similar measured abilities; they should not be inferred from the observer's eventual response alone.

The limitations of human information processing have longstanding empirical and conceptual foundations. Simon's analysis of bounded rationality treats actual choice as a problem involving computational and informational simplifications rather than the frictionless evaluation of a complete option set [72]. Kahneman's account of attention and effort develops the limited allocation of processing resources to concurrent and competing demands [40]. Baddeley and Hitch's working-memory framework connects limitations of short-term processing to reasoning, comprehension, and learning [5]. These literatures concern distinguishable mechanisms and experimental settings. Their joint relevance is methodological: a model that assigns every accessible document an equal probability of full human comprehension, or treats all encountered details as permanently retained, lacks plausible cognitive foundations. The present paper uses these sources as constraints on candidate models and does not presume that one architecture exhausts the diversity of human cognitive strategies.

A further consideration is the purposeful use of external cognitive resources. Risko and Gilbert survey *cognitive offloading*, through which people alter the demands of a task by using actions or supports outside unaided internal processing, and discuss the role of metacognitive judgments in deciding when to offload [65]. Archive search, note-taking, source-link inspection, and summary consultation can be analyzed as potential extensions of an observer's inquiry rather than being coded automatically as cognitive failure. Offloading can preserve scarce attention for difficult inference, or it can create dependence on a representation that omits information needed for the target task. The relevant comparison concerns both the benefits of external support and the residual ability to identify source lineage, ambiguity, and error. The review does not itself test generative-AI evaluation of historical contribution, so the application remains a testable extension.

An observer's processing begins with a task-specific encounter. A hypothetical evaluator may receive an archive containing a participant's earlier work, a concise description of it, and a cross-context endorsement. The archive contains the material needed to verify an intellectual contribution, yet the evaluator may first inspect only the endorsement. Another evaluator may inspect the original artifact,

discover an inconsistency in its description, and discount the endorsement. A third may lack the technical knowledge necessary to assess the artifact and reasonably preserve uncertainty. These are alternative interpretations of an anonymous constructed situation, not observations of an actual institution. The example separates the available documentary history from the selected items of attention, the reconstructed account of authorship, the content inspected, and the resulting judgment. It also exhibits the possibility that a better-informed evaluator becomes less confident in a positive claim.

This paper accordingly treats *cognitive recognition* as a task-conditioned family of operations by which an observer detects, discriminates, recalls, and interprets evidentially relevant aspects of a represented relational history. The term includes recognition of an artifact's existence, reconstruction of contribution and provenance relations, and identification of the propositions to which evidence bears. It does not stipulate favorable evaluation, social popularity, or institutional standing. An observer might recognize a contribution accurately and find that its authorship remains uncertain; may recognize the source of a false statement; or may recognize the limits of their own competence to interpret a technical record. Recognition of a relationship is also distinguishable from recognition of the moral legitimacy of that relationship. The analytical scope concerns the cognition of evidentially represented trajectories, while acknowledging that such representations may be incomplete, contested, and revised.

The target proposition must also be specified before performance is evaluated. The claim that a document exists, the claim that a named actor authored a particular method, an estimate of that actor's competence, a prediction of future reliability, and a judgment about whether an institution ought to reward the actor have distinct evidentiary and normative requirements. A record can supply strong support for one proposition while remaining uninformative about another. For that reason, a study should state its response domain, the information that would count for or against each proposition, and whether the answer key permits uncertainty or multiple admissible interpretations. This prevents successful source recognition from being mistaken for demonstrated skill, and prevents a favorable assessment of skill from being mistaken for an entitlement to an institutional benefit. The paper's cognitive variables describe how evaluators process relevant representations; the legitimacy of an evaluative rule and the distribution of benefits under it require additional arguments.

The object called a *relational trajectory* is not assumed to consist of an ontologically fixed sequence of predefined event types. Its analytic representation may include records of productive activity, testimonies, cross-context contact, revisions, institutional decisions, and other context-relevant relations. Event individuation, ordering, and the evidentiary meaning of a relation may depend on the observer's research question and available sources. In some applications, dependencies admit a partial ordering; in others, reconstructing an ordering is itself an epistemic task. A cognitive observer may receive a narrative suggesting an unambiguous succession although the underlying records support multiple compatible reconstructions. The model must therefore distinguish the provenance-supported historical representation, the observer's internal model of that representation, and the evidence that warrants favoring one reconstruction over alternatives. This distinction avoids interpreting a remembered narrative as a transparent copy of social reality.

A central psychological problem is how attention relates to informational novelty. Statistical surprisal, measured relative to a specified distribution, can describe an event that the observer regards as unexpected. Yet attending to a stimulus, recognizing its relevance, understanding its causal implications, and regarding it as credible depend on different conditions. An unusual item may be ignored because its content is obscure or the task demands attention elsewhere. A familiar item may be highly diagnostic because it corrects a crucial attribution or provides independent corroboration. The classic study of heuristics under uncertainty identifies representativeness, availability, and anchoring as economical judgment procedures that can also yield systematic error in particular settings [79]. The proposed inquiry therefore allows attention to respond to relevance, expected usefulness, processing cost, salience, and prior expectations in different directions. A universal positive relationship between surprise and trust would collapse distinctions essential to the analysis.

Memory introduces another transformation. An observer who once examined a record may later remember a simplified conclusion, retain the conspicuous name of a recommender, or misattribute a

proposition to a more familiar source. Such reconstruction can occur even without new relational events affecting the original historical material. Schacter’s review of memory errors organizes forgetting and distortions, including misattribution, suggestibility, and bias, while emphasizing the adaptive functions associated with memory processes [68]. The relevant inference is circumscribed: remembering an earlier assessment does not establish that the observer can retrieve its documentary grounds, and high confidence in a recollection does not guarantee the accuracy of its provenance. Our formal development must allow source information, ordering information, and substantive content to be compressed or lost in different ways.

The resulting evaluation cannot be understood solely as a storage-and-retrieval process. Interpretive schemas help observers decide which features of evidence count as relevant, which counterfactual explanations merit consideration, and how a contribution relates to the competence of its producer. The same technical artifact can be assessed under different disciplinary standards or translated through descriptions that preserve distinct subsets of its properties. Disagreement may follow from differences in legitimate evaluative aims, incomplete evidence, or the availability of relevant interpretive knowledge. Cognitive-ecological approaches to judgment emphasize how sampled instances can condition judgment and produce systematic distortions even when each observed instance is accurately represented [22]. An empirical inquiry should accordingly track the evidence sampling process as well as the eventual judgment, distinguishing consequences of unequal exposure from consequences of inferential processing conditional on the same exposure.

Prior beliefs and motivations may influence which interpretations receive attention and how ambiguities are resolved. Kunda’s review of motivated reasoning investigates the effects of directional motivations on reasoning while recognizing constraints imposed by the availability of plausible justifications [48]. This literature warrants hypotheses about context-dependent processing, not an automatic attribution of motivated distortion to observers who disagree. A researcher can have a legitimate prior belief supported by domain evidence; an evaluator can give a source limited weight because its provenance is genuinely weak. The present analysis therefore reserves the term *cognitive bias* for deviations from an explicitly stated task-relative benchmark or a well-supported causal mechanism of processing distortion. Differences in judgments alone cannot identify a bias without specifying what information was available, which loss or inferential criterion was appropriate, and which competing interpretations were reasonable.

The distinction between a normative or epistemic benchmark and a descriptive cognitive process is especially important in trust research. An observer may report a high degree of confidence while failing to appreciate the dependence of two purportedly separate witnesses. Conversely, an observer may decline to trust a claim because the available evidence is not sufficiently diagnostic, even when the claim happens to be true. The benchmark must be stated with care: a calibrated probabilistic judgment can be evaluated against a well-defined target and data-generating setting, but many substantive claims about competence, integrity, or relational contribution resist reduction to a single objective probability without operational decisions. Bayesian inference can supply a conditional reference model when prior distributions and likelihoods are defensible. The observer’s actually expressed belief remains an empirical object whose dynamics need not follow the benchmark exactly.

The paper proposes a provisional processing architecture for specifying this distinction. Let H denote an evidence-bearing representation of a history, C a task and observer context, E the material evidence accessible at an interface, A the information attended to, M its encoded or retrieved cognitive representation, Z the observer’s interpretation, and J a specified judgment. For an observer o , one candidate factorization is

$$P_o(J = j \mid H = h, C = c) = \int K_o^{\text{jud}}(j \mid z, c) K_o^{\text{int}}(dz \mid m, c) K_o^{\text{mem}}(dm \mid a, e, c) K_o^{\text{att}}(da \mid e, c) K_o^{\text{acc}}(de \mid h, c). \quad (1)$$

Equation (1) is a *candidate restricted factorization*, not a discovered neural sequence. Each K is a conditional probability kernel on locally specified measurable representations, and its interpretation depends on a common model of the target task. The factorization assumes that the context C and the adjacent

represented states capture the dependencies omitted from each displayed conditional. It can fail when retrieval, judgment, attention, or external information feed back into earlier processing, when latent observer states are omitted, or when the representation itself changes during assessment. The full joint law could then require expanded state conditioning, coupled kernels, or a nonsequential model. The integration signs indicate summation or integration over appropriate intermediate states. Null-access and non-encounter outcomes must be represented explicitly rather than silently dropping observations. More extensive formal treatment belongs to Section 2.

The mathematical expression also has a substantive identification limit. Multiple systems of conditional kernels can induce the same observed distribution of final judgments. A fitted composition therefore cannot, by itself, establish that a changed response was mediated by attention, memory, or interpretation rather than by another unobserved route. Its factorization is an explicit restriction on represented dependence, and the labels attached to its components are hypotheses about measurement, not experimentally recovered psychological entities. Comparisons need intermediate observations and interventions suited to the task, while remaining attentive to the fact that asking for recall, justifications, or confidence reports can alter the processes being measured. In this sense the equation is a map of possible explanations and possible research operations; it acquires explanatory force only through empirically discriminating restrictions and tests.

This construction retains an interface with the structural theory while giving cognition its own research object. The access kernel K_o^{acc} is a contextual description of what can reach a specified observer, in the spirit of Paper II's observation model. The attention, memory, interpretation, and judgment kernels concern processes after and during cognitive encounter. That division is functional, not a commitment to a universal temporal order. For example, a prior belief can influence attention before the presented evidence is fully processed, and a provisional interpretation can determine which part of an archive is examined next. The conditioning context may itself change through the interaction. The purpose of the staged proposal is to isolate interventions and measurements, identify where a stronger model is required, and prevent a change in accessible information from being misclassified as a change in inferential competence.

The observer's interpretation of a history also depends on source assessment. Human communication frequently provides evidence through testimony, which permits collective knowledge while creating a risk of error and manipulation. Sperber and colleagues' account of epistemic vigilance examines how sensitivity to sources and contents can condition the acceptance of communicated information [73]. Relational histories may be reported by persons with firsthand observations, by intermediaries repeating a source, by automated summaries, or by a mixture of those pathways. An evaluator can be vigilant in assessing whether a source was competent and independent without thereby establishing that the source's claim is correct. Equally, a testimony may be substantively informative even when the intermediary lacks high institutional status. Source criticism should therefore distinguish provenance, actual information independence, perceived authority, and the source's capacity to support the target proposition.

Contemporary misinformation research adds another reason to study the internal representation of a claim rather than treating exposure and correction as equivalent to belief change. A review by Ecker and colleagues synthesizes cognitive, social, and affective mechanisms implicated in misinformation belief and the persistence of influence after correction [20]. The present theory does not claim that every correction fails or that every observer carries the same vulnerability. It uses the literature to motivate the possibility that an observer can update an explicit judgment while continuing to draw on a previous narrative or source association in later reasoning. A suitable experiment must measure immediate evaluation, delayed retrieval, provenance attribution, and subsequent decisions separately when such persistence is of interest. Otherwise the mechanism of recognition will be inferred from a single response that may conflate multiple cognitive states.

The meaning of *trust* also requires explicit separation. Organizational trust models distinguish features of the trustor and trustee, the situation, vulnerability, and behavioral outcomes [55]. Research on automation further distinguishes subjective trust from appropriate reliance on a system whose performance is imperfect and context-dependent [51]. These sources do not supply a universal definition of

Table 1: Observer-level analytical objects and discriminating measurements.

Analytical object	Representative measurement	Distinction preserved
Evidence availability	Document or provenance route accessible under a specified protocol	Availability differs from encounter
Attention and encounter	Inspection logs, time allocation, noticed evidence identifiers	Encounter differs from comprehension
Encoding and memory	Content recall, ordering reconstruction, source attribution	Recall confidence differs from fidelity
Interpretive assessment	Explanation of evidence relevance and competing hypotheses	Coherence differs from warrant
Subjective judgment	Belief rating, confidence and uncertainty interval	Favorable assessment differs from calibration
Declared trust and reliance	Stated trust, chosen verification, risk-bearing behavior	Declared belief differs from action

epistemic trust in relational histories. For the current project, a target-relative epistemic judgment concerns an observer’s support for a specified proposition, for example whether a historical contribution was made by a particular actor or whether a source has relevant expertise. Expressed trust concerns the observer’s reported willingness to credit a person or source. Behavioral reliance concerns an action taken under uncertainty and with potential consequences. Cooperation, credential recognition, and material return are additional outcomes affected by their own rules and opportunities. They should be operationalized separately from the observer’s confidence in an evidentiary proposition.

Table 1 summarizes the distinct objects that must be measured to test the proposed architecture. Its categories are analytic variables rather than a fixed typology of human cognitive events. A given task may require finer subdivisions, fewer stages, or a different representational basis. The separation is useful because an intervention can change one variable while leaving subsequent variables unchanged, or can affect them in opposing directions.

The table also shows why a favorable trust update is an inadequate general objective for research on human recognition. An observer who receives reliable counterevidence against an attribution may lower confidence, and that change may represent improved epistemic performance. An observer who becomes less certain because two apparently separate witnesses are revealed to depend on a common source may become better calibrated despite losing subjective confidence. Conversely, a persuasive narrative can produce greater confidence without adding diagnostic information. Proper evaluation must specify the target of judgment, the inferential benchmark, the cost of inquiry, and the possibility of justified suspension. It also requires measures of error that distinguish false positive trust from false negative distrust where the target proposition and decision setting permit such classification.

The proposed model generates hypotheses that are conditional rather than universally directional. First, holding accessible evidence constant, a change in salience or evaluation task may alter which portions are attended to and thereby affect recalled provenance and judgment. Second, holding examined information constant, a change in source labeling may affect inferential weighting when the observer believes the label carries information about reliability or standing. Third, a reminder that two accounts share a common origin may reduce inappropriate double-counting even if the number of visible messages remains fixed. Fourth, additional contextual explanation may improve recognition of a contribution for some observers while increasing processing cost or reducing attention for others. Fifth, receiving reliable negative evidence may improve the accuracy of evaluation while reducing declared trust. These directions reflect candidate effects under explicit task conditions; their magnitude, sign, and generalizability await tests.

A practical comparison can separate a change in what the interface supplies from a change in the observer’s treatment of the same supplied evidence. An experiment may independently manipulate the salience of a provenance cue, the availability of a directly inspectable source trail, and the time allotted to examine the corpus. Source-trail availability changes the potential evidence set; salience changes the presentation of an otherwise fixed evidentiary item; time constraints alter the cost of inquiry. Such dis-

tinctions matter because the same behavioral outcome may originate from improved information, altered selection, or changes in the perceived value of investigating. Observer-level variation in language proficiency, relevant expertise, accessibility requirements, prior familiarity, and verification goals should be considered as candidate moderators rather than folded into a generic “attention deficit.” Negative and inconclusive findings would inform whether a proposed stage is analytically useful, including cases where source transparency increases inspection without improving calibrated judgment, or where summary assistance improves accuracy without increasing source recall.

A research design should distinguish the effects of acquiring new information from the effects of processing existing information differently. An experimental comparison might provide participants with the same underlying hypothetical contribution history and randomize features of how its evidence is arranged: direct documents versus summaries, source-chain transparency, order of presentation, or opportunities to inspect primary materials. Outcomes should include actual inspection, recalled detail, inferred provenance, accuracy under an independently defined answer key, confidence calibration, declared trust, and voluntary information-seeking behavior. Such manipulation can reveal presentation-conditioned processing differences, but conclusions about external institutional practice require attention to participant expertise, task incentives, language, time budget, and the reliability of the materials. Randomizing visual salience does not randomize the validity of underlying testimony, and measuring response confidence does not measure warranted trust without a suitable criterion.

For causal interpretation, the assigned interface condition, the material actually inspected, the judgment elicited, and any subsequent action must be distinguished in the recorded data. Random allocation of a display feature may identify its average effect under a valid study design; it ordinarily does not identify the proportion of that effect transmitted by a particular cognitive mechanism. Optional inspection, remembered exposure, and post-treatment reports can depend on unobserved traits and on the intervention itself. A study wishing to estimate cognitive mediation therefore needs a separate causal model and a defensible strategy for the relevant intermediate variables, as developed in Section 9 [35]. Alternative accounts should include task demand, perceived institutional prestige, interface familiarity, and changes in effort. A successful design may yield only a reduced-form comparison, but that is more informative than treating an attractive sequence of cognitive labels as proof of mechanism.

Observational studies introduce further difficulties. A person who searches more extensively may have stronger prior interest, more specialist knowledge, greater skepticism, or better access to resources. Increased examination and improved judgment may therefore be associated without one causing the other. Retrospective recall of having encountered a document can itself depend on current beliefs and later communication. When participants select which records to examine, analyzing only their final judgments conditions on a post-access selection process. The methodological strategy should therefore retain denominators for accessible, presented, noticed, inspected, and assessed items; preserve source-lineage information; and, where feasible, record the sequence of voluntary examinations. The companion structural study’s distinction between access and examination is retained, while the cognitive study adds direct observations of encoding, interpretation, and judgment.

A stronger model must also allow feedback from judgment to later encounters. An initial tentative interpretation may change what a person subsequently searches for, how the person questions witnesses, or which summaries seem relevant. A recalled narrative can influence the further production of testimony, including through discussion, recommendation, and correction. Such feedback makes the observer an active participant in a relational trajectory, rather than a passive endpoint of information transmission. The present paper acknowledges this possibility without assuming an ontologically privileged present moment or universal temporal direction. Any indexed sequence used in a behavioral experiment describes an observation protocol and a specified ordering of task operations. Claims about history-dependent cognition require independently defined traces, relations of dependence, or justified partial orderings. A model may be revised when its chosen ordering fails to represent the interaction.

This feedback also exposes the limits of treating every recognition problem as an individual deficit. If essential context is absent, an observer may be unable to make a responsible inference even while processing the available material competently. If a highly compressed summary removes provenance,

the observer’s uncertainty may reflect the design of the communication interface. If access rules prevent verification, personal vigilance has limited power to recover missing records. Conversely, greater infrastructural openness does not guarantee accurate individual cognition. The division between the structurally possible and the cognitively realized is therefore analytically necessary. Paper II investigates access conditions, while the present paper investigates conditional processing under those conditions. The fourth paper is reserved for endogenous power over participation, authority, admissibility, attention and resources, including how evaluative institutions can constrain which observations count as legitimate evidence.

The fifth companion investigation addresses the circulation and conversion of references, attestations, endorsements, and other representations through which historical records can support further interactions and productive opportunities. Such circulation may depend on an observer’s trust assessment, but the act of judging evidence is distinct from converting that judgment into an employment decision, credential, collaboration, allocation, or material return. The present paper therefore limits its behavioral outcomes to those necessary to identify the cognitive relation between examination and reliance. It does not treat a positive assessment as an entitlement to a position, infer justice from favorable recognition, or suppose that credibility circulates as a uniform exchange token. The purpose of these boundaries is to make each mechanism researchable without dissolving cognition, power, and political economy into a single scalar concept.

The study’s proposed novelty should be assessed in relation to established research. Bounded rationality, selective attention, working-memory limitations, heuristic judgment, motivated reasoning, epistemic vigilance, trust in automation, and correction-resistant misinformation have substantial histories [72, 40, 5, 79, 73, 20]. The paper makes no priority claim to discovering these mechanisms. Its tentative contribution is a representation-sensitive integration directed at the assessment of relational histories: it aims to define how structural evidence access feeds into observer-specific cognitive processing, and to specify interventions that distinguish encounter, memory, source attribution, diagnostic inference, and consequential trust. Whether such integration produces an empirically distinct explanatory model is a question for theoretical comparison and experimental testing. Failure to improve on simpler attention, memory, or Bayesian models must remain a possible result.

The subsequent sections develop these commitments in a progressive but revisable analytical order. Section 2 specifies observer-relative evidence representations and the scope of candidate processing kernels. Section 3 examines attention allocation; Section 4 develops memory compression and provenance reconstruction; and Section 5 considers interpretive schemas, attribution, and relevant bias mechanisms. Section 6 treats diagnostic evidence and belief revision, while Section 7 investigates epistemic vigilance and source dependence. Section 8 distinguishes epistemic judgments, expressed trust, and behavioral reliance. Section 9 presents controlled and observational research designs with their identification limits, and Section 10 evaluates the conditional contributions and remaining uncertainties. Together, these investigations seek to explain how humans recognize and evaluate evidence-bearing relational histories under historically situated constraints on attention, cognition, and inquiry. They offer a research programme for testing such mechanisms, not an empirically demonstrated general law of trust formation.

2 Cognitive Interfaces to Relational Evidence

The analytical task of this section is to specify how an observer’s access to represented historical evidence is connected to the cognition through which that evidence is assessed. Section 1 introduced a candidate composition of access, attention, encoding, interpretation, and judgment kernels. That construction is useful only if the represented objects, conditioning information, and limits of the decomposition are explicit. A person may be able to obtain a record without attending to it, may remember its substantive claim while misidentifying its source, or may accurately reconstruct the evidence while reaching a different conclusion for defensible task-specific reasons. The present section therefore defines a local observer model, distinguishes failures of encounter from failures of representation, and states when the proposed factorization is defensible. Its formal claims concern a specified observation protocol rather

than a universal architecture of human cognition. The model also identifies what an investigator would have to measure before assigning a particular discrepancy to attention, memory, or inference.

2.1 The Observer, Task, and Evidentiary Interface

The first modeling decision is the unit of analysis. Let o index an observer and let q specify a target question, such as whether an artifact precedes a later claim, whether two testimonies share a source, or whether an individual contributed to a documented procedure. The index o identifies a participant in a specified study, not an invariant psychological essence. The question q identifies a proposition or decision criterion and may differ from the questions for which the original documents were created. A represented history H consists of artifacts, source indications, relation claims, and uncertainties that the investigator elects to preserve for the task. It can include contradictory evidence or several admissible partial orderings. Its units of description are provisional: an investigator may revise how an interaction is individuated, separate one composite record into several records, or merge redundant attestations as the evidentiary analysis develops.

The distinction between historical relations and their documentary reconstruction is fundamental. H names the investigator's held, evidence-bearing description for the comparison at hand; it is not equivalent to the complete set of relations that occurred. Source material may have been destroyed, withheld, produced strategically, or never recorded. Consequently, randomization of a display derived from H can identify responses to the represented material without identifying the truth of an incompletely documented history. Nor does a longer source chain automatically supply an independent corroboration chain. The relation between source records and the historical proposition under examination is part of what the study must substantiate. The target q should therefore be registered at the level of the actual claim evaluated, including a feasible response domain, admissible uncertainty, and the loss associated with different errors. A chronological question, a credit-allocation question, and an inquiry about the reliability of future work can use the same documents yet require different inferential standards.

The context C collects the conditions necessary to interpret the observer's response. These can include the observer's task instructions, relevant background knowledge, language competence, available processing time, interface configuration, incentives, and experimentally observed baseline beliefs. Some of these conditions are measured; others remain latent. When the context contains a representation policy ρ , it specifies which distinctions in H are expressed in a particular evidentiary display. That policy need not be a lossy compression: a direct archival interface may preserve extensive provenance, whereas a summary may suppress relations considered peripheral by its designer. Equally, a representation can introduce organizational structure that was not explicitly recorded in the source documents. The fact that it is easy to navigate does not establish that its structure is historically warranted.

Material circumstances belong within this context. Available time, cost of access, language mediation, accessibility, professional training, paid evaluative obligations, and reliance on employer-approved information systems influence what examination is feasible. Two people may have equal formal permission to open a repository while facing substantially different costs of interpreting its contents. In an AI-mediated interface, machine-generated synthesis may reduce search effort yet make the boundary between independent testimony and repeated reporting less visible. This is a possible mechanism for experimental comparison, not a general empirical conclusion about every model or user. Whenever evaluative labor is distributed unequally, conditioning on observed expertise alone may conceal the conditions under which expertise, time, and documentary access were acquired. The analyst should therefore state which aspects of C are manipulated, recorded, assumed stable, or unobserved.

The interface variable E records the evidentiary material available at the specified point of potential encounter. It should distinguish a document that the observer can retrieve, a record automatically displayed, and a recommendation encountered as a substitute for the original source. An archive held behind a separate search command differs operationally from a quotation already present on screen. Where necessary, E therefore contains both a content descriptor and its mode of presentation. A designated null value $\perp_E$ indicates that the relevant evidence does not become available within the protocol. This null

value should be distinguished from a record that exists and is available but receives no attention. If an experiment supplies the same documentary material to every participant, its access component can be fixed by design, leaving variation in cognitive processing and experimentally manipulated presentation features. If availability varies because of infrastructure, language, cost, or authorization, then the access component must remain in the model and may require analysis from Paper II.

More precisely, assign H , E , A , M , Z , and J measurable state spaces, provisionally taken to be standard Borel spaces when regular conditional probabilities and conventional stochastic-kernel constructions are required. Let A describe what an observer actually attends to, M an encoded or retrieved representation, Z an interpretation relevant to q , and J an observable task-specific judgment. The state spaces can be finite for carefully designed experiments; they may be structured or variable-dimensional when the task concerns multiple possible histories. Standard Borel assumptions support many probability-theoretic operations, but they are mathematical regularity conditions, not claims that personal cognition has a countable or predefined vocabulary [42]. The relevant empirical question is whether a chosen encoding resolves the differences required by the target task.

The access component of the framework can be written as a conditional distribution over possible presentations. Equation 2 makes the representation and observer context explicit.

$$K_{o,\rho}^{\text{acc}}(B \mid h, c) := \Pr_o(E \in B \mid H = h, C = c, \rho), \quad B \in \mathcal{E}. \quad (2)$$

Here $\mathcal{E}$ is the sigma-algebra on the chosen presentation space. The conditional probability is a descriptive distribution of encounter opportunities in a specified protocol. It does not imply that a change of ρ alone identifies the causal effect of an institutional reform. In an experimental setting, it can instead be induced by randomized allocation to well-defined interfaces, subject to compliance, interference, and measurement restrictions. This scope distinction is important: one can hold a corpus constant and change which parts are displayed, yet such a manipulation establishes the effect of the interface assignment under the experiment, not the social value of the infrastructure that ordinarily produced it.

Three descriptions should be kept separate in applications of Equation 2. The *held corpus* identifies the records against which the investigator compares an interface; the *opportunity set* describes what a participant could retrieve under the protocol; and the *presentation* specifies what was displayed or actually returned following an information request. A single recorded E can represent the last of these only if the other two are retained in design metadata. A search result without its underlying document, for example, records an opportunity to inspect the source while conveying only a mediated description of its contents. It would be misleading to use that search result to code successful exposure to every relation in the original document. Analytically, extending E with metadata can preserve the distinction, although more detailed coding generally entails additional measurement costs.

Observer heterogeneity also requires care. Two observers with the same interface may reasonably produce different judgments if one can interpret a technical method and the other lacks that competence. Their responses cannot be ranked epistemically without specifying the target proposition, relevant background information, and a suitable criterion of correctness or calibration. A participant who expresses uncertainty in response to inaccessible provenance may perform the task more responsibly than a participant who supplies a confident attribution. Equally, unequal processing time does not by itself explain which interpretation is correct. The formal role of o and C is to permit such differences to be represented and studied without prematurely classifying them as cognitive failures.

2.2 The Composition and Limits of Processing Kernels

Equation 1 supplies the restricted composition used throughout the initial investigation. Conditional on a held history and context, the access kernel determines a possible evidentiary presentation; attention selects an aspect of that presentation; memory or encoding generates a cognitive representation; interpretation gives the representation task-relevant meaning; and a judgment kernel produces an assessable response. The sequence is an economical decomposition of a joint probability law, not a claim that the mind invariably performs these operations in that order. In particular, the recorded M may depend on

information already retrieved at the start of the trial, and the attention allocation A may respond to a provisional judgment that is not separately included in the restricted model.

For a one-pass protocol, the mathematical chain rule yields a factorization of a sufficiently specified joint law, yet the restricted product in Equation 1 requires the omission of specific conditioning variables. To state the restriction, take the target question as part of C , hold the representation policy ρ fixed within a modeled protocol, and require the displayed E to include presentation characteristics that may influence later processing. One sufficient collection of conditional-independence assumptions is given in Equation 3. The symbol $\perp\!\!\!\perp$ denotes statistical conditional independence under the specified observer law.

$$\begin{aligned} A &\perp\!\!\!\perp H \mid (E, C, \rho), \\ M &\perp\!\!\!\perp H \mid (A, E, C, \rho), \\ Z &\perp\!\!\!\perp (H, E, A) \mid (M, C, \rho), \\ J &\perp\!\!\!\perp (H, E, A, M) \mid (Z, C, \rho). \end{aligned} \tag{3}$$

Equation 3 is a substantive modeling restriction rather than a psychological theorem, following the statistical use of conditional independence surveyed by Dawid [16]. When a source’s identity affects a judgment even after the coded interpretation is held fixed, for example, Z may omit a relevant component of what the observer considered. When unrecorded expertise shapes both attention and reconstruction, the context C may be insufficient. The conditional laws can consequently depend on variables excluded from the displayed kernels. Additional state variables may restore a useful approximation, although they do not establish the independence conditions merely by being named. An unrestricted chain-rule decomposition remains valid at the level of the measured joint distribution under regularity conditions; the shortened, interpretable factors warrant independent empirical scrutiny.

The null outcomes in the processing states are consequential. Write $\perp_A$ for the absence of relevant attention and $\perp_M$ for the absence of a retrievable task-relevant trace. An observer may receive $E = e$ and nevertheless have $A = \perp_A$. An observer may attend to a relevant segment and later have $M = \perp_M$ at the scheduled recall test. These states are analytically different from a trace that retains content but misattributes its origin. For instance, a participant who never sees a provenance footnote cannot be evaluated for source recall in the same manner as a participant who read the footnote and subsequently names the wrong source. A design that records only whether an attribution is correct will merge these pathways unless its protocol measures their intermediate states.

The difference between opportunity and cognitive encounter is summarized by Equation 4. For a specified task and history, the probability of attending to something relevant is the average of the attention probability over the possible presentations.

$$\Pr_o(A \neq \perp_A \mid h, c, \rho) = \int_E [1 - K_o^{\text{att}}(\{\perp_A\} \mid e, c)] K_{o,\rho}^{\text{acc}}(de \mid h, c). \tag{4}$$

Equation 4 follows from the restricted factorization and the definition of the null-attention state. It is not a psychological discovery. If earlier evaluations affect attention or if the stored cognitive state supplies relevant information that is omitted from c , then K_o^{att} requires additional conditioning and the equality may fail for the intended interpretation. The equation nevertheless provides a useful accounting distinction: a change in observed attention can result from a change in presentations, from changed attention conditional on the same presentation, or from both. Without separately recording presentation and attention, the components cannot be assigned unique empirical magnitudes.

An experimentally coded “relevant encounter” should refer to a prespecified task and an explicit opportunity boundary. The complement $A = \perp_A$ may include skimming unrelated material, inspecting a link but failing to open its target, or looking at the relevant passage without performing the specific task-relevant discrimination. These mechanisms may be indistinguishable under a coarse attention instrument. Equation 4 averages over the presentation distribution; it should not be read as a universal estimate of a person’s attentional capacity. If the same participant undertakes a series of tasks, context and preceding experience can change at each observation. The resulting measurements require repeated-measures or

state-dependent analysis rather than treating the participant as supplying independent draws from one stationary kernel.

The analyst must also distinguish different ways of representing the same evidentiary materials. Suppose an original artifact is accessible through a repository and a mediated interface describes it in a short paragraph. Both may be treated as representations of a held historical material h , but the observations E are different. Even a completely accurate short paragraph may omit details necessary for another target question. Thus, *representation invariance* should be asserted only relative to a stated proposition and a stated sufficiency criterion. A summary that is sufficient for determining whether a document exists may be insufficient for judging the independence of two cited findings. Apparent equality of summary content does not justify equality of source lineage, examination burden, or conditional likelihoods.

The relevant notion of representational adequacy is therefore proposition-relative. Two presentation policies may be equivalent for detecting the existence of an artifact yet differ for reconstructing a causal dependence or establishing an authorship claim. Where a benchmark relation is defined by an independently inspectable synthetic corpus, equivalence can be tested using a predeclared scoring rule. Where histories are contested, an equivalence claim should specify which unresolved alternatives both displays preserve. In neither case does equality of final mean ratings demonstrate equivalence of the cognitive pathways, since compensating changes in attention and interpretation can conceal a difference in evidence contact. The analyst can compare tasks and presentations without assuming that one representation is globally superior for every epistemic purpose.

This treatment also creates an interface to the institutional and technical analysis of the previous paper while preserving the psychological object's autonomy. A portal that changes K^{acc} can expand opportunities to encounter evidence, and a well-designed presentation may also change K^{att} or K^{mem} . A critic who attributes every change in cognition to the portal would conflate these components; a critic who attributes every unsuccessful judgment to an individual's memory would ignore the possibility that crucial evidence was never exposed. The proposed layered account instead identifies the measurement conditions under which a more specific explanation becomes plausible. Subsequent sections will examine mechanisms within the attention and memory components without assuming that the complete decomposition is already validated.

2.3 Missing Evidence, Distorted Encoding, and Uncertain Reconstruction

Cognitive recognition is particularly vulnerable to a confusion between *absence of evidence* and *inaccurate representation of encountered evidence*. Both can yield an incorrect answer, yet they demand different explanations and remedies. In the first case, a participant lacks a relevant document or did not examine a portion of the available material. In the second, the participant encountered it but later changes a name, reverses a dependence relation, merges two witnesses, or infers a chronological ordering not warranted by the records. A third case concerns the accurate encoding of a record followed by an inferential interpretation that exceeds its support. An observer may remember correctly that a recommendation was issued by a prestigious office while treating the recommendation as evidence that the office independently verified the underlying work. The inference could be unwarranted even though the memory is faithful.

A careful representation of M therefore contains more than a text-like summary. For the local task, its coordinates can distinguish remembered substantive content, attributed source, relational dependence, and uncertainty about ordering. An investigator need not assert that the biological memory system stores those coordinates as literal fields. Rather, the coordinates specify what the research instruments aim to elicit. Different testing procedures may recover different aspects of the underlying representation, and measurement itself may alter later recall.

An immediate probe and a delayed reconstruction should also be distinguished from the latent representation at the moment of initial encounter. A written recall request may operate as another retrieval event, alter future source attribution, and become part of subsequent evidence. A design that measures memory repeatedly therefore observes a sequence of conditioned reports, not independent readings of a

fixed internal object. The proposed M can be indexed by the operation at which it is operationalized, with measurement-induced transitions analyzed separately when the protocol supports them. Reports of confidence in memory, certainty about a source, and agreement with the underlying claim should be elicited as distinct responses, because an accurate origin judgment need not imply endorsement of the claim. The classic source-monitoring account treats judgments about origins of remembered information as active attributions based on available mnemonic qualities rather than infallible retrieval of source tags [39]. That account motivates separately assessing source claims, but it does not entail the proposed relational encoding scheme or establish how its parameters vary across disciplines.

An explicit distortion measure is useful only when a task-specific comparison standard is justified. Let $g_{q,\rho}(a, e)$ denote a prespecified, evidence-supported feature or relation that a faithful encoding of attended material would preserve for question q , defined for an evaluable subset of attention–presentation pairs. Let $\phi_q(m)$ denote the feature inferred from the observer’s reported memory state under a stated measurement operation. The conditional encoding–discrepancy probability in Equation 5 is defined only on that evaluable subset.

$$d_o(a, e, c; q, \rho) := K_o^{\text{mem}}(\{m : \phi_q(m) \neq g_{q,\rho}(a, e)\} \mid a, e, c). \quad (5)$$

Equation 5 presupposes that $g_{q,\rho}$ is defined and justified for the tested feature, including how uncertainty and multiple admissible answers are scored. Where records are contested, g may be a set of supported responses rather than a unique answer, and the discrepancy event must be defined accordingly. If $A = \perp_A$, the observer may have no relevant material to encode; the definition should then be treated as inapplicable rather than assigning an ordinary error score. Thus, a failure to encounter evidence and an inaccurate account of evidence actually encountered are separate outcomes. Nor should a difference between reported interpretation and a researcher’s favored narrative automatically be counted as a memory error.

A multiple-answer standard may replace equality to one designated $g_{q,\rho}(a, e)$ with membership in a prespecified set of admissible responses. Its construction should be independently reviewed where possible, because a benchmark derived from one controversial interpretation can create apparent cognitive errors by design. Furthermore, a difference between $\phi_q(m)$ and the benchmark concerns the elicited report; it cannot by itself separate faulty initial encoding from inaccurate retrieval or reporting. Such separation requires additional observations or experimental restrictions. The quantity d_o is therefore best described as an encoding–report discrepancy under the indicated measurement model, unless the protocol independently establishes that reported M faithfully measures encoding.

Relational histories require special attention to uncertainty about source dependence and partial order. Consider three anonymous records: a technical note, a subsequent report citing it, and an endorsement that describes the report. The represented documentation may establish that the report refers to the note, while leaving uncertain whether the endorsement relies on independent inspection or merely repeats the report. One participant may infer independent corroboration, another may infer dependence, and a third may preserve the ambiguity. The epistemic status of those answers depends on the available traces. An experimental investigator can construct a setting in which source dependence is known by design, but a historical inquiry often cannot. The model must therefore distinguish the participant’s inferred relational reconstruction from the justified uncertainty set supported by the source corpus. A fully connected causal graph inferred from a suggestive narrative may be less faithful than a sparse partial representation that keeps uncertain relations unresolved.

One proposed representation of this reconstruction is a probability distribution $\Pi_o(du \mid m, c)$ over a locally defined family of candidate history descriptions $u \in \mathcal{U}$. The family $\mathcal{U}$ is an analytical device, and its membership may change as the record is examined; it is not a fixed catalog of social events. The observer’s actual distribution may be inaccessible or only partly elicitable. For comparison, a normative Bayesian benchmark can be defined where a defensible prior $\pi^*(du \mid c)$ and an evidence-generating likelihood $L^*(m \mid u, c)$ are available. Equation 6 states the benchmark over a countable candidate family for clarity.

$$\Pi^*(u \mid m, c) = \frac{L^*(m \mid u, c) \pi^*(u \mid c)}{\sum_{v \in \mathcal{U}} L^*(m \mid v, c) \pi^*(v \mid c)}, \quad \text{when the denominator is positive.} \quad (6)$$

Equation 6 defines a conditional inferential comparator rather than a descriptive law of human reconstruction. It requires a meaningful candidate family and an evidence model that accounts for common sources and selection. The actual interpretation kernel K_o^{int} need not implement Bayes' rule, and an observed response need not identify a unique internal probability distribution. Structured probabilistic accounts of cognition show how priors, evidence, and abstract representations can be connected in explanatory models [76]; their existence does not demonstrate that this particular relational model describes all human observers. When the documentary record cannot distinguish candidate histories, responsible judgment can preserve uncertainty instead of selecting the most narratively complete reconstruction.

Dependence among documents matters for this comparison. If two descriptions reproduce one earlier source, treating them as conditionally independent observations can multiply apparent evidentiary support without adding independent information. The likelihood L^* must represent the common origin, or explicitly declare why the approximation is adequate for the task. Likewise, the prior π^* should not quietly encode as a default the investigator's desired credit distribution. A disagreement between Π_o and Π^* is interpretable as a cognitive discrepancy only relative to these declared modeling choices; alternative defensible priors or candidate families may support different posterior rankings. This makes sensitivity analysis and documentation of provenance structure part of the evaluative procedure.

The distinction has an additional ethical consequence. An observer's failure to recognize a contribution can result from a defective memory, a presentation that omitted crucial provenance, legitimate uncertainty about contradictory traces, or a task whose requested attribution cannot be supported. These possibilities should be reported separately where data permit. A test that rewards confident identification of one person despite genuinely ambiguous source material can penalize epistemically careful participants and manufacture apparent bias. Conversely, offering a fully documented answer key does not ensure that participants understand its evidentiary basis. Measurement of memory, inference, and trust must therefore specify which statements are verifiable, which remain contested, and what opportunities the observer had to inspect the supporting records.

2.4 Feedback, Observer State, and the Conditional Order of Operations

The restricted one-pass sequence suppresses a familiar feature of inquiry: an observer can search again after an initial judgment. A participant reading a summary may suspect that two endorsements share an origin, open an original record, discover an omitted caveat, and revise a judgment. Another participant may accept the summary and conduct no additional inspection. In both cases, interpretation shapes subsequent exposure. A model in which attention is conditionally independent of the observer's current beliefs becomes inappropriate unless those beliefs and the decision to search are included in the conditioning context. More generally, the meanings of "encounter" and "judgment" depend on the protocol's stopping rule. A final response after free exploration is not comparable to an immediate response after compulsory exposure unless the designs explicitly account for the intervening activity.

Let S_k denote the observer's task-relevant state at the k th recorded operation, including whatever prior beliefs, recalled observations, or uncertainties the research design can support. Let $R_k \in \mathcal{R}$ denote a retrieval or inquiry action, which may consist of opening a primary document, consulting a translation, requesting an explanation, or ending the inquiry. This action variable is distinct from the target proposition q . One simple feedback specification samples the next retrieval action from the current cognitive state and uses it to determine a subsequent evidentiary encounter. Equation 7 makes this conditional relation explicit.

$$\Pr(E_{k+1} \in B \mid S_k = s, H = h, C = c, \rho) = \int_{\mathcal{R}} K_{o,\rho}^{\text{acc}}(B \mid h, c, r) K_o^{\text{qry}}(dr \mid s, c). \quad (7)$$

In Equation 7, K_o^{qry} describes information-seeking behavior and the extended access kernel conditions on the realized action r . The displayed equality assumes that, conditional on the modeled state and context, retrieval-action selection has no remaining dependence on H , and that the subsequent presentation is conditionally independent of S_k once H, C, ρ, R_k are given. If unrecorded experience about the history

affects the next request, or the interface tailors its response using features of the user state not captured in C , those assumptions require revision. An explicit terminal action may be represented by an absorbing observation state rather than silently removing stopped inquiries from the analysis. The index k marks operations in a specified observation protocol; it does not posit that social or cognitive processes admit a unique global historical clock. If the observer’s action changes the source corpus itself, for example by eliciting a new testimony or altering a shared archive, the state of H must also be updated rather than held fixed. If actions and retrieved records jointly influence one another within an operation, a joint transition law may be preferable to the simple product shown here.

Observer state makes the modeling commitment testable in principle but also reveals its limitations. Task instructions may elicit a report of a current belief, yet a participant’s actual cognitive state remains partly latent. A state S_k designed only from earlier judgments may fail to preserve relevant information about remembered sources or changes of attention. Treating the process as Markovian would then be an additional, possibly false, restriction. Repeated exposure, learning from corrections, and changes in task strategy can produce long-range dependencies that cannot be summarized by a short record of preceding responses. A flexible state-space model can sometimes approximate such effects, but state expansion does not by itself solve the problem of identifying hidden mechanisms from limited behavioral outcomes.

For empirical protocols, a minimal operational extension would record (R_k, E_{k+1}) alongside the preceding access opportunity, the observer’s stated task, and any intermediate judgment or uncertainty report. The resulting data can distinguish an unoffered source from an available source that was not requested. They generally cannot reveal why a person did not search without additional assumptions or an elicitation that may itself change behavior. In deployed systems, personalization also makes the presentation conditional on earlier traces of interaction. The experiment should record whether apparent improvements in recognition stem from changing observer processing, adaptive retrieval, or a changed documentary corpus. These are different candidate mechanisms with potentially different ethical and institutional implications.

The choice of operation order should also remain empirically revisable. In a structured experiment, the investigator may ensure that each participant first sees a display, then answers a recall question, and then reports a judgment. Outside that protocol, remembering may guide searching, interpreting may change attention before display inspection ends, and a verbalized judgment may reshape the subsequent memory report. The same variables can be arranged differently in alternative observational models, and a good fit to the final judgments is insufficient to establish which order occurred. The scientifically relevant order is the order justified by task intervention, recorded dependence, and behavioral traces. This methodological restraint is especially important for relational evidence, where apparent chronological coherence may be an artifact of narrative presentation.

These considerations explain why the paper uses the staged architecture as a modular research design rather than a closed cognitive ontology. Each conditional component suggests possible measurement and intervention, but the composition is valid only at the resolution supported by the task. A stateful extension can represent some feedback, while a more general interactive or partially observed model may be needed for complex inquiry. The decision among models should depend on explanatory gain, parameter identifiability, and predictive adequacy rather than preference for one formal language. Importantly, such extensions remain within this paper’s object of human cognition. Institutional choices about whose testimony is admissible and who controls the interface require the complementary power analysis; changes in those institutions need not be reduced to changes in an individual’s state.

2.5 Non-Identification and Conditions for Empirical Comparison

An explicit decomposition does not guarantee that the observer’s internal mechanisms can be inferred from the final response. Suppose a binary task records only whether an observer endorses a proposition and produces an observed endorsement probability r_h under history representation h . One hypothetical model exposes the evidence to all participants, while an encoding stage produces an affirmative internal trace with probability r_h ; a deterministic judgment maps that trace to endorsement. A second hypothet-

ical model attends to the relevant evidence with probability r_h , faithfully encodes whenever attention occurs, and gives a negative response otherwise. Both generate the same observable endorsement rate under the fixed protocol. If the only measured variable is the final judgment, the two explanatory pathways are observationally equivalent, although one attributes variation to encoding and the other to encounter. The illustration is a constructive example of non-identification, not evidence that either process dominates real behavior.

Proposition 2.1 (Final judgments do not identify the stages). *Suppose a protocol records only a binary final judgment J and a held representation H , with $\Pr(J = 1 \mid H = h) = r_h$ and $0 < r_h < 1$. If the intermediate attention and encoding states are unobserved and their conditional kernels are otherwise unrestricted, at least two distinct attention–encoding decompositions can generate the same joint law of (H, J) .*

To verify Proposition 2.1, construct a first model in which attention to the relevant display is certain, the encoded binary trace equals one with probability r_h , and judgment copies that trace. Construct a second model in which attention occurs with probability r_h , encoding is faithful when attention occurs, the trace equals zero when attention is absent, and judgment again copies the trace. The marginal probability of endorsement is r_h in both models for each fixed h , while the attention probabilities differ. The construction establishes observational non-uniqueness under its assumptions; it does not exclude identification when additional measurements, interventions, or structural restrictions are introduced. A zero or unit endpoint can be treated separately without affecting the basic lesson. Neither model is offered as a general psychological theory.

The construction also shows why a reported increase in perceived credibility after a display redesign is causally ambiguous. Random assignment can support an average effect of the assigned display on the recorded judgment, under the experiment’s usual design assumptions. It does not automatically recover the share of that effect transmitted through noticed provenance, corrected source recall, or changed inferential weighting. Those are mediation questions requiring additional restrictions on confounding and measurement, including care when the treatment itself changes later mediator–outcome confounding [35]. Accordingly, the causal language of the paper should refer to a specific randomized intervention when present and to observational associations when it is absent.

The result has direct methodological implications. A study reporting higher trust ratings after source disclosure cannot conclude from that outcome alone that participants noticed the disclosure, encoded its provenance correctly, or understood its evidentiary implications. The manipulation may also affect perceived authority, task demand, or willingness to express uncertainty. A stronger design should separately record which materials were actually made available, which were displayed, which were inspected, what participants recalled, how they reconstructed dependence and order, and how they justified their judgments. These measurements remain imperfect. Eye fixation is a useful indicator of visual selection in some tasks but does not establish semantic comprehension; recognition questionnaires can cue retrieval; and requests for written explanations can alter subsequent interpretations. Measurement instruments must be calibrated against the particular cognitive construct they claim to observe.

The construct-validation problem is especially pressing where participants provide a single score called “trust.” The score might reflect perceived truth of a sentence, credibility of its source, belief about contributor competence, institutional deference, willingness to delegate, or a demand effect. Calling these outcomes by a shared name does not give them equivalent measurement properties. The construct-validation literature emphasizes the need for evidence about whether an instrument measures the proposed psychological construct [23]. For the present programme, a multicomponent instrument should therefore distinguish substantive judgment, provenance reconstruction, uncertainty calibration, and behavior wherever the research question requires them. Its measurement invariance across languages, professional communities, and presentation policies should be examined rather than assumed.

A controlled comparison can assign participants to two presentation conditions while holding an evidentiary corpus constant. In one condition an endorsement and a quoted claim appear without a visible link to their shared source; in the other, the underlying source path is presented with a means to inspect

Table 2: Observable distinctions required to investigate the candidate cognitive interfaces. The table describes research-design targets rather than established psychological stages.

Analytical component	Potential observation	Important unresolved ambiguity
Evidence availability	Authorized retrieval logs; assigned display condition; recorded corpus snapshot	Availability need not imply encounter or comprehension.
Task and evaluation rule	Registered target proposition; admissible uncertainty; independent source-corpus key where feasible	Different propositions may require different criteria and losses.
Attention and examination	Interface event logs; observed opening and navigation; task-specific selection probes	A click or fixation may not imply substantive attention.
Encoding and memory	Immediate and delayed content recall; independently scored source attribution	Recall tests may cue, reshape, or contaminate the trace.
Trajectory reconstruction	Source-lineage reconstruction; supported partial-order judgments; uncertainty reports	A single answer key may be inappropriate for contested records.
Interpretation and judgment	Argument explanation; probability estimate; assessment of diagnosticity	Reports may be motivated, noisy, and dependent on task framing.
Behavioral reliance	Optional verification, delegation, or resource-limited follow-up within a stated task	Behavior depends on perceived costs, incentives, and external constraints.

it. The investigator should distinguish the intention-to-treat effect of assignment from the association between voluntary inspection and a final judgment. The latter is confounded when a participant’s prior skepticism or specialist knowledge influences the decision to inspect.

If assignment affects the probability of inspection, comparing only people who inspected in both arms can introduce selection bias, even if the original presentation assignment was randomized. Repeated measures may improve precision yet also induce carryover as participants learn the source structure. A preregistered contrast between assigned opportunity and requested inspection, followed by sensitivity analyses for unobserved motivation, offers a more defensible starting point than claiming a clean attention-mediated effect. Treatment effects may also vary with expertise, language, and the documented cost of examination, which should be reported as substantive effect modifiers rather than treated solely as nuisance variation. It is also possible that the addition of source information changes perceived difficulty, fatigue, or the salience of the claim. An interpretation that attributes all observed judgment changes to improved source monitoring would consequently require supporting intermediate measures and suitable alternative conditions.

Research designs should preserve the distinction between objectively verifiable source relations and contested historical claims. When the investigator constructs a synthetic corpus with a known source chain, provenance discrimination can be scored against a defined answer key. When the corpus consists of real documents whose independence or authorship is disputed, an answer key may have to represent admissible alternatives, annotated uncertainty, and the reasons for excluding particular interpretations. A design that forces a definitive response from underdetermined evidence measures willingness to guess alongside understanding. The model should instead permit uncertainty reports, decisions to seek further information, and recognition of insufficient evidentiary grounds. Such outcomes may be particularly valuable when distinguishing warranted skepticism from simple distrust.

Table 2 summarizes the central design boundaries and their measurement limitations. Its rows should not be read as seven independent causal modules. A participant’s inspection can alter later source memory; an interpretation probe can influence a judgment; and a reliance decision can depend on institutional incentives beyond the experiment. Such dependencies are part of the phenomenon rather than nuisance terms to be hidden. The table also clarifies why the framework should be tested against simpler competing accounts. A single salience effect, a task-demand explanation, or a conventional Bayesian inter-

pretation of received information may fit the data as well as a more elaborate multi-kernel model. The contribution of the additional structure must be established through discrimination among mechanisms or improved out-of-sample predictions, not through the ability to assign a variable to each theoretical stage.

Finally, the meaning of warranted recognition requires explicit local specification. An observer's subjective confidence, the accuracy of an independently checkable reconstruction, the extent of examined evidence, and the prudence of a subsequent decision may move in different directions. An improved display can make a participant less confident precisely because it exposes conflicting testimony. An observer may accurately identify a source while declining to infer the competence of the person associated with it. A source-monitoring intervention can correct a false attribution without yielding a measurable change in declared trust. These are substantively possible outcomes, not anomalies that must be forced into a monotonic recognition metric. The later analysis of belief revision, epistemic vigilance, and reliance will preserve these distinctions and specify the evidentiary assumptions relevant to each.

The formal objective of the present section has thus been limited deliberately. It identifies a conditional interface between evidence-bearing relational representations and human cognition, provides a disciplined vocabulary for missingness, distorted encoding, and uncertain reconstruction, and shows why final ratings alone are insufficient to recover the cognitive mechanism that produced them. It also states the statistical restrictions underlying that factorization and identifies which feedback relations require state-dependent extensions. The following section develops selective attention and evidentiary encounter within these bounds, examining how task relevance, salience, processing costs, and competing demands condition what an observer actually examines. That analysis will treat surprise and informational diagnosticity separately, allowing structurally ordinary evidence to receive attention when it is important and unexpected material to remain peripheral when its relevance is weak.

3 Selective Attention and Evidentiary Encounter

The purpose of this section is to determine how evidentiary material that is available within a represented relational history becomes the object of an observer's selective examination. The distinction matters because the structural access conditions studied in Paper II, and represented by the interface kernel in Section 2, leave open the question of which materials receive cognitive processing. Under limited time and competing demands, an observer may inspect a conspicuous yet weakly informative record while passing over an ordinary document that bears directly on the question under examination. This section develops a conditional model of attentional allocation, separates statistical surprise from salience, relevance, and expected informational value, and examines the effects of workload, competing evidence, and active inquiry. It then derives limited formal results and specifies empirical comparisons that could distinguish those mechanisms. The model is offered as a testable interface description, with particular attention to the material and temporal conditions under which evidence is encountered.

3.1 Evidentiary Opportunity and the Allocation of Attention

A relevant evidentiary opportunity consists of more than the presence of an item in an archive. In a specified observation task, the observer may face a display containing original documents, excerpts, annotations, source references, suggested links, or algorithmic descriptions. Some items are immediately visible; others require a search, translation, authorization, or navigation step. The term *available evidence* refers here to material reachable within the protocol. *Encountered evidence* refers to material actually selected for sufficient inspection to provide a chance of encoding its relevant content. *Examined evidence* requires a task-specific criterion concerning the degree of substantive processing. These distinctions allow a researcher to state, for example, that a document was reachable and opened while remaining uncertain whether the observer evaluated its source lineage.

Work on selective visual attention establishes several relevant precedents. The account of competition and biased selection developed by Desimone and Duncan addresses how some visual represen-

tations are favored when simultaneous processing is limited [17]. Computational research on saliency demonstrates that an image-processing system can select conspicuous locations for additional analysis without independently establishing their task significance [36]. These studies motivate, but do not directly validate, the extension to documentary evidence. An archive involves semantic reading, language competence, historical reconstruction, and institutional context in ways that a visual search display may not. It is therefore necessary to specify which aspects of attention are represented by a model rather than treating the visual-attention literature as a complete account of research judgment.

Let D_t be the set of evidentiary items reachable at encounter occasion t for observer o , represented under policy ρ and task q . The set is locally finite in a controlled experiment, but its members need not be instances of fixed, ontologically given event categories. Each item i can be a document, a possible source link, a jointly presented record, or another task-specified evidentiary unit. Its individuation may be revised if a composite record is separated into distinct testimonies. Let $\mathcal{H}_{o,t}$ denote the observer’s measurable interaction history and state available to the research design; it may include previous displays, observed choices, instructions, and an imperfect representation of prior beliefs. Let $I_t \in D_t \cup \{\emptyset\}$ describe the next selection, where $\emptyset$ denotes stopping, deferring, or declining further examination. A candidate attention policy is specified in Equation 8.

$$\pi_{o,t}(i \mid D_t, \mathcal{H}_{o,t}, q, \rho) = \Pr(I_t = i \mid D_t, \mathcal{H}_{o,t}, q, \rho), \quad i \in D_t \cup \{\emptyset\}. \quad (8)$$

The policy is a conditional probability law, not an assertion that the observer deliberately computes a selection probability. It can represent impulsive orienting, deliberate searching, fatigue-dependent stopping, or mixtures of these processes. To interpret it empirically, the protocol must identify what was in D_t , which features were perceptually accessible before selection, and what information was acquired only after an item was opened. This last distinction prevents a methodological error: a later-assessed evidentiary property cannot automatically be treated as a feature that caused an earlier choice unless the observer had a suitable preview or expectation concerning it.

The choice variable I_t should be distinguished from both *orientation* and *substantive encounter*. A glance at a title, a fixation on an icon, an opening action, and successful recognition of a source relation are empirically separable episodes. A first-glance process can alter which previews are cognitively registered; later choice can determine which source becomes available for reading; comprehension may then depend on language fluency and the content actually delivered. Some interfaces present the entire item from the outset, whereas others expose only a preview until the participant clicks. To avoid imposing an artificial sequence, the investigator should annotate the first moment at which each attribute is actually perceptible, distinguish automatic presentation from requested retrieval, and state which observable behavior counts as the next choice. For a task whose scoring target is source independence, a document opening alone cannot constitute successful encounter with its dependence structure.

The distinction creates a specific risk of selection bias. Let an observer-specific variable, such as previously acquired source familiarity, affect both the likelihood of opening a record and the ability to interpret that record. A comparison of interpreters conditional on having opened it can then select observers with different familiarity and different prior uncertainty, even when opening probabilities are experimentally manipulated. A design that records only opened documents can also fail to measure errors arising from the systematic nonencounter of relevant evidence. The appropriate denominator depends on the question: all assigned opportunities for an interface-level encounter rate; all displayed previews for a preview-selection rate; and all substantively examined sources for the conditional accuracy of a provenance judgment. Such rates answer different questions and should remain separate in analysis.

In sequential inquiry, D_t changes as an observer follows links, discovers alternative terminology, or learns that two apparently independent claims originate in the same record. Selection at occasion t may change both the accessible opportunities and the observer state at occasion $t + 1$. An experimental design that treats all choices as independent draws from a fixed menu consequently suppresses mechanisms central to historical-source investigation. Conversely, a static menu can be an appropriate deliberately restricted approximation when the research question concerns initial allocation to simultaneous summaries. The validity of a model depends on the contrast for which it is being used.

The observer’s finite resources create an allocation problem even when documentary access is equal across participants. Time, reading effort, working-memory load, translation effort, navigational friction, and prior fatigue can affect what is selected. The same document may require minutes for a specialist and considerably longer for an observer unfamiliar with its notation. A simple nominal limit on the number of opened items therefore inadequately measures attention costs. In a defined experimental context, an investigator might instead specify the time or task-specific effort associated with each inspection, estimate those costs from independent trials, and permit variation across observers. Such parameters describe operational constraints and should not be interpreted as immutable cognitive capacities.

3.2 Surprise, Saliency, Relevance, and Expected Information Value

Several quantities commonly described as *interestingness* play different roles in the study of evidentiary encounter. The realized surprisal of an observation depends on a specified reference distribution. Sensory or interface saliency concerns its likelihood of attracting perceptual selection under particular display conditions. Relevance concerns the relation between the material and a target question. Diagnosticity concerns how an observation discriminates among hypotheses. Expected information gain describes a prospective reduction in uncertainty under a chosen probabilistic model, whereas decision value also depends on the consequences and costs of action. Curiosity and exploratory preference may influence selection in addition to the requirements of an immediately assigned task. The distinction between goal-directed information sampling and more open-ended information search has been developed in the literature on active sampling and curiosity [27, 44]. These terms consequently cannot be substituted for one another without specifying the analysis.

For a represented observation y_i under reference law $P_o(\cdot \mid \mathcal{H}_{o,t}, q, \rho)$, define its realized surprisal as Equation 9.

$$S_o(y_i \mid \mathcal{H}_{o,t}, q, \rho) := -\log P_o(Y_i = y_i \mid \mathcal{H}_{o,t}, q, \rho), \quad (9)$$

where the formula applies to a discrete observation with positive probability; continuous outcomes require a suitable density-relative treatment, and the reference measure and its interpretation must be made explicit. The subscript o indicates an observer-specific or analyst-assigned predictive law, rather than an objectively privileged measure of novelty. A low-probability result may be conspicuous because of prior expectations, but surprisal does not indicate whether the result concerns the target proposition. Furthermore, realized surprisal is defined after observing y_i . An observer selecting an unopened document typically has an expectation over possible contents rather than access to the contents’ realized surprisal. Experiments that regress initial selection on realized surprisal must consequently specify what preview information permitted that expectation.

A further division concerns the information already conveyed by the preview. Let $T_{i,t}$ describe the title, recommendation text, source badge, or other signal genuinely available before item i is opened. A preview may itself supply evidence about the target proposition; the additional value of opening the source should then be evaluated *conditional on that preview*. In a coherent model with a specified observer history, the remaining expected gain is the conditional information that possible inspection outcomes supply beyond the visible preview. Equation 10 expresses the corresponding benchmark at a preview value t_i with positive predictive probability.

$$\text{EIG}_o^{\text{res}}(i \mid t_i, \mathcal{H}_{o,t}, q) := I_o(\Theta; Y_i \mid T_{i,t} = t_i, \mathcal{H}_{o,t}, q). \quad (10)$$

This quantity can be zero even when the preview was highly informative, since reopening a verbatim copy adds no further model-based information. It can also be substantial when the preview says almost nothing about provenance and the source contains an independent dated record. The term *residual* designates new information relative to the conditioned preview; it is not a claim about an observer’s subjective ability to foresee inspection outcomes. In empirical work, eliciting an anticipated value rating after a preview changes the assessment process and may itself alter subsequent search. Expected-value predictors should therefore be elicited in independent calibration samples or through a clearly specified, noninterfering design whenever initial unprompted selection is the outcome of interest.

The distinction from prospective epistemic value can be formalized. Suppose Θ is a task-relevant uncertain proposition or parameter, Y_i is a possible result from inspecting item i , and $p_o(\theta \mid \mathcal{H}_{o,t})$ is a predictive belief under a fully specified model. Equation 11 defines a Bayesian benchmark for expected information gain.

$$\text{EIG}_o(i \mid \mathcal{H}_{o,t}, q) := \mathbb{E}_{Y_i \mid \mathcal{H}_{o,t}}[D_{\text{KL}}(p_o(\Theta \mid Y_i, \mathcal{H}_{o,t}) \parallel p_o(\Theta \mid \mathcal{H}_{o,t}))]. \quad (11)$$

The expression is meaningful when the relevant posterior laws, reference measure, and integrability assumptions are defined. In a properly specified probabilistic model, it is conditional mutual information between Θ and Y_i given the observer's information history. It measures an expectation under that model. A descriptively bounded observer might not represent the required alternatives correctly and might choose an item for reasons uncorrelated with this benchmark. Nor does expected information gain alone specify the value of subsequent action. Evidence capable of resolving a low-stakes uncertainty could have high information gain and little consequence for a specified decision; modest information concerning a consequential decision may have greater practical value.

For a decision problem with allowable actions $a \in \mathcal{A}$, utility or loss representation $u(a, \theta)$, and an available option to ignore new material, a corresponding gross expected value of inspection is given in Equation 12.

$$\text{GVI}_o(i \mid \mathcal{H}_{o,t}) := \mathbb{E}_{Y_i \mid \mathcal{H}_{o,t}} \left[\sup_{a \in \mathcal{A}} \mathbb{E}_o[u(a, \Theta) \mid Y_i, \mathcal{H}_{o,t}] \right] - \sup_{a \in \mathcal{A}} \mathbb{E}_o[u(a, \Theta) \mid \mathcal{H}_{o,t}]. \quad (12)$$

This benchmark assumes that expectations and optimal values exist, that the predictive model is coherent, and that the action set remains available after inspecting the item. Under those assumptions, the gross value is nonnegative because the decision maker may choose to ignore the observation. It can be zero even when inspection produces substantial belief changes if those changes do not affect a preferred action. The *net* value may be negative after subtracting search and processing costs. This normative comparison should be distinguished from the descriptive benefit an observer actually realizes, since incomplete understanding or miscalibration can make observed performance worse after exposure to more material.

Proposition 3.1 (Surprisal without diagnostic information). *For a discrete observation Y independent of a target proposition Θ conditional on observer history $\mathcal{H}$, assume the conditional probabilities and posterior are defined. For an observed y with positive conditional probability, the posterior of Θ given $(Y = y, \mathcal{H})$ equals the posterior given $\mathcal{H}$. Consequently, the realized belief change and the expected information gain about Θ are zero, although the realized surprisal of y can be arbitrarily large across admissible distributions.*

The conditional independence assumption in Proposition 3.1 gives the first conclusion immediately. For the second, consider a binary Y independent of Θ for which $P(Y = 1 \mid \mathcal{H}) = \varepsilon$ with $0 < \varepsilon < 1$. Observing $Y = 1$ yields surprisal $-\log \varepsilon$, which increases without a fixed finite bound as ε approaches zero, while the posterior for Θ remains unchanged. The example concerns a family of hypothetical models and does not establish that real observers disregard irrelevant rarity. It establishes the narrower analytical point that rarity and diagnostic information are logically separable. This is especially relevant to relational histories in which an unusual cross-context encounter may attract an observer while a routine source document determines whether an attribution is supported.

The distinction is equally important for correlated and dependent sources. Suppose two summaries Y_1 and Y_2 reproduce one underlying record. If the second summary is an exact copy of the first under the modeled evidence description, then after Y_1 is known, the additional information $I(\Theta; Y_2 \mid Y_1, \mathcal{H})$ is zero. If the second provides an independent check, a translation correction, or a new contextual clue, its marginal value need not be zero. Nor is there a general theorem that conditional information gain must decrease monotonically after observing any other source: interaction and dependence can produce complementary information. The task is to model the particular provenance relation and observation

protocol, rather than assigning a fixed value to the number of documents. Consequently, a criterion based on sheer citation frequency may strongly favor repeated testimony even when the additional documents convey little independent evidence.

Saliency introduces another distinction. A source may be visually prominent because of position, contrast, typography, familiarity, recommendation ranking, or emotional framing. A document may be substantively important yet located several clicks below a more colorful summary. The visual-saliency construction of Itti and colleagues provides a well-specified example of feature-based selection [36], but documentary saliency also has linguistic and interpretive dimensions. A label such as *original source* may be salient because an observer understands its evidentiary role, rather than because it is perceptually bright. An interface intervention can alter both forms of saliency at once. The study therefore requires distinct manipulations or measurements if it aims to identify either mechanism.

The policy of Equation 8 can be operationalized using a restricted choice approximation. Let $x_{o,i,t}$ contain information that is actually available before selection, including position, preview appearance, anticipated relevance, and expected access cost. A descriptive multinomial model could use an estimated choice score $v_o(x_{o,i,t}; q, \mathcal{H}_{o,t})$, together with a stopping score. Equation 13 is one such family, rather than a psychologically established universal rule.

$$\pi_{o,t}(i \mid D_t, \mathcal{H}_{o,t}, q, \rho) = \frac{\exp\{v_o(x_{o,i,t}; q, \mathcal{H}_{o,t})\}}{\exp\{v_o(x_{o,\emptyset,t}; q, \mathcal{H}_{o,t})\} + \sum_{j \in D_t} \exp\{v_o(x_{o,j,t}; q, \mathcal{H}_{o,t})\}}, \quad i \in D_t. \quad (13)$$

The score can be parameterized by independently manipulated design features, and the model can be compared with deterministic threshold policies, order-based selection, or search strategies with memory. Its use does not license a claim that observers are rational utility maximizers or that its coefficients recover stable psychological traits. Correlated interface features, unobserved learning, and state-dependent stopping may leave the score weakly identified. In particular, the investigator should avoid inserting the realized information gain of an opened item as though it were known before inspection. An anticipated value signal must be operationalized through preview information or a participant-specific predictive model.

The softmax approximation also yields a limited, transparent statement about menu competition. Let r be an existing source and let Z be the strictly positive sum of exponentiated fixed scores for the original items and stopping option in Equation 13. Add a candidate k without altering any original score. The change in the first-choice probability for r obeys Equation 14.

$$\pi_{o,t}^+(r) = \frac{\exp(v_r)}{Z + \exp(v_k)} = \pi_{o,t}(r) \frac{Z}{Z + \exp(v_k)} < \pi_{o,t}(r). \quad (14)$$

The inequality follows solely from the chosen one-choice normalization. It is a mathematical property of this restricted model, not an empirical law stating that new information always distracts human observers. The result may fail to describe subsequent source examinations, changes to preview quality, adaptation of other scores, a display that expands the inspection budget, or an item that teaches the observer where to search. It supplies a useful falsifiable baseline: when irrelevant new items change inspection of an existing record beyond the predicted menu effect, investigators should consider changed rankings, interaction among sources, sequence dependence, or resource constraints rather than automatically attributing every displacement to sensory capture.

At least four rival accounts can produce similar initial choices: simple position-based scanning; selection by visually prominent features; an approximate task-relative value-of-information strategy; and a curiosity-driven policy that gives weight to unexplored states. A fifth strategy may prioritize familiar or institutionally endorsed sources, reflecting learned credibility cues rather than informativeness. Some resource-rational approaches explicitly study the fit between information-processing strategies and their computational costs [52]; the principle supplies a comparator, not a warrant to assume that any observed choice is optimal. Moreover, experiments on reward-learning horizons have separated information-directed exploration from increased choice variability in their particular setting [83]. Their

results support testing distinct exploratory mechanisms, while transfer to documentary inquiry requires dedicated study. Model comparison should accordingly include a simple order or location heuristic as well as richer models, report predictive performance outside the fitted trials, and recognize that several cognitive strategies can remain observationally equivalent with the available behavioral record.

3.3 Competing Demands and the Conditions for Distraction

Limited examination resources make the composition of the evidentiary menu consequential. In one task, conspicuous irrelevant material may draw attention away from a source citation. In another, an unexpected annotation may cue examination of a previously neglected original document. The effect of distraction therefore depends on task instructions, available control, the perceptual and cognitive demands of the primary task, and the stage at which competing information is introduced. Lavie’s review of load theory reports a distinction between high perceptual load, which can reduce distractor processing in studied paradigms, and burdens on cognitive-control processes, which can increase distraction [50]. The distinction cautions against treating *greater load* as a single monotonic predictor across evidentiary tasks.

For a one-round design, let $c_{o,i} \geq 0$ be an independently specified inspection cost and let B_o be the budget made available to observer o . A candidate feasible set is defined in Equation 15.

$$\mathcal{F}_o(D, B_o) = \left\{ A \subseteq D : \sum_{i \in A} c_{o,i} \leq B_o \right\}. \quad (15)$$

Equation 15 captures a particular operational constraint rather than a fixed neural capacity. The set A records selected items under a hypothetical additive cost rule. Costs may fail to be additive when inspection of one document makes another easier to understand or when switching between sources imposes additional burdens. Sequential and adaptive designs must then replace the static constraint with state-dependent costs and a stopping rule. Nevertheless, the simplified model helps isolate an important possibility: an added item may increase the number of reachable records while reducing the chance that one relevant record receives the limited time available.

Consider a hypothetical task concerning the origin of a method described in a dossier. A source registry contains a mundane entry recording the relevant provenance, whereas an accompanying graphic gives an unusual but logically unrelated numerical pattern. Under a single-item budget, an observer who selects the visually most prominent candidate may inspect the graphic and leave the registry unopened. A second observer who has a source-attribution task cue may select the registry. Both have the same formal evidence access, and neither example requires the archival material itself to change. The comparison demonstrates the possibility of allocation-dependent evidence loss. It does not establish that the first observer is irrational, since the perceived task, credibility of the registry, and hidden costs may differ. With a larger budget, both materials could be inspected, so the displacement mechanism may disappear even while the graphic remains salient.

A dynamic selection process can also create positive complementarities. Opening an initially tangential item may expose a previously unknown source link, making later investigation easier. An unusual cross-disciplinary translation may change which search terms are available; a concise source map may lower the effort of navigating an otherwise opaque corpus. The resulting attentional benefit depends on subsequent opportunities and on what the item actually permits the observer to learn. Thus, assessments of attention should preserve the distinction between the immediate selection of an item and the effect of that selection on later inquiry. In a sequential model, the prospective value of an item may partly reflect the additional search opportunities it creates, even when its immediate content has low diagnosticity for the target proposition.

The additive-budget feasibility set in Equation 15 is best understood as the investigator’s declared restriction on a task. For descriptive inference, the cost of inspection can depend on the selected order, the state reached after earlier reading, and whether two records share technical vocabulary. A second source may be cheap after the first has introduced its terminology, or expensive when a language change

or source-format conversion interrupts work. Fatigue and motivation may change over a session. If $c_{o,i,t}$ is allowed to depend on an evolving state, a valid feasibility condition must be evaluated over the realized sequence, including switching and translation costs, rather than using the static sum alone. This matters for experimental interpretation: participants who inspect few sources may be constrained by access friction rather than attention narrowly construed.

An inequality in available time should not silently become a statement of deficient intellectual ability. A participant who can only inspect a dossier during a short paid evaluation interval faces a different opportunity from someone who can reread the record without deadline pressure. Similarly, a summary service may redistribute the labor of triage, alter who pays the cost of verification, and make it difficult to recover which source relations were removed upstream. Those differences can be tested by holding the corpus fixed, systematically varying the time and provenance-display conditions, and measuring both accuracy and the work required to produce it. A cognitive model lacking such conditions risks treating structurally caused nonencounter as a personal preference for ignorance.

Evidence overload is also historically and institutionally conditioned. The volume and ranking of displayed material, the availability of translation, the time allocated to peer review, and the need to meet reporting deadlines affect the realized attention budget. The observer's competence and preferences interact with those conditions; neither individual psychology nor institutional access alone supplies a complete explanation. AI-generated summaries may assist by reducing navigational effort or obstruct by obscuring provenance or creating additional salient but weakly diagnostic claims. These are competing hypotheses for specified workflows, not properties that can be inferred from the presence of AI mediation alone. The mechanism of attention selection developed here addresses the individual encounter, while the political-economic production and governance of these interfaces remain the objects of the companion investigations.

3.4 Observer Heterogeneity and Relational Search

An observer's task and background knowledge condition the perceived significance of an evidentiary opportunity. A historian concerned with priority may seek dated source records; an engineer reproducing a procedure may prioritize procedural details; an evaluator considering source independence may examine which apparently separate reports share an origin. These are differences in legitimate research tasks as well as possible differences in cognitive capacity. A study of attentional encounter should record the target question before treating distinct item choices as an error or an epistemic deficiency. A document that is diagnostic for one proposition may be peripheral for another. Even within a single task, differences in language knowledge, experience with archives, and anticipated verification costs can produce different inspection trajectories.

Relational evidence further challenges the conception of attention as selection among self-contained documents. The diagnostic content of an item may lie in its relation to another record. A citation, for example, can be read simply as a bibliographic string, as evidence of influence, or as a potential clue to shared informational dependence. The relevance of that item can be recognized only after consulting another part of the corpus. An analyst might represent this as attention to a pair of documents or as a sequence in which reading the first changes the observer's anticipated value of the second. Both are provisional representations of interaction. They need not imply that the relational event exists as a timeless and independently bounded object. Where temporal order is only partially evidenced, the observer's search may concern alternative admissible orderings, with uncertainty preserved instead of forcing a single historical narrative.

The literature on active sampling provides a complementary rationale for allowing the observer to determine subsequent questions [27]. An observer may examine unfamiliar material to discover potential relevance that could not be specified fully before inspection. Such behavior may be epistemically productive, particularly in exploratory research, but an open-ended search may also compete with a narrow task. The investigator should distinguish an assigned verification goal from exploratory objectives and from participants' private interests. A selection that appears inefficient for an imposed laboratory task

may help an observer learn something valuable that the task’s scoring rubric does not capture. A suitable design can measure both the prespecified target outcome and a clearly delimited exploration outcome, avoiding post hoc relabeling of every unexpected selection as either optimal or irrelevant.

The unit of search may itself be relational. An observer might inspect a document because an archive presents a title, yet the sought evidence may be a relation between that document and another entry. In a task concerning duplicate testimony, successful examination might require opening both records and comparing a particular source passage. An observer who investigates only one record can detect a local claim but cannot establish the dependence relation. Thus, the investigator should distinguish item-level coverage from relation-level coverage and construct a task-specific map of the minimum evidentiary conjunctions required to answer the question. The map can include alternative adequate paths: a directly inspectable dated record may be sufficient in one designed task, while a conflict among witnesses may require comparison of several independent records in another. Its construction should be audited independently of participant choices.

The adaptive character of such tasks limits post hoc judgments about search efficiency. A participant might recognize a source dependence only after discovering a third document, at which point the informational value of revisiting a previously ignored item rises. Conversely, the discovery that several items reproduce one press release may render further inspection redundant for a particular attribution question. A defensible evaluation of search strategy must therefore condition the benchmark on information that had actually become available at each occasion. Judging the participant against a fully revealed corpus without recording their earlier knowledge can manufacture an apparent failure of attention through hindsight.

This observer-relative perspective also constrains use of statistical novelty. A cross-context source link may be rare under one observer’s prior experience while routine under another’s. An unusual event in a newly adopted representation may reflect changes in classification rather than changes in the historical record. Novelty may create a motive to inspect evidence, but the warranted implications depend on the proposition, source independence, and the evidence obtained after inspection. The separation established in Proposition 3.1 therefore preserves the distinction between possible attention attraction and the evidentiary force of the eventual observation.

3.5 Experimental Comparisons and Identification Boundaries

An empirical research program should discriminate among mechanisms rather than infer them from the final rating. One design presents a fixed synthetic corpus concerning a precisely defined source-attribution proposition. Before exposing the evaluative interface, participants can receive controlled information establishing base-rate expectations about a particular incidental feature. The actual evidentiary relevance of each source is fixed by the constructed corpus, while preview prominence and contextual surprise cues are varied according to independently specified conditions. This makes it possible to investigate whether attention follows perceptual salience, trained rarity expectations, or cues regarding source diagnosticity. The manipulation of rarity must be checked independently; simply inserting a visually striking item does not identify an effect of statistical surprisal.

A second design varies the time available for inspection while manipulating the number and prominence of competing items. It can measure how often a provenance-bearing document is opened, how long it remains accessible on screen, and whether the participant identifies the shared origin of several claims. It should also include memory and inference probes administered after the relevant choices, since an opened item may be skimmed without being understood. The direction of effects is an empirical matter. Increased time could raise source inspection rates or permit more distraction; stronger provenance highlighting could improve attribution accuracy or induce superficial confidence in the interface. The outcome measures need to distinguish these possibilities.

A third comparison allows sequential inquiry and makes the link structure itself experimentally variable while keeping the substantive documents constant. One interface displays a source map; another provides a conventional ranked list; a third may present an algorithmic summary with individually in-

Table 3: Proposed distinctions and measurement targets for studying selective attention to relational evidence. The measures describe observable proxies rather than direct readings of mental states.

Construct	Operational comparison	Interpretive constraint
Evidence availability	Reachable materials, assigned display, retrieval permissions	Access can remain unused and may differ from automatic presentation.
Perceptual salience	Randomized position, contrast, preview prominence	May also influence comprehension or perceived credibility.
Expected rarity	Trained base rates and surprise-cue manipulation checks	Realized surprise is generally observed after selection.
Task relevance	Prespecified target proposition and blinded relevance coding	Relevance depends on question and source interpretation.
Expected information value	Predictions conditional on available preview signals	Model-based residual gain differs from experienced utility and from hindsight.
Processing cost and load	Time budgets, calibrated reading effort, competing tasks	Perceptual load and cognitive-control demands can differ in effect.
Evidentiary encounter	Opening, dwell, navigational sequence, comprehension probe	Behavioral traces alone do not establish semantic processing.

spectable provenance links. Participants can inspect further records subject to a common time allocation. The primary outcome could be accurate source-dependence reconstruction, while secondary outcomes include inspection order, calibrated uncertainty, and time spent verifying. The study should separately report assignment effects and observed-path associations. A choice to follow a link is influenced by prior beliefs and interest, so analyses conditional on voluntary inspection generally do not identify causal effects of inspection without additional assumptions. These comparisons would help determine whether more transparent relational representations improve evidentiary reasoning through attention, understanding, or both.

Table 3 separates the candidate explanatory constructs from their possible measurements. Its purpose is to prevent an investigator from operationalizing *attention* by one proxy and subsequently interpreting a change in final judgment as though it had established all intermediate mechanisms. The choice of outcome also affects the claim that a change constitutes an improvement. If an experimental manipulation leads observers to spend more time on a source and then express uncertainty because they recognize conflicts, the direction of the final confidence change may be negative while the quality of source assessment improves. Conversely, a more conspicuous citation icon may increase clicks without increasing correct provenance reconstruction. Epistemic evaluation requires a separately specified criterion for the target task.

Two conditional hypotheses follow from the proposed account. The first concerns competition between prospective information value and conspicuity; the second concerns budget-sensitive displacement. They are formulated as empirical contrasts and are not presented as established general psychological laws.

Hypothesis 3.2 (Relevance-sensitive selection under controlled salience). In a bounded inspection task with independently manipulated preview salience, documented prior expectations, and objectively specified evidence dependencies, increasing valid cues to the diagnostic relevance of a source will raise its probability of substantive examination in at least some task and observer strata, conditional on the cost of inspection and on the availability of competing evidence.

Hypothesis 3.3 (Budget-conditioned distraction from weakly diagnostic material). When inspection time is experimentally restricted and a competing item is salient but conditionally uninformative about the target proposition, increasing its prominence can reduce examination of a provenance-bearing source in some display conditions. The effect is expected to depend on the relation between inspection costs, task instructions, and load, and may diminish or reverse when the selection process permits further inquiry.

The primary causal contrast should be defined at assignment, before participants have chosen an inspection path. For instance, random assignment to a provenance-rich versus ordinary preview interface supports an intention-to-treat comparison of the probability of substantively examining a prespecified source, provided assignment and outcomes are measured appropriately. The comparison captures the total effect of the interface as implemented; it does not identify the isolated effect of attention when the interface also changes how evidence is understood. The number of opened documents, gaze duration, and reported subjective difficulty may all lie on paths affected by assignment. Conditioning on them can distort the original randomized comparison, and a mediation claim requires assumptions and supporting measurement beyond randomization of the interface itself. Section 9 considers identification more comprehensively.

A practical factorial design should separately randomize the visibility of a target source, preview-level cues of its expected diagnosticity, and available inspection time, while keeping the underlying evidentiary claims constant. Where multiple documents from one source family are included, the assignment unit and standard errors should reflect the relevant dependence structure and possible spillovers. A familiarity or rarity training manipulation requires a separate check that participants acquired the intended prior expectations; surprise inferred from the rarity of the realized item is insufficient. Task instructions, wording of provenance labels, reading difficulty, and translation quality also require equivalence checks. Source comprehension should be scored on preregistered, evidence-supported features, with uncertainty or multiple admissible answers credited where appropriate. To distinguish perceptual noticing from substantive assessment, a subset of trials may permit immediate, low-intrusion observation of opening sequence, followed only after the target choices by probes asking what source relation the participant understood. Interpretation of those probes must account for retrieval and testing effects.

Two additional comparisons guard against convenient but uninformative model fits. First, present source candidates that are visually matched yet differ in expected task relevance, and in separate trials candidates matched for relevance yet differing in position or visual contrast. Second, alter the order in which a dependent pair of records becomes available while keeping their contents unchanged. These comparisons can reveal whether the putative information-value model predicts within-participant adaptation better than a fixed ranking or a position heuristic. Cross-validated model predictions and preregistered exclusions are preferable to treating fitted attention weights as psychological facts. A failure of any hypothesized positive association, including a null or a sign reversal, is an informative result when the measurement process and the proposed causal contrast have been specified in advance.

A valid test requires manipulation checks, a prespecified causal contrast, and protection against selection on post-assignment behavior. Participants may differ in competence, reading speed, and prior expectations even after randomization; these variables can be measured before assignment for precision and heterogeneity analysis. Eye-tracking, cursor movement, and opening logs can indicate selection but cannot independently establish correct interpretation. Conversely, post-task questions may influence retention or invite retrospective rationalization. A design concerned with initial choice should therefore measure immediate selection unobtrusively and place more intrusive source-comprehension tasks after the relevant encounter. Replication across document genres, languages, and observer communities would be necessary before extending results to scholarly review, public knowledge infrastructures, or human–AI collaboration.

The proposed factorization also has a failure condition. Some observers read and interpret continuously while searching; the supposedly separate attention stage can depend on partial semantic comprehension acquired before conscious selection, and the boundary between preview and inspection may be empirically unstable. A more realistic temporal model can alternate short acts of sampling, interpretation, and new selection, conditioning each step on an evolving observer state. Different tasks may require different temporal resolutions. A one-round attention policy is a helpful experimental approximation only when the chosen resolution preserves the causal distinctions of interest. The present formulation therefore establishes an agenda for measurement rather than a claim that cognitive attention follows a fixed number of steps.

The section’s contribution to the larger argument is to identify a conditional transformation between

evidence availability and evidentiary encounter. A historically preserved source may be accessible while receiving little examination; a structurally unusual document can attract attention without carrying relevant information; and an ordinary source may acquire heightened salience when a new task or relational pathway makes its importance intelligible. These possibilities can be represented without equating novelty, salience, and justified trust. The following section examines the next interface: how attended materials are compressed into memory traces and subsequently reconstructed, with particular attention to provenance, relational order, and uncertainty in source attribution.

4 Memory Compression and Trajectory Reconstruction

This section investigates the cognitive transformation between attending to relational evidence and later reconstructing the history to which that evidence appears to refer. Its objective is to distinguish retention of substantive content, retention of source and provenance information, and reconstruction of relations among contributions. These operations can support different judgments, and their errors need not coincide. The analysis combines established findings on working memory, recognition, source monitoring, and reconstructive remembering with a deliberately limited probabilistic framework. It introduces task-dependent measures of compression, a source-attribution kernel, and uncertain reconstruction of partially ordered histories. It then specifies experimental comparisons capable of distinguishing loss of trace detail from source confusion and from post-encoding reinterpretation. The formal constructions represent analytical alternatives available to an investigator; they do not assert that cognition realizes one invariant compression algorithm or internally stores a fully specified temporal graph.

The distinction matters for the present paper's research object. A reviewer may recall the central argument of a document while assigning it to a later commentator; a reader may remember an author's name but fail to recover which evidence supported the claim; a participant may reconstruct the correct contributors yet impose an unsupported sequence on interactions that the archive only partly orders. Each outcome can affect epistemic trust through a different pathway. In a scientific or institutional setting, the preservation of a public record also differs from the preservation of that record in any observer's memory. Paper II studies the former conditions of access. Here the question concerns the observer's cognitively retained, retrieved, and sometimes reconstructed history once a specified presentation has been encountered.

A relational archive makes three kinds of retention particularly consequential. *Item retention* concerns whether a proposition, artifact, or contribution is recognized. *Provenance retention* concerns the relations that connect the item to a speaker, creator, source document, or institutional intermediary. *Relational reconstruction* concerns whether the observer can recover supported dependence, succession, co-production, or unresolved relations among recorded episodes. These targets have distinct denominators and potentially different documentary standards of correctness. A study that collects an overall memory score without specifying the target will therefore mix empirically different errors. The division is analytical: it does not require the corresponding psychological processes to be anatomically or functionally independent.

The observer's position in the material reproduction of knowledge also matters. Scholars, reviewers, students, administrators, and readers have unequal time, language resources, disciplinary familiarity, and opportunities to reopen source documents. A selective or apparently incomplete recollection can be partly a consequence of the task assigned and of the infrastructure through which records are presented. This historical and institutional variation is represented here through the encounter conditions and task-specific scoring rules. It should be analyzed directly in applications rather than converted into an undifferentiated individual memory deficit.

4.1 Encoding, Retention, and Task-Dependent Compression

The model of Section 2 treats A as attended evidentiary material and M as the resulting encoded or retrieved state. That notation can describe an immediate test, a delayed reconstruction, or a repeated-

reading task, provided the observation protocol specifies which of these situations is intended. The distinction between an initial trace and a later report is useful because remembering itself can introduce new selections and interpretations. A research design concerned with delayed attribution should therefore allow M^{enc} to represent an unobserved or partially measured state after an initial encounter and M^{ret} to represent a state available at the later test. These variables need not coincide, and a reported recollection is not a direct measurement of the original trace.

Working-memory research establishes a substantial tradition of studying resource constraints in the temporary maintenance and manipulation of information [5]. Research on recognition has also distinguished recollection of contextual detail from familiarity-based responding, although the decomposition into distinct processes is itself model-dependent [84]. Work on memory-guided behavior has also examined partially distinguishable systems concerned with entities and the construction of richer contexts [64]; such neurocognitive accounts do not map directly onto the present observer kernels. For a relational history, the practical implication is that noticing or recognizing an artifact does not ensure retention of all relations necessary for a later provenance or priority judgment. Content, authorship, date, method, cross-reference, and the conditions under which a claim was made can vary in their task relevance and memorability. The present framework therefore treats a memory state as a context-dependent representation whose fidelity must be assessed relative to a question, rather than by assuming an all-purpose quantity of remembered information.

A candidate within-person model specifies the encoding and later retrieval laws as two conditional kernels. Equation 16 separates these stages while keeping context and testing conditions visible.

$$\begin{aligned} K_o^{\text{enc}}(dm \mid a, e, c) &:= \Pr_o(M^{\text{enc}} \in dm \mid A = a, E = e, C = c), \\ K_o^{\text{ret}}(dr \mid m, c, d) &:= \Pr_o(M^{\text{ret}} \in dr \mid M^{\text{enc}} = m, C = c, D = d). \end{aligned} \quad (16)$$

The symbol D denotes a specified retrieval condition, which can include the interval since encounter, reminders, task instructions, intervening exposure, and available cues. Equation 16 is a proposed decomposition of a conditional law. It is most defensible when the investigator knows which exposures and interventions occur between encoding and testing. If a participant repeatedly reviews an external archive, retrieval becomes partly a new access-and-attention episode and the two-stage decomposition is insufficient. If prior memory influences the original attention selection, then a dynamic representation conditioning on earlier observer state is also required. The point of writing two kernels is to identify what must be measured or experimentally separated, not to infer an unobserved trace uniquely from a later answer.

A useful task-specific description of compression is the loss incurred when a cognitive representation supports an answer to an evidentiary question. Let q index a question selected from a stated set $\mathcal{Q}$, let $T_q(H)$ denote the document-grounded target answer under an explicit scoring protocol, and let $d_q(M)$ be the answer reconstructed from memory. For weights $w_q \geq 0$ summing to one, define the expected task loss in Equation 17.

$$\mathcal{L}_o(M; \mathcal{Q}, C) := \sum_{q \in \mathcal{Q}} w_q \mathbb{E}_o[\ell_q\{d_q(M), T_q(H)\} \mid C]. \quad (17)$$

Here ℓ_q is a declared scoring loss: it could record incorrect identification of a documented contributor, omission of uncertainty in a source claim, or unwarranted assertion of an order relation. The expectation in Equation 17 is defined relative to a particular study population and joint distribution. The functional does not imply that human observers consciously optimize it. It simply makes the meaning of a “compressed” representation operational. A state that preserves a publication’s principal finding may have low loss on content questions and high loss on provenance questions. Another state containing a source map but few technical details may show the reverse pattern. The weights reflect the study’s evaluative priorities, so any aggregated score is conditional on those priorities and cannot serve as a universal scale of historical understanding.

Compression can also be modeled at different resolutions. An observer may retain a gist of a collaborative episode, a selected causal explanation, a list of participants, or a coarse narrative sequence.

Schacter and colleagues review evidence that some memory distortions are related to otherwise useful reconstructive and generalizing processes [69]. Their review motivates hypotheses about task-dependent omission and reconstruction; it does not establish a universal tradeoff between compression and accuracy. In the present application, omission may be epistemically harmless if a question concerns the existence of a result, while the same omission can be decisive when the question concerns who generated the underlying technique. An investigator must accordingly record the task’s necessary relational distinctions before deciding whether compression represents an error.

A subtle issue arises when participants possess relevant prior knowledge. A participant might reconstruct a provenance relation correctly from disciplinary expertise without retaining the original footnote; another might remember the footnote without understanding what it implies. Accuracy of a final answer therefore does not prove faithful episodic retention, just as an inaccurate answer does not uniquely identify deficient initial encoding. A good study will measure recognition of exact presented details, performance on novel relational inferences, and confidence about the source of each inference separately. Multiple well-designed tasks can constrain alternative mechanisms even when latent trace states remain unidentified.

The distinction between observation and latent memory requires a sharper identification statement. In a restricted, fixed-encounter setting, suppose the later retrieval condition $D = d$ does not alter the distribution of M^{enc} conditional on the encounter and context, and suppose the delayed state is conditionally generated by K_o^{ret} without unrecorded intervening exposures. The composed relation then takes the form of Equation 18.

$$\begin{aligned} \Pr(M^{\text{ret}} \in B \mid A = a, E = e, C = c, D = d) \\ = \int K_o^{\text{ret}}(B \mid m, c, d) K_o^{\text{enc}}(dm \mid a, e, c). \end{aligned} \quad (18)$$

Equation 18 is a conditional factorization under its stated assumptions. Even if the left-hand law could be measured precisely, it generally would identify only the *composition* of the kernels. Distinct encoding and retrieval mechanisms can induce the same observed distribution, so a final recall score does not establish whether information was never encoded, became inaccessible, was overwritten, or was reconstructed differently. Randomizing retrieval cues after the encounter and collecting an immediate, carefully chosen measurement may narrow these alternatives, while the immediate test can itself affect subsequent remembering. Unrecorded outside searches and recurrent encounters require a richer longitudinal model. This nonuniqueness is an identification limitation of the proposed representation, independent of whether a particular memory theory is accepted.

Evidence on selective attention and the later binding of contextual information reinforces the need to measure target dimensions separately. Uncapher and Rugg studied how attentional selection can influence the mnemonic binding of contextual information under a specified laboratory paradigm [80]. The result is relevant to the hypothesis that source-specific details and document content can be differentially encoded, but it does not identify how scholars remember chains of contribution. At the level of method, a task should therefore distinguish what a participant actually read, what was attended as source context, and what was later reconstructed from prior knowledge. Contextual details that are never displayed should not be scored as failures to remember them.

4.2 Source Monitoring and Provenance Attribution

Source monitoring concerns how a person attributes remembered information to an origin. Johnson, Hashtroudi, and Lindsay’s account emphasizes evaluative processes that use qualities of remembered experience and contextual cues; remembered content is not ordinarily accompanied by a perfectly reliable, directly retrieved abstract source label [39]. This distinction is especially relevant for relational evidence, since the source of a proposition, the producer of a cited artifact, and the institution through which it was encountered are different candidate referents. A well-formed provenance question has to

specify whether “source” means the original contributor, a witness, a repository, a secondary commentator, or the immediate interface encountered by the observer.

Consider a hypothetical archive containing an original procedural note, an independent assessment of that procedure, and a later synthesis that quotes the note. After reading an interface displaying all three, a participant accurately recalls the method but attributes the original proposal to the synthesis author. This is a source-attribution error relative to the retained records, even though the recalled method is correct. Alternatively, the participant may accurately recall that the synthesis quoted the earlier note but remain uncertain whether the note demonstrated practical competence. That second outcome represents a more cautious evidentiary judgment rather than a source-memory failure. Distinguishing these cases requires the experimenter to specify the target relation among texts, contributors, and asserted claims.

A local attribution kernel can express the distinction. Let $S_q(H)$ denote the source relation defined by question q and supported by the held records, and let $\hat{S}_q$ denote an observer’s elicited attribution. The distribution of elicited source judgments is given in Equation 19.

$$K_o^{\text{src}}(B \mid m, c, q) := \Pr(\hat{S}_q \in B \mid M^{\text{ret}} = m, C = c, q). \quad (19)$$

The codomain of $\hat{S}_q$ should include explicit uncertainty and, where relevant, several jointly supported sources. It need not be a fixed list of actor types. A collaborative artifact may have multiple evidentially warranted contributors, and the role of a contributor may differ from that of a publisher or subsequent disseminator. A single-choice test can impose unwarranted exclusivity and make the participant’s answer appear more definite than the underlying records justify. The kernel in Equation 19 is descriptive; its values are not a measure of warranted attribution unless compared to an appropriately established record-based target.

Provenance is often many-to-many. One contributor may supply the method, another its independent validation, and several people a subsequent representation. The interface can present a synthesis that visibly names one editor while concealing upstream documents, or it can reveal a citation chain without explaining what kind of contribution each reference supports. An experiment should record at least the relation type as well as the alleged source identity: proposed by, performed by, documented by, quoted by, distributed by, and evaluated by are possible task-specific relations. The empirical question is whether memory preserves a *typed relation* rather than whether an observer selects the socially most salient name. The ontology of roles and the permitted multiplicity of answers must follow the evidence and the study’s stated purpose.

An additional difficulty is the observer’s post hoc inference. A respondent who does not remember a source label may infer that a technically sophisticated result probably came from a senior institution; another may infer that a repeated or stylistically familiar passage must have come from a frequently encountered commentator. The same elicited answer can arise from a retained attribution or from a new inference over disciplinary and institutional priors. The distinction matters when interpreting a source-memory test, because an elicited judgment need not be a report of an already stored label. Process-dissociation methods provide one relevant methodological tradition for separating hypothesized components of responding under explicit model assumptions [37]. That tradition motivates rival process models here; it does not license assigning pure familiarity and recollection components to any particular survey response without validation.

Proposition 4.1 (Source identity under indistinguishable memory evidence). *Consider a binary source-attribution task with candidate sources s_1 and s_2 . Conditional on a held experimental context $C = c$ and eligible observations, suppose both sources have equal prior probability and produce the same conditional distribution of a memory measurement M^{ret} . Then a decision rule based solely on that measurement and the held context has expected source-identification accuracy at most one half under zero–one loss, even if content recognition is perfect.*

Proof. Under the assumptions, Bayes’ rule gives equal posterior probabilities for s_1 and s_2 for almost every memory measurement in their common support. Any possibly randomized classification rule based

only on that measurement therefore has expected success probability one half. A memory report containing the full substantive claim remains uninformative about source identity when its conditional distribution is identical across sources. $\square$

Proposition 4.1 is a simple identification result, not a newly discovered property of human memory. It clarifies why accurate content recall alone cannot establish provenance accuracy. When sources have unequal priors, the best rule can favor the more common source without acquiring discriminating memory evidence; when the observer has an independent record or contextual cue, the premises fail and better attribution becomes possible. The proposition also assumes that the available memory measurement captures what the decision rule can use. Unmeasured contextual impressions or a fresh search can supply further information. For these reasons, its empirical role is to guide measurement of source-specific traces rather than to place a categorical ceiling on real-world identification.

The indistinguishable-source case has a useful quantitative extension. Let P_1 and P_2 be the conditional laws of the available memory measurement under the two candidate sources, with equal source priors and no extra discriminatory cue. Under the optimal zero-one source classifier, Equation 20 relates attainable accuracy to the total-variation distance between those laws.

$$\text{Acc}^*(P_1, P_2) = \frac{1}{2}(1 + \|P_1 - P_2\|_{\text{TV}}). \quad (20)$$

Here $\|P_1 - P_2\|_{\text{TV}} := \sup_B |P_1(B) - P_2(B)|$. Equation 20 is a decision-theoretic identity for the stipulated binary experiment, obtained by assigning each measured outcome to the source with higher likelihood. When the conditional distributions agree, accuracy is one half; when their supports are disjoint, an ideal classifier can identify the source. Intermediate overlap gives an explicit limit for discrimination from that measurement alone. The identity is *not* an empirical estimate of human source memory, a model of the observer's actual decision strategy, or an assertion that existing archives have known equal source priors. Its value is to show that source-discriminating information, rather than mere detailed recall of claim content, is the analytically relevant ingredient.

A further complication is provenance depth. Remembering the latest speaker is different from remembering the origin of the cited claim. Suppose a participant hears a method described by three intermediaries, only one of whom inspected the original document. A source-memory task may correctly identify the immediate speaker while failing to reconstruct the upstream dependency. Conversely, an observer might know the original document from prior study but be unable to name the most recent intermediary. The distinction between direct encounter and inferred source lineage is important for Paper II's evidence-dependency problem and for the testimony evaluations planned in Section 7. At the cognitive level it asks which lineage relations were encoded, what can be reconstructed, and where uncertainty persists.

Source-monitoring experiments should therefore compare at least three observable responses: recognition that a proposition was encountered, identification of where it was encountered, and attribution of its originating evidentiary basis. A fourth task can ask whether two reports were independent. The tasks may be scored relative to documentary provenance and compared across presentation formats. Separating them prevents a single accuracy percentage from obscuring whether a display helped content retention while undermining origin discrimination. It also preserves a critical ethical distinction between measured misattribution and a claim about the participant's motivations, competence, or integrity.

4.3 Reconstructing Relational Order Under Incomplete Evidence

Relational histories often contain incompletely ordered observations. A record may establish that one methodological note influenced two later reports without establishing a temporal relation between those reports. A collaborative process may include overlapping exchanges, revisions, and contributions whose boundaries are defined differently in different archives. A cognitively reconstructed narrative can nonetheless impose a single sequence because linear narration is convenient for explanation. The task

here is to model the difference between reconstructing the order supported by documentary evidence and reconstructing an order selected by an observer’s expectations.

The use of partial order is analytical and local. Let V_h be a finite set of provisionally individuated evidentiary events in a chosen historical representation and let $\mathcal{R}(h)$ be the set of order relations admitted by the stated documentary constraints. This set may contain several different partial orders; the investigator must explain which records and uncertainty rules determine admissibility. An observer’s report about a candidate relation, such as $u \prec v$, may be coded separately from the occurrence, duration, or causal significance of the corresponding activities. A difference in order judgments should not automatically be interpreted as a difference in the historical world, since several chronologies may remain consistent with the available documentary traces.

A useful analyst-side device is a distribution over admissible reconstructions, shown in Equation 21. The distribution can represent explicitly elicited uncertainty or a fitted descriptive model; it does not assert that the observer internally enumerates graph orders.

$$\pi_{o,q}(r \mid m, c) := \Pr(\hat{R} = r \mid M^{\text{ret}} = m, C = c, q), \quad r \in \mathcal{R}_*(q). \quad (21)$$

The set $\mathcal{R}_*(q)$ in Equation 21 is the investigator’s declared response space, which can include admissible, partially specified, and unsupported alternatives as well as an “insufficient evidence” option. It is therefore deliberately distinguished from $\mathcal{R}(h)$, the constraints supported by the held documentary representation. A study can measure whether the respondent concentrates probability on relations established by the source material, distributes probability across genuinely unresolved alternatives, or confidently asserts relations incompatible with available records. Such comparisons are relative to the documentation protocol; they do not solve the philosophical question of a unique and globally given historical ordering.

Pairwise judgments offer one way of examining the result. If a respondent assigns probabilities to $u \prec v$ and $v \prec u$, the residual probability can be reserved for overlap, unresolved relation, or another explicitly defined option. A robust elicitation should avoid forcing exhaustive binary rankings when the records permit incomparability. When several pairwise assessments are combined into a reconstructed partial order, consistency constraints such as antisymmetry and transitivity matter. Independently elicited probabilities may support mutually inconsistent most-likely pairwise decisions, so a fitted order should include a criterion for reconciliation rather than silently projecting them onto a convenient narrative. An analyst may compare a constrained order model with a representation allowing unresolved relations; the comparison tests the usefulness of the imposed structure for the chosen task, not its ontological primacy.

The distinction is illustrated by a hypothetical three-artifact trajectory. Artifact a documents an initial technique. Artifacts b and c each cite a , but the archive records no direct relation between b and c . A respondent who remembers the substantive technique and both citations may nevertheless report that b preceded and caused c because the display listed b first. Another respondent might correctly state that the relative order remains unresolved. A third might claim c came first after reading an unrelated summary that foregrounds it. These outcomes should be distinguished from the independent scientific question of whether b and c represent separate contributions or derivative accounts of one result. The relational representation, memory report, and evaluative inference each introduce different uncertainty.

The meaning of an apparent “missing edge” requires special care. An unrecorded relation may be false, possible but undocumented, or outside the archive’s temporal and conceptual resolution. A partial order can encode a represented relation of precedence under a stipulated event individuation, while a citation graph encodes documentary reference and a dependency graph encodes a proposed generative relation. These graphs need not have the same edges. For example, a second document may cite a first after both authors have independently developed related procedures. Citation establishes a relation in the published record without by itself establishing the origin or magnitude of either contribution. The observer’s task must therefore name the *kind* of relation to reconstruct and record whether its absence signifies an explicit negative finding or a lack of information.

A practical scoring protocol should separate three types of answer relative to the documentary standard: relations positively supported by the curated corpus, assertions contradicted by documented constraints, and claims for which the corpus remains indeterminate. Those categories are epistemic and

protocol-relative, not a comprehensive partition of historical reality. For instance, confidently reporting $b \prec c$ when the archive leaves b and c unresolved is an *unsupported ordering assertion*, while reporting $c \prec a$ when $a \prec c$ is established is a *contradicted ordering assertion*. Both can be examined, but merging them into a single false-memory class would discard an important difference. Scores for acknowledging uncertainty should have prespecified costs so that a respondent is neither rewarded automatically for indiscriminate abstention nor penalized for recognizing genuine gaps. Where sources conflict, researchers can use sensitivity analyses across alternative reasonable codings of documentary support.

Reconstructive memory research provides grounds to investigate such distortions without treating them as uniform defects. Evidence from associative false-memory tasks shows that participants can report unpresented but semantically related material, including confident recollection in specific experimental settings [66]. Reviews of reconstructive remembering describe ways in which semantic organization and contextual processing can support useful generalization while sometimes yielding errors [69]. Those findings motivate experiments on inferred missing links and unsupported sequences, but they do not demonstrate that archival readers systematically invent contribution histories. The relational-historical extension remains a hypothesis to test across document genres, expertise levels, and presentation formats.

A distinction is also needed between a mistaken relation and a reasonable inference. If a record establishes that one manuscript was uploaded before another, the earlier upload date does not alone establish who first developed a method. If a document cites an earlier paper, this is evidence of a documented citation relation; it does not prove independent comprehension, personal acquaintance, or the absence of unrecorded influence. In other cases, domain knowledge may legitimately support an inference about the likely ordering of technical steps. Evaluating a respondent’s answer thus requires separating directly recorded relations, inferences supported by defensible background assumptions, and relations for which relevant evidence is unavailable. This separation is indispensable when the task concerns recognition of intellectual contributions rather than mere date memorization.

4.4 Repeated Exposure, Familiarity, and Confidence in Reconstruction

Scientific and public communication often exposes observers repeatedly to a claim while revealing its evidentiary history only intermittently. Recurrence may arise through reviews, summaries, automated recommendations, quotations, or repeated contact with the same original document. A repeated presentation could strengthen recognition of the claim while leaving the observer unable to reconstruct its origins. In other circumstances, repeated engagement with explicit provenance could help source memory. These are competing possibilities that depend on which material is repeated, how it is presented, and which later question the observer must answer.

Recognition-memory research distinguishes familiarity-based responses from recollection of contextual detail under several theoretical accounts [84]. Metacognitive research likewise distinguishes people’s predictions and feelings about their knowledge from performance, showing in specific paradigms that judgments rely on cues and can be imperfectly calibrated [46]. Neither literature licenses treating confidence as universally unreliable or repetition as uniformly harmful. The useful research distinction is among familiarity with a statement, confidence in being able to reproduce it, confidence about who produced it, and the externally assessable accuracy of those judgments.

For a defined source task, a simple calibration target can be written in Equation 22. Here Q^{src} is a participant’s reported probability that the elicited source attribution is correct, and Y^{src} is an independently scored binary accuracy indicator under a task with an adequately documented answer.

$$\text{Cal}_o^{\text{src}}(p) := \mathbb{E}_o[Y^{\text{src}} \mid Q^{\text{src}} = p] - p. \quad (22)$$

Equation 22 gives a descriptive population-conditional calibration discrepancy where the conditional expectation is estimable, often through predeclared probability bins rather than pointwise estimation. A zero discrepancy within a bin is compatible with many different degrees of individual discriminative ability and does not by itself establish that a respondent considered the appropriate evidence. Source

judgments also may have several partly correct answers; in such cases the scoring rule must allow graded or set-valued outcomes and Equation 22 is an oversimplification. Calibration can therefore complement, but cannot replace, documented provenance checking and analysis of how a conclusion was reached.

Repeated exposure creates another identification difficulty. A participant may become more confident because a claim is easier to recognize, because the participant consulted the original document, because an independent corroboration was encountered, or because the participant was asked to make the same judgment several times. These processes predict different evidentiary warrants. Their observational signatures may nonetheless overlap if a study measures only a final confidence score. In particular, repeated circulation through several channels may appear to be corroboration unless the respondent recognizes that the channels share an origin. That latter phenomenon concerns memory for dependence as well as memory for content and leads naturally to the testimony analysis of Section 7.

An experiment could randomize exposure repetitions and the presence of provenance cues while holding the substantive documentary text fixed. For example, one condition could repeat a short summary three times, another could present the summary once alongside a clear source lineage, and a third could repeat the original document with its citation relations intact. These conditions have different total reading burdens, so a cleaner factorial design would separately balance cumulative reading time, number of presentations, and provenance content. Investigators should avoid interpreting an effect of added source cues as a pure memory improvement if the cues also change initial attention, interface navigation, or comprehension. A multiple-time-point design can help identify when changes occur, although measurement itself may alter subsequent retention.

Hypothesis 4.2 (Dissociation between claim familiarity and source discrimination). When repetitions increase familiarity with a claim while leaving source-specific cues relatively weak, item recognition may improve more than source discrimination. The difference is expected to depend on task demands, prior expertise, and whether the repeated presentation contains explicit source relationships.

Hypothesis 4.2 is proposed for relational documentary settings and follows as a testable extension of memory and source-monitoring research, not as a universal consequence of repetition. It could be falsified in settings where the source is distinctive, repeatedly encoded, or central to the observer's task. The appropriate test requires comparable materials and separate item-recognition and provenance-reconstruction measures.

Recall may also change through the act of retrieval. Karpicke and Roediger found that retrieval practice improved later retention of foreign-language vocabulary relative to some repeated-study conditions in their experimental design [43]. That result provides a reason to treat testing as a potential learning intervention. It does not establish that repeated retrieval improves complex provenance memory, for which source details and relations might receive little retrieval practice even as claim content becomes more accessible. Similarly, Hubbach and colleagues reported reminder-dependent incorporation of subsequently learned items into recall of an earlier list in a multi-session laboratory task [34]; a follow-up using explicit source questions showed asymmetric misattribution of items to the earlier learning session in the studied reminder condition [33]. These results motivate hypotheses about contextual reactivation and source confusion in documentary tasks while leaving open whether the underlying processes transfer to sustained scholarly reading.

Three longitudinal mechanisms should be distinguished in a relational-evidence experiment. A *practice effect* may increase access to the details actually retrieved. A *source-updating effect* may alter how new material is linked to a remembered encounter. A *decision-policy effect* may change willingness to assert an attribution without changing what can be recalled. These mechanisms can produce similar aggregate improvements in recognition accuracy, and none can be read off a single delayed report. A design should record when a source was first encountered, when each retrieval prompt occurred, whether feedback was supplied, and what new materials became accessible between assessments. Repeated exposure that only restates an assertion should also be distinguished from repeated independent evidence for that assertion. Failing to track this dependence can lead observers and analysts alike to overcount corroboration.

A memory experiment could use a crossed design with source-rich and source-poor retrieval prompts, retrieval attempts and matched restudy controls, and immediate versus delayed testing. A contrast between assigned prompts supports an estimate of the *total assignment effect* on later task performance when randomized and properly implemented. It cannot automatically decompose that total into enhanced encoding, reactivation, updating, and a change in decision threshold. The distinction between an observer's stated confidence and their ability to differentiate genuinely distinct source relations therefore requires both accuracy and response-bias measurements. Participants should be permitted to inspect records again in a separate condition, since a source-checking strategy can be epistemically preferable to unaided recollection even when the latter appears faster.

4.5 Measurement Designs and Limits of Historical Attribution

An empirical study of relational memory should begin with a documentary corpus whose relations can be scored with explicit uncertainty. Researchers can construct short hypothetical evidence histories with known relationships among artifacts and controlled omissions, then assign participants to presentations that vary provenance visibility, relational complexity, study time, or repeated exposure. The corpus should include genuine ambiguities rather than a fully determined chronology disguised by an arbitrary test. For example, a target relation can be established by two independent records, established by one source only, contradicted by another record, or left unresolved. Participants should be able to respond with calibrated uncertainty when the documents warrant it.

Table 4 sets out a minimum measurement programme that can distinguish several pathways between encountered evidence and a later source or trust judgment. The table deliberately separates mechanisms from their observable proxies, since no single measure establishes the underlying cognitive process uniquely.

The measurement programme has a further dependency: the researcher needs a provenance annotation scheme for the records used as targets. A preregistered coding manual should distinguish documentary statements of authorship, source-to-source citation, temporal precedence, procedural influence, and evaluative claims about competence. Two independent annotators can code a subset of materials, record disagreement, and resolve the scoring standard before outcome assessment. Some disputes will remain substantively unresolved and must be represented in the permissible answer space. Agreement between annotators is informative about reproducibility of the coding operation, while it cannot establish that the archive exhaustively represents the history. The use of wholly fictional or consented archives minimizes the consequences of falsely imputing credit or error to identifiable people.

Measurement order deserves special attention. An explicit source-recall question can direct attention to attribution, change subsequent reading strategies, or encourage searches that would not otherwise occur. Repeated tests can themselves become new learning events. Investigators can address some of these complications through parallel participant groups, counterbalanced task order, delayed measures, and preregistered primary outcomes. They should also report missing responses rather than discarding participants who select "insufficient information," since that option may reflect appropriate epistemic caution. Eye movement or interface logs provide useful evidence of encounter, while immediate comprehension probes and delayed recall assess different states; their combination can constrain explanatory possibilities without directly revealing a unique cognitive mechanism.

A second design would compare narrative presentations with relation-explicit presentations. One group might read a continuous prose account, another a provenance-linked document map, and a third the unmodified documentary archive. These displays could be matched for information content but will rarely be equivalent in processing effort. It is therefore important to define the intervention operationally rather than treating a graph as intrinsically more accurate or cognitively superior. A graph can make documented relations easier to inspect and can also introduce unjustified event boundaries, omit context, or give unsupported causal links an appearance of certainty. A narrative can retain context and can equally encourage an unwarranted linear sequence. Representational appropriateness depends on what the documents establish, which task is being performed, and the cognitive abilities of the population

Table 4: Distinct cognitive outcomes in reconstructing relational evidence, with candidate measures and remaining ambiguities.

Target construct	Candidate measure	Interpretive limitation
Claim or item retention	Old–new recognition and free recall of documentary content	Familiarity and reconstruction may yield the same response
Source attribution	Identify origin, encounter interface, and cited intermediary separately	Available cues and source distinctiveness affect identification
Relational reconstruction	Report documented, unresolved, and inferred links among artifacts	Task priors and imposed response formats can create false certainty
Order uncertainty	Confidence or probability judgments over partial-order relations	Documentary ground truth may itself be incomplete
Source-confidence calibration	Probabilistic accuracy reports scored against verified relations	Aggregate calibration can conceal discrimination and selection
Repeated exposure	Crossed variation in repetition, source cues, and assessment delays	Reading time and testing can alter both attention and retention
Retrieval-induced updating	Delay, reminder assignment, source-specific intrusions across sessions	Recollection reports cannot uniquely identify reconsolidation or a latent trace

examined.

Hypothesis 4.3 (Partial-order preservation under provenance-explicit representation). For tasks requiring reconstruction of documented contribution relations, interfaces that explicitly mark unresolved orderings and source dependencies may reduce confident unsupported ordering relative to otherwise comparable presentations that visually or narratively impose a single sequence. The predicted effect is conditional on comprehension, task familiarity, and the reliability of the displayed source annotations.

Hypothesis 4.3 concerns a testable comparison of presentations, not a general ranking of visual and textual media. A null result may indicate that participants already use source relations successfully, that the manipulation is too weak, or that explicit representations impose cognitive load that offsets their informational benefit. An adverse result would be substantively informative because it would reveal a mismatch between formal transparency and usable cognitive access.

The study’s ethical and interpretive boundaries are similarly important. Experimental archives should use consented or wholly hypothetical documents when evaluating possible misattribution, and should avoid assigning purported mistakes to identifiable real individuals. Historical cases with contested contributions require source criticism, explicit uncertainty about missing records, and respect for the potential consequences of naming particular actors. Participants’ source-memory scores should be interpreted in relation to language, expertise, task design, accessibility accommodations, and available time. A failure to reconstruct a citation chain may reflect an unreasonable task demand or an unreliable interface; it is not inherently evidence of a stable personal deficiency.

The strongest comparison concerns whether cognitive limitations in a documented environment are mitigated by changing the infrastructure of access. An observer may be asked to recognize a complex contribution history unaided, or allowed to consult a stable source-linked record. The latter may reduce avoidable source confusion at the cost of time or navigation effort, and the size and direction of that change will depend on the quality of the archive and interface. In an AI-mediated workflow, citation markers and automatically generated source descriptions require validation against the underlying documents, since plausible descriptions can be erroneous or detached from the supporting passage. External availability should therefore be scored separately from successful independent verification, and both should be distinguished from unaided memory accuracy. These distinctions connect the cognitive analysis to the institutional question of what conditions permit responsible recognition at realistic levels of attention and labor.

A final safeguard concerns the direction of causal interpretation. Interface assignment can modify

selection, encoding, retrieval opportunities, confidence, and later reliance in a connected process. A measured improvement in one source-memory score does not establish that changes in subsequent trust are mediated exclusively by memory: comprehension, perceived legitimacy, and the costs of checking can change concurrently. Studies that attempt to separate these paths need temporal measurement, independently justified mediator assumptions, and sensitivity analyses. Because different observer communities may value content fidelity, contributor recognition, or documentary caution differently, a study should preregister which scores bear on which theoretical claims rather than selecting a favorable aggregate after observing results.

The theoretical contribution of this section is a differentiated account of memory's role in relational recognition. Accurate recall of a claim can coexist with mistaken origin attribution; accurate recognition of a source can coexist with unsupported beliefs about temporal and causal order; and increased familiarity can coexist with unresolved provenance. Equations 16–22, including the composition and source-distinguishability identities, Proposition 4.1, and Hypotheses 4.2 and 4.3 provide complementary ways to specify and investigate these dissociations. They remain conditional models and proposed tests, awaiting direct evidence from relational-documentary recognition tasks. The following section examines the interpretive schemas and contribution-attribution criteria applied to reconstructed evidence, including expertise, language, and motivated reasoning. That analysis is necessary because a correctly retained relational history can still support several defensible or mistaken interpretations depending on the observer's question and background assumptions.

5 Interpretive Schemas and Attribution of Contribution

This section investigates the transformation of an accurately encountered or reconstructed relational history into an observer's interpretation of contribution, competence, responsibility, and evidentiary significance. Its analytical objective is to distinguish differences produced by incomplete memory from differences produced by the criteria through which remembered evidence is understood. A participant may reproduce the content and provenance of a document correctly yet disagree with another participant about whether the document establishes an original method, a competent application, a decisive criticism, or a contribution dependent on earlier work. Such disagreement can arise because evaluators ask different questions, possess different relevant expertise, apply divergent admissibility standards, or weight the same features differently. Some differences are defensible responses to distinct tasks; others are errors relative to a shared evidentiary question and an appropriately justified assessment rule. The section develops a task-conditioned interpretation kernel, a distinction between observed performance and inferred competence, and a formal illustration of why attribution cannot in general be identified from a final rating alone. It then examines motivated interpretation, reputation cues, contribution-role documentation, and cross-disciplinary and linguistic mediation through hypothetical, non-identifiable cases. The proposed tests separate interpretable task differences from errors relative to a stated standard, and separate randomized label effects from claims about their psychological mechanism.

The point of departure is the cognitive architecture introduced in Section 2. There, E designates available evidentiary material, A the portion attended to, M an encoded or retrieved representation, Z its task-relevant interpretation, and J an observable judgment. Section 3 established the importance of distinguishing selective encounter from informational value; Section 4 distinguished remembering content from correctly reconstructing provenance and relational order. The present question begins downstream of those operations: conditional on the same documented evidence and an approximately equivalent memory state, what does the observer regard as established, and by which inferential criteria? An observational comparison of reviewers is only informative about interpretation if differences in access, encounter, and retention have been measured or otherwise constrained. The theory accordingly treats interpretation as a candidate explanatory process rather than as a label for whatever remains unexplained in a judgment discrepancy.

The research problem has a historically situated dimension. Scientific contributions circulate through disciplinary descriptions, abstracts, citations, public archives, peer commentary, and increas-

ingly machine-mediated summaries. A claim generated in one technical vocabulary may be assessed by readers trained in another, sometimes after a representation has compressed its dependencies or translated its terms. Recognition therefore requires more than the recovery of correct names and documents. It may require establishing what a method accomplished, which relations underwrote the result, what alternatives were available, and which evaluative question the observer is entitled to answer. These are interpretive tasks embedded in changing practices of research, documentation, and communication. The section studies cognitive interpretation under specified practices; it leaves the endogenous distribution of authority and the institutional power to impose evaluative criteria to Paper IV.

5.1 Schemas, Expertise, and the Question Being Asked

An interpretive schema is treated here as a provisional representation of task-relevant distinctions, expectations, and inferential relations used by an observer. It is an analytical category whose operationalization may involve elicited explanations, categorization decisions, probabilistic assessments, or performance on controlled transfer tasks. The proposal does not assume a fixed inventory of schemas or require that an individual consciously applies an identifiable rule. Different descriptions of the same history can preserve different features because the observer’s task and background knowledge shape which relations appear meaningful. An expert trained to inspect the validity of a proof may emphasize inferential dependencies that another evaluator, concerned with implementation feasibility, reasonably leaves outside the assigned question. Neither discrepancy establishes incompetence without an appropriate criterion.

Classic experimental work by Chi, Feltovich, and Glaser examined how experts and novices categorized and represented physics problems, finding differences between categorizations based on underlying principles and more literal features in that domain [14]. This result motivates the possibility that domain knowledge influences how an evidentiary history is represented, while its scope remains physics problem solving rather than historical attribution as such. The proposal also draws on broader work concerning structured knowledge and abstraction [76]. Expertise may permit an observer to recognize that two techniques use the same underlying principle despite different surface terminology. It may equally encourage premature analogy when the observer treats a relevant difference as inessential. Greater specialization can therefore improve interpretation on well-matched tasks and fail to improve it on tasks demanding unfamiliar methodological or linguistic distinctions.

Consider a hypothetical history containing a preliminary algorithm, a subsequent implementation, an independent performance comparison, and a later conceptual synthesis. An observer asked who *first documented a specified procedure* should attend to the preserved records and their relation to the procedure’s criteria. An observer asked who *demonstrated a robust implementation* should assess the implementation and experimental standards. A third observer asked which work *explained the underlying mechanism* can reach another answer without contradicting either of the first two. The term “contribution” is underdetermined until the relevant evaluative predicate is identified. Conversely, if each observer is asked the same well-specified priority question and given the same admissible evidence, disagreement becomes a substantive phenomenon requiring examination of interpretation, memory, statistical uncertainty, or prior assumptions.

For a defined task q , denote by S_o a measurable or latent description of the observer’s interpretive state, including relevant domain knowledge, expectations, and task rules, without treating any one psychological taxonomy as exhaustive. A candidate conditional interpretation kernel is introduced in Equation 23.

$$K_o^{\text{int}}(B \mid m, q, s, c) := \Pr(Z \in B \mid M = m, Q = q, S_o = s, C = c), \quad B \in \mathcal{Z}. \quad (23)$$

Equation 23 is a restricted description for protocols in which M captures the trace available at the interpretive stage. The notation separates the task Q from the schema S_o , because a schema that is apt for one question may be misleading for another. The joint law may require conditioning on E , earlier judgments, or a task-specific history when a simplified kernel fails to capture observed dependence. In that event, the conditional description should be enlarged; there is no general warrant for attributing all unexplained

dependence to a latent “schema.” The support of Z can include multiple mutually compatible interpretations, qualified conclusions, or refusal to resolve an underdetermined question. An investigator who forces a unique answer where the source documents warrant uncertainty can create the appearance of cognitive error through the measurement design itself.

A further distinction concerns the *level* at which a schema operates. A reader may use disciplinary expectations to decide which passages to inspect, to organize the passages after encoding, to infer the meaning of a technical claim, or to map an understood claim into a rating category. These operations can generate the same observed association between expertise and final judgments. When a protocol intends to study interpretation specifically, its outcome should be tied to an explicit evidentiary inference, such as the participant’s reconstruction of the conditions under which a method was applicable. A generic global judgment of originality or prestige is insufficiently diagnostic. Short source-linked justification tasks can be helpful, provided investigators recognize that the request for justification may itself induce additional source search and reconsideration.

A schema should also be assessed against the observer’s *assigned question*. A principle-based category learned in one specialty may help identify a shared method yet erase a crucial difference in sampling, inferential scope, or task constraints. Conversely, a surface distinction can be indispensable when an attribution task asks about a specific implementation, language version, or recorded date. The relevant benchmark is thus a task-defined mapping between source-supported distinctions and candidate responses. A mismatch between a schema and that mapping may justify an error classification. A difference between two legitimate tasks or two adequately defended interpretive criteria should be reported as criterion heterogeneity rather than folded into a generic accuracy deficit.

Schema effects can be studied through task and expertise manipulations, provided competence measures do not merely recapitulate the dependent judgment. For example, participants might first complete independently scored exercises on the relevant technical principle and then examine the same hypothetical archive. Randomized task instructions could direct one group to identify the documentary origin and another to evaluate methodological adequacy. A difference between task groups would then reflect a change in the question under conditions controlled by the experiment; a difference between expertise groups would be observational unless expertise itself were meaningfully manipulated. In both cases, accurate recall and source monitoring should be tested separately to determine whether interpretation, rather than differential exposure, supplies the plausible mediating mechanism.

5.2 Attributing Actions, Skills, and Epistemic Contributions

Attribution requires distinguishing several objects of assessment. A record may show that a specified action occurred, that a result was obtained under certain conditions, or that a person contributed to a documented artifact. Those observations do not by themselves determine a person’s enduring competence, responsibility for an entire later outcome, or the value of the work under every relevant conception of scientific achievement. In particular, an observed performance can result from training, collaboration, available instrumentation, favorable conditions, guidance, or chance, as well as from task-relevant skill. Similarly, an original insight can be an important contribution even when its initial implementation remains incomplete. Attribution should be conditioned on the target predicate and on the causal and evidentiary granularity permitted by the documents.

Let R denote an explicitly defined relational account of a hypothetical project and x a particular participant in that account. Define four distinct targets: $T^{\text{act}}(x, R)$ records a documented action; $T^{\text{perf}}(x, R, q)$ evaluates a specified performance under task q ; $T^{\text{comp}}(x, R, q)$ denotes a latent competence relevant to that task; and $T^{\text{contr}}(x, R, q)$ denotes a task-defined contribution assessment. The first two can sometimes be scored directly against records or performance benchmarks. The latter two depend on further modeling decisions, counterfactual comparisons, inferential assumptions, or institutional criteria. The domain and coding of each target should be justified locally. No continuous common scale connecting them is presumed.

A simple descriptive performance decomposition clarifies why competence attribution requires as-

sumptions beyond observation. Equation 24 represents an observed performance response through task-relevant ability and situational conditions, with the response model left unspecified.

$$\Pr(Y = y \mid x, q, c) = \int \Pr(Y = y \mid \theta, u, q, c) P_x(d\theta, du \mid q, c). \quad (24)$$

Here θ is a provisional latent representation of task-specific capability and u represents contextual assistance, collaboration, tool support, opportunity, and other features not summarized by θ . Equation 24 is an identity obtained by marginalizing a stipulated joint distribution, and therefore supplies no identification by itself. Observing one successful result does not uniquely determine the distributions of θ or u . Even repeated performance can leave uncertainty if task variation, selection into trials, or background assistance is unmeasured. The equation is useful insofar as it forces the investigator to describe what would count as evidence for enduring competence, rather than turning an output into an unsupported trait attribution. Its role is complementary to the analysis of source identity in Section 4.2: correct knowledge of who produced a record is a necessary but insufficient basis for many competence claims.

Contribution attribution brings further relational dependence. A paper may incorporate a conceptual distinction from one source, a computational technique from another, and a public dataset maintained by a third group. The question of what each contributor accomplished requires specifying which relations are documented, which are inferred, and how the chosen criterion treats enabling work. An observer who credits only the latest artifact can underdescribe precursor contributions; another who credits every cited precursor equally can ignore differences in role, evidentiary importance, and task relevance. These are potential mistakes with respect to a declared attribution standard, rather than grounds for assigning a single global numerical contribution value to all contexts.

The development of structured contributorship descriptions supplies a practical point of reference. Allen and colleagues described proposals to make diverse research contributions more explicit, and the later Contributor Role Taxonomy (CRediT) distinguishes roles such as conceptualization, software, validation, data curation, and writing [2, 57]. These role labels can improve the inspectability of a collaborative record if attached to reliable, sufficiently specific assertions. Their presence does not establish the relative importance of different roles, the historical priority of an idea, the truth of an individual self-report, or eligibility for authorship under a particular venue’s rules. A role can be shared among contributors; an individual can undertake multiple roles; and a role description may be incomplete or disputed. The analytical value of the taxonomy is its capacity to help formulate questions, not its authority to settle the disputed normative distribution of credit.

For example, one hypothetical group may record that a contributor conceived a research question, while another coded the analysis and a third verified the tests. These role statements become epistemically useful when the record links each assertion to a source, a time interval, the claimed scope of activity, and a means of correction. An auditor can ask whether the statement concerns a completed action, an independently tested output, or a retrospective summary by a project participant. A citation, timestamp, repository commit, experimental log, and signed contribution statement may have different evidentiary capacities. A repository commit can support claims about a recorded code change without automatically revealing the intellectual origin of the design. A verified benchmark can support a performance claim without establishing that a particular person achieved it independently. The relation between recorded events and defensible contribution predicates should remain explicit.

A well-designed task may elicit an assessment vector rather than one “recognition” score. Equation 25 expresses this requirement as an operational measurement choice.

$$\hat{\mathbf{T}}_{o,q}(x) := (\hat{T}^{\text{act}}, \hat{T}^{\text{perf}}, \hat{T}^{\text{comp}}, \hat{T}^{\text{contr}})_{o,q,x}. \quad (25)$$

Equation 25 is not a statement that the components are psychologically independent, cardinally measurable, or normatively commensurable. It is a reporting discipline: distinguish an observer’s estimate of an action, performance, competence, and contribution, and record which components have documentary answer keys. A response that correctly recognizes an individual’s documented action but remains uncertain about individual competence can be epistemically stronger than a uniformly confident response. Later

analysis of belief updating in Section 6 should retain the target proposition attached to each estimate; a favorable change in one component does not imply a favorable change in the others.

For contribution histories with incomplete documentary records, an explicit range of defensible answers can be more informative than a single forced ranking. Let $\mathfrak{R}_q$ be a preregistered collection of admissible, fully stated source-evaluation rubrics for question q , and let $a_r(D; q)$ denote the attribution answer produced by a specified rubric r applied to documentary record D . Equation 26 defines a *rubric-indexed answer set*.

$$\mathcal{A}_q(D; \mathfrak{R}_q) := \{a_r(D; q) : r \in \mathfrak{R}_q\}. \quad (26)$$

Equation 26 is an operational sensitivity object rather than a claim that all rubrics have equal standing or that every answer is correct. The investigator must justify the inclusion and exclusion of rubrics, specify their evidentiary requirements, and disclose changes made after seeing the outcomes. A singleton set indicates agreement only within the declared class and its recorded evidence. A larger set can identify the location of criterion sensitivity, distinguish unresolved source gaps from disputed values, and motivate targeted evidence collection. When documentary support itself is disputed, the study may need multiple plausible record reconstructions in addition to multiple rubrics. The approach preserves the possibility of a justified partial ordering or an abstention while preventing uncertainty from disappearing inside an arbitrary mean score.

The distinction also protects against retrospective certainty. If a record demonstrates that a collaborative project succeeded, allocating success to particular persons can require information about division of labor, absent alternatives, and conditions of possible substitution. Documentary citations and interpersonal testimony may help reconstruct those relations but do not always resolve them. A careful analysis may therefore yield a justified partial description: an action is documented, a performance benchmark is satisfied, competence remains uncertain, and the relative contribution of collaborators is underdetermined. The framework treats such an outcome as a substantive result of inquiry.

The inferential boundary becomes particularly important when production is collaborative and mediated by software or generative systems. Evaluation can ask whether a person specified objectives, chose between candidate designs, wrote a verified implementation, curated data, tested assumptions, or merely possessed an output. An output artifact may provide weak information about the allocation of these activities. A model-assisted workflow might make some acts unusually visible through interaction logs and leave others, such as framing questions, rejecting poor solutions, or recognizing errors, largely invisible. Treating an inferred share of labor as a directly observed property reproduces the very evidentiary omission the trajectory framework aims to study. Conversely, recording every software interaction as a separate intellectual contribution could inflate attribution by mistaking activity traces for evidentially meaningful work. The appropriate level of event granularity depends on the question and a transparent theory of relevance.

Attribution also has a temporal dimension. The appearance of a later and more polished artifact does not by itself identify which earlier work was necessary, substitutable, or causally decisive. Those are counterfactual and historical questions that demand additional assumptions. Acknowledging upstream influence, identifying priority, describing indispensable enabling infrastructure, and assigning a share of present reward are distinct operations. The cognitive experiment proposed here can test whether participants conflate them. The political and institutional choice of which operation determines reward is reserved for Paper IV.

5.3 Belief-Consistent Interpretation and Epistemic Bias

Interpretive differences may also reflect prior beliefs, group affiliations, incentives, and goals. Psychological research has long examined how expectations can affect the selection and interpretation of information. Tversky and Kahneman describe several heuristics that influence judgment under uncertainty [79]; Kunda’s analysis develops the case for motivated reasoning [48]. Lord, Ross, and Lepper report a controlled study in which participants with opposing prior attitudes evaluated purportedly mixed evidence differently [54]; Nickerson reviews a wider range of confirmation-bias phenomena [58]. These

literatures justify considering asymmetric interpretation as a possible mechanism. Their existence does not mean that any two observers who disagree are biased, or that prior belief is intrinsically an illegitimate evidentiary input.

Prior information may be relevant. A reviewer might correctly know that a particular technique requires special instrumentation; a novice may misinterpret a routine operation as evidence of exceptional skill. A reviewer might also possess justified reasons to question a source’s reliability. Bias becomes a researchable hypothesis when interpretation differs in a manner inconsistent with the relevant evidentiary standards or when the weighting of equivalent evidence depends on an experimentally manipulated, task-irrelevant cue. The benchmark must be specified rather than assumed. When a historical record is genuinely incomplete, divergent posterior beliefs may be rational consequences of different but defensible priors. An empirical study must therefore distinguish justified domain knowledge, preference-dependent standards, and belief-consistent distortion.

Laboratory studies of expectation can help motivate suitable contrasts while remaining sensitive to task differences. Darley and Gross experimentally varied social-class information and exposure to subsequent performance evidence when participants assessed a child’s academic abilities, observing expectation-consistent interpretation in the specified setting [15]. Their result neither identifies a universal response to social labels nor establishes that a laboratory school judgment generalizes to credit for research artifacts. Bordalo and colleagues offer a formal account of how representativeness can distort group judgments, with context-sensitive reference groups [10]. The two accounts suggest complementary possibilities: a cue can affect which interpretation is constructed from evidence, and the comparative categories an observer uses can change the salience of an otherwise unchanged cue. In a scholarly-attribution experiment, such possibilities require direct measurement of label comprehension, target-question understanding, and whether the cue supplies independently relevant information.

A participant who reports uncertainty when evidence is genuinely insufficient should not be scored as biased merely for declining a forced dichotomy. Likewise, a status cue can sometimes be a legitimate, though imperfect, proxy for features of source quality when direct assessment is costly. Whether its use is defensible depends on the question, availability of stronger evidence, and costs of false inference. To isolate a status effect on otherwise equivalent documentary merits, it is preferable to use fictional identities and non-identifying descriptions with assigned cues known to be orthogonal to the evidence in the experimental design. The intervention’s effect then pertains to those constructed cues and materials; claims about actual institutions require additional research.

One approach is to ask whether a change of an extraneous label affects an observer’s interpretation while the documentary basis and declared assessment target are held fixed. Let $L \in \{\ell_0, \ell_1\}$ designate two presentation labels, such as alternative non-identifying institutional descriptions, assigned independently of documentary merits. Let $Z(\ell)$ denote the task-specific interpretation a participant would report under label ℓ in a well-defined experimental intervention. Equation 27 defines an average label-conditioned contrast for a bounded coding function g .

$$\tau_{q,g}^{\text{label}} := \mathbb{E}[g\{Z(\ell_1)\} - g\{Z(\ell_0)\} \mid Q = q]. \quad (27)$$

Equation 27 is a potential-outcome estimand for a specified population and randomized protocol, not an identified effect from ordinary retrospective ratings. Even with random assignment, interpretation of $\tau_{q,g}^{\text{label}}$ requires a stable treatment, compliance assumptions, a coding rule that captures the intended aspect of interpretation, and attention to communication between participants. A nonzero contrast can reveal label sensitivity without establishing that the effect is inappropriate. A label that conveys valid information about a source’s evidentiary record may rationally alter an assessment. To investigate unjustified status-driven interpretation, the experimental task must make the label irrelevant or independently control its legitimate informational role.

The same distinction applies to language and disciplinary markers. If a document is marked as belonging to an unfamiliar field, the observer might discount it because the method is genuinely less applicable to the task, because the observer lacks interpretive competence, or because of an unsupported expectation about that field. Those explanations can yield similar ratings. A counterbalanced design

could measure prior familiarity, provide a transparent technical glossary, and hold the underlying evidentiary claim constant. It might then inspect whether judgments track actual methodological relevance, recognition of provenance, or assigned surface labels. In such a design, expertise and stereotype-based associations remain distinct candidate variables; they should not be collapsed into a single “prestige” coefficient.

5.4 Reputation Cues and Cognitive Attribution

Research recognition occurs in contexts where evaluators may know the reputation of people, laboratories, journals, or institutions. Merton’s analysis of the Matthew effect considers the interaction of scientific rewards and communication with the prominence of established scientists [56]. This is a historically situated account of scientific recognition, not a parameter estimate for the present cognitive model. Reputation may influence the likelihood that a document is encountered, the credibility provisionally assigned to it, the interpretation of its technical quality, or the final allocation of credit. These possibilities sit at different points in the observer pipeline. The present paper examines how a cue is cognitively used when encountered; a separate analysis of who controls exposure, valuation rules, and reward distributions belongs to Paper IV.

A controlled study of single- and double-blind conference review by Tomkins, Zhang, and Heavlin found advantages associated with author and institutional prominence under the particular WSDM reviewing arrangement [78]. That study provides more directly relevant evidence of review-condition sensitivity than experiments on generic social stereotypes. It does not establish that every individual review decision was biased or that the effect arose exclusively from interpretation after reading: author information was available at bidding, assignment, and evaluation stages in the experimental setting. The distinction is central. A change in final recommendation could reflect selective assignment or exposure, interpretation of scientific substance, judgment thresholds, or combinations of these mechanisms.

Three measurement contrasts help separate these possibilities. First, access and allocation can be recorded before the paper is read, so that bidding and review participation are not conflated with within-document interpretation. Second, a randomized label or identity disclosure can be introduced only *after* an identical evidence encounter, with attention and recall measured before disclosure; the resulting change estimates a late-stage intervention under that protocol, although asking for repeated ratings may itself alter responses. Third, independent measures can elicit what conclusion the participant believed the methods supported, whether that interpretation relied on the documentary content, and which threshold determined a recommendation. This third contrast enables a richer diagnostic account but does not in itself identify latent mediation. Conditioning directly on post-treatment comprehension or recall can create selection problems when label disclosure also affects those variables. A prespecified temporal design and mediation assumptions are required for causal claims about the intervening stages.

The distinction between *evidentially warranted credibility* and *reputational substitution* also requires a declared target. Some decisions legitimately use external reliability data when primary documents are incomplete; others explicitly ask whether an archived contribution occurred, for which a prestige label is irrelevant once the relevant provenance is verified. A participant may reach an accurate answer using an inappropriate shortcut, or an uncertain answer after carefully examining a genuine ambiguity. Outcome accuracy, quality of justification, and sensitivity to irrelevant status information should therefore be reported separately.

Proposition 5.1 (Observational equivalence of interpretive pathways). *Suppose a study observes only a final binary judgment J and finds $\Pr(J = 1) = p$ for some $0 < p < 1$. Without additional observations or restrictions on the interpretation and judgment kernels, the study cannot determine from this marginal law alone whether response variation originated in interpretation Z or in the mapping of an unchanged interpretation into the reported judgment.*

Proof. Consider one model with a binary interpretation $Z \sim \text{Bernoulli}(p)$ and a deterministic judgment $J = Z$. Construct a second model with constant $Z = 0$ and a judgment kernel $\Pr(J = 1 \mid Z = 0) =$

p . Both produce the same observed distribution of J while allocating its variation to different stages. Therefore the marginal judgment law alone does not identify the interpretive pathway. The construction is a mathematical counterexample; its purpose is to demonstrate non-identification under the stipulated observation regime, rather than to claim that real observers randomly generate judgments independently of meaning. $\square$

Proposition 5.1 extends the earlier non-identification result of Section 2.5 to a different interface. It warns against diagnosing motivated cognition or inadequate expertise on the basis of final scores alone. A direct interpretation measure, such as an explanation of the evidentiary link being asserted, can help distinguish mechanisms, but eliciting explanations may itself change task performance. Better identification calls for multiple measurement modes, interventions that isolate stage-relevant features, and explicit sensitivity analysis for the possibility that explanations are rationalizations produced after the judgment.

5.5 Cross-Language and Interdisciplinary Interpretation

Relational evidence often moves between language communities and technical specialties. A preserved document can become newly accessible through translation, yet the translation may change which causal, temporal, or normative distinctions are available to an observer. A sentence stating that a procedure *adapted* an earlier method differs from one claiming that it *introduced* the method, and the difference matters for attribution. Similarly, descriptions of collaboration can distinguish providing data, designing an instrument, formulating a hypothesis, and validating an outcome. Translating them into a single broad term such as “developed” may create an appearance of equivalent contribution that the source records do not warrant. The issue is semantic and inferential as well as lexical: disciplinary concepts have locally established criteria that may resist one-to-one substitution.

Brislin’s study of translation and back-translation investigated factors affecting translation quality and criteria for equivalence across languages and content conditions [12]. Later scholarship on translation and interpreting emphasizes the role of contextual assumptions and situated professional judgment [61]. Neither source implies that a single back-translation check guarantees scientific equivalence, and neither supplies a direct estimate of error in AI-mediated scholarly summaries. They motivate the need to specify what equivalence is required for a particular attribution task. If the task asks who presented the first documented procedure, terminological differences relevant to what counts as “the same procedure” deserve explicit scrutiny. If the task asks whether two descriptions are compatible at a broader level of abstraction, a less granular translation may be sufficient.

The accessibility stakes of translation are also empirical. In biodiversity-conservation literature, Amano and colleagues identified a substantial share of relevant documents written in languages other than English, illustrating the possibility that monolingual search can omit pertinent evidence [3]. Their estimates concern the sampled conservation literature and do not quantify a universal distribution across scientific fields. Steigerwald and colleagues examine potential benefits and limitations of machine translation for multilingual scholarship [75]. These contributions motivate attention to a stage preceding interpretation: if records in one language never become accessible or encountered, an otherwise accurate translation policy cannot recover what the observer did not locate. This section addresses semantic and inferential preservation conditional on access; Sections 2 and 3 supply the broader encounter architecture.

Let D be the source-language documentary expression, V_λ a representation delivered under translation or disciplinary mediation protocol λ , and $T_q(D)$ a task-relevant answer determined by source criticism under an explicit adjudication rule. A practical measure of task-conditioned interpretive preservation compares answers supported by the source and by the mediated representation. Equation 28 defines a corresponding expected loss.

$$\mathcal{R}_q(\lambda) := \mathbb{E}[\ell_q\{d_q(V_\lambda), T_q(D)\}], \quad (28)$$

where d_q is an assessment rule applied to the mediated text and ℓ_q is a declared scoring loss. Equation 28 is an evaluation framework, not a model of the translator’s intentions or a claim that there is always one

uniquely correct target-language equivalent. For ambiguous documentary questions, $T_q(D)$ may be a set of acceptable answers or an uncertainty-preserving target. The scoring rule must accommodate those cases. Differences in $\mathcal{R}_q(\lambda)$ can reflect translation content, format, evaluator expertise, or downstream interpretation; designs that seek to isolate translation should hold the latter features constant or measure them separately.

A hypothetical experiment could supply two language versions of a short archive with explicit provenance annotations, then compare participants proficient in both relevant languages with participants relying on one translated version. Critical terms would concern whether a contribution was independently originated, modified from a documented predecessor, or merely described retrospectively. Independent subject-matter adjudicators could identify where translations alter evidence relevant to the question, while blinded evaluators record attribution responses and confidence. Additional conditions might provide bilingual quotations, linked primary records, or an explanatory glossary. Differences between conditions could then be interpreted relative to a specified equivalence target and participant populations. The design must avoid assuming that bilingual participants always represent an unbiased criterion or that machine translation is inherently inferior or superior to human translation. Disagreements among qualified translators may expose genuine semantic indeterminacy rather than a correct-versus-incorrect dichotomy.

AI-generated abstracts, interactive search answers, and automated translations add another layer of mediation. The relevant analytical question concerns which documented relations the representation exposes and which inferences it licenses. A summary stating that a group “created” a tool may suppress whether it adapted a public implementation, whether the method was specified earlier by another contributor, or whether the reported achievement was a replication. This risk is analytically distinct from whether the summary is fluent, and it can arise in human-written as well as machine-generated prose. Without an evaluation dataset, the present paper makes no estimate of the prevalence of such distortions or the relative performance of particular systems.

For comparative testing, protocols could randomize between an unmediated short archive, a faithful provenance-linked summary, and a summary with an explicitly controlled omission of one evidential relation. The omitted relation should be relevant to one prespecified attribution predicate but immaterial to another. Such a design tests whether the representation changes task-sensitive judgments in the expected pattern. Ideally, the same participants would not be exposed to mutually revealing versions before the outcome measure, and source-criticism adjudicators would record residual disagreement. The investigator should also test whether visibly marked uncertainty, clickable quotations, or correction links change interpretation, confidence, and willingness to check the primary documents. A successful interface would support inspectable inference and corrigible attribution; neither high confidence nor agreement alone suffices.

Cross-disciplinary interpretation raises analogous issues even when the source and target share a language. A model regarded as a theoretical contribution in one field may be treated as a routine adaptation in another, based on different histories and purposes of use. Some disagreement can be resolved by specifying the claim and locating earlier documented work; some remains a difference in warranted evaluative standards. The investigator should record which standards were stated before assessment and which were introduced only afterward. This helps distinguish an observer’s inability to recognize an applicable method from legitimate disagreement about the aims of an inquiry. It also avoids assigning a single field’s evaluative vocabulary the status of a universal epistemic ontology.

5.6 Operational Comparisons and Limits of Interpretation

The proposed mechanisms require a study design that separates at least four stages: comprehension of the relevant documentary content, retrieval of its provenance, interpretation of its relevance to a named question, and the subsequent reported judgment. Table 5 records potential indicators and the associated non-identification problems. The table is not a fixed taxonomy of observer types. Its categories describe separable analytic tasks under a study protocol and may be revised when the empirical materials demand

Table 5: Task-defined interpretation and attribution measures, with competing explanations requiring separate tests.

Assessment object	Candidate observable	Important alternative explanation
Schema and domain expertise	Principle-based sorting; scored transfer tasks; elicited explanations	Expertise may alter attention and encoding before interpretation
Documented action and performance	Identification of source-supported actions and criterion-based outputs	Missing records or varying performance conditions can limit scoring
Inferred competence	Confidence across standardized task repetitions and assistance disclosures	Successful output may reflect collaboration, tools, selection or opportunity
Contribution attribution	Separate estimates of origin, methodological role, validation and enabling work	Different legitimate contribution predicates can give different answers
Belief-consistent interpretation	Responses to randomized, substantively irrelevant labels	Labels may carry legitimate information unless controlled
Cross-language interpretation	Task-specific equivalence judgments and reconstruction of translated provenance	Semantic ambiguity and evaluator proficiency can confound the comparison

finer distinctions.

A factorial study might cross the presence of an explicit provenance relation with task instructions and a non-diagnostic disciplinary label, while balancing reading time and document content. Outcome measures could include source identification, recall of the relevant evidence, ability to state the assessment criterion, calibrated uncertainty, and separate ratings for action, performance, competence, and contribution. A delayed phase might examine whether participants can revise their interpretation when a primary document or bilingual explanation becomes available. To reduce demand effects, investigators may compare direct explanation questions with indirect transfer tasks and preregister which outcomes are primary. No single method reveals the latent schema uniquely: even explanation protocols can elicit retrospective constructions that differ from the processes governing the initial judgment.

For the experimental design to support a causal interpretation, random assignment needs a precisely declared intervention and observation window. A factorial protocol can vary source-linked role descriptions, status labels, and mediation format, but high-order interactions will require sample sizes and hypotheses defined in advance. Investigators should separately record source exposure, time spent on relevant passages, delayed recall, task-specific paraphrases, attribution judgments, uncertainty or abstention, and response changes after verification. Expertise groups should be stratified using assessments that are independent of the dependent archive task. Role descriptions introduced during a treatment must be audited for truthfulness so that a manipulation of legibility is not confounded with a manipulation of the underlying facts.

Several common analyses require restraint. Regressing a final judgment on measured comprehension after randomization does not automatically estimate the effect of the label “through interpretation,” because comprehension can itself respond to the treatment and share unmeasured determinants with the final rating. Counterbalancing source order addresses order effects only under assumptions about carryover; it cannot erase learning from previous trials. A design that gives participants primary records on request must distinguish treatment effects on *opportunity to verify* from self-selected verification behavior. Preregistered measurement models, manipulation checks, reasonable sensitivity analyses, and anonymized hypothetical archives can improve reproducibility without transforming latent processes into directly observed facts.

Hypothesis 5.2 (Task-specific benefit and possible cost of expertise). Given equivalent documentary access and adequate comprehension, expertise aligned with the assessed methodological principle may improve principle-relevant attribution relative to novice interpretation. When the task depends on prove-

nance distinctions outside that expertise, the advantage may shrink or reverse. The prediction depends on task validity, measured knowledge and source quality.

Hypothesis 5.2 makes expertise a conditional explanatory variable rather than a status ranking. It may be rejected where the materials are transparent without specialist knowledge or where the experts' prior framework is poorly matched to the actual contribution history. A design capable of observing such failures should avoid recruiting experts only for tasks on which their advantage is assumed in advance.

Hypothesis 5.3 (Interpretive benefit of provenance-preserving mediation). For tasks that depend on distinctions among originating, adapting and summarizing a method, a translation or interdisciplinary summary that makes those distinctions inspectable may reduce unsupported conflation compared with an otherwise similar representation that erases them. The direction and size of the effect depend on terminology, study time, interface complexity and observer competence.

Hypothesis 5.3 concerns specified representation contrasts rather than an unrestricted claim that adding context improves cognition. More context can increase processing demands or introduce misleading interpretations; an explicit annotation can itself be inaccurate. Consequently, the investigator should retain access to underlying records and allow an evaluator to contest the proposed mapping. The experiment should also distinguish improvement in documentary attribution from greater confidence, since the latter can increase even when interpretive accuracy deteriorates.

Hypothesis 5.4 (Stage-specific sensitivity to status cues). When an experimentally irrelevant reputation cue is disclosed only after a fixed, independently scored interpretation task, its effect on a subsequent global recommendation may differ from its effect when disclosed before free source selection. Differences across these protocols may indicate the importance of encounter timing or decision thresholds, but do not alone identify a unique psychological mediator. The direction and magnitude of the contrast are empirical questions.

Hypothesis 5.4 requires random allocation, common documentary merits, equivalent scoring opportunities, and checks on whether the timing intervention changes what participants read or remember. Its value lies in separating possible points of intervention, including cases in which status information has no detectable effect.

The interpretive framework has limits. An observer's reconstruction of the facts can be accurate while the normative importance of a contribution remains legitimately contested. Competence can be difficult to infer when opportunities to perform are unequal. Translation can preserve task-critical evidence without preserving every rhetorical or cultural nuance, and its adequacy cannot be evaluated independently of the question it is supposed to support. Motivated reasoning can be present, absent, or opposed by countervailing incentives in specific settings. These limits require investigators to avoid turning every unresolved judgment into a psychological diagnosis or every interpretive difference into an institutional injustice. The fairness and legitimacy of historically imposed evaluation standards are indispensable topics, but their endogenous determination belongs principally to Paper IV.

The section's contribution is therefore a set of distinctions and testable interfaces, rather than a completed explanation of recognition. It separates retained evidence from interpreted relevance, observed performance from inferred competence, and document-grounded contribution from broader judgments of personal merit. It also specifies how prior schemas, motivations, and translation protocols may alter assessments while leaving the underlying documentary history constant. Equations 23–28, together with the rubric sensitivity construction in Equation 26, Proposition 5.1, and Hypotheses 5.2, 5.3, and 5.4, are conditional tools for investigating these distinctions. The next section examines how interpreted evidence can change belief or confidence, separating realized surprise and evidentiary diagnosticity from the assumptions required for Bayesian and bounded updating. That transition matters because a participant may interpret the same document correctly yet revise a belief differently when prior probabilities, error costs, or inferential constraints differ.

6 Belief Revision Under Surprisal and Diagnostic Evidence

This section investigates what happens after an observer has encountered, reconstructed, and interpreted evidence bearing on a relational contribution history. Its task is to specify conditions under which an observation warrants a change in belief, and to distinguish such a warranted change from the numerical or verbal response a person actually gives. The preceding sections treated the allocation of attention, the reconstruction of provenance, and task-conditioned interpretation. Each can alter the material supplied to an updating process. The present inquiry asks how that material bears on a definite proposition, how an observer might integrate it with prior uncertainty, and when confidence becomes disproportionate to evidentiary support. Its principal distinction is between *surprisal about an observation* and *diagnostic evidence concerning a specified hypothesis*. A rare relational encounter may produce strong curiosity but little justification for revising an attribution, while an ordinary archival entry may be decisive when competing hypotheses predict it differently. This distinction follows the logic of information theory and probabilistic inference, while the psychological question concerns how real observers approximate or depart from that logic [71, 28, 79].

The section retains the programme’s provisional account of events and trajectories. A document, an act of verification, a translation, and an account of authorship can occupy different positions in a task-specific representation; their evidentiary roles depend on the proposition under examination. The same documentary item may support an attribution claim, weaken a competence claim, and remain irrelevant to an unrelated claim. Consequently, an updating rule must identify what is being assessed. Neither novelty nor persuasive intensity can serve as a universal measure of trust. Human confidence in a proposition, confidence in a source, and readiness to rely on a person are distinct constructs. Here the formal target is a proposition-specific assessment; the detailed evaluation of testimony belongs to Section 7, and the relation between epistemic judgment and behavioral reliance belongs to Section 8.

6.1 Reference-Conditioned Surprise and Evidentiary Diagnosticity

Suppose $\Theta \in \{0, 1\}$ denotes a specified claim relevant to an observer’s task. For example, $\Theta = 1$ might mean that a particular method was documented independently in an earlier archival record. Let C be a recorded context containing the task definition, accessible prior information, and the observation protocol. Let Y denote an interpreted evidentiary observation, such as the content of a dated record together with the provenance information that the task permits the observer to examine. For the moment, take these as random variables in a finite or countable reference model; extending the construction to more complex spaces requires corresponding regular conditional distributions and technical assumptions. The variable Y is deliberately indexed by the observation protocol. A summary alone and an inspected original record need not instantiate the same observation.

Surprisal measures an observation’s unexpectedness relative to a predictive distribution. Diagnosticity measures its ability to differentiate hypotheses. In Equation 29, p_o denotes the model representing observer o ’s expectations; probabilities refer to the specified context and measurement procedure. The right-hand quantity is a signed log-likelihood ratio, with positive values favoring $\Theta = 1$ and negative values favoring $\Theta = 0$.

$$\begin{aligned} s_o(y | c) &:= -\log p_o(Y = y | C = c), \\ \lambda_o(y | c) &:= \log \frac{p_o(Y = y | \Theta = 1, C = c)}{p_o(Y = y | \Theta = 0, C = c)}. \end{aligned} \tag{29}$$

The ratio assumes strictly positive probabilities for the displayed observation under both hypotheses; zero probabilities can be addressed through extended-valued likelihood ratios when the model justifies them. The unconditional predictive probability in the first line is a mixture over possible hypotheses. A high value of s_o therefore need not imply a high absolute value of λ_o . The latter also depends on the alternative under consideration. Changing the alternative can alter diagnosticity while leaving the visible document unchanged. The subscript o indicates that empirical observers may possess distinct

predictive models; a defensible analytical benchmark can instead employ a separately specified p^* , not an unexamined claim about which participant’s prior is correct.

A simple numerical construction illustrates the separation. Consider two hypothetical hypotheses that each assign probability 0.001 to a rare form of annotation. An observer who encounters the annotation obtains approximately 6.91 nats of surprisal, yet its likelihood ratio is one and its log-likelihood ratio is zero. The annotation is unusual under the mixture but has no capacity to distinguish these two hypotheses. In a second hypothetical case, a commonplace registry entry is expected under the first hypothesis with probability 0.8 and under the second with probability 0.2. Its log-likelihood ratio is $\log 4$, giving a clear direction of evidentiary support despite the absence of unusual visual or semantic content. These are stipulated likelihoods offered to demonstrate logical possibilities; they are not estimates of actual archival systems or cognitive behavior.

Entropy adds an expectation over the specified predictive distribution and thus answers another question. Shannon’s formulation gives a probability-weighted measure of uncertainty before the realization of a particular symbol or observation [71]. An observer might face high entropy concerning which evidence item will appear, encounter one highly surprising item, and still learn virtually nothing about the authorship proposition. Conversely, a low-entropy observation channel can be extremely informative about a binary claim if its outcomes discriminate the relevant alternatives. The use of information-theoretic quantities here is conditional on an explicit model of the observation space; a change in the classification of events changes the predictive distribution and may change the numerical surprisal. This dependence is particularly important in our treatment of evolving relational histories, whose admissible descriptions are provisional rather than fixed event types.

The distinction also prevents an attractive but unsupported extrapolation from Section 3. An unusual encounter between disciplines might increase the probability that a document receives attention. That is a selection effect concerning which evidence becomes cognitively available. The encounter’s contribution to justified belief change requires an additional account of the inspected evidence, its alternatives, and its provenance. If the unusual encounter only repeats a previously examined assertion, it need not add evidentiary weight. If it reveals a primary record that was previously absent from the observation protocol, it may change the available likelihood comparison. Attention and diagnosticity can therefore interact without becoming identical variables.

Observation, search, and the conditioning set. The apparent separation of surprisal from diagnosticity can fail to be empirically operational if the analyst neglects how an item entered the dossier. Let R denote the retrieval or display policy, D an indicator that a target record was discovered, and Y its content when discovered. The meaningful benchmark for an investigator who knows the policy is $p^*(D, Y_{\text{obs}} \mid \Theta, C, R)$, where Y_{obs} includes an explicit missing value when the item is not retrieved. An observer who is shown only the retrieved content may instead use a working distribution $p_o(Y \mid \Theta, C, D = 1)$ and overlook what selection on D implies. These are distinct inference problems even if both observers read precisely the same displayed sentence. A search-engine ranking, institutional filtering procedure, or AI-generated synthesis changes the opportunities for encounter, and hence the observation likelihood, without changing the underlying primary record. Fiedler’s sampling analysis offers a relevant, established account of judgment biases induced by the selective samples available to the observer [22]. It does not establish the magnitude of these biases in any particular research registry.

It is useful to separate the *event of retrieval* from the *meaning attributed to retrieved content*. A record may be retrievable yet uninterpretable because its language, encoding, or terminology lies outside the observer’s competence. Conversely, a faithful summary may be cognitively accessible while concealing which records were searched unsuccessfully. In either case, estimated likelihoods require an observation protocol rather than a bare document description. This is an interface between the structural account of legibility in Paper II and the cognitive account here, rather than a replacement of either with a universal numerical measure of evidence.

6.2 Bayesian Updating as a Conditional Normative Benchmark

Under a well-specified probabilistic model with positive prior mass on both alternatives, updating a binary proposition by an observed y can be represented through posterior odds. Equation 30 states the ordinary Bayesian identity, with the same conditioning context and evidence model used on both sides.

$$\log \frac{p_o(\Theta = 1 \mid y, c)}{p_o(\Theta = 0 \mid y, c)} = \log \frac{p_o(\Theta = 1 \mid c)}{p_o(\Theta = 0 \mid c)} + \lambda_o(y \mid c). \quad (30)$$

This is an identity within the stipulated model. It does not demonstrate that real cognition implements such an operation, that the participant possesses accurate likelihoods, or that a single prior is normatively obligatory across heterogeneous tasks. Priors must be open to scrutiny: a prior inferred from field prevalence, a prior supplied experimentally, and an observer's lived expectations represent different evidentiary situations. Furthermore, historical records are subject to selection, omission, and delayed discovery. The likelihood relevant to the observed item must account for the procedure through which it became observable, including any selection that depends on the hypothesis. An analyst cannot substitute the frequency of a document in a database for the probability of its encounter under a task-dependent retrieval protocol without further justification.

The direction of the Bayesian update follows the likelihood ratio. A value $\lambda_o > 0$ increases posterior odds for the first hypothesis, a value $\lambda_o < 0$ decreases them, and $\lambda_o = 0$ leaves the prior odds unchanged. This permits positive, negative, and null responses without treating trust as a quantity that must increase when access expands. The identities concern changes in a proposition-specific posterior. They do not imply that all later consequences, including source evaluation and practical cooperation, inherit that direction. A primary document may increase confidence in an archival date while simultaneously decreasing confidence in an earlier commentator's attribution. Different targets can therefore change in different directions under the same evidence encounter.

Proposition 6.1 (Equal likelihoods yield a null evidentiary update). *Let $0 < p(\Theta = 1 \mid c) < 1$ and suppose an observed y has equal positive conditional probability under $\Theta = 1$ and $\Theta = 0$, given the same context c and observation protocol. Then $p(\Theta = 1 \mid y, c) = p(\Theta = 1 \mid c)$, independently of the predictive surprisal of y . If the likelihood under $\Theta = 1$ is strictly larger (respectively, smaller), the corresponding posterior probability is strictly larger (respectively, smaller) than the prior.*

Proof. Bayes' rule multiplies the prior odds by the ratio of the two conditional likelihoods. Equal likelihoods produce a ratio of one; strictly larger or smaller likelihoods give a ratio respectively above or below one. With interior prior probabilities, the mapping from odds to probability is strictly increasing. Predictive surprisal does not enter this comparison independently of the likelihoods. $\square$

Proposition 6.1 recovers, in a proposition-specific updating form, the independence result discussed in Section 3. It is mathematically elementary; its intended role is to discipline the paper's conceptual claims, not to present Bayes' theorem as a new discovery. In particular, it exposes the extra assumptions required to infer epistemic improvement from a structural encounter. The encounter must yield evidence that distinguishes at least some relevant alternatives, and the observer's model must represent that distinction sufficiently well. A null update can be correct where the available evidence is nondiagnostic. An increase in confidence can be incorrect where the model mistakes a copied claim for independent corroboration.

Evidence arrives under observation protocols that may impose dependencies. Suppose two displayed endorsements Y_1 and Y_2 derive from the same source. Even if each endorsement is individually diagnostic when compared with an alternative, the second may add little conditional information after the first. The sequential Bayesian benchmark in Equation 31 makes the conditioning explicit.

$$\log \frac{p_o(\Theta = 1 \mid y_1, y_2, c)}{p_o(\Theta = 0 \mid y_1, y_2, c)} = \log \frac{p_o(\Theta = 1 \mid c)}{p_o(\Theta = 0 \mid c)} + \lambda_o(y_1 \mid c) + \lambda_o(y_2 \mid y_1, c). \quad (31)$$

Here the second increment is the conditional log-likelihood ratio for Y_2 given Y_1 . Writing two identical unconditional increments would require an independence assumption that may be false in citation chains,

repeated summaries, or shared archival provenance. If the second item is fully determined by the first in the stipulated measurement model, its conditional increment is zero wherever the conditional likelihoods coincide. Conversely, a second item may supply significant diagnostic information when it independently verifies the content or reveals a limitation missing from the first. This is a formal issue of conditional information rather than a presumption that multiple reports are necessarily redundant. Section 7 examines the additional psychological and social problems of source vigilance in testimony chains.

It is possible for a later observation to reverse a previous direction of updating. Under Equation 31, the first increment may be positive while the second is negative and larger in magnitude. The resulting posterior can fall below the original prior. Such a path is coherent when the second observation is more diagnostic of the alternative hypothesis after conditioning on the first. Relational histories are particularly susceptible to this possibility because a newly discovered primary record may recontextualize an earlier secondary account. The claim concerns a permitted probabilistic trajectory under a specified protocol, rather than an ontological premise that historical time proceeds in a universal sequence. The indices 1 and 2 identify the order of observation in the experiment, not a globally fixed order of social events.

Conditional dependence and the meaning of corroboration. A compact generative representation illustrates how apparent multiplicity can arise from one origin. Let S be a latent source state that influences two circulated reports. For a source-mediated model, the joint likelihood may be written

$$p(y_1, y_2 \mid \theta, c) = \sum_s p(s \mid \theta, c) p(y_1, y_2 \mid s, \theta, c). \quad (32)$$

This factorization does not assume conditional independence of the reports given S ; that additional assumption would need to be defended. Nor does the latent variable have to describe a human actor. In an AI-mediated citation chain it might distinguish a primary record, a syndicated extract, and derivative summaries, insofar as those categories are actually observable or modeled. If the second report is an exact transformation of the first under both hypotheses, then $p(Y_2 \mid Y_1, \Theta = 1, c)$ and $p(Y_2 \mid Y_1, \Theta = 0, c)$ coincide and its conditional log-likelihood contribution is zero. Adding it as an independent likelihood factor artificially multiplies the evidentiary strength of the original record. When a second archival copy provides genuinely independent metadata, however, it can improve the conditional likelihood comparison. The relevant question concerns the dependence structure, not the number of visible records.

A further qualification concerns the order in which an observer sees evidence. For a fixed joint distribution over Y_1, Y_2 and a fixed initial context, the final Bayesian posterior given both observations is invariant to which observation is processed first: both orderings represent the same joint likelihood by the chain rule. In the psychological process, however, the *encounter* sequence can change comprehension, memory, and even the hypotheses under consideration. Hogarth and Einhorn’s belief-adjustment account characterizes circumstances in which order and response mode matter to reported judgments [32]. A measured order effect cannot by itself show a violation of probability axioms: it may instead signal that the effective information encoded in the two conditions differs. To test order-dependent updating, an experiment must establish which records were comprehended and retained, while recognizing that testing comprehension can itself alter processing.

Model adequacy and unresolved alternatives. Bayesian coherence is an internal property of a specified model. If the proposed hypotheses omit a credible alternative, a perfectly coherent calculation can still give unwarranted confidence. A dossier that frames a contribution dispute as a choice between two named claimants may be incomplete if another contributor, an earlier independent development, or an institutional process generated the work. Model criticism therefore precedes any strong claim about the right magnitude of an update. Gelman and Shalizi emphasize model checking and revision as essential to applied Bayesian inference [24]. Within this manuscript, the appropriate implication is modest: a single posterior number should remain revisable in light of errors in the source process, the representation of alternatives, or the data-generating assumptions. The paper does not assume that an observer can enumerate all future hypotheses in advance.

6.3 Information Gain, Decision Value, and the Strength of Evidence

An observer may learn from evidence without being able to apply it to a consequential decision, and a decision may improve without a conspicuous change in reported confidence. To connect the updating benchmark with information theory, fix a task context c and a specified observation design π , including any selection or display mechanism. For a discrete target and evidence variable, Equation 33 defines expected posterior information gain as a conditional expectation of a divergence between posterior and prior distributions.

$$\mathcal{I}(\Theta; Y \mid c, \pi) := \mathbb{E}_{Y|c, \pi} [D_{\text{KL}}(p(\Theta \mid Y, c, \pi) \parallel p(\Theta \mid c, \pi))] . \quad (33)$$

This is conditional mutual information for the chosen joint distribution when regularity and finiteness conditions hold. The divergence is nonnegative as a formal property [47]; the expectation depends on what is counted as an observation and on how alternatives are represented. Equation 33 extends the Section 3 discussion from evidence-seeking decisions to realized belief revision under a fixed protocol. Its nonnegativity gives no guarantee that a particular observed proposition's posterior probability increases. It also does not guarantee that humans who encounter the evidence will integrate it correctly. The expected reduction in uncertainty and the subjective experience of novelty are separately measurable quantities.

Expected information gain can be related to expected utility when an experimental goal is specified [8]. That connection requires identifying the target of inference, the permissible decisions or assessments, and a utility or scoring rule; it does not supply a general measure of human value or social justice. An experiment may yield information about a minor factual distinction while consuming substantial attention needed for a more consequential judgment. A source registry may offer relatively little entropy reduction because the alternatives are already constrained, yet matter greatly if it corrects a consequential misattribution. Thus, maximizing Shannon information about an arbitrarily chosen parameter is not necessarily equivalent to improving a historically meaningful contribution assessment. The broader question of how institutions distribute the consequences of such assessments is reserved principally for Paper IV.

Griffin and Tversky distinguish strength or extremeness of evidence from its weight or credence, and present evidence of patterns in human confidence that can reflect insufficient sensitivity to weight [28]. Applied cautiously to relational histories, the distinction is useful for understanding why a vivid testimonial statement might receive more weight than its underlying source supports, while a mundane dated record may be discounted because it lacks rhetorical force. Their experiments concern specified judgment tasks, however. They do not establish that every observer, scientific field, or culturally situated evaluation displays the same underweighting. Furthermore, in our setting the weight of evidence can depend on its provenance and the protocol that made it accessible. A faithful translation, an independently archived version, and several automated rewritings of one source may have different conditional likelihood structures even when their surface claims resemble one another.

Consider three hypothetical research dossiers. In the first, a highly unusual collaboration produces a striking graph that is independent of the question of who originated a method. The graph may raise attention and perceived novelty but has a likelihood ratio of one for the attribution hypothesis. In the second, a routine archived message includes dated derivations and verified source metadata, producing a likelihood comparison that favors an earlier contributor. The observer may justifiably revise the attribution despite the record's ordinary presentation. In the third, a seemingly corroborative summary is shown to depend entirely on the initial claim rather than on a separate observation. Correctly incorporating source dependence can reduce confidence relative to a naive interpretation that counted two independent sources. None of these cases requires personally identifiable examples, imputations of misconduct, or conclusions about the ultimate moral worth of contributors.

The distinction between evidence strength and weight also matters for judgments about the absence of expected records. Failure to retrieve a record is informative only if the protocol would probably have retrieved it had the relevant hypothesis been true. An absent response from an incomplete archive, an inaccessible repository, or a failed translation system may reveal little about original authorship. If the

observation mechanism has high sensitivity and documented coverage, absence can have a nontrivial likelihood ratio. Such comparisons require modeling what the observer had a real opportunity to see. They also require recognizing that the observation process itself may evolve as search tools, classifications, or language resources change, which is a central linkage back to Paper II.

When an absent record is evidence. Negative evidence is sometimes confused with merely missing evidence. Suppose a specified search policy π would discover a particular class of record with probability q_1 under $\Theta = 1$ and q_0 under $\Theta = 0$, after conditioning on the same context and database coverage. If $D = 1$ denotes a successful discovery, the absence of a record within that protocol has the conditional log-likelihood ratio

$$\lambda_{\neg D}(c, \pi) := \log \frac{p(D = 0 \mid \Theta = 1, c, \pi)}{p(D = 0 \mid \Theta = 0, c, \pi)} = \log \frac{1 - q_1}{1 - q_0}, \quad 0 \leq q_0, q_1 < 1. \quad (34)$$

This identity follows immediately from the definitions of q_1 and q_0 . When $q_1 = q_0$, unsuccessful retrieval is nondiagnostic. When $q_1 > q_0$, the absence of the expected record lowers the odds of $\Theta = 1$; when $q_1 < q_0$, it raises those odds. These are consequences of stipulated detection probabilities, not assertions that an absence of documentation demonstrates an absence of contribution. Search sensitivity and specificity may be unknown, and selective preservation, archival access rules, naming conventions, translation quality, and repository decay can affect them. Reporting the search policy and uncertainty about its coverage is accordingly part of responsible inference.

There is a second, distinct interpretation of *negative evidence*: an observed record directly supports the alternative hypothesis. That observation has its own signed log-likelihood ratio and need not resemble a failed search. An explicit contradiction in a dated record, a missing document after a high-coverage search, and an inability to read a source are three different observations. Collapsing them into a single category of “no support” can mask selection biases and produce unwarranted certainty. The appropriate comparison is between models that predict different chances of these observations, not between a favored narrative and an undifferentiated absence of confirming examples.

Decision consequences and asymmetric losses. The relevance of information to an action depends on consequences in addition to posterior confidence. For a narrowly defined binary decision, suppose the available actions are to accept or withhold an attribution claim, with false-positive loss $c_{FP} > 0$, false-negative loss $c_{FN} > 0$, and zero loss for correct decisions. A risk-minimizing benchmark accepts the claim when

$$p(\Theta = 1 \mid \mathcal{E}, c) > \frac{c_{FP}}{c_{FP} + c_{FN}}, \quad (35)$$

with indifference at equality, where $\mathcal{E}$ denotes the evidence included in the declared protocol. This stylized threshold makes explicit why identical warranted credences can support different permissible decisions under different loss specifications. It does *not* define justice through a single numerical penalty: the distribution of losses, who has authority to set them, and the availability of contestation are normative and institutional questions beyond this narrow decision exercise. In particular, an irreversible public accusation and a reversible request for additional documentation should not be assigned the same costs without argument.

For completeness, the expected value of sample information under a cost-free, correctly specified decision model may be represented as the difference between optimal expected loss without and with an offered observation:

$$\begin{aligned} \text{EVSI}(Y \mid c, \pi) &:= \min_a \mathbb{E}[\ell(a, \Theta) \mid c, \pi] \\ &\quad - \mathbb{E}_{Y \mid c, \pi} \left[\min_a \mathbb{E}[\ell(a, \Theta) \mid Y, c, \pi] \right] \geq 0. \end{aligned} \quad (36)$$

The inequality holds because the decision maker could ignore Y and use the initial decision. It assumes an identical admissible action set before and after observation, well-defined expectations and optimization,

and no cost of obtaining or processing the observation. If searching, verification, exposure, or cognitive effort carries cost, *net* value need not be positive. Nor does the inequality show that actual observers use the optimal rule. It distinguishes an informational resource under a model from a guaranteed benefit of circulating additional records.

6.4 Bounded Updating and Calibration of Expressed Confidence

Bayesian identities provide a benchmark for coherent probability assessments given specified inputs. Descriptive psychology asks whether human observers possess those inputs, approximate the computations, or use different strategies altogether. Phillips and Edwards studied conservatism in a simple probability inference task [62]; other work on heuristics and motivated reasoning documents systematic ways in which judgment can diverge from normative reference models in particular contexts [79, 48, 58]. These literatures should not be collapsed into an assertion that people invariably underreact to evidence. Patterns can depend on the elicitation method, source credibility, task difficulty, prior commitments, and availability of feedback. The theoretical task is to specify alternative descriptive hypotheses that can be compared using evidence rather than to explain every discrepancy with one unspecified bias parameter.

One restricted descriptive family is given in Equation 37. Let $L_{o,k}$ be an elicited log-odds judgment at protocol step k ; let $\hat{\lambda}_{o,k}$ be an investigator-defined or separately elicited estimate of the diagnostic increment accessible to observer o ; and let $\varepsilon_{o,k}$ represent residual measurement and process variation.

$$L_{o,k+1} - L_{o,k} = \alpha_{o,k} \hat{\lambda}_{o,k} + b_{o,k} + \varepsilon_{o,k}. \quad (37)$$

The sensitivity parameter $\alpha_{o,k}$ may be less than one, equal to one, or greater than one, depending on the observed task. The term $b_{o,k}$ can represent systematic deviations associated with framing, prior commitments, or source-label effects in a specified design. It is not an independently identified latent psychological mechanism merely because it appears in the equation. In particular, an underestimated likelihood ratio and a low α can produce similar responses. Identification requires independent interventions or measurements, and any model comparison should report this ambiguity. The family can also fail altogether when responses are categorical, observers revise their hypothesis space, or interpretation changes the meaning of the target proposition during examination.

This model distinguishes epistemic norms from descriptive states. A reported confidence score need not equal an observer's entire credence state, and an observer need not represent uncertainty numerically at all. Linguistically expressed certainty, a point probability elicited with incentives, and an internal disposition to rely on an answer may respond differently to social demands. An institutional assessment rubric may also prescribe a confidence threshold for a decision without representing a participant's belief about the proposition's truth. Thus, the same outward score can reflect different cognitive and practical processes. The non-identification arguments developed in Sections 2 and 5 remain applicable to the updating stage and caution against labeling residual variance as "irrationality" without stronger evidence.

Calibration offers an external check when repeated, objectively scorable outcomes are available. If an observer makes probabilistic predictions $\hat{p}_i$ about binary outcomes $z_i \in \{0, 1\}$, one proper scoring rule is the mean squared probability error or Brier score in Equation 38, historically introduced for probability forecasts [11].

$$BS_o := \frac{1}{n} \sum_{i=1}^n (\hat{p}_{o,i} - z_i)^2. \quad (38)$$

A lower expected Brier score indicates better probabilistic forecasting under the scoring rule's assumptions. Calibration in the narrower reliability sense additionally concerns the agreement between forecast probabilities and empirical outcome frequencies within adequately populated comparable groups. Accuracy under a scoring rule, reliability calibration, and precision of discrimination are related but distinct properties. A lower Brier score does not alone identify which internal mechanism improved, and calibration cannot be inferred from a single statement of certainty. For questions of originality or responsibility,

a gold-standard binary outcome may be unavailable or contestable, so assigning z_i requires publicly defensible evidence and careful specification of the task. It would be methodologically unsound to treat institutional majority opinion as an infallible truth label.

The psychology of confidence is especially relevant where social circulation makes assertions familiar. Repeated presentation may increase fluency or perceived credibility without adding independent evidence, while subsequent verification might justify either a higher or lower posterior. Section 4 already distinguished memory confidence from source accuracy; the present section adds that confidence can also be miscalibrated with respect to a target proposition. Griffin and Tversky’s experiments provide one established route from evidence characteristics to confidence errors [28]. The proposed relational application is an extension requiring its own validation. The later study of epistemic vigilance will address how an observer assesses a communicated source and distinguishes repeated assertions from independent testimony.

Descriptive asymmetry, observer variation, and measurement equivalence. A common but incomplete empirical practice is to compare the observed change in log odds with an analyst’s benchmark and label every smaller change “conservatism.” Suppose one group reads the original document while another sees a translated summary with uncertain provenance. The nominal evidence package can have one analyst-defined likelihood ratio yet induce different *subjective* likelihoods. An observed $\alpha_{o,k} < 1$ in Equation 37 could then reflect limited updating, less credible evidence as the observer understood it, or an imperfect mapping between reported scores and internal belief. The data do not uniquely identify these explanations. An experiment that elicits perceived evidence quality before and after exposure can constrain some interpretations, but it may also prime participants to reason differently. Multiple measurement regimes and preregistered sensitivity analyses are preferable to treating a fitted coefficient as a psychological essence.

A further distinction separates asymmetry that is normatively warranted from asymmetry caused by selective evaluation. It may be reasonable to scrutinize highly consequential allegations more carefully than routine claims when evidentiary standards and loss asymmetries are explicit. It can also be a bias to subject disconfirming evidence to substantially more skeptical standards solely because it challenges a favored position. Lord, Ross, and Lepper’s study of attitude polarization illustrates that opposing groups can evaluate purported evidence differently [54]; it does not supply effect-size estimates for contribution disputes or prove that all conflicting assessments reflect motivated cognition. An observer may properly change an earlier belief in either direction. Evaluation should therefore separate the observer’s prior commitment, comprehension of the methodological quality of each source, the conditional likelihood model, and the consequences of error.

Calibration, discrimination, and the limits of outcome verification. The Brier score in Equation 38 is a *proper* forecasting loss for binary events, with a unique expected minimum at the forecaster’s true probability under the scoring model. In particular, if an outcome has probability p and a reported forecast is q , its expected squared loss decomposes as

$$\mathbb{E}_{Z \sim \text{Bernoulli}(p)}(q - Z)^2 = p(1 - p) + (q - p)^2. \quad (39)$$

The decomposition explains propriety and also its limitation: repeated scoring presupposes an interpretable outcome distribution and a defensible realization of the target variable. The general theory of strictly proper scoring rules supports this separation between eliciting probabilities and examining their predictive quality [25]. One should report calibration curves with bin-size uncertainty, discrimination or resolution measures, and aggregate scores rather than infer a fully calibrated observer from an improved single summary metric. In some archival controversies, an observer can articulate uncertainty responsibly even though the ground truth required for scoring is inaccessible. Such cases call for procedural assessment of evidence quality and careful adjudication, not the manufacture of binary labels to simplify evaluation.

Table 6: Operational distinctions for testing evidence-driven belief revision. The contrasts describe possible experimental manipulations or measurements rather than established effects.

Construct	Example operationalization	Principal inferential caveat
Surprisal	Pre-elicited or model-based predictive rarity of a displayed item	Depends on an explicit reference distribution and event description
Diagnosticity	Stipulated or independently estimated likelihood contrast for a target claim	Can be distorted by hidden selection and an unsuitable alternative
Belief revision	Within-observer change in elicited proposition-specific probability	Rating scales and response demand need not reveal internal credence
Source dependence	Shared versus independently generated evidentiary records	Apparent multiplicity can conceal a common informational origin
Calibration	Repeated predictions compared with validated binary outcomes	Contested historical labels may lack a defensible ground truth
Updating cost	Inspection time, switching effort, and experimentally varied workload	More effort can reflect difficulty, conscientiousness, or both

6.5 Empirical Comparisons and the Limits of Trust Inference

An empirical test of belief revision must separate the variation in evidence that participants receive from the variation in their interpretations and responses. A useful controlled protocol begins with a specified binary or multiclass claim, a documented set of admissible hypotheses, and an elicited initial belief. Participants are randomly assigned to receive evidence packages varied independently on at least three dimensions: surface rarity or visual surprise, task-specific likelihood diagnosticity, and provenance dependence. The content of some packages can be held constant while its visual presentation changes; in complementary conditions, the display format can be matched while the evidentiary likelihood comparison changes. The investigator measures actual inspection, comprehension, posterior probability reports, and delayed reconstruction. Such a design avoids assuming that randomized display assignment guarantees that the same material was cognitively received by every observer.

Table 6 emphasizes that the same apparent result can have several explanations. Consider participants who revise toward an attribution after viewing a provenance table. Some may have understood the dated source, some may have followed an authority cue embedded in the table header, and some may have guessed the experimenter's preferred answer. Adding a direct provenance-comprehension test and a source-label manipulation can distinguish part of this variation, but those additions create their own possible demand effects. Repeated questioning can change what a participant remembers or attends to. A design therefore needs an explicit observation schedule and, where feasible, groups in which cognitive-state measurements are withheld until after the main judgment. Comparing immediate and delayed assessments can provide evidence about stability, while imposing no assumption that delayed judgments necessarily reflect more mature or more accurate thought.

Identification strategy and negative controls. The proposed manipulations require an explicit factorial structure. One factor varies visual or narrative salience while holding underlying diagnostic content fixed; another varies diagnostic content with presentation characteristics matched; a third makes source dependence more or less transparent while keeping the primary evidence record constant. Assignment effects on reported posterior judgments can be estimated under successful randomization and appropriate handling of missing outcomes. However, the mechanism “salience changed attention, which changed belief” is not identified merely by observing both an attention measure and a subsequent judgment. Attention may be affected by comprehension, prior conviction, fatigue, and source familiarity. A neutral control record whose likelihood ratio is specified to equal one can test whether a condition drives nonspecific confidence shifts, though a null response under the model is not guaranteed empirically. A second control may vary cosmetic duplication of one statement without changing the informational source, testing whether repeated exposure is mistaken for corroboration.

The study should distinguish a *target-specific posterior* from confidence in the investigator, willingness to share the material, or institutional endorsement. If the investigator seeks effects under real-world retrieval policies, an archive audit and a sampling design must supplement laboratory vignettes. Comparing observers across linguistic or disciplinary communities requires tests of response-scale equivalence before treating raw differences as differences in epistemic quality. The appropriate preregistration reports which priors and likelihood assumptions are stipulated by the task, which are elicited from observers, and which are estimated by an analyst. When these levels disagree, that disagreement is substantive data rather than noise to be silently discarded.

Hypothesis 6.2 (Diagnostic evidence under matched surprise). Under a task with reliably understood hypotheses, matched evidence-access conditions, and independently validated likelihood contrasts, manipulations increasing the diagnosticity of evidence should predict larger directionally appropriate changes in proposition-specific beliefs than manipulations increasing surface surprise alone. The comparison is conditional on comprehension and may be weakened by attention failure, prior saturation, or response noise.

Hypothesis 6.3 (Source-dependent repetition and apparent confidence). When participants receive multiple mutually consistent statements with experimentally established common provenance, a condition that obscures that dependence may elicit greater confidence than a condition that makes it legible. The predicted difference can vanish or reverse when observers independently verify the source relation, distrust the display, or use evidence-quality criteria that appropriately discount repetition.

Hypothesis 6.4 (Processing order under a fixed evidence set). Where the underlying pair of evidentiary items, task proposition, and stipulated joint likelihood model are held fixed, changing presentation order or stepwise versus end-of-sequence reporting may alter expressed intermediate or final confidence. Such effects should be evaluated jointly with source recall and item comprehension rather than taken as direct proof that the observer violated a Bayesian identity.

These hypotheses are falsifiable through specified contrasts. For the first, an experiment can cross high versus low surface salience with high versus low likelihood diagnosticity while balancing display time and measuring inspection. It should report effects on prior-to-posterior change separately from effects on recall and from final response accuracy. For the second, the underlying primary observations can be kept identical while one interface transparently links summaries to their common origin and another uses unlinked attributions. Researchers should avoid creating misleading real-world claims; hypothetical records with clearly disclosed experimental status can achieve the design goal. The difference between experimental conditions should be interpreted as an effect of the display and provenance information jointly supplied, unless the protocol separates them.

A further comparison concerns the direction of rational updating. To distinguish warranted distrust from an indiscriminate tendency to reject evidence, investigators may include accurate counterevidence under a source protocol known in advance. Participants who reduce confidence in a poorly supported claim can be responding appropriately, even if a simple score of “trust” falls. At the same time, reductions in self-reported trust cannot be equated with improved calibration without an externally defensible criterion and measurements across multiple tasks. The contribution of belief revision to justified recognition must therefore be evaluated relative to the explicit proposition and evidentiary alternatives. Decisions to collaborate, share resources, or accept institutional authority involve further losses, incentives, and social relations that Section 8 examines separately.

The identification problems extend beyond experimental demand. An observer’s priors can be difficult to elicit without affecting them; likelihood estimates may reflect learned expertise or an unobserved task model; and source selection can depend on both the hypothesis and the information-processing strategy. Attrition from long evidence tasks may be associated with confidence, cognitive load, or disagreement with the experiment. Causal effects estimated among participants who complete all examinations may therefore differ from effects among those originally assigned to the protocol. Empirical researchers should preregister operational definitions, retain an intention-to-treat contrast where appropriate, and

2699 treat mechanistic mediation through attention or comprehension as requiring further assumptions. No
2700 single design resolves the difference between accessible, attended, interpreted, and appropriately inte-
2701 grated evidence.

2702 A normative evaluation of an individual judgment must also allow for rational disagreement. Two
2703 observers with different but defensible prior evidence may retain different posteriors after encountering
2704 the same record. Their disagreement may further concern the hypothesis partition or uncertainty about the
2705 retrieval process. Conversely, agreement in final reported confidence need not indicate equal recognition
2706 of evidence: one observer may have updated substantially from a skeptical prior, another barely at all
2707 from an already favorable position. For this reason, a serious study of cognitive recognition should
2708 preserve elicited pre-observation states, uncertainty in the assessment model, and where ethically and
2709 practically feasible, the sources available before the experimental encounter. Observer heterogeneity is
2710 a condition for analysis, not an error term to be eliminated by assumption.

2711 The section's analytical result is a qualified architecture for belief revision. It distinguishes
2712 probability-weighted uncertainty, realized surprisal, diagnostic likelihood evidence, posterior change,
2713 decision value, expressed confidence, and empirical calibration. The formal relations provide bench-
2714 marks whose meaning depends on their conditioning structure, and the descriptive model supplies
2715 testable but potentially non-identifiable departures. Positive, null, negative, and reversing belief
2716 changes are permissible outcomes of inquiry; none is intrinsically more trustworthy without a specified
2717 claim and evidence basis. These considerations explain why the same novel relational encounter can
2718 leave justified assessment unchanged, correct an earlier favorable judgment, or support a previously
2719 overlooked contribution. Section 7 now examines the credibility and independence of communicated
2720 testimony, which must be evaluated before evidence-bearing statements can be integrated into epistemic
2721 trust judgments.

2722 7 Epistemic Vigilance and Testimony Evaluation

2723 This section examines how a human observer evaluates evidence that arrives through another person's re-
2724 port, a mediated document, or a sequence of communications involving automated systems. The research
2725 objective is to identify cognitive conditions under which testimony can warrant a proposition-specific re-
2726 vision, and conditions under which apparent corroboration reflects the circulation of an existing assertion.
2727 Earlier sections separated evidentiary access from attention, memory, interpretation, and belief revision.
2728 The present section locates testimony within that architecture by distinguishing scrutiny of a communi-
2729 cated *claim*, assessment of its *communicative source*, and reconstruction of its *evidentiary lineage*. The
2730 argument combines the philosophy and psychology of epistemic vigilance with a conditional probabili-
2731 ty representation and hypothetical, non-identifiable examples. It closes with measurement proposals,
2732 interpretive limits, and three falsifiable hypotheses. The cognitive mechanisms remain provisional; the
2733 mathematical identities hold only for their explicitly stated probability models.

2734 Testimony is indispensable to inquiry conducted across disciplinary, linguistic, geographical, and
2735 institutional boundaries. An individual cannot personally reproduce every experiment, inspect every
2736 instrument, or establish each element of a complex historical record. Nevertheless, an encountered state-
2737 ment can have several distinct origins. It may report direct observation, describe another person's record,
2738 offer an inference from established measurements, summarize several independent investigations, or re-
2739 produce an earlier assertion. The recognition of a relational history therefore requires a way to distinguish
2740 *communicative plurality* from *evidentiary plurality*. The distinction is particularly important when the
2741 observer has limited time and an item seems credible because it arrives through a familiar institution, an
2742 apparently knowledgeable intermediary, or several mutually agreeing channels. Sperber and colleagues'
2743 account of epistemic vigilance identifies scrutiny directed toward both speakers and communicated con-
2744 tents, while Goldman addresses the practical difficulty of assessing expert testimony from a position of
2745 limited expertise [73, 26]. The framework proposed here specifies how these inquiries depend on source
2746 lineage and conditional informational contribution; neither source establishes the proposed representa-
2747 tion as a comprehensive theory of trust.

7.1 Content Scrutiny and Source Assessment

A communicated claim can be evaluated along at least three analytically different dimensions. First, *content scrutiny* concerns the coherence, specificity, evidentiary implications, and independently checkable features of the proposition expressed. Second, *source assessment* concerns the observer's reasons for treating the communicator as informed, competent in the relevant domain, and disposed to transmit information accurately in the particular task. Third, *lineage assessment* concerns the relation between the communicated claim and the observations, records, and earlier communications from which it arose. A source can accurately transmit an unsupported claim, an expert can misunderstand evidence outside the domain of expertise, and an automated system can faithfully summarize a document whose original claims remain unverified. These possibilities create different tasks for cognitive vigilance.

For purposes of this section, *competence* denotes task-relevant ability to obtain, understand, and transmit information; *integrity* denotes the consistency of relevant conduct with communicative commitments such as accurate attribution and disclosure; and *credibility* denotes an observer's proposition- and context-dependent assessment of a particular source or report. These are working analytical distinctions. Organizational-trust research provides related but non-identical dimensions of ability, benevolence, and integrity [55]. Credibility must therefore be specified relative to a communication task and claim rather than inferred from a person's general status or favorable reputation. The present analysis also separates evidence concerning *a source's reliability in a task* from evidence concerning *the target proposition itself*. A reliable messenger can report authentic negative evidence, while the repetition of a reassuring message can remain poorly supported.

These distinctions also separate what a communicator *knows*, what a communicator has an opportunity to observe, what a communicator *says*, and what the recipient *takes the statement to mean*. In practical inquiry, a source may possess reliable data yet communicate a poorly qualified conclusion; another may offer an accurate account of a predecessor's testimony without being able to validate its original measurements. A recipient may misunderstand a cautious proposition as categorical, particularly after translation or summarization. Treating these stages as one source-quality score obscures the very errors that an explanation of testimonial evaluation should be able to locate. Reliability is indexed by a claim, task, period, informational access, communication form, and potential contestation procedure. General prestige can enter an observer's prior as a cue, but it neither substitutes for the evidence-generation mechanism nor establishes that a particular inference is warranted.

The observer's own capacities belong in the analysis. Expert disagreement may be opaque to a novice who cannot distinguish a substantive methodological objection from a terminological difference. A reader may have access to a citation but lack the language or technical preparation to adjudicate what it supports. The extended epistemic-vigilance framework investigated in science education explicitly includes attention to the receiver, alongside scrutiny of source and claim; its pilot findings pertain to the studied students and instructional context, not a validated cognitive mechanism across all populations [9]. In the present proposal, receiver capacities enter as task-conditioned moderators of whether access becomes comprehension and whether an apparently successful check actually tests the relevant proposition.

Let H be the proposition under consideration, S a source, R the report observed from that source, and C the observer's specified task and background conditions. For a simplified binary content model, Equation 40 introduces a latent state T representing whether a source's relevant informational basis correctly tracks the state of H in a particular reporting circumstance.

$$P(R \mid H, S, C) = \sum_{t \in \{0,1\}} P(R \mid T = t, H, S, C)P(T = t \mid H, S, C). \quad (40)$$

This is an application of the law of total probability, rather than an estimated psychological law. The latent variable T is a modeling convenience whose definition must be tailored to the particular reporting mechanism. It can describe source access to a valid instrument record, successful quotation of an archive, or accurate transmission from a prior witness. Source competence may affect components of the conditional distribution, while communication incentives, interpretation, and memory errors may affect other

components. Equation 40 is meaningful only when these variables can be specified and the conditionals are coherent. It should not be interpreted as assigning people a stable numerical truthfulness trait or as warranting the extraction of such a trait from sensitive historical data.

A useful hypothetical contrast concerns two communications about the priority of a research method. One reports an accessible dated protocol with an archive identifier and the specific passage on which its claim rests. Another offers an equally confident evaluation based on the communicator’s professional standing but supplies no inspectable documentary basis. Expertise can remain relevant to interpreting the technical relation between methods, but the first communication gives the observer a different route to assessing historical priority. If the archival identifier is fabricated, the first message may be misleading despite its detailed appearance. If the second source has documented direct access to a missing original record, its testimony may be consequential even when immediate documentary inspection is impossible. These cases motivate contextual assessment rather than a simple rule that document links outweigh expertise or that experts outweigh documentary evidence.

Epistemic vigilance can also be socially costly. Verification takes time, language proficiency, access permissions, and sometimes specialized instruments. Requesting a source for a statement may be necessary for a high-stakes attribution while disproportionately burdening an unfamiliar contributor if established actors are exempted. A procedurally consistent examination standard is therefore relevant to the ethics of evaluation. The observer should also distinguish uncertainty from misconduct: incomplete provenance can result from historical losses, translation constraints, changing documentation practices, or justified confidentiality. A missing link alone does not identify deception. At this level, the paper analyzes human assessment under constrained evidence; the governance of institutional disclosure obligations and power over access is primarily reserved for Paper IV.

A comparable procedural burden can still have unequal effects. If one evaluator demands raw primary records from a newcomer while accepting a senior colleague’s undocumented assertions, any later difference in recorded credibility may partly reflect the verification regime itself. Conversely, a uniform requirement to publicly release confidential field notes may exclude research conducted under legitimate privacy obligations. An empirical account should therefore record which kinds of evidence were feasible to supply, the effort and risk of inspection, and the procedural options for qualified private review. Such observations are important for explaining recognition without converting incomplete documentation into presumed moral deficiency. They also distinguish the psychological acceptance of an authority cue from the justificatory adequacy of the underlying evidence.

7.2 Independent Origins and Repeated Assertions

The number of encountered reports provides a poor measure of evidentiary independence unless their origins and dependence structure are known. Two publications may rely on the same field notebook, two witnesses may have discussed an event before reporting it, and a series of online descriptions may trace to one uncertain source. Conversely, apparently similar accounts may arise from independent observations. Section 6 introduced a proposition-specific likelihood benchmark, and Section 4 described difficulties of remembering the sources of familiar statements. Those earlier results combine here: the observer must reconstruct the dependency relation among reports before determining how much additional evidence a new report supplies.

Let $R_1, \dots, R_n$ denote received reports and C all relevant background conditions. For positive conditional probabilities, the sequential likelihood factorization in Equation 41 makes the role of dependence explicit.

$$\frac{P(R_{1:n} \mid H, C)}{P(R_{1:n} \mid \neg H, C)} = \prod_{j=1}^n \frac{P(R_j \mid R_{1:j-1}, H, C)}{P(R_j \mid R_{1:j-1}, \neg H, C)}. \quad (41)$$

The equality is a chain-rule identity. An independence-based product of marginal report likelihoods follows only under suitable conditional-independence assumptions for the actual report-generating process, separately under H and its alternative. Such conditions cannot be inferred from the number of apparent authors, publishing organizations, or user-interface cards. Dawid’s formulation of conditional

independence supplies the statistical background for these distinctions [16]. The evidential interpretation further requires appropriate modeling of which reports were selected for display, including omissions and shared access constraints. When an observer encounters only reports that a ranking system has selected, the likelihoods in Equation 41 should represent that selection regime or explicitly condition upon it.

Dependence can be represented by an unobserved common origin rather than a deterministic copying relation. Let O denote a hypothesized origin structure, such as one archive repeatedly quoted, two interviews derived from a shared briefing, or several independent laboratory records. The origin-state space $\mathcal{O}$ must be specified for the inquiry. By marginalizing over uncertain structures, one obtains

$$P(R_{1:n} \mid H, C) = \sum_{o \in \mathcal{O}} P(R_{1:n} \mid O = o, H, C) P(O = o \mid H, C). \quad (42)$$

Here the mixing distribution may itself depend on H and the environment. Conditional independence within a selected origin model, if assumed, does not generally imply independence after integrating over O . Agreement can be generated by shared dependence even when communicators did not coordinate with one another. In addition, independent origin does not entail reliable measurement: independently produced reports can share calibration errors, common definitions, or systematic incentives. The investigator should distinguish documentary lineage, statistical independence conditional on an explicit model, and independence of substantive error mechanisms.

A further qualification concerns *selection into the observer's evidence*. Let $V = 1$ mean that a particular report configuration is presented by a ranking, retrieval, or publication mechanism. For events with positive relevant probabilities, Bayes' rule gives a selection-conditioned likelihood ratio:

$$\frac{P(R_{1:n} \mid H, V = 1, C)}{P(R_{1:n} \mid \neg H, V = 1, C)} = \frac{P(R_{1:n} \mid H, C)}{P(R_{1:n} \mid \neg H, C)} \frac{P(V = 1 \mid R_{1:n}, H, C)}{P(V = 1 \mid R_{1:n}, \neg H, C)} \frac{P(V = 1 \mid \neg H, C)}{P(V = 1 \mid H, C)}. \quad (43)$$

This identity does not assert that every interface produces bias. It shows where the sampling or display regime enters the evidentiary calculation. An observer may see five agreeing reports because the world produced five independent observations, because one claim was retold five times, or because an interface preferentially surfaced agreeing material. Differentiating these explanations requires information beyond the visible count. Even exact knowledge of an interface's ranking rule may be insufficient if the distribution of unobserved records is unknown.

The distinction is especially important for evolving relational histories. Historical archives, bibliographies, and online indices are produced through labor-intensive processes with heterogeneous resources. Translation, transcription, metadata maintenance, and institutional preservation affect which records become searchable. An apparent absence of testimony about a contributor or method may therefore reflect archival loss or index coverage, as well as actual lack of supporting evidence. Section 6 treated informative non-retrieval conditional on search sensitivity; the present extension emphasizes that the sensitivity itself can vary across language, institution, period, and document type. These are observable research dimensions, subject to identification limits, rather than grounds for imputing a missing contribution.

A simple limiting case demonstrates the problem with treating echoed statements as corroboration. Suppose a second report is a deterministic transformation of the first, $R_2 = f(R_1)$, and that transformation is identical across the competing hypotheses once R_1 is known. A second copy can vary in formatting or wording without adding an independent observation. The qualified result appears below.

Proposition 7.1 (Deterministic retelling has no incremental likelihood evidence). *Let H and $\neg H$ be hypotheses with positive probability, and let $R_2 = f(R_1)$ almost surely under both hypotheses conditional on the same background information C . For every jointly possible (r_1, r_2) with $r_2 = f(r_1)$ and positive likelihood under both hypotheses, the joint likelihood ratio of (R_1, R_2) equals that of R_1 alone.*

Proof. Conditional on $R_1 = r_1$ and C , the probability of $R_2 = f(r_1)$ equals one under both H and $\neg H$. The conditional likelihood ratio contributed by R_2 therefore equals one. Applying Equation 41 for two reports gives the joint likelihood ratio as the first report's likelihood ratio multiplied by one. The result depends on observing the first report and knowing the transformation relation; it does not describe a situation where the retelling alone is the accessible artifact. $\square$

The proposition has a narrow but useful consequence. If a research blog quotes a paper and a news article repeats the blog’s quotation, neither repetition provides a new independent test of the paper’s factual assertion solely by reproducing it. Nevertheless, those derivative records can convey useful evidence about communication history, dissemination, or which interpretations circulated. Their informational relevance changes when the target proposition changes. For example, the news article may be direct evidence that a claim was publicly communicated at a specified occasion, even if it adds nothing to the truth of that claim. The model therefore requires target-relative evaluations rather than globally labelling a source as redundant.

Partial dependence creates a richer problem. A report may preserve an original assertion while adding a new observation, independent measurement, error, editorial interpretation, or contradictory quotation. A summary can introduce a previously unavailable document identifier or disclose a previously hidden conflict of interest. Such elements may contribute incremental evidence conditional on earlier reports. The appropriate question is whether a new report contains information that distinguishes H from its alternatives after the earlier reports have been accounted for. Equation 44 expresses this point for one additional report.

$$\ell_{j|<j} = \log \frac{P(R_j | R_{1:j-1}, H, C)}{P(R_j | R_{1:j-1}, \neg H, C)}. \quad (44)$$

A positive, zero, or negative value of $\ell_{j|<j}$ can arise according to the report’s direction and its conditional reliability. It is a model-dependent log likelihood ratio, not an observable intrinsic feature of a document. An observer who has not seen the original source may find a faithful summary informative, while an observer who has already examined the original may obtain no new information concerning the same proposition. The latter case requires familiarity with the dependency relation, which real observers may fail to reconstruct. Experimental work on repeated exposure and illusory truth provides a reason to investigate that failure rather than assume it away [21]. Such studies document psychological effects in their tested paradigms; they do not establish the magnitude of the proposed relational mechanism in scholarly review.

A useful counterexample prevents an overstrong reading of the deterministic-retelling proposition. A translated report may supply the first usable account to a reader who cannot read the original language. Its *incremental content relative to the original record* may be zero under the specified ideal transmission model, yet its *incremental evidence relative to the recipient’s actually accessible observations* can be substantial. Conversely, a reader who already knows the original may nevertheless update inappropriately when it returns through a second prestigious outlet. These are different reference conditions. The expected informational increment depends jointly on the report-generating process and the observer’s existing evidentiary state. Neither usefulness nor redundancy can be assigned to a message without specifying this comparison.

Even the category “independent corroboration” is proposition-specific. An independent observer can corroborate that a public lecture occurred while having no independent basis for judging a technical theorem stated during the lecture. A photograph of a dated notebook can corroborate the existence of the page while leaving authorship and the originality of its contents unresolved. Similarly, a bibliographic database may reliably establish that a DOI resolves to a publication, while providing no direct proof that the cited publication supports the citing sentence. Experimental tasks that collapse existence, attribution, content entailment, and underlying validity into a single correct/incorrect measure confound multiple stages of verification. The research design should separately score each stage and document adjudication disagreements.

7.3 Mediated Testimony Through Human and AI Systems

AI-mediated communication introduces further transformations between an originating record and the report encountered by an observer. A researcher may publish a technical protocol; a model may produce a summary; another person may quote the summary as an explanation; a subsequent interface may display several derived answers as separate recommendations. This arrangement can change source accessibil-

ity and perceived fluency without adding independent evidence about the originating event. It can also create genuinely useful access by translating a previously unreadable language, extracting a verifiable quotation, or locating a primary record. The question is therefore the evidentiary function performed at each transformation, together with the observer’s ability to recover its provenance. The relevant historical context includes the demonstrated limitations of natural-language generation and risks arising from uncritical interpretation of language-model output [38, 7].

Consider a hypothetical chain containing a primary record X_0 , a generated representation X_1 , and a subsequent human report X_2 . Let M_1 denote the particular generation procedure and M_2 the human reporting or editorial procedure. Under a simplified sequential generative model, Equation 45 states the joint conditional law when the entire transformation history is specified.

$$P(x_0, x_1, x_2 \mid H, M_1, M_2, C) = P(x_0 \mid H, C)K_1(x_1 \mid x_0, M_1, C)K_2(x_2 \mid x_1, M_2, C). \quad (45)$$

Here K_1 and K_2 are stochastic kernels for the transformations and should be estimated or stipulated for the relevant task. The factorization presupposes that the displayed conditioning variables contain what is needed to express the stated conditional dependence: a human author might consult another record after reading the AI summary, and a model might retrieve external documents not represented in X_0 . Those circumstances call for expanded kernels or a different factorization. The labels “human” and “AI” alone determine neither accuracy nor independence. A system’s polished wording does not supply evidence that it examined the claimed source, and a human narrator’s confident attribution does not establish that the original document exists.

A mediated chain also contains distinct questions of *fidelity* and *warrant*. Fidelity asks whether a transformation preserves the relevant meaning, qualifications, and attribution of its input. Warrant asks whether those inputs adequately support the target claim. A perfectly faithful paraphrase of an unsupported assertion cannot make the assertion supported by transmission alone. Conversely, an inaccurate paraphrase can obscure valid evidence. Evaluating the two dimensions requires comparison across levels, ideally with access to the originating passages and a task-specific account of what counts as support. An exact bibliographic identifier contributes to source findability, while a verified quotation and an interpretation of its scope contribute different forms of assessment. The existence of a reference is therefore an intermediate check, not the completion of evidentiary verification.

Recent research provides unusually direct reasons to measure the difference. In a 2025 experiment, Ding and colleagues varied citation presence and relevance in AI-generated answers and reported increased self-reported trust when citations appeared, including when citations were random; citation checking in that study was associated with lower trust [19]. Those findings concern particular experimental materials, outcomes, and participant behavior; they do not imply that displaying references always increases trust or that verification necessarily lowers justified confidence. A separate 2025 study of generator-aware retrieval found that document authorship metadata could alter model attribution performance, indicating a possible model-side effect of source labels in the tested retrieval systems [1]. Neither experiment establishes a single general psychological mechanism. Together they motivate independent measurement of citation display, actual support, model-generated attribution, reader checking, and final judgment.

Under the special assumptions of Equation 45, the sequence H, X_0, X_1, X_2 forms an information-processing Markov chain conditional on the procedures and background. The data-processing inequality then limits the ability of downstream transformations to add information about H when they use only their stated inputs. It does not rule out a downstream transformation making evidence more accessible to a bounded human observer, nor does it rule out a summary being easier to understand. An observer whose attention or linguistic competence is constrained may benefit from a carefully constructed translation even when the translation contains no new information relative to the full primary record. Such a benefit belongs to the observer’s effective processing channel, distinct from an increase in the intrinsic information contained in the original evidence. If a model uses an independent retrieval tool or a reviewer performs a new experiment, new observations must be represented explicitly before this limiting argument can be applied.

The corresponding conditional information comparison can be expressed through the data-processing inequality. Under the Markov and hypothesis-invariant transmission assumptions above, with well-defined marginal laws for both hypotheses,

$$D_{\text{KL}}(P_{X_2|H,C} \parallel P_{X_2|\neg H,C}) \leq D_{\text{KL}}(P_{X_0|H,C} \parallel P_{X_0|\neg H,C}). \quad (46)$$

The statement is an information-theoretic comparison over the complete distributions of observations, not an upper bound on a particular person’s comprehension or on the probability of a correct answer after a useful translation. If the reporting kernels are themselves correlated with H through unmodeled access, or incorporate new hypothesis-relevant external evidence, the displayed inequality need not apply. In particular, a search-augmented system that discovers a previously unknown experiment is not adequately represented by a fixed transformation of X_0 alone. Explicitly distinguishing retrieval from recitation is therefore a substantive modeling obligation rather than a technical footnote.

The distinction suggests a practical evaluative procedure. An observer can trace a mediated statement backward to its apparent source, distinguish reproduced quotations from synthesized conclusions, inspect the claimed evidentiary link when feasible, and ask whether a later report contains genuinely independent observations. If the source is unavailable, the observer may represent the uncertainty rather than manufacturing a complete lineage. Multiple summaries with identical unusual phrasing or identical unverified references may suggest common origin but cannot prove it without further evidence. Conversely, two summaries of the same document can reveal different relevant passages or methodological interpretations. Their interpretive diversity can have value while their underlying observational origin remains shared.

Generated misinformation introduces particular risks in scientific and historical recognition. A language model may supply a plausible-sounding citation to a nonexistent document, merge distinct contributors, omit a qualification from a technical protocol, or attach a statement to an incorrect date. Those are possible error modes consistent with the natural-language-generation survey literature, not claims that a given system commits them with a particular frequency [38]. When these statements are repeated by other people, an observer may face both a provenance-reconstruction problem and a familiarity-induced confidence problem. The first requires source tracing; the second concerns the psychology of evaluating repeated content. Separating them supports interventions that can target each mechanism. A linked source viewer may assist provenance checking, while an explicit origin-count label may assist the evaluation of apparent corroboration. Neither guarantees that the observer will allocate sufficient attention or interpret the original evidence accurately.

In scholarly communication, a bibliographic chain is itself a research object. Recent empirical work has developed cross-database checks for references that cannot be resolved to identifiable sources; such audits indicate that citation existence can be investigated systematically, while unverifiability under a specific protocol must not be equated with proven fabrication [29]. A DOI lookup can flag a mismatch, a repository record can establish that a work is deposited, and a manual reading can assess whether it supports a particular attribution. These operations address different error classes. A generated summary that includes an accurate link yet misstates the conclusion can pass one check and fail another. An incomplete citation to a legitimate older or non-digitized source can fail automated resolution while remaining recoverable by specialist archival work. A rigorous model of epistemic vigilance should preserve these distinctions and document unresolved cases.

7.4 Competence, Integrity, Contestation, and Calibrated Reliance

A relational account of testimony needs to preserve the possibility of justified reliance on another actor’s expertise. Constant personal re-verification is often infeasible, particularly when evidence depends on specialized instruments or unfamiliar languages. Epistemic vigilance therefore involves the calibrated allocation of scrutiny in light of available alternatives, stakes, and expected verification costs. Goldman’s discussion of expert selection emphasizes the distinctive difficulty that nonexperts encounter when evaluating disagreements among putative specialists [26]. The manuscript’s proposed extension is to ask

what information about a source’s relational history can improve that assessment without treating publicly visible reputation as a proxy for all relevant competencies.

The evaluation of competence, integrity, and task-specific credibility should be grounded in demonstrable features where possible. Competence might be evidenced by a record of accurate predictions on similar tasks or the ability to explain technical constraints in a reproducible way. Integrity-related evidence might concern whether a communicator accurately reports uncertainty, discloses relevant transformations, and corrects attributable errors. In both cases, observable records are fallible and context-dependent. Institutional certification can offer useful evidence about training or procedures while leaving particular claims open to examination. Public anonymity can constrain some assessments without logically precluding reliable testimony. Likewise, an unfamiliar contributor can supply a verifiable primary record. These distinctions avoid converting established authority into a universal epistemic ranking.

The evaluation of competence also requires denominator information. A communicator who publicly displays several accurate predictions may have made many unavailable predictions; counting the visible successes alone yields no calibrated task-specific performance estimate. The selection problem may be more pronounced for highly visible actors whose work is extensively summarized, while less visible contributors lack the infrastructure to make their complete records searchable. Conversely, public accountability can produce a richer correction history that is diagnostically valuable. An observer can learn from verifiable calibration histories, explicit uncertainty, documented correction behavior, and independent checks, but must consider how those observations entered the accessible record. This observation-level concern complements the historical-materialist focus on the material and institutional conditions of knowledge circulation.

Integrity judgments present a separate inferential difficulty. A quotation error may arise from deliberate misrepresentation, careless transmission, translation, an ambiguous original, or automated extraction. A single inaccurate statement is therefore generally insufficient to identify a communicator’s motive. Where the investigative task requires moral or institutional judgments, the evidentiary standard must distinguish what is demonstrated, what remains uncertain, and which opportunities for explanation or correction were offered. Such procedural distinctions protect both recipients of misleading testimony and contributors whose work is represented incompletely. The manuscript’s cognitive inquiry can model an observer’s attribution of motives, but it should not treat that attribution as an established fact about an identifiable person.

The act of *contesting* a report may itself generate evidence. A reviewer might challenge a translation, identify a circular citation, or request that an attribution be connected to the originating record. An appropriate response can support confidence in the communicative process, while newly disclosed contradictory material can warrant decreased confidence in the original proposition. An observer should record which target is being assessed: the correctness of a finding, the adequacy of a historical attribution, the reliability of a reporting process, or the trustworthiness of a given source for future tasks. A successful correction can reduce confidence in a particular claim while increasing confidence that an inquiry process is responsive to criticism. This target separation extends Section 6’s proposition-specific approach without assuming that contestability alone guarantees epistemic justice or social legitimacy.

For an observer facing a time budget, testimonial evaluation also has a decision aspect. Let a denote an available examination action such as opening a cited record, requesting provenance, consulting a specialist, or deferring assessment. Let Y denote a task-relevant outcome, d a downstream decision, and $U(d, Y)$ its context-specified utility or loss-sensitive evaluation. Equation 47 describes a stylized value-of-examination comparison against continuing with the current evidence E .

$$\text{VE}(a \mid E, C) = \mathbb{E} \left[\max_{d \in \mathcal{D}} \mathbb{E}[U(d, Y) \mid E, Z_a, C] \mid E, C \right] - \max_{d \in \mathcal{D}} \mathbb{E}[U(d, Y) \mid E, C] - \text{cost}(a, C). \quad (47)$$

The random variable Z_a denotes the evidence that would become available after action a , and $\mathcal{D}$ is a specified set of downstream decisions. This is an idealized expected-value expression for a decision problem, not a claim that people consciously solve it. Feasible actions and the value of information may depend on privacy, legitimate access restrictions, and the opportunity cost of verification. Under the usual

idealized conditions with free option to ignore the new evidence, the gross expected information value is nonnegative; subtracting examination cost can yield either sign. An inquiry that increases information is therefore not necessarily the best use of limited attention in every task.

Practical reliance introduces an additional distinction. A person might judge an automated summary likely correct while declining to act on it in a high-stakes setting because the consequence of error is severe. Another person might consult a provisionally trusted summary to locate an original document while withholding endorsement of the summary's conclusions. These are rationally different responses to information, stakes, and available alternatives. Research on appropriate reliance in automation supports studying such decision-context interactions [51]. The detailed analysis of behavioral reliance, action thresholds, and feedback remains the research object of Section 8. In the present section, calibrated vigilance refers specifically to evidence-seeking and testimony evaluation under stated constraints.

Vigilance can also be excessively costly or counterproductive when it is applied without regard to informational value. A recipient might demand repeated full audits of a reliably validated instrument while declining to inspect a novel high-impact assertion supported only by an unverifiable quotation. This contrast is not adequately captured by labeling one individual trusting and the other skeptical. The informational target, verification cost, alternative evidence, and consequences of remaining uncertain must all be specified. The idealized examination value in Equation 47 offers a conditional normative comparison; actual human checking can instead be driven by perceived status, disagreement aversion, curiosity, time pressure, or a wish to comply with an investigator. Section 8 will model how these assessments enter risk-bearing choices rather than equating epistemic scrutiny with a decision to cooperate.

7.5 Measurement Designs and Inferential Limitations

The proposed mechanism is testable only when the investigator can distinguish the origins of reports from their rendered presentation and the observer's perceptions of those reports. A controlled design could construct fictional but internally consistent research histories containing an originating protocol, two independently produced verification records, derivative quotations, translated summaries, and selectively omitted qualifications. The documentary source structure would be known to the investigator while remaining partially concealed from participants under pre-registered conditions. Participants could then evaluate a specified proposition about attribution or method performance, report their confidence, and decide whether to inspect the original records. The use of clearly marked hypothetical materials would limit risks of falsely attributing misconduct to real researchers.

A first factorial contrast would vary the *actual dependence* among reports separately from the *visibility of their shared origin*. For example, three visibly distinct summaries could be generated from one source, or three reports could be based on independently produced observations. Across each origin condition, a provenance interface could either disclose the lineage clearly or withhold the relationship while preserving the report content. The investigator would measure perceived number of independent sources, ability to identify the originating record, confidence change, choice of verification action, and factual accuracy under an agreed task criterion. Repetition itself can be held constant when the experiment tests provenance effects. This design can separate some consequences of origin dependence from consequences of the display, but any actual independence manipulation may also alter content quality; matched stimulus construction and independent pretesting are required.

A second contrast would compare human-written and AI-generated intermediary descriptions under controlled factual content and presentation. One condition could provide an attributable quotation and a verifiable source link; another could provide a faithful paraphrase without a link; a third could contain a deliberately introduced but explicitly experimental attribution error. Participants should not be told that one provenance label is synonymous with accuracy. Actual factuality should be validated separately, and source labels should be randomized where feasible so that preferences toward technology or professional roles do not automatically explain the outcome. Investigators must avoid inferring a general human-versus-machine accuracy ranking from a limited comparison of artificially selected texts. The purpose is to determine how observers use, misread, or ignore cues about evidence lineage and reporting

competence in a specified task.

To separate a superficial authority cue from verification, an experiment could cross two different interventions: an *appearance condition* showing an identical claim with no citation, a plausible citation, or a link; and a *support condition* in which the cited record genuinely supports, contradicts, or fails to address the target proposition. A third randomized dimension could determine whether a source viewer exposes the relevant passage directly or requires additional navigational effort. This design distinguishes the effect of seeing a familiar bibliographic form from the effect of gaining usable evidence. It should not intentionally create deceptive citations in real public information systems; fabricated or mismatched items would exist only inside consented, clearly fictional study materials and would be explained in debriefing. The investigator should also pretest plausible link formatting and citation visibility so that presentation quality does not unintentionally reveal which case has been assigned.

One useful estimand is the assignment effect of a provenance-disclosure interface on proposition-specific probability revision. Let $Z \in \{0, 1\}$ denote randomized assignment to the interface, and let $\Delta p = p_{\text{post}} - p_{\text{pre}}$ denote the participant's reported belief change for the specified claim. Then

$$\tau_{\text{assign}} = \mathbb{E}[\Delta p \mid Z = 1] - \mathbb{E}[\Delta p \mid Z = 0]. \quad (48)$$

In a properly randomized design with valid outcome measurement, the contrast identifies an average effect of *assignment to the interface* within the study population. It does not identify the effect of actually opening a record, understanding it, or correcting a mistaken belief, since those are post-assignment events. A sign on τ_{assign} is not intrinsically a sign of epistemic improvement: the appropriate direction depends on the target proposition and on whether its presented evidence supports or contradicts it. Study preregistration can specify truth-conditional calibration or documentary-adjudication criteria separately from confidence direction.

Receiver heterogeneity should be specified before analysis rather than recovered only through post hoc subgroup inspection. Relevant contrasts include domain familiarity, source-language proficiency, technical literacy for interpreting citations, availability of assistive access, and prior familiarity with the claims. The 2024 pilot study by Bielik and Krell illustrates that source-, claim-, and receiver-oriented vigilance can be operationalized as distinct assessment categories [9]; its educational setting limits the transfer of its empirical findings to other adult populations. In the proposed design, researcher-coded source understanding, participant-reported confidence, and actual checking behavior should be recorded separately. Reporting moderators with uncertainty intervals, rather than ranking demographic groups by presumed gullibility, would better match the provisional status of the model.

Table 7 distinguishes the manipulation of reports from the observer-level construct that the experiment is intended to measure. A participant who clicks a source link may be curious, skeptical, or simply compliant with perceived experimental expectations; a click alone is not evidence of effective verification. The investigator should therefore include comprehension checks about what the original document establishes and preserve a measurement condition in which the final judgment is elicited before auxiliary questions. Similarly, a participant who correctly identifies a shared source might still overcount the reports when updating confidence. The experiment should elicit independent-origin estimates and belief assessments separately, with randomized question order when its effects are of interest.

Hypothesis 7.2 (Provenance disclosure and redundant corroboration). For observers who understand the meaning of the disclosed lineage, making a common evidentiary origin explicit should reduce the extent to which multiple derivative reports are treated as independent corroboration, relative to an otherwise matched condition with obscured origin. The effect may be absent where participants already recognize the dependence, disregard source information, or interpret the disclosure as an unrelated credibility cue.

Hypothesis 7.3 (Task-conditioned benefits of mediated source access). In tasks where a human or AI intermediary provides accurate, usable access to an otherwise difficult primary record, transparent source links and appropriate comprehension support should improve observers' ability to discriminate claims supported by that record from unsupported variants, compared with equally fluent unsupported summaries. The predicted benefit is conditional on inspection opportunities, translation quality, task difficulty, and participants' capacity to evaluate the primary evidence.

Table 7: Operational contrasts for studying epistemic vigilance in relational testimony. Every proposed measure requires validation for the relevant task and population.

Construct	Potential manipulation or observation	Primary identification concern
Provenance visibility	Explicit shared-origin links versus concealed links across matching reports	Link disclosure may alter salience and credibility together
Origin independence	Separate verified observations versus retellings of one observation	Content quality and selection may change with origin
Source competence	Prevalidated task-specific performance history	Reputation labels can cue status independently of competence
Communicative integrity	Corrections, uncertainty disclosures, or accurate quotation records	Observer may evaluate politeness rather than accuracy
Mediated transmission	Human or AI paraphrase with matched source documents	Familiarity, language, and style may differ across outputs
Vigilant examination	Opening primary records, requesting clarification, tracking origin	Additional effort does not necessarily imply correct interpretation
Judgment calibration	Probability reports against validated outcomes across many cases	Histories with contested labels may lack defensible truth criteria

Hypothesis 7.4 (Citation appearance and verified claim support). When citation appearance is varied independently of whether an accessible record supports the proposition, some observers will show a confidence response to citation presence even for references that do not support the claim. Providing a usable passage-level verification opportunity should improve discrimination between supporting and non-supporting citations among participants who engage with and understand the record. Either component may have a null effect when participants already know the evidence, ignore links, or cannot interpret the sources.

This hypothesis is consistent with the motivation supplied by citation-display experiments [19], but its specific prediction about passage-level discrimination is proposed for future testing. It must be distinguished from model-side authorship-cue sensitivity [1], which changes a generated artifact before the reader evaluates it. The two stages can interact, but an experiment that randomizes only a label on a fixed artifact cannot establish how metadata affects a generative system’s upstream attribution decisions.

These hypotheses distinguish accessible relational evidence from its actual cognitive use. A potentially important outcome is a reduction in confidence after effective vigilance uncovers counterevidence. Another is a null change where the original claims were already well justified. Confidence increases count as favorable outcomes only when they move assessment toward the appropriate task-specific evidentiary standard. When such a standard is contested, investigators can report agreement with independent documentary adjudication procedures, sensitivity analyses under alternative interpretations, and inter-rater disagreement. The study should not convert a single institution’s preferred attribution into unquestionable ground truth.

The most serious inferential difficulty is that vigilance, credibility judgment, and reliance influence one another. A suspicious participant may inspect more records, but inspection can change both source impressions and substantive beliefs. Randomly providing an interface does not randomize who actually uses it, and analyses restricted to users who clicked the link may suffer selection bias. A credible causal estimand should therefore distinguish the assigned disclosure intervention from uptake, comprehension, and subsequent judgment. Mechanism-specific inferences require further controls, assumptions, or instruments. Beyond the laboratory, testimony changes through repeated social interaction, algorithmic ranking, and incentives to preserve or contest a received narrative. Those processes can make the documentary sample observed by a researcher unrepresentative. The current models offer interpretable hypotheses for investigation rather than an empirical account of any particular community.

A final limitation concerns the ground truth of source structure. In a constructed scenario the experi-

3217 menter can specify which reports share origins; in an actual archive the same structure may be disputable,
3218 partly lost, or dependent on how an evidentiary unit is defined. Inherited citations, independent replica-
3219 tions, reanalyses of shared data, and translations may occupy different positions under different research
3220 questions. The analyst must make the target question and origin criteria explicit before classifying re-
3221 ports as redundant. A transparent workflow should preserve contrary documentary readings and residual
3222 uncertainty, especially when its classification could influence the recognition of living contributors or
3223 contested intellectual histories.

3224 It is equally important not to conflate checking effort with accuracy or skepticism with virtue. A
3225 participant may open several sources and still misconstrue the relevant evidence, while another may
3226 correctly use well-grounded testimony without direct inspection. The proposed architecture therefore
3227 evaluates vigilance in relation to calibrated, proposition-specific judgments, documented source inter-
3228 pretation, and feasible alternatives. It does not prescribe exhaustive checking, generalized suspicion, or
3229 reliance upon a privileged social category. The structural organization of access, the cognitive recog-
3230 nition of source relationships, and the later decision to rely on testimony remain analytically separable
3231 even though they influence one another in practice.

3232 This section's contribution is a conditional analysis of human epistemic vigilance over relational
3233 testimony. Source credibility depends on the claim and task; communicative plurality and evidentiary
3234 independence require separate accounting; and human or AI mediation may improve accessibility while
3235 introducing representational errors or concealing common origins. The formal results concern likelihood
3236 factorization, deterministic retelling, and specified transmission kernels, not a universal ranking of wit-
3237 nesses or media. The observer-level predictions concern recognition of source dependence, selective
3238 checking, and calibrated belief changes that may have positive, null, or negative signs. Section 8 next
3239 examines how these epistemic judgments relate to declared trust, practical decisions, and behavioral
3240 reliance under differing stakes and institutional conditions.

3241 8 Epistemic Trust and Behavioral Reliance

3242 This section investigates the relation between an observer's assessment of evidence and the observer's
3243 decision to act on a claim, accept advice, delegate a task, or cooperate with a source. The research object
3244 is a set of separable outcomes: proposition-specific evidentiary warrant, subjective confidence, declared
3245 trust in a communicator, and practical reliance under specified consequences. Sections 6 and 7 examined
3246 justified belief revision and the evaluation of testimony, respectively. The present section extends that
3247 analysis to decisions made in conditions of uncertainty, resource scarcity, and potentially asymmetric
3248 losses. Its objective is to explain why a justified belief can coexist with prudent non-reliance, why a
3249 person can rely on an uncertain source for instrumental reasons, and why a favorable trust declaration can
3250 fail to predict behavior. The proposed formal constructs are decision-analytic benchmarks and research
3251 hypotheses, rather than an established general psychology of cooperation.

3252 The distinction matters for the recognition of relational contributions. An observer may judge that
3253 an earlier scientist probably introduced a method, report trust in the documentary archive preserving that
3254 claim, and nevertheless refrain from making a consequential attribution until an original record can be
3255 inspected. A different observer with the same documentary evidence may cite the claim provisionally
3256 because the expected cost of deferral is high and the consequences of correction are manageable. The
3257 two observers need not disagree about what the record supports. Conversely, two observers who take the
3258 same action may hold different beliefs, face different alternatives, or operate under different permissions.
3259 Actual reliance therefore cannot be treated as a direct readout of epistemic trust. Classic analyses of
3260 organizational trust distinguish a willingness to accept vulnerability from assessments of trustworthiness,
3261 while human-automation research separates subjective trust from appropriately calibrated reliance [55,
3262 51]. The framework developed below engages those distinctions within the paper's observer-specific
3263 model of relational evidence.

8.1 Epistemic Warrant, Confidence, Declared Trust, and Reliance

Let $\Theta \in \{0, 1\}$ denote a particular proposition under evaluation, such as the claim that a cited method was already documented in an earlier record. Let $Y = y$ designate evidence actually examined under an explicit observation protocol, and let c describe the epistemic task, assessment criteria, and feasible decision setting. In keeping with Section 6, the analyst may stipulate a benchmark $q^*(y, c) = p^*(\Theta = 1 \mid y, c)$ when its reference model, priors, likelihoods, and evidentiary selection mechanism can be defended. An observer's subjective judgment is denoted $q_o(y, c)$; it may differ from q^* because of memory reconstruction, incomplete examination, divergent priors, or model misspecification. Neither number represents a context-free quality of the person or source whose history is being assessed.

Four outcomes should be recorded separately. *Evidentiary warrant* concerns the degree of support afforded by the task-relevant evidence under an articulated standard. *Subjective confidence* concerns the observer's own graded commitment to a proposition, which may be expressed on a probability scale or through a verbal report. *Declared trust* concerns a response to an elicitation about a source, institution, system, or relationship; the target and meaning of the question must be specified. *Behavioral reliance* concerns the selection or continuation of an action whose consequences depend in some way on the claim, source, or predicted performance. Reported trust in a communicator can be high while confidence in a particular communicated claim remains low; trust in one domain can coexist with refusal to delegate in another. Where the experiment includes several claims or interlocutors, these quantities require separate indices rather than being collapsed into a single scale.

For a fixed task, define Q_o as the observer's elicited proposition-specific confidence, T_o as declared trust in the designated source, and A_o as the observed decision from an action set $\mathcal{A}(c)$. Their joint conditional distribution can be represented without presuming any deterministic conversion from confidence to action. In Equation 49, Z_o denotes measured decision conditions, including anticipated consequences and constraints, and Y_o is the evidence available to the observer's cognitive process. The factorization is an ordinary probability identity, not a causal ordering; in particular, the middle conditional cannot be interpreted as the effect of experimentally setting reported trust.

$$p_o(q, t, a \mid y, z, c) = p_o(a \mid q, t, y, z, c) p_o(t \mid q, y, z, c) p_o(q \mid y, z, c). \quad (49)$$

This decomposition has a methodological benefit. An experimental intervention that increases expressed trust might leave calibrated confidence unchanged and might leave action unchanged because reliance is prohibited or costly. A second intervention could change actual behavior by reducing the cost of evidence checking without altering the person's judgment of source trustworthiness. The two experiments would establish different findings. The same warning applies to institutional attribution decisions: quotation, endorsement, funding, delegation, acceptance into a catalogue, and permission to reproduce a result have different decision meanings. Behavioral reliance on a document as a *lead* for further investigation also differs from reliance on it as settled evidence. Lee and See's review is relevant to the distinction between trust and appropriate use of automation, but does not establish the present paper's multistage factorization as a cognitive mechanism [51].

A person's confidence report may itself depend on incentive-compatible elicitation, available numeracy, the formulation of the question, and the timing of the request. Participants can express greater certainty to communicate social confidence or to meet perceived expectations even when their underlying evidential state has changed little. Declared trust ratings are likewise sensitive to whether the question concerns honesty, competence, predictability, legitimate authority, or affective comfort. A source can be honest and mistaken, competent and unavailable, or correct on the present task despite a history of poor performance elsewhere. Hence the four constructs should be used as a measurement discipline rather than interpreted as naturally separable psychological modules. The constructs may interact continuously within a real cognitive episode. Section 2 already allows such feedback, subject to identifiable observation and intervention protocols.

A further distinction is required between *using information* and *authorizing a decision*. The same observer can accept a source's report as a useful description, assign that source the task of finding additional documents, rely on its calculations under supervision, and refuse to grant it authority to ad-

judicate priority. These actions imply different transfers of discretion and different opportunities for reversal. In particular, conditional use of an automated summary can occur alongside skepticism about its claimed provenance. An experimental dependence measure should consequently specify the point at which source output constrains an observer's subsequent options. A mere click, quotation, or citation is too coarse an indicator of acceptance if the downstream activity remains verification or critical review.

This distinction becomes more important when a reported trust measure has an unspecified referent. Trust in an archive's preservation procedures does not entail trust in the interpretation supplied by an annotator; trust in a language model's fluency does not entail trust in the dates or source origins embedded in its answer. The relevant targets can be separately indexed as $T_{o,s,d}$, for observer o , source or intermediary s , and evaluative domain d . Interpreting changes in this indexed report requires specifying whether the observer was asked about truthfulness, competence, completeness, restraint, or willingness to admit uncertainty. A strong response on one dimension may coexist with a weak response on another. The indexing scheme is a research convenience, and it does not commit the paper to a universal decomposition of interpersonal trust.

The act of eliciting confidence can itself intervene in a cognitive sequence. Asking for a numerical forecast before presenting the action menu may encourage participants to examine evidence or anchor their choice on their own prior answer. Eliciting trust only after a consequential decision may invite rationalization. Accordingly, the temporal order of elicitation, verification opportunity, and action should be recorded and, where suitable, randomized. Models of observed behavior that condition on a post-decision confidence report would then describe a different distribution from those conditioning on a pre-decision report. A researcher should not infer that the reported variable occupied the hypothesized causal position merely because it appears as a factor in Equation 49.

8.2 Consequences, Alternatives, and Thresholds for Reliance

An observer's decision concerns feasible alternatives and their consequences. It need not mirror the sign of the latest belief update. Consider a deliberately restricted example in which the observer either relies on a proposition, $a = 1$, or declines reliance, $a = 0$. Let the subjective probability of correctness be $q \in [0, 1]$. Suppose relying yields a benefit $B > 0$ when the proposition is correct and incurs a cost $C > 0$ when it is false; the alternative of declining reliance is normalized to zero utility. Under expected-utility maximization with these stipulated consequences, Equation 50 gives both the expected advantage of relying and its decision threshold.

$$\begin{aligned} \Delta U_{\text{rely}}(q; B, C) &= qB - (1 - q)C, \\ a^*(q; B, C) &= 1 \iff q \geq \frac{C}{B + C}, \end{aligned} \tag{50}$$

where ties are resolved in favor of reliance by convention. This is a normative calculation for a two-action, risk-neutral model. Actual decision makers may weight losses differently, have multiple actions, face social obligations, or lack reliable access to the probabilities and consequences. The assumption that declining reliance has zero utility can also fail when inaction creates a loss. The model is useful precisely because its assumptions are transparent and can be varied experimentally.

For illustration, consider an observer assigning $q = 0.8$ to the accuracy of a technical claim. If $B = C = 1$, the threshold is 0.5 and relying is preferred under the stipulated model. If $B = 1$ and $C = 9$, the threshold is 0.9 and declining reliance is preferred, even though confidence is unchanged. The latter may be a sensible choice when an erroneous attribution could produce serious, difficult-to-remedy consequences; the former may be appropriate when a provisional attribution is readily revisable. Both examples are hypothetical and do not imply that actual organizations assign these cardinal utilities to intellectual contributions.

Proposition 8.1 (Identical confidence permits different optimal reliance). *Fix any subjective confidence $q \in (0, 1)$. Within the binary choice model of Equation 50, there exist strictly positive benefit–cost pairs*

(B_1, C_1) and (B_2, C_2) such that reliance is strictly optimal for the first pair and declining reliance is strictly optimal for the second.

Proof. Reliance is strictly preferred if $C/B < q/(1-q)$; decline is strictly preferred if $C/B > q/(1-q)$. Since $q/(1-q)$ is positive and finite, choose a positive ratio below it for the first pair and one above it for the second. The two choices arise with the same q , establishing the claim within the model. $\square$

The proposition is not a novel theorem of decision theory. Its role is to prohibit an invalid interpretation of observed behavior within our relational-recognition programme: a refusal to rely cannot by itself establish low confidence, distrust, or inadequate recognition. The decision may instead reflect expected harm, available remedies, institutional obligations, or unequal ability to absorb losses. The reverse inference is also invalid. A person may act on weak information because every available alternative is worse, time for verification is scarce, or the action can be reversed at little cost. An observer's willingness to use a claim as a search lead does not imply endorsement of that claim as established knowledge.

The simple threshold also depends on who bears the benefit and cost. An individual evaluator may face a small personal error cost even when misattribution imposes substantial burdens on a contributor. This creates a potential divergence between private decision incentives and collectively defensible evaluation procedures. Historical-materialist questions about the distribution of the relevant costs can thus motivate empirical investigation, but the present cognitive model does not derive the ownership, institutional authority, or distributional processes that generate those consequences. Their endogenous power relations remain within the scope of Paper IV. Here, the observer's perceived and stipulated consequences are treated as contextual inputs whose relationship to objective harms must be independently assessed.

Real choices also involve uncertainty about the consequences themselves. A participant might judge a report likely correct but remain uncertain about whether the result applies to a particular application, whether a claimed skill generalizes, or whether an apparently open archive will remain accessible. Such uncertainty calls for richer decision models with separate state variables. Collapsing claim correctness, source competence, reliability of future performance, and policy acceptability into one binary event would overstate the inferential reach of a trust rating. The proposed binary benchmark is deliberately local and cannot be assumed to generalize across tasks.

The consequence model becomes more realistic when actions are indexed by an observer's authority and available resources. Let $\mathcal{A}_o(h, c)$ be the feasible actions for observer o given an admissible information state h and institutional setting c . Let $\mathcal{S}$ collect decision-relevant states, which may include proposition accuracy, conditions of deployment, the actual availability of a remedy, and downstream harms. A contextual decision benchmark is

$$a_o^*(h, c) \in \arg \max_{a \in \mathcal{A}_o(h, c)} \sum_{s \in \mathcal{S}} q_o(s \mid h, c) U_o(a, s, h, c). \quad (51)$$

The symbol $\arg \max$ allows more than one optimal action, and the model assumes the state distribution and utility comparisons are specified. In empirical work, the feasible set and perceived utilities are not generally observable without supplementary measurement. A reviewer who is contractually required to submit a judgment may have no option to abstain; an independent scholar may be able to defer; an automated pipeline may quietly issue a public record before any individual exercises discretion. Identical observed endorsements can therefore arise from sharply different counterfactual choices.

The binary result in Proposition 8.1 should also be separated from a moral or juridical warrant to impose risks on others. An agent may find a personally optimal action in the stipulated utility function while lacking legitimate authority to perform it. The utilitarian benchmark cannot settle procedural obligations concerning consent, contributor recognition, appeal, or public explanation. Those obligations may be represented as constraints on $\mathcal{A}_o$ in a local exercise, but this representation does not derive their validity from expected utility. As a matter of research design, investigators should ask separately whether the participant perceived an action as profitable, institutionally permitted, and normatively acceptable. Combining all three in a single score would erase precisely the distinctions the formalism seeks to protect.

A heterogeneous trajectory further complicates the attribution of consequences. A public statement that an individual contributed a research method can benefit a present evaluator by resolving an administrative task while exposing the person named, a collaborator, or a previously overlooked contributor to reputational consequences. The relevant party may lack access to the evidence that determined the statement and may bear costs that are not reflected in the evaluator's private payoff. A conditional extension could assign party-specific utilities U_j and investigate the distribution of decision costs across j , while postponing aggregation and authority questions to a separately defended normative framework. This extension is especially relevant for low-visibility creators whose records are incomplete through conditions that they did not control.

Reliance also varies in its reversibility. Provisional use under a clearly communicated uncertainty label can preserve opportunities for withdrawal or correction. A publication, credential, financial commitment, or algorithmically propagated citation may be difficult to retract even when the original claim can later be refuted. An observer may therefore rationally choose supervised, staged, or time-limited reliance at a confidence level that would not support irrevocable delegation. The experimental choice set should contain such intermediate actions whenever they are real options. A forced binary choice can manufacture apparent distrust or overreliance by suppressing the actions participants would otherwise prefer.

8.3 Verification Choices and Appropriate Reliance

Between unconditional reliance and total refusal lie actions that preserve opportunities for scrutiny: checking a primary record, requesting a second independent source, seeking a translation, issuing a provisional citation, or accepting assistance only under human review. These are important in historically distributed knowledge production because the observer's resources and the available procedures for correction can strongly affect the appropriate choice. In an environment with a low-cost verification interface, an observer may decline immediate reliance yet pursue evidence efficiently. In a restricted archive, the same observer might proceed provisionally with explicit uncertainty or suspend judgment. Neither choice is a simple measurement of their respect for a source.

Let $\mathcal{A}$ be a feasible action set, $U(a, \theta, c)$ a stipulated utility for action a when the relevant state is θ , and Z the additional observation produced by a verification procedure. Taking $Y = y$ as the evidence already examined, Equation 52 defines a one-step expected net value of verification under a specified probability model and a fixed verification cost $k_v \geq 0$. The expectation integrates over the possible verification results and subsequent posterior assessments. The representation assumes that verifying does not itself change the underlying proposition during the decision interval and that the utility scale is comparable across the available actions.

$$V_{\text{verify}}(y, c) = \mathbb{E}_{Z|y,c} \left[\max_{a \in \mathcal{A}} \mathbb{E}\{U(a, \Theta, c) \mid y, Z, c\} \right] - \max_{a \in \mathcal{A}} \mathbb{E}\{U(a, \Theta, c) \mid y, c\} - k_v. \quad (52)$$

Absent the cost and with unconstrained future action choice, the expected value of information is nonnegative under a coherent Bayesian decision model, because the observer can ignore an unhelpful verification outcome. Positive cost, restricted actions, processing burdens, or a misspecified information channel can remove the practical incentive to verify. This distinguishes a logically valuable observation from an observation worth purchasing under present circumstances. An additional nuance is that verifying *who said something* need not verify *whether it is true*; the appropriate observation variable Z must follow the target proposition. Where the claim concerns original authorship, a primary dated artifact may be diagnostic; where the claim concerns reliability of a practical method, successful replication or performance evidence may be more appropriate.

The cognitive processes described in earlier sections enter this comparison in several places. An observer can underestimate the probability of finding diagnostic evidence, misremember a record's origin, expect too much from a fluent summary, or exaggerate the cost of checking. Conversely, an observer may

accurately anticipate that an archive search would be inconclusive. These possibilities require measurement of perceived search costs and perceived information quality, not only subsequent clicking behavior. A recorded decision to open a link may represent curiosity, verification, compliance with an instruction, or routine workflow; a decision not to click may result from prior knowledge, distrust, overload, or inaccessibility. Examination logs are informative about encounter opportunities, but they do not directly reveal an internal trust state.

In automated environments, the design of the reliance interface matters. A system can support calibrated use by communicating uncertainty, preserving links to underlying sources, distinguishing quoted evidence from generated interpretation, and allowing revision when a record changes. These are design hypotheses, rather than guarantees that transparency raises trust or improves performance. Lee and See develop the notion of appropriate reliance in human–automation interaction, and Hoff and Bashir review contextual and individual determinants of trust in automation [51, 31]. Their analyses caution against prescribing universally higher trust: the objective is responsiveness to task-relevant reliability and consequences. Systems may sometimes be too readily believed and sometimes rejected despite useful performance. Experimental research reports both avoidance of algorithms after observing errors and preference for algorithmic over human advice under different comparison conditions [18, 53]. These findings concern the studied judgment tasks and should not be generalized to every contemporary generative system, domain, or population.

The difference between choosing an adviser and determining how much influence its advice receives offers another operational perspective. In a numerical estimation task with an initial estimate x_0 , an advice estimate x_s , and a final estimate x_1 , one conventional behavioral statistic is the weight of advice,

$$w_{o,s} = \frac{x_1 - x_0}{x_s - x_0}, \quad x_s \neq x_0. \quad (53)$$

A value near one corresponds to adopting the source’s numerical advice in this restricted setup, while a value near zero corresponds to retaining the initial estimate. Values outside $[0, 1]$ are possible and should not automatically be discarded. The statistic is undefined when the advice equals the initial estimate, unstable when the denominator is small, and inapplicable to many categorical, normative, and cooperative actions. Even within an estimation task it measures movement of a response, which can reflect anchoring, persuasive framing, strategic behavior, or genuine information integration. Logg and colleagues use advice weighting to study algorithm appreciation in specific numerical judgment settings; this measurement does not directly establish generalized epistemic trust in a person or AI system [53].

The distinction between verification *availability* and verification *use* should be made explicit in interface studies. A design may give every participant a link to a primary record, but only some will notice it, interpret it, or have sufficient time to inspect it. A nominally transparent system may therefore create little additional cognitive access. Conversely, an interface that requires a brief justification before adopting a machine suggestion may alter the distribution of effort as well as the decision threshold. An experiment by Bućinca and colleagues found lower overreliance in particular cognitive-forcing conditions than in simpler explanation-based interfaces, alongside less favorable subjective ratings for the strongest interventions; moderation by a measured cognitive-motivation variable further limited uniform conclusions [13]. Those findings motivate studying effort and comprehension, not assuming that an additional click is universally beneficial.

The direction of a design effect can depend on the accuracy of advice. In mixed-method studies of selected classification tasks, Bansal and colleagues found that providing model explanations increased participants’ willingness to accept AI recommendations regardless of their correctness, without an accompanying increase in complementary human–AI team performance in their tested conditions [6]. This observation makes it essential to distinguish global acceptance rates from selective acceptance conditional on the recommendation’s correctness and from actual final-task performance. The experiments concern particular task interfaces and model configurations, and they cannot by themselves show how documentary researchers would behave with historically uncertain evidence.

Verification can itself have a socially distributed cost. Maintaining provenance, making original materials interpretable, and staffing dispute-resolution processes require work. An instruction to “verify

everything” can transfer those costs onto users with little time, disability accommodations, access, or institutional support. Observed low verification rates therefore should not be classified as cognitive negligence without establishing feasible alternatives. A rights-respecting design may combine source-preserving interfaces with institutional obligations to provide adequate records and accessible routes of challenge. This consideration connects the cognitive framework with the larger research programme’s distinction between observer behavior and infrastructure-level opportunity.

8.4 Behavioral Feedback and the Limits of Observational Inference

Reliance can change the later evidence environment. If an observer consults one source, the source’s vocabulary and references may shape subsequent searches. If an automated assistant generates an intermediate account, later reports may reuse that account as an apparent independent witness. If a tentative attribution is circulated, other people may quote it and give it retrospective visibility. Such feedback can alter the history the observer later encounters while leaving the underlying question of initial contribution unresolved. The relevant mechanism is thus a coupling between cognitive evaluation and information-seeking behavior, rather than a one-way conversion of posterior confidence into a behavioral response.

A minimally specified sequential representation distinguishes an action-conditioned observation law from an observation-only association. Let H_k denote the admissible documentary state at investigation step k , let A_k be an intervention on the choice to search, verify, cite, or delegate, and let Y_{k+1} denote subsequent available evidence. The causal target in Equation 54 is the distribution of subsequent evidence under a defined feasible action intervention. It is not identified simply by comparing people who happened to take different actions.

$$\mathcal{K}_k^a(B \mid h, c) := P(Y_{k+1} \in B \mid \text{do}(A_k = a), H_k = h, C = c). \quad (54)$$

The intervention requires a well-defined action, a coherent temporal comparison, and suitable assumptions regarding selection, confounding, consistency, and interference. In real scientific networks, one observer’s citation or verification request may affect what another observer later sees. The documentary state H_k can also change through subsequent production or cataloguing. The notation does not establish that potential action outcomes can be recovered from passive logs, nor does it assume that a settled past and an independently evolving future are given ontological primitives. The index k orders specified investigation steps within the research design; histories and event boundaries remain provisional descriptions of an evolving relational system.

A particular methodological hazard arises when experimenters condition on participants who chose to examine a source. That subgroup is jointly selected by prior beliefs, perceived stakes, curiosity, time availability, and interface accessibility. Its higher subsequent trust or accuracy cannot be interpreted as the effect of examination without further assumptions. A randomized encouragement to inspect evidence estimates the effect of *offering or encouraging* verification unless compliance and exclusion assumptions justify a more specific estimand. Similarly, random assignment to a source label can establish an effect of that label on decisions while leaving the psychological mediator ambiguous. The same observed action might arise from source affinity, perceived competence, expected reward, familiarity, or institutional pressure. These identification concerns will be developed systematically in Section 9.

Another limitation concerns the meaning of reliance in cooperative relationships. A collaborator may depend on someone else’s work because their skills are complementary, because resources are pooled, or because a contractual or institutional arrangement requires the allocation of particular tasks. Such cooperation can create mutual vulnerabilities and new evidence about reliability, but it is not reducible to a single probability that a stated proposition is correct. Some cooperative acts express solidarity, reciprocity, courtesy, or obligation even when epistemic evaluation remains cautious. The model treats these as possible action motivations that should be elicited or controlled, rather than presumed to be epistemic endorsements. An investigator should avoid projecting a uniform trust meaning onto heterogeneous forms of social cooperation.

Historical-materialist analysis adds another boundary. The costs of checking, capacity to access records, and consequences of accepting or rejecting claims are influenced by historically developed labor arrangements, technologies, and institutions. The present section recognizes those conditions as potential determinants of the observer's feasible actions, without claiming that choice theory explains their origins. For example, a properly trained reviewer with access to a maintained repository faces different verification costs from an independent participant relying on an incomplete public interface. This difference may alter reliance independently of personal trust. Relational inequalities can thus appear in behavioral data through constrained option sets. Their explanation requires attention to institutions and resource distribution beyond the cognitive decision model.

Two different forms of dependence should be distinguished in longitudinal datasets. A report may appear to corroborate an earlier one because it was independently generated from a common external event, or because it simply copied the earlier report. A reliance action can alter the mix of these pathways: recommending an AI-generated synthesis may cause its language to be replicated by subsequent human accounts. Later observers can then encounter greater apparent agreement without any increase in independent evidence. The mechanism is plausible, but its magnitude depends on the source graph, copying behavior, and what interfaces expose. In an experiment, a complete synthetic source graph permits direct assessment; in an historical archive, source lineage is usually incomplete and should be recorded with uncertainty rather than assigned from surface similarity alone.

Interventions on a single user's actions may also affect others' subsequent opportunities. A publicly circulated attribution can modify search rankings, the vocabulary used by later cataloguers, or the perceived credibility of a source. The conventional no-interference assumption is therefore questionable in networked or collaborative settings. The causal object in Equation 54 remains intelligible if an intervention is specified for a bounded unit and timeframe; extrapolating its effect to a network requires an explicit exposure mapping or a design that randomizes at an appropriate social level. A single-user randomized interface trial identifies an effect on that trial's observable outcomes, not a universal effect on the future circulation of contributions.

Another boundary is the observer's ability to exit a reliance relationship. A worker may be directed to use an institutional system; a student may be required to follow a prescribed documentation platform; and an automated workflow may distribute a decision before the affected person can inspect it. Calling those events "trust" would collapse voluntary acceptance into compliance. Research on automation has long differentiated use, misuse, disuse, and problematic deployment arrangements that shape the operator's role [60]. For the purposes of this paper, reliance by instruction should be represented as a constrained action outcome, with the authority assigning the instruction separately identified. The distinction permits studying individual cognition without treating institutionally produced dependence as a personal endorsement.

Because reliance can shape later evidence, a stable association between present confidence and later choice cannot be assumed without describing intervening experience. A participant who initially distrusts a generated summary may encounter a confirming primary source, increase confidence, and then rely on the summary's remaining portions. Another may follow a suggestion, discover an error, and become cautious; a third may experience an interface failure unrelated to source veracity. Longitudinal analysis should register the content, ordering, and diagnostic strength of these encounters. It should also treat documented search and communication steps as operational observations rather than as a metaphysically privileged sequence of all generative events.

8.5 Measurement, Calibration, and Research Hypotheses

Empirical tests should jointly observe proposition-specific judgment and subsequent behavior under variation in stakes and verification possibilities. One workable design provides participants with an identical synthetic documentary history, elicits a probability that an earlier contribution satisfies a specified attribution criterion, and separately measures source competence and integrity judgments. Participants then decide whether to make a provisional citation, seek more information, or issue a consequential public

Table 8: Task-specific measurements for confidence, declared trust, verification, and reliance.

Target	Possible observable	Principal ambiguity or safeguard
Evidentiary warrant	Documentary adjudication, likelihood benchmark, competing readings	Reference standard and source selection require independent scrutiny
Subjective confidence	Probability report for a specific proposition	Elicitation, numeracy, incentives and misunderstood target
Declared source trust	Separate ratings of competence, integrity and task credibility	Social desirability, broad wording and context dependence
Verification behavior	Inspection logs, queries, requested primary records	Searching does not establish comprehension or belief change
Reliance selectivity	Correct- and incorrect-advice uptake; final task outcome	Adjudication standard, costs, denominators and alternatives must be reported
Decision authority	Available actions, ability to abstain, mandatory use or delegation rules	Recorded system use may reflect obligation rather than endorsement
Behavioral reliance	Provisional citation, delegation, endorsement or withheld action	Stakes, constraints, obligations and availability of alternatives
Decision quality	Specified payoff, regret or calibrated performance over tasks	Task utility differs from public or institutional justification

endorsement. Benefits, error costs, action reversibility, and checking costs can be experimentally varied. This design permits comparisons of behavior conditional on separately elicited beliefs, while keeping the documentary stimulus and the assessment target as constant as feasible. The design does not assume that elicited probabilities perfectly measure participants' underlying uncertainty.

Performance should be assessed with more than agreement between stated trust and actual reliance. If a task has an adjudicable claim outcome, forecast calibration and proper scoring can assess the quality of probabilistic judgments under a carefully defended reference standard; Brier's formulation supplies one well-known squared-error criterion [11]. Behavioral decisions can be assessed for regret or utility relative to the stated experimental payoff structure. Reliance calibration means responsiveness to task-relevant reliability and stakes, and should be distinguished from a simple correlation between a trust rating and the frequency of delegation. The reference standard might be straightforward in a synthetic experiment and historically disputed in a real attribution controversy. In the latter case, uncertainty and disagreement concerning the benchmark must remain visible.

Table 8 maps the proposed analytical distinctions to observable variables and principal limitations. The table is a design aid and does not imply that the constructs are experimentally independent.

Hypothesis 8.2 (Stake-sensitive reliance under matched confidence). When the proposition-specific evidence, source description, and elicited confidence are comparable, increasing the explicitly stated cost of an erroneous reliance decision should reduce reliance or increase verification seeking in tasks where those alternatives remain feasible. The prediction is conditional on participant understanding of the stakes, incentives, and available actions; its magnitude and possible reversals are empirical questions.

Hypothesis 8.3 (Provenance-sensitive calibration of mediated advice). In tasks with separately checkable source origins, interfaces that make independent and derivative report lineages inspectable should help observers differentiate corroborated advice from repeatedly circulated advice, compared with equally fluent interfaces that conceal lineage. The expected downstream effect is better calibrated confidence or action relative to an explicit reference standard, with null or negative results possible under interface overload, poor comprehension, or unreliable source metadata.

To assess *selective* reliance rather than indiscriminate acceptance, consider an experiment with an AI or human recommendation $D \in \{0, 1\}$, adjudicated recommendation correctness $R \in \{0, 1\}$, and a behavioral adoption indicator $A \in \{0, 1\}$. The simple conditional pair

$$\rho_1 = P(A = 1 \mid R = 1, c), \quad \rho_0 = P(A = 1 \mid R = 0, c) \quad (55)$$

separates acceptance of correct from incorrect recommendations within task context c . A larger difference $\rho_1 - \rho_0$ can be informative when the costs, opportunities, and correct-versus-incorrect item mix are controlled. It is not a context-free score of appropriate reliance: justified verification, abstention, and differential consequences may warrant rejecting even a correct recommendation, and some recommendations labeled incorrect may still contain useful leads. When adjudicated correctness is unavailable or disputed, each quantity must be interpreted under a stated reference rule rather than silently treated as ground truth. Researchers should report final decision accuracy and task consequences alongside the acceptance rates.

A proposed factorial experiment would independently randomize (i) visible provenance versus a presentation with the same content but hidden lineage, (ii) an advisory versus an institutionally required interface mode, and (iii) reversible versus difficult-to-reverse consequences. Advice correctness or the synthetic source’s reliability can be controlled by the stimulus generator and stratified in the analysis. For a binary randomized provenance display Z_P , a bounded intention-to-treat contrast conditional on assigned design factors s, r is

$$\tau_P(s, r) = \mathbb{E}[A \mid Z_P = 1, Z_S = s, Z_R = r] - \mathbb{E}[A \mid Z_P = 0, Z_S = s, Z_R = r]. \quad (56)$$

Here Z_S is an assigned stakes condition and Z_R is a prespecified recommendation-correctness stratum or assigned stimulus condition. The contrast estimates the effect of being *assigned* the display under standard randomized-design conditions; its sign need not be positive because the proper action depends on the stimulus and consequence. If participants can circumvent an interface mandate or encounter others’ assignments, adherence and interference must be evaluated. Conditioning the same contrast on post-treatment source inspection or post-treatment confidence can introduce selection bias and will not generally identify a direct cognitive pathway.

The factorial design also enables separate study of *epistemic* and *action* moderators. A display may change how accurately people identify independent witnesses while having little effect on reliance when every available action is compulsory. Changing whether the participant can defer or seek another source can alter actions without changing the assessed probability of the proposition. Group-specific effects may vary with prior expertise, task language, available time, disability accommodations, or familiarity with the type of evidence. These variables should be measured ethically and interpreted as context-sensitive opportunities and competencies, not as fixed personal deficits. Claims about subgroup heterogeneity require adequate sample sizes and preplanned contrasts.

Hypothesis 8.4 (Compelled use and the trust–reliance association). For comparable synthetic records, advice sources, and explicit costs, making use of a system institutionally compulsory should weaken the extent to which observed system use distinguishes participants with different reported source trust, relative to a condition with an accessible, genuine choice to use or decline the system. Effects depend on participants understanding the mandate, compliance, and the practical alternatives offered. A null or reversed contrast would restrict the scope of the proposed mechanism.

The hypothesis is about the interpretation of behavior in defined settings. It does not propose that a mandate has no cognitive effects: requiring use may alter familiarity, resentment, perceived legitimacy, attention, and subsequent beliefs. An intervention may therefore affect both the action menu and the observer’s evaluative state. Separating these pathways requires intermediate measurements and stronger designs than a single end-of-task trust scale. Evidence about automation use and human–AI teaming provides motivations for such experiments, while the model’s precise quantitative predictions remain untested [60, 6, 13].

The first hypothesis tests a consequence-sensitive action mechanism rather than assuming that caution implies distrust. The second links the paper’s cognitive processes with the evidence-lineage concerns developed in Section 7 and the structural-access framework of Paper II. Neither hypothesis has been tested in the present manuscript. Appropriate studies would randomize feasible aspects of the task, measure attention and memory alongside final choices, and register alternative explanations in advance. Follow-up studies could examine whether source trust declarations change after experience, whether

reliance precedes or follows inspection, and whether the effect of source labels differs when advice is correct, ambiguous, or demonstrably mistaken. These designs require particular care where participants might infer that an experimental interface itself endorses the underlying source.

The conditional predictions also accommodate ethical considerations. Experiments should use synthetic records or non-identifiable cases where possible, avoid manufacturing harmful reputational allegations, and provide explanations of induced uncertainty. A claim's epistemic warrant can be independently significant even when an institution refuses to authorize reliance; conversely, the fact of public endorsement does not retroactively prove the soundness of the underlying evidence. Different actors may bear the harms of misinformation and of excessive withholding. A serious assessment of trust in relational histories should therefore report who decides, who can inspect the supporting record, which alternatives were feasible, and how mistakes may be corrected.

The contribution of this section is a structured account of the gap between warranted belief, experienced or declared trust, and conduct. A simple threshold proposition makes one source of that gap transparent: identical confidence can rationally support different actions under different consequences. A value-of-verification model shows how information-search opportunities create further choices; an action-conditioned observation law indicates that reliance can subsequently alter the material available for assessment. The empirical hypotheses make these mechanisms investigable while preserving the distinction between conditional normative benchmarks and actual human cognition. Section 9 will draw together measurement strategies for the full cognitive sequence, including the problem of identifying its intermediate stages from limited observational data.

9 Research Designs and Identification Limits

This section develops an empirical programme for examining the cognitive recognition of relational histories. Its primary research object is the distinction between observable effects of presenting evidence and the intermediate processes by which human observers attend to, remember, interpret, evaluate, and act upon that evidence. Sections 2–8 proposed a conditional processing architecture and separate hypotheses for attention, reconstruction, attribution, belief updating, testimony evaluation, and practical reliance. The present section translates those propositions into experimentally manipulable conditions, stage-sensitive measurements, causal estimands, and historical comparison designs. Its method combines randomized experimental contrasts, explicitly limited causal models, repeated-observer protocols, and analysis of representational validity. These designs are proposals; the manuscript contains no newly collected participant data, preregistered experiment, or estimated cognitive effect.

A controlled experiment can establish that one interface changes the average observed assessment while leaving the psychological explanation underdetermined. This distinction is especially important for relational evidence. A participant may reach an inaccurate attribution because an original record was hidden, because an available record was overlooked, because the attribution was remembered without its provenance, because the observer applied a different legitimate criterion, or because a consequential choice was constrained by the task. The several outcomes can coincide in a final rating. Treating a single favorable assessment or reliance decision as evidence of successful recognition would consequently confound the distinct research objects of the preceding sections. The programme below therefore gives priority to identifying an interpretable total effect, measuring intermediate states without assuming perfect access to them, and evaluating which stronger explanatory claims the resulting observations can actually support.

9.1 Controlled Comparisons of Access, Salience, Provenance, and Prior Information

A first experiment can use fictional, non-identifiable knowledge-production histories with an adjudicable documentary record. Each history might contain an original method description, an independently documented implementation, a later review of the method, and a chain of summaries that vary in their source dependence. The documents should be constructed so that their substantive content and attribu-

tion criteria can be specified in advance. Their temporal organization need not be forced into a total order: independent or only partially ordered activities can be represented as such, with a scoring protocol permitting an explicitly unresolved relation. The experimental manipulation should concern the presentation or accessibility of these records, rather than the intrinsic historical truth of an actual person’s contribution.

A preregistered factorial design can vary four families of assigned conditions. *Access* determines whether the original document can be opened, is reachable through an additional step, or is unavailable during an initial decision. *Salience* determines how visually or linguistically prominent a specified item appears without changing its content; salience must be distinguished from document order and task relevance as far as feasible. *Provenance* varies whether paths from summaries to their source documents are inspectable or are presented as unlabeled links. *Prior information* varies an explicitly disclosed, experimentally supplied background claim about an author’s earlier work, with the manipulation separated from real-world identity cues. The resulting conditions should include baseline levels that make the four manipulations interpretable. An actual study would select a feasible subset of combinations when the full factorial design exceeds participant or materials resources.

Define $Z_i \in \mathcal{Z}$ as a randomly assigned interface condition for participant–task unit i , $Y_i(z)$ as the outcome that would be observed under condition z , and q as a fixed evaluative question, such as whether a specific documentary claim is supported. For a prespecified contrast between two complete interface packages z_1 and z_0 , Equation 57 defines the intention-to-treat effect on an outcome scale that must be named before analysis.

$$\tau_Y(z_1, z_0; q) = \mathbb{E}[Y_i(z_1; q) - Y_i(z_0; q)]. \quad (57)$$

The expectation in Equation 57 is over the study’s defined participant–task population. Under randomized assignment, appropriate measurement, treatment consistency, nonzero allocation probabilities, and absence of relevant interference, it admits estimation from the difference in assigned-group mean outcomes. The estimand covers the effect of the *assigned interface package*, including any changes in behavior that the package induces. If a participant elects to ignore a displayed record, the intention-to-treat effect remains defined. Conditioning analysis on the participants who actually opened that record changes the comparison and may introduce selection bias. A substantive effect on source-identification accuracy does not, by itself, reveal whether the causal path operated through attention, source memory, interpretation, or a combination of these processes.

Assignment is also an intervention on a *version* of an interface, not merely on an abstract property called provenance or transparency. Two provenance displays with identical labels may differ in scrolling requirements, screen-reader support, how visibly they distinguish original from derivative records, and the effort needed to inspect a citation. If versions are behaviorally consequential, consistency cannot be invoked for a poorly specified treatment category. The experimental protocol should record executable interface versions, document accessibility behavior, and archive synthetic stimuli or their reproducible generation rules. A feature-level claim requires replicated implementations or an explicit argument about why the feature is the only relevant difference. Randomizing a polished, provenance-rich interface against a less usable comparison would estimate a package effect and would not isolate provenance legibility.

A useful extension identifies effect heterogeneity by characteristics measured prior to assignment. Let G_i be a prespecified, pretreatment observer grouping, such as documented experience with bibliographic research, and suppose both treatment versions are possible within the relevant group. Equation 58 defines the corresponding conditional intention-to-treat contrast.

$$\tau_Y(z_1, z_0; q \mid g) = \mathbb{E}[Y_i(z_1; q) - Y_i(z_0; q) \mid G_i = g]. \quad (58)$$

Equation 58 does not justify searching many post hoc groups for striking differences. Eligibility, language competence, disability accommodations, historical familiarity, and access to training can influence both task meaning and intervention responsiveness. Comparisons of effects across such groups require appropriate assignment overlap, sufficient group-specific precision, and explicit attention to whether the outcome instrument retains comparable meaning. Conditioning on source-opening behavior, confidence after the task, or an experimentally changed literacy score does not produce this pretreatment contrast.

Treatment-effect heterogeneity is evidence about a specified population and design, not an intrinsic ranking of the epistemic ability of individuals or social groups.

Factorial assignment can test whether a provenance disclosure changes the impact of a salient summary. For two binary assigned factors S (salience) and P (provenance visibility), with other assignments held at a specified reference condition, Equation 59 defines their interaction on the chosen outcome scale.

$$\tau_{SP} = \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)] - \mathbb{E}[Y_i(1, 0) - Y_i(0, 0)]. \quad (59)$$

Here the ordered treatment pair specifies (S, P) , and the equation suppresses the held-fixed factors for readability. The interaction is scale-dependent: an interaction in response probability need not be present on a log-odds scale, and vice versa. Randomization can support this comparison under the same design assumptions as Equation 57; it does not confer psychological meaning on the interaction automatically. For instance, an effect of provenance visibility that becomes larger under high salience might reflect improved comparison of records, additional curiosity, or the interface drawing attention to a source button. Stage measurements and alternative interface implementations are needed to differentiate these explanations.

A factorial design should therefore be justified by the contrasts it can estimate with useful precision rather than by the visual neatness of enumerating combinations. The minimum substantively meaningful effect and the primary unit of randomization must be specified before recruiting participants. Repeated tasks per person increase observations but do not supply the same independent information as additional observers; source histories may also form a second sampling dimension. Simulation-based planning can represent within-participant and within-document correlations, plausible nonadherence, unequal accessibility-related attrition, and the uncertainty of interaction estimates. A study that can estimate a main effect reasonably precisely may be underpowered for a factor interaction or a small subgroup contrast. Sample-size justification should be tied to the inferential target, resource limitations, and plausible precision, as emphasized in methodological guidance on informative study planning [49]. Preregistration can specify primary contrasts and their uncertainty summaries, rather than implying that failure to reject an interaction demonstrates the absence of dependence between features [59].

A preregistration should select a small family of primary outcomes. Candidate outcomes include accuracy of a proposition-specific attribution against an adjudicated synthetic record, the probability of choosing an explicit “insufficient evidence” response, a proper-score measure for probability judgments, and a task-specific reliance decision under defined incentives. Source trust ratings and self-reported confidence should be secondary constructs with separately documented target questions. If accuracy is used, the adjudication rules must allow uncertainty about partially ordered histories and genuine ambiguity in contribution standards. Treating reasonable abstention as an error would artificially reward overconfident attribution. The study should report intention-to-treat contrasts even when actual exposure rates differ between conditions, together with assigned-condition adherence and comprehension measures.

9.2 Stage-Sensitive Measurement and the Limits of Cognitive Mediation

A single dependent variable cannot establish a mechanism connecting an interface intervention to a final decision. The experiment should therefore obtain several kinds of evidence, each with an explicit temporal and conceptual role. Interface logs can record which records became available, were opened, and remained displayed; dwell time and pointer movements can provide coarse measures of encounter behavior. Gaze measurement, where technically feasible and consensually collected, may offer another indicator of visual orientation. These measures concern behavioral allocation of attention under an interface. They do not guarantee reading, semantic comprehension, or later retention. Eye location also varies with individual viewing practices and accessibility technology, creating measurement differences that should be evaluated rather than treated as cognitive deficits.

Memory should be assessed through separate tasks. A content-recognition task asks whether the participant recalls a documentary statement. A source-discrimination task requires distinguishing the

originating record from a subsequent summary, including an “unknown” option. A provenance task asks which reports derived from the same original document. A partial-order reconstruction task permits participants to represent “preceded,” “followed,” and “not determined by the record” relations. Immediate retrieval probes can themselves redirect attention or change subsequent encoding. Accordingly, a design that measures every proposed stage in the same trial may interfere with the natural sequence being studied. One remedy is to randomize measurement schedules, assigning some participants to minimally probed tasks and others to stage-specific probes, and to compare their final performance. Such a comparison estimates measurement reactivity as an experimental effect; it does not eliminate the possibility that unobserved mental processing differs across schedules.

To formalize this comparison, let $B_i \in \{0, 1\}$ indicate a randomized intermediate-probe regime, where $B = 1$ introduces a source-memory question before the final judgment and $B = 0$ defers that question until afterward. Let $Y_i(z, b)$ be the final outcome under interface assignment z and probe regime b . A factorial experiment can estimate the effect of the *probe policy* within interface condition z :

$$\delta_B(z) = \mathbb{E}[Y_i(z, 1) - Y_i(z, 0)]. \quad (60)$$

This contrast is distinguishable from the effect of improving memory itself. A pre-outcome probe may redirect search, rehearse an attribution, suggest a criterion, or reveal the experimenter’s interest. In a follow-up experiment, investigators could vary probe content and placement to bound some of these alternative explanations. Such comparisons require that probe regimes be assigned and described before outcome observation. They cannot retroactively identify an unmeasured psychological state by observing a correlation between recalled provenance and later trust.

Time-indexed observations should likewise be tied to the study’s operational event records, not elevated into a unique psychological order. An opening log timestamps an interface event; it does not establish when a representation formed or when one cognitive process ceased. Source understanding and remembered impressions may reciprocally modify one another during the task. Accordingly, analyses may compare different process models and measurement schedules instead of requiring that all observers instantiate a single linear chain. An observed order of button presses is a feature of the task environment, while the proposed dependence among cognitive states remains a separately testable hypothesis.

Interpretation should be probed using task-specific questions that distinguish documentary recognition from normative attribution. After a participant identifies an original source, the study might ask about novelty of a conceptual method, reliability of an implementation, and responsibility for a subsequent application in separate questions. Before labeling responses as biased, investigators should determine whether observers understood the same attribution criteria. Within-observer comparisons can hold the document history fixed while varying a stated criterion, such as priority of public disclosure versus independent technical replication. Differences in resulting ratings may reflect coherent criterion use. Contradictions or inconsistent applications of a stated criterion form a different outcome, and their detection requires repeated or closely matched instances rather than one disagreement with an investigator’s preferred answer.

Let M_i denote an unobserved cognitive state or collection of states relevant to a theory, and let W_i denote a vector of recorded proxies, such as opening logs, recognition responses, and gaze-derived indicators. A measurement model is expressed by Equation 61, conditional on the experimental task c and the assigned interface Z_i .

$$P(W_i = w \mid M_i = m, Z_i = z, c) = K_W(w \mid m, z, c). \quad (61)$$

Equation 61 names a conditional observation law and does not claim the latent cognitive state is uniquely recoverable. Distinct combinations of attention, memory, and interpretation can induce the same distribution of proxy responses. Measurement invariance across assigned interfaces, languages, and observer groups must be evaluated if common parameter values are imposed. A visually complex interface could change both the underlying attention process and the relation between that process and a dwell-time measurement. Construct validity therefore requires additional evidence beyond the internal consistency of questionnaire items or a convenient factorization [23].

The distinction between a total effect and a path-specific causal claim is especially consequential. A hypothetical sequence might be described as interface assignment Z , attention A , encoding M , interpretation I , and final judgment Y . The sequence can be a useful bookkeeping representation, and the conditional laws in Section 2 specify an exploratory processing model. Yet randomization of Z does not randomize A , M , or I . Participants' fatigue, expertise, earlier beliefs, and motivation can jointly affect encounter behavior and judgments. Even perfect measurement of the proposed intermediates would therefore leave mediator–outcome confounding as a potential problem. Identification of particular mediation effects depends on additional assumptions and sensitivity analysis, as the general causal-mediation literature emphasizes [35]. If an interface changes an intermediate variable that then confounds a later mediator–outcome relationship, conventional natural effect decompositions encounter still stronger identification problems [82]. Such decompositions should be treated as optional, separately justified extensions rather than automatically reported percentages of a total effect.

The practical implication is to prioritize interventions on specific stages when feasible. An independently randomized interruption can test the effect of allocating an additional inspection opportunity; a source-memory cue can test an effect on source discrimination; an unambiguous statement of the attribution criterion can test a framing effect. Each intervention may still influence several processes. Claims about the effect of *providing the cue* are usually more defensible than claims that its influence flowed exclusively through a single latent cognitive variable. Sequential randomization or carefully designed repeated experiments can refine those comparisons, but the timing and consequences of earlier tasks must be represented explicitly. The study should report the effects of interventions in language matching the variables actually assigned and the outcomes actually observed.

A more ambitious longitudinal experiment may assign interface opportunities in stages according to an investigator-specified rule: for example, an initial source-access policy followed by a randomly assigned offer of supplementary lineage information among all participants meeting a prespecified eligibility rule. Let H_{ik} denote the measured record available *before* assignment at study decision k , let Z_{ik} be that assignment, and let $d_k(H_{ik})$ be a proposed policy. Comparisons between dynamic regimes then concern the outcomes under the *implemented policy sequence*. They require sequential consistency, appropriate positivity, well-defined assignment histories, and conditional exchangeability of each assigned decision; randomized sequential decisions can help defend the latter within an experiment. Analyses of nonrandom historical policy sequences require stronger justification and measured time-varying confounders [30]. A history recorded by the experiment should be understood as a conditional ordering of measured decisions and observations. It is not evidence that past, present, and future constitute fixed ontological primitives in the relational system under study.

A diagram of hypothesized observables and confounders should accompany any mechanism analysis. Investigators should specify which variables are interventions, which are outcomes, which are merely observed after assignment, and which variables represent previous conditions or selection into the study. Candidate negative-control measures or neutral interface manipulations may reveal some residual biases; they do not conclusively validate the absence of all confounding. Competing explanations, including differences in linguistic comprehension, active source search, criterion selection, and compulsory reliance, should be articulated before the design is interpreted as evidence for a particular cognitive stage. The study may support a local empirical regularity while leaving its mechanistic explanation plural.

9.3 Within-Observer Tasks, Source Dependence, and Interference

Recognition of relational evidence unfolds through sequences of exposures. A design can present several fictional histories to the same observer, vary the dependencies among their reported sources, and revisit some materials after a delay. This permits estimation of learning, fatigue, source confusion, and carryover under controlled task histories. The order of tasks should be randomized or counterbalanced as appropriate, and the assignment mechanism documented. A response to a repeated history may reflect earlier memory or a changed strategy rather than the current display alone. The analyst should therefore index the condition by an exposure history when prior assignments can affect later potential outcomes.

Random intercepts and participant-level clustering can account for some dependence in analysis, but a statistical correlation structure cannot repair a design whose causal intervention is ill-defined.

An especially informative contrast varies evidence lineage while holding the surface number of reports constant. One condition presents three separately produced observations of a documented claim; another presents three fluent summaries derived from a single observation. The wording, length, and basic salience of the reports should be matched as far as the intended manipulation permits. Participants can then be asked to identify the number of independent origins, estimate confidence in the claim, and choose whether to inspect a primary record. Such a study directly targets the conceptual distinction of Section 7. It can also test whether displaying a lineage diagram improves source discrimination at the cost of increased processing time. A favorable average effect is only one possible result; poorly designed visualizations may distract observers or generate false assurances of independence.

A manipulation must distinguish documentary independence from merely different speaker identities. Three human narrators may have consulted the same archive; a human reporter and an AI-generated summary can share one textual source; two distinct documents may copy earlier material without making their dependency explicit. Conversely, separate observations can contain dependent errors arising from common instruments, shared measurement conditions, or common interpretive assumptions. The analyst should stipulate what independence means under the synthetic data-generating mechanism, and should test the participant’s ability to track that definition. Counting visibly distinct interfaces or distinct nominal communicators as independent evidence would reproduce the very inferential error the experiment is intended to study.

Socially interactive studies introduce another complication. Participants may exchange findings, correct one another’s source identifications, or pass along claims before their own judgment is measured. An individual assignment can then affect another participant’s exposure. For an assignment vector $\mathbf{Z}$ and an explicitly chosen exposure mapping $g_i(\mathbf{Z})$, Equation 62 states a possible restriction on the potential outcomes under interference.

$$Y_i(\mathbf{Z}) = Y_i(g_i(\mathbf{Z})) . \quad (62)$$

Equation 62 requires that the selected exposure mapping capture every aspect of the assignment vector relevant to the outcome. This is an assumption to investigate, not a property of social influence guaranteed by defining g_i . The mapping could classify participants by their own interface assignment and the number or configuration of neighbors who received provenance information. If the mapping is credible, network-aware randomization and estimation methods can make corresponding causal contrasts interpretable; the validity of the mapping, allocation probabilities, and dependence assumptions remains central [4]. Where investigators cannot justify an exposure mapping, descriptive network comparisons should be presented as such. Cluster randomization may reduce cross-condition contamination while changing the estimand and requiring enough independently randomized clusters.

The selection of a cluster is itself a theoretical decision. Randomizing by discussion group can limit immediate information sharing, yet participants may communicate across groups through external networks, shared archives, or generative-AI tools. Conversely, artificially forbidding all communication can remove the collective practices of verification that an institutional intervention is intended to support. Researchers should specify a credible *partial interference* boundary only when communication pathways warrant it, record permitted collaboration and observed sharing, and distinguish the direct assignment effect from spillover-related contrasts. A cluster-level policy effect includes the collective environment that the assignment changes; an individual-level effect in isolation may not transport to that environment. Network evidence is particularly sensitive to missing relationships and the investigator’s choice of relational representation, so neither a graph visualization nor a network regression supplies identification by itself.

Historically occurring social recognition is more open-ended than a laboratory network. A new index, translation, publication, or intermediary may change one observer’s search behavior and, indirectly, the evidence others encounter. There can be multiple feedback loops and evolving network boundaries. Rather than forcing those dynamics into a predetermined chronology, a longitudinal design can document dated record availability, exposure opportunities, actual consultations, observed attribution statements,

and later corrections. The inferential target should be specified at the scale justified by the data: a change in access conditions, a participant's judgment following an encounter, or a diffusion pattern across a social network. Historical sequences alone rarely isolate a uniquely cognitive causal pathway. Observational case studies can enrich mechanism formation and help identify plausible boundary conditions, with causal conclusions qualified accordingly.

9.4 Missing Outcomes, Construct Validity, and Researcher Degrees of Freedom

Data on cognitive stages will generally be incomplete. Some participants will fail to open a source, some will decline a memory probe, and some will discontinue a task. Such patterns may be substantively important: refusal to examine evidence can itself be a behavioral response to the intervention. Other missing outcomes can reflect fatigue, comprehension difficulties, technical incompatibility, accessibility constraints, or ethical withdrawal. Reporting only participants with a complete series of responses may therefore change the participant population and create differences between assigned conditions. The missing-data literature distinguishes circumstances in which a response process may be treated as ignorable from circumstances requiring explicit modeling and sensitivity analysis [67]. Random assignment does not, by itself, justify the assumption that post-assignment missingness is independent of the unobserved outcome.

Let $R_i(z)$ indicate whether the principal outcome would be measured under assignment z , and $Y_i(z)$ its value whether observed or missing. Equation 63 expresses the mean that a complete-case analysis estimates in an assigned condition.

$$\mathbb{E}[Y_i \mid Z_i = z, R_i = 1] = \mathbb{E}[Y_i(z) \mid R_i(z) = 1]. \quad (63)$$

The equality in Equation 63 assumes randomized assignment, treatment consistency, and no relevant interference; it does not equate the selected conditional mean to the full-population quantity $\mathbb{E}[Y_i(z)]$. If the intervention changes which participants finish the study, the complete-case contrast can compare different selected populations. A defensible analysis should report retention by assignment and outcome, characterize observable predictors of missingness, and provide sensitivity analyses for unobserved outcomes. For a binary outcome, transparent worst-case bounds can sometimes be obtained using observed counts and the possible values of missing cases. Narrower conclusions require further restrictions justified by the actual design and evidence; they cannot be manufactured by treating the unanswered questionnaire as a negative answer.

For an outcome known to lie in $[0, 1]$, a simple design-based bound makes the inferential loss concrete. Write $p_z = P(R_i(z) = 1)$ and $\mu_{z,\text{obs}} = \mathbb{E}[Y_i(z) \mid R_i(z) = 1]$ when $p_z > 0$. With no information about the unmeasured outcomes other than their range, Equation 64 bounds the full-condition mean:

$$p_z \mu_{z,\text{obs}} \leq \mathbb{E}[Y_i(z)] \leq p_z \mu_{z,\text{obs}} + (1 - p_z). \quad (64)$$

Under the trial's randomization and measurement conditions, observed assignment-specific completion rates and responses can estimate the quantities on the right-hand side. Subtracting the upper bound for z_0 from the lower bound for z_1 , and conversely, gives a bounded interval for an assignment contrast. Those bounds may be wide; their width is itself informative about what was not measured. Parametric missingness models and multiple imputation are not automatic improvements. Their assumptions should be recorded, and sensitivity scenarios should cover withdrawal related to unmeasured comprehension, source skepticism, or fatigue.

In observational studies, a complementary difficulty arises before missing-outcome analysis: access to documentary evidence can be related to prior expertise, institutional privilege, motivation, and the eventual recognition outcome. A comparison of observers with and without original-source access then mixes the consequences of access with selection into access. A credible design should describe the target intervention, the backdoor paths it seeks to block, observed overlap, and how poorly measured covariates could affect the result. Structured sensitivity analysis for residual confounding can be informative, but summary metrics apply only on the scales and under the assumptions for which they were defined

[81]. No scalar robustness statistic replaces a substantive study of selection into archives, institutions, or evaluation opportunities.

Construct validity presents a parallel problem. Source-recognition accuracy, confidence calibration, trust in a communicator, and reliance under an incentive scheme are distinguishable constructs. Their responses may nevertheless be correlated, and any particular measurement instrument may partly capture several of them. A question labeled “trust” can be interpreted as reliability, honesty, familiarity, willingness to collaborate, or confidence in one claim. Surveys should identify the response target, specify a reference time and task, and assess comprehension across participant groups. Translation of test materials should preserve the functional meaning of instructions as well as lexical content, with ambiguity explicitly documented. Linguistic, cultural, disciplinary, and accessibility differences may alter measurement conditions. Comparative results need evidence that their instruments measure sufficiently comparable constructs; translated wording alone does not establish equivalence.

Where multi-item scales are used, analyses can investigate whether item structure and loadings behave comparably across languages, accessible presentation modalities, expert strata, and repeated encounters. Configural, metric, and scalar conditions answer different comparison questions; the strongest condition is not needed for every estimand, but a claim about latent means requires a defensible link between the response scale and the construct across groups [63]. For direct behavioral outcomes, invariance questions do not disappear: the meaning of a source-opening action can change when one interface requires extra steps or a participant uses assistive technology. Researchers should document differential item functioning, equivalent task affordances, response-mapping ambiguity, and the limits of any translated benchmark. Evidence that a manipulation changes one group’s observed score more than another group’s score does not alone license a claim about group differences in underlying competence.

Data-processing decisions also change apparent results. Investigators must decide how to score partial-order answers, combine repeated tasks, handle response times, exclude failed comprehension trials, and aggregate confidence ratings. Each choice can alter the statistical contrast. A preregistered primary analysis should state its outcome scales, handling of unanswered questions, exclusion rules, family of planned hypotheses, and uncertainty intervals. Supplementary analyses may investigate sensible alternative scoring and selection rules. The principle of examining a range of defensible analysis decisions is supported by methodological work on multiverse analyses [74]. Such analysis should not be converted into a search for a preferred significant result. Report the whole prespecified range, identify consequential decisions, and distinguish exploratory results from confirmatory tests.

A reproducible report should contain the synthetic materials or a reproducible generator, condition-assignment probabilities, interface versions, an analysis plan with time-stamped deviations, documented scoring rules for partial orders and abstentions, de-identified data where disclosure is ethically possible, and code sufficient to reproduce figures and estimates. A preregistration is an explicit record of prior analytic commitments, rather than a guarantee of theoretical validity [59]. Investigators should report effect sizes and uncertainty intervals, substantively informative null or reversed estimates, multiplicity handling for families of outcomes, and failed assumptions. If the target is a bound or a descriptive pattern rather than a point estimate, that target should be reported as such. Findings should not be reformulated as universal psychological laws when they are supported only by a particular sample, document generator, task, and instrument.

Table 9 summarizes a feasible stage-resolved measurement programme. Its columns deliberately separate target constructs from the quantities an experiment can directly observe and the ambiguities that persist.

Hypothesis 9.1 (Stage-sensitive discrimination under provenance disclosure). In a controlled documentary task that matches substantive evidence and report fluency, accurate and comprehensible disclosure of report lineage should improve discrimination between independent and derivative testimony, relative to an interface that conceals lineage. The predicted gain concerns source discrimination; improvements in confidence calibration, overall trust, or reliance require separate tests. High information density, poor interface comprehension, and misleading metadata may attenuate or reverse the proposed effect.

Table 9: Stage-sensitive research designs for cognitive recognition and their inferential boundaries.

Analytical target	Manipulation and observable	Identification safeguard
Evidence encounter	Randomized access and salience; source-opening logs	Opening does not demonstrate comprehension; report assignment effects
Content and source memory	Delayed content recognition and origin discrimination	Probe reactivity, guessing and reconstructive error
Relational reconstruction	Partially ordered history task; provenance dependency response	Permit unresolved orders; stipulate documentary ground truth
Interpretive attribution	Randomized criteria; criterion-specific contribution rating	Distinguish warranted criterion differences from inconsistency
Belief revision	Elicited priors and posteriors; diagnostic evidence variation	Elicitation effects and the defended benchmark model
Testimony evaluation	Independent versus derivative source lineages	Preserve common-origin structure; separate fluency from independence
Declared trust and reliance	Target-specific trust question; stakes and action choice	Feasible alternatives, incentives, and obligations affect behavior
Historical extension	Index or access change; dated consultations and responses	Time-varying selection, confounding, spillovers and record changes

Hypothesis 9.2 (Dissociation between evaluation and practical response). When proposition-specific evidence and the elicitation of belief are held comparable, experimentally changing the consequences and reversibility of a reliance decision should change action patterns even if mean probability judgments remain stable. The claim is conditional on comprehension of the payoff structure and on viable alternative actions; any observed belief changes should be reported rather than assumed absent.

The first hypothesis is directed at the distinction between document lineage and repeated appearance. The second separates decision-theoretic effects from epistemic belief revision. Neither requires a fixed inventory of event types or a single canonical model of historical relations. The experimental objects can be constructed under multiple representations, and an analyst can test how changing the coding of a relational history alters measured effects.

Hypothesis 9.3 (Probe-induced change in source discrimination). For a source-discrimination task whose documentary evidence and assigned interface remain comparable, the introduction of a source-origin probe before final assessment will change the distribution of final source-origin classifications relative to a deferred-probe condition. The direction of change is intentionally unspecified: probing can facilitate retrieval, heighten scrutiny, distract from substantive evidence, or introduce an unwanted hint. A null effect with informative precision or a result limited to particular tasks would delimit the measurement-reactivity account. Randomized probe timing estimates an intervention effect and cannot, on its own, identify a unique cognitive mediator.

This is particularly important when new evidence changes the appropriate individuation of events or reveals previously unrecorded dependencies. Stability across alternative justified representations would strengthen an empirical claim; sensitivity would indicate a boundary of its generalization.

9.5 Ethical Conditions and Historical Generalization

Controlled evidence experiments make it possible to stipulate an adjudicable record and to avoid exposing identifiable people to unsupported allegations about originality, competence, or integrity. Synthetic documentary histories should be realistic enough to test retrieval and source attribution, but they should be expressly identified as experimental materials and constructed without appropriating a real individual's disputed contribution history. Participant information should explain relevant tasks, data collection, and any authorized use of temporary deception, with a proportionate debriefing procedure when deception is methodologically justified. Personal identifiers and any collected gaze or behavioral traces require

minimization, security safeguards, retention limits, and consent procedures consistent with applicable institutional and legal requirements. Claims about source competence and integrity should be framed as task-specific judgments, not generalized moral labels.

Researchers should also distinguish ethically relevant consequences borne by different participants and groups. An intervention that improves mean source discrimination may increase effort, fatigue, or accessibility burdens for certain observers. A highly detailed provenance display can help experts yet overwhelm people with limited time, screen-reader compatibility, or domain familiarity. Research design should therefore evaluate burden, comprehension, and opportunity costs as outcomes alongside accuracy. The practical objective is not to maximize skepticism indiscriminately. Some forms of necessary reliance permit cooperation when exhaustive individual verification is infeasible. An ethically defensible interface supports proportionate examination, uncertainty expression, correction, and accessible routes to pertinent evidence.

Historical-materialist interpretation calls for attention to the conditions under which cognitive performance is realized. Training, language resources, time for checking, maintained repositories, access rights, and the distribution of verification labor affect which tasks an observer can complete. Two observers with similar cognitive abilities may exhibit different measured recognition because their available resources and permissions differ. Experimental randomization can isolate some of these factors locally; the broader historical processes producing them remain outside the identification of a laboratory treatment contrast. A multi-site programme can compare different institutional arrangements and linguistic settings, yet cross-site measurement equivalence, case selection, and differences in the meaning of recognition tasks would still require direct investigation.

A laboratory study's average assignment effect is defined for the participants and documentary tasks entering that study. Extending it to librarians, public researchers, multilingual professional communities, or evaluators operating under restricted-access institutions requires a distinct transport question. For a pretreatment covariate vector X sufficient to characterize relevant effect heterogeneity, a proposed target-population effect might be expressed as Equation 65:

$$\tau_T(z_1, z_0) = \int \tau_S(z_1, z_0 \mid x) dF_T(x). \quad (65)$$

Here $\tau_S(\cdot \mid x)$ is the trial-population conditional effect and F_T the target-population covariate distribution. The expression is justified as a transport formula only under suitable invariance of those conditional effects, sufficient covariate coverage and overlap, compatible definitions of the treatment and outcome, and defensible sampling or adjustment methods. Equation 65 is not a general law of recognition and does not imply that a laboratory task's scoring rubric retains its meaning in a different institution. Work on generalizability of experiments illustrates both the relevance of differences between study samples and intended populations and the limits of adjustment when their characteristics scarcely overlap [77].

Historical comparisons introduce further nontransportable mechanisms. An index change can occur alongside a new funding rule, a change in staffing, a platform redesign, or shifts in the value of producing documentary evidence. Organizations also may respond to anticipated evaluation rules, while those rules are themselves revised after disputes. Each causal target must specify the intervention version, its implementation setting, the observational interval or partial order of relevant records, and the comparison that remains plausible. Documentary traces preserve evidence about processes that occurred; they rarely capture every opportunity not taken or every suppressed attempt to contest an attribution. A controlled finding concerning recognition of a fictional inventor cannot alone establish the social distribution of credit, the legitimacy of an institution, or long-term changes in knowledge production. These require complementary historical, institutional, and participatory evidence and a separate normative argument [70].

The interface with the companion papers should remain explicit. Paper II concerns historically changing structures through which evidence becomes accessible, while this paper specifies cognitive transitions conditional on that access. Paper IV will examine power over opportunities to define and govern episodic arrangements; the present section treats those opportunities as contextual variables or externally

assigned interventions when empirical identification permits. Paper V concerns broader circulation of relational warrants and the associated consequences for collective value recognition. The designs proposed here can inform those investigations, but a measured laboratory improvement in source discrimination does not automatically establish political legitimacy, distributive justice, or beneficial social circulation. Such outcomes require independent concepts and evidence.

A feasible research programme should proceed cumulatively. An initial preregistered experiment can test a limited two-factor comparison with a small number of primary outcomes and a synthetic ground truth. A second study can add independent measurements of content memory, source attribution, and interpretation, explicitly testing the reactivity of those measurements. A third study can investigate repeated encounters and source dependency under randomized task histories. Subsequent observational work can examine historical changes in discovery infrastructures, with cautious claims about causation and extensive documentation of what remained unobserved. Across studies, reports should retain null findings, unexpected reversals, alternative justified codings, and unresolved causal pathways. These outcomes are intellectually significant: they establish where the conditional framework fails, where it remains too coarse, or where a rival account more adequately describes cognitive recognition.

This section provides an empirical interpretation of the full architecture without converting its formal order into an assumed psychological chronology. Controlled assignment can estimate effects of specified evidentiary interfaces. Stage-sensitive measures can narrow the family of plausible cognitive explanations while remaining imperfect proxies. Causal mediation, missingness, repeated observation, and social interference impose additional assumptions that must be checked or acknowledged. Historically situated comparisons can enrich the analysis of material conditions for recognition without deriving institutional consequences directly from laboratory behavior. Section 10 will synthesize these contributions, their limits, and their relation to the larger research programme.

10 Conclusion

The inquiry developed in this paper concerns the cognitive conditions under which an observer recognizes and evaluates an evidence-bearing relational history. Its objective has been to make the interval between documentary access and human assessment analytically tractable, while preserving the plurality of cognitive mechanisms through which evidence is selected, remembered, interpreted, and used. The resulting account is a conditional research framework. It identifies distinctions, formal benchmarks, and empirically discriminable hypotheses, with recognition, justified assessment, and practical reliance treated as related but separate outcomes. Its contribution is best understood as an integration directed toward relational histories whose provenance, dependencies, and historical ordering may be incomplete or contested. The framework neither establishes a universal sequence of mental operations nor reports that its proposed pathways have been confirmed in participants.

The theoretical claim concerns a relation among levels of description, rather than a new unitary variable called recognition. An archival system can represent evidence, an observer can form a belief about that evidence, and an organization can treat the resulting judgment as a reason for action. Success at one level is not sufficient for success at the others. In particular, an accurate answer on a constrained attribution task can coexist with an inability to explain the source chain, just as an accurate reconstruction of a source chain can coexist with a disputed evaluation of a contribution. The paper therefore recommends defining outcomes against a specified proposition, documentary record, interpretive criterion, observer population, and practical decision. Without that specification, aggregate “recognition” measures may mix different accomplishments and different errors. This is a conceptual and methodological recommendation, not a claim that one measurement instrument can exhaust the meaning of recognition.

The first conclusion concerns the discontinuity between the availability of evidence and its cognitive realization. A repository or evidentiary interface may make a record accessible to an observer who never inspects it. The observer may inspect the record without retaining its source, remember its content while misidentifying its provenance, or reconstruct the provenance accurately while applying a different evaluative criterion. These possibilities have different explanations and admit different corrective interventions.

Section 2 specified an observer-, task-, and representation-conditioned family of processing kernels for describing those transitions. The proposed factorization provides a way of recording dependencies and making assumptions explicit. It leaves open whether a given experiment or historical setting requires recurrent, parallel, partially ordered, or jointly determined processing models. An externally recorded order of presentation does not, by itself, establish a uniquely ordered sequence of psychological states.

This distinction matters for the study of intellectual contributions and other relational histories. The historical record made available to a researcher is usually assembled through documentary practices, descriptions, classifications, and selective preservation. What qualifies as a distinct contribution or relevant relation depends in part on the question being asked. The representation of an archive as a single narrative can conceal uncertain provenance, joint activity, contested interpretation, or evidence supporting only a partial ordering. Cognitive recognition therefore includes the capacity to preserve uncertainty about origins and relations where the available record warrants it. A participant who selects an “undetermined” source or refuses to assign a unique chronology may perform the epistemic task more appropriately than one who supplies an apparently complete but unsupported history. Measurement systems that reward confident completion alone risk misclassifying responsible uncertainty as failed recognition.

The available representation also shapes what can count as a correction. A graph of citations, a chronological narrative, and an interactive archive can expose different relations from an identical underlying set of documents. Their apparent agreement cannot automatically establish a representation-independent history, because the documents themselves may be missing, selectively preserved, or assigned different evidentiary roles. An observer may rationally defer an attribution until competing source hypotheses can be compared, yet this act of deferral can be invisible to an evaluation rubric that demands a single author, date, or originating node. An appropriate task design should therefore distinguish unsupported completion, contradiction by the stipulated record, accurate reconstruction, and defensible nonresolution. A system that leaves provenance visibly contestable can be more informative than one that resolves uncertainty through silent interpolation. Whether such a design improves actual human judgment remains an experimental question.

A second conclusion concerns selective encounter. Section 3 distinguished statistical surprisal relative to an observer’s reference distribution from perceptual salience, task relevance, expected informational benefit, and processing cost. These attributes can vary independently. An unusual encounter may redirect attention toward a pertinent earlier record, while a familiar and unremarkable document may carry the most diagnostic information for a source-attribution task. High surprise is therefore a candidate influence on attention allocation, subject to available resources, rather than a general measure of truth, contribution quality, or warranted trust. Equally, an interface that succeeds in capturing attention may reduce the attention available for corroborating materials. Whether its net effect is beneficial depends on the evidentiary task, the alternatives displaced, and the consequences of incomplete examination. Empirical work should compare such outcomes under specified workloads rather than infer epistemic improvement from engagement or novelty alone.

An important consequence is that the epistemic effect of highlighting one record depends on the records it displaces. Search ordering, alert timing, citation previews, and the number of verification steps can change a finite observer’s opportunity set even when the underlying archive is held fixed. Improvements in initial encounter may therefore be accompanied by worse source discrimination, reduced scrutiny of less conspicuous alternatives, or unequal gains between observers with different prior expertise. This possibility calls for comparisons of complete task performance, including omissions and verification costs, rather than attention captured per item. It also motivates recording the reference distribution relative to which “novelty” or “surprise” was computed; a surprising event for one observer may be familiar to another, and surprise alone carries no verdict about epistemic value.

The treatment of memory extends this analysis into the reconstruction of relationships. Section 4 differentiated remembering a documentary proposition from remembering its origin and reconstructing relations among its producers, summaries, and subsequent applications. A coherent recollection can preserve a claim while substituting a familiar commentator for its originating source. Repetition may increase accessibility and subjective familiarity without increasing provenance accuracy. The problem

becomes especially difficult when an observer receives derivative representations that omit source identifiers or compress distinctions among actors and contributions. The formal source-identification result established only a conditional limit for indistinguishable evidence under specified prior assumptions; it did not establish a universal bound on human memory. The associated experiments need independent measures of content recognition, source discrimination, and uncertainty about the relational ordering. They also need to consider that the act of measuring memory may itself change subsequent recall and inquiry.

Memory-based recognition makes epistemic repair a distinct research object. A provenance reminder might recover an omitted source, but the same reminder can direct attention, supply new information, or prime a response believed to be desirable. Source-tag displays and retrieval exercises should consequently be evaluated against carefully matched alternatives, with measurements of both immediate answers and later independent reconstruction. The paper's formal indistinguishability argument does not imply that documentary verification is futile. It specifies a reason that additional discriminating records may matter, while leaving open whether an individual observer will locate, understand, and incorporate them. In practical settings, a durable link to a primary record may help an observer re-open a claim without presuming that human memory faithfully preserves the prior encounter. What should be retained is the possibility of correction, alongside an explicit account of what remained uncertain.

Interpretation provides a further source of heterogeneity among observers who have encountered and remembered similar records. Section 5 separated recognizing what someone demonstrably did from evaluating the significance of that action for originality, competence, responsibility, or technical reliability. Distinct tasks may legitimately require different standards. A researcher evaluating conceptual priority may inspect different features from a reviewer assessing whether an implementation is reproducible. Disciplinary expertise and language can alter which distinctions are intelligible and which ambiguities remain unresolved. Motivated reasoning and schema-consistent assimilation are additional candidates for distortion, but disagreement alone does not establish their presence. A credible investigation must specify the evaluative question, identify admissible evidence, and compare the observer's application of a criterion with an appropriate task-specific standard. The cognitive framework thereby accommodates both justified plural assessments and errors arising from incomplete or selectively construed evidence.

Attribution also has a counterfactual dimension that a static accuracy score may miss. If the same documented action is assigned to a well-known scholar rather than an unfamiliar contributor, an altered assessment could arise from a legitimate belief about domain expertise, an irrelevant prestige cue, or a changed interpretation of the task. These explanations require different controls. Concealed and disclosed identity conditions can reveal an average cue effect on a specified outcome, but they cannot, without further measures, identify the participant's underlying reason. A criterion-sensitive analysis should ask whether reputation provides evidence pertinent to the exact claim, whether the rubric licenses that evidence, and whether comparable information is available for differently situated contributors. Similar caution applies to translation and automated summaries: a fluent account can aid comprehension while deleting distinctions that the substantive attribution task requires.

The paper's account of belief revision provides a normative comparison without converting normative consistency into a descriptive psychology. Section 6 distinguished an observation's unlikelihood from its likelihood ratio with respect to a specified proposition. An event can be unexpected under a reference distribution yet have little bearing on whether an attribution claim is correct. Evidence can also be unsurprising while supplying strong reasons to revise an established judgment. Bayesian updating offers a conditional benchmark when hypotheses, observation models, and prior information are well defined. Actual observers may approximate, simplify, or depart from this benchmark because of limited resources, mistaken dependencies, inaccessible alternatives, or other mechanisms. Positive, null, and negative revisions are therefore possible outcomes of newly examined evidence. A decrease in confidence after encountering a previously hidden inconsistency can represent improved epistemic calibration, even if it appears as diminished trust in a simple rating instrument.

One further implication concerns the evidentiary meaning of absence. An unsuccessful search for an earlier document may update belief about a source claim only to the extent that the search procedure

would probably have found the document had it existed in the relevant collection. Coverage, indexing decisions, search languages, archival loss, and the observer's query strategy matter to this likelihood. A failure to locate material is thus different from a documented contradiction, and repeated unsuccessful searches are not automatically independent negative evidence. Stating these conditions guards against a common slide from "not encountered" to "not produced." The present framework cannot reconstruct an unseen historical contribution from a formal probability alone; it can, however, make the assumptions behind an inference of absence inspectable and open to challenge.

The distinction between corroboration and circulation sharpens the treatment of communicated evidence. Section 7 considered testimony about the same relational history as potentially dependent on earlier reports. Several polished explanations, endorsements, or machine-generated summaries may derive from one evidentiary origin. Their numerical multiplication does not, without additional assumptions, multiply the underlying likelihood evidence. The deterministic-retelling result expresses a particularly clear case: a fully known transformation of an already observed report supplies no additional likelihood information about the original target proposition. More complex mediated chains may omit, combine, or add information and must be analyzed with their transmission and source-dependence assumptions made explicit. Derivative representations may improve intelligibility or enable access to a document that otherwise remains unexamined, even while providing no independent corroboration of its substantive assertion. These distinct benefits and risks explain why source lineage is relevant to cognitive vigilance in both interpersonal and human-AI-mediated communication.

There are also distinct meanings of verification along a mediated chain. Checking that an article exists, checking that a citation points to that article, checking that the article states the attributed claim, and assessing whether the claim is supported by primary evidence are different activities. A citation badge or machine-generated reference may raise perceived credibility while completing none of the latter tasks. A derived account can nevertheless assist legitimate inquiry when it preserves links, distinguishes quotation from interpretation, and allows an observer to inspect conflicting sources. Thus the appropriate comparison for automated mediation is neither indiscriminate acceptance nor indiscriminate rejection. It is a task-dependent analysis of the information added, lost, selected, or made practically examinable, together with the costs imposed on the person asked to verify it.

A further conclusion separates judgments about evidence from decisions undertaken on their basis. Section 8 distinguished epistemic warrant, subjective confidence, declared trust, and observable reliance. Beliefs influence decisions in conjunction with action costs, available alternatives, the feasibility of verification, and institutional constraints. Identical subjective confidence can rationally accompany divergent decisions when the consequences of an error differ. The decision-threshold result established this implication within a deliberately restricted expected-utility model, and its generality is limited by that model's assumptions. A refusal to delegate can express reasonable caution rather than generalized distrust; a decision to proceed can reflect scarce alternatives rather than an unqualified endorsement. Moreover, reliance can change the evidence subsequently encountered, including through feedback, experience, and selective further consultation. Consequently, final actions and confidence reports do not uniquely reveal the earlier cognitive pathways that produced them.

Reliance outcomes need equally careful interpretation. In a numerical-advice task, the degree to which a participant moves toward a recommendation can be separated from whether the recommendation was correct under a stipulated criterion. For institutional decisions, however, reliance may consist of supervised use, selective delegation, verification before action, or use required by policy. Observed uptake under compulsion is weak evidence of voluntary trust; refusal under high error costs is weak evidence of poor epistemic judgment. Experiments on human-AI advice should consequently measure acceptance of both correct and incorrect recommendations and specify what alternatives, stakes, and supervision were available. An increase in average acceptance can be undesirable when it reflects more acceptance of incorrect advice, while reduced acceptance may be desirable when it is better targeted. These judgments depend on task-specific loss and warrant, not a universal preference for higher trust.

Taken together, the formal propositions in Sections 2–8 support methodological cautions of a limited kind. Observational equivalence can arise when different latent processes lead to the same final

rating. Source-identification performance can be limited by evidence that fails to distinguish candidate origins. Repeated reports can share a common evidentiary basis. Reliance thresholds can shift with task consequences while reported probabilities remain constant. Such results follow from their stated formal assumptions. They cannot independently establish that a particular observer experienced an identified memory error, that an interface caused a specific bias, or that one institutional arrangement systematically increases warranted trust. These empirical claims require measurements, interventions, and model comparisons beyond those contained in the present theoretical manuscript.

The formal results should also be read with their quantifiers intact. A conditional independence claim holds relative to chosen variables and a specified generative or statistical model; an information-processing inequality concerns the defined transformation and observations; a decision threshold follows from the loss function and action set under study. None provides an unrestricted ontology of mental states or a general proof about institutions. The same caution applies to any precedence relation imposed on an evidence trajectory: documentary timestamps, recorded dependencies, hypothesized production relations, and experimentally observed cognitive events are not interchangeable orderings. A useful representation may organize inquiry without being uniquely determined by the phenomenon represented. Making this conditionality visible is a source of theoretical precision rather than a weakness to conceal.

Section 9 therefore proposed a cumulative research programme oriented toward discriminating among rival explanations. Experiments could vary presentation salience, documented provenance, or the availability of primary records while keeping a stipulated documentary history stable. Stage-sensitive measures could record document encounters, source-memory discrimination, interpretation of specified claims, belief revisions, and decisions under explicitly varied consequences. The distinction between random assignment and inferred mediation remains essential: a randomized interface may identify a total effect on a defined outcome while leaving its allocation among attention, memory, interpretation, and decision pathways unidentified. Intermediate measurements are themselves imperfect, potentially reactive, and sometimes inaccessible to people using different reading or assistive practices. A robust analysis would test measurement equivalence where possible, report response missingness, retain accessible denominators, and acknowledge dependent observations or interference among participants.

Identification requires particular care where intermediate outcomes are measured after treatment assignment. Conditioning an interface comparison on documents that participants chose to open can introduce selection, because opening is itself potentially influenced by the assigned display and by unmeasured motivation. Similarly, an intermediate source-memory question may change the final assessment it is intended to explain. Section 9 therefore distinguishes randomized assignment effects from indirect effects requiring additional assumptions and from observational patterns that support multiple cognitive accounts. For bounded but missing outcomes, partial-identification intervals can be more defensible than a single estimate obtained by unverifiable imputations. For transport beyond the experimental sample, coverage, measurement comparability, and stability of relevant conditional responses require separate justification. These constraints are especially salient when comparing experts, nonexperts, languages, accessibility modes, and institutions.

A convincing empirical result would go beyond showing that additional contextual information changes an average trust score. One more discriminating pattern would combine verified source-identification improvement with better calibrated proposition-specific belief revision, while separately documenting the effort and costs required to obtain that improvement. A different pattern could show improved encounter rates but unchanged interpretation accuracy; this would support an attention effect while constraining claims about later cognition. An interface could raise confidence and reliance without improving discrimination, indicating a potentially important divergence between subjective persuasion and epistemic warrant. Equally informative would be null effects, reversals under higher processing load, or cases where a simpler model based on prior expertise explains the observations as well as the proposed multi-stage architecture. These outcomes should be treated as opportunities to revise the framework rather than exceptions explained away by an unmeasured general appeal to complexity.

A stronger test of this framework would specify in advance what findings would count against a proposed decomposition. If source-memory measures add no predictive or explanatory value after con-

trolling for record encounter and prior expertise, the corresponding stage may be dispensable for that task. If a staged model fails to predict how one intervention changes another outcome, a recurrent or simpler joint model may be more adequate. If a provenance-rich interface increases the ability to report citations but leaves claim discrimination unchanged, the evidence supports a narrower advantage than the one initially sought. Cross-task consistency and out-of-sample performance matter because a flexible collection of latent explanations can fit an isolated result without specifying a distinctive mechanism. Such tests should compare plausible alternatives, include negative and ambiguous outcomes, and separate predictive adequacy from causal interpretation.

The historical setting gives these questions significance beyond laboratory information-processing tasks. Contemporary knowledge practices involve expanding archives, search systems, translated descriptions, automated summaries, and chains of institutional or interpersonal recommendation. These infrastructures are maintained through labor and distribute the opportunities and costs of verification unevenly. A record can be technically retrievable while remaining practically inaccessible to an observer who lacks the relevant language, expertise, time, or authority to examine it. Conversely, an apparently rich flow of derivative descriptions may leave a source history difficult to reconstruct. The paper's historical-materialist orientation asks how such material arrangements condition the realization of cognitive abilities, without deriving a fixed cognitive outcome from a particular ownership form or assuming that the expansion of information production produces epistemic progress automatically. Historical interpretation will require independent evidence about specific institutions, technologies, and forms of evaluative labor.

A historical-materialist inquiry adds a question concerning who has the capacity to undertake this work. The labor of writing documentation, maintaining archives, translating terminology, verifying origins, and reviewing claims is distributed across persons and organizations with unequal time, remuneration, infrastructure, and authority. The cognitive opportunity to recognize a prior contribution is consequently linked to materially produced conditions of encounter and evaluation. Yet the historical fact of unequal access does not by itself determine the psychology of a given observer, nor does a controlled attention effect establish the causes of an institution's hierarchy of recognition. These explanatory levels should be investigated together without collapsing one into another. A productive empirical extension would document both source-level cognitive outcomes and the institutional work needed to make those sources available, legible, revisable, and contestable.

These concerns also impose ethical constraints on the investigation of recognition. A study should avoid treating participants' hesitancy to make unsupported judgments as deficiency, or treating agreement with an investigator's preferred interpretation as proof of intellectual competence. Source and contribution claims can affect reputations, opportunities, and trust among people who have not consented to experimental judgment. Hypothetical, non-identifiable documentary histories are appropriate for initial controlled comparisons. Researchers should record how records were selected, whether alternative sources were available, how uncertainty could be expressed, and which burdens a proposed verification interface imposed on different observers. Improving source discrimination is a relevant epistemic objective, but a socially defensible interface must also be usable, corrigible, and attentive to the costs of unnecessary scrutiny. A higher mean accuracy score does not, by itself, settle questions of distributive justice, institutional legitimacy, or the appropriate allocation of risk.

One ethical design implication is to treat the option to verify, decline, or suspend judgment as an outcome worth preserving. A provenance interface may increase cognitive effort for some observers while reducing it for others; mandatory disclosure of exhaustive search trails could compromise privacy or generate burdens disproportionate to a decision's importance. Accuracy, response time, uncertainty reporting, and effort are therefore complementary indicators rather than a single optimization target. Study designers should distinguish a participant's inability to access material from unwillingness to expend additional effort under the offered incentives. Claims about inclusion require attention to accessible formats, linguistic plurality, and who bears the cost of correction. Questions about the legitimacy of rules for distributing those costs extend beyond the cognitive scope of this paper, but the cognitive measurements should not obscure them.

The boundaries of the wider research programme help specify the present contribution. Paper II investigates the relational and infrastructural conditions through which historical evidence becomes observable, including changes that make earlier work accessible to new observers. The present paper examines conditional human processing once evidence enters an observer's task. Paper IV addresses how authority and unequal power can shape evidence access, evaluative rules, and opportunities to contest an assessment. Paper V considers how relational warrants, attestations, and assessments circulate and acquire consequences in wider processes of value creation and institutional interaction. The same empirical setting may implicate all four dimensions, but the concepts retain distinct explanatory roles. Evidence becoming retrievable is not equivalent to its being remembered; an accurate judgment does not automatically authorize an institution to act; and a circulated endorsement does not establish that an independent examination has occurred.

The interfaces among the papers are most useful when they preserve distinct counterfactual questions. One can ask whether changing an archive's accessibility would alter what evidence people encounter, whether changing a display would alter their interpretations given accessible evidence, whether changing an organization's authority structure would alter which assessments become effective, or whether changing a circulation channel would alter whose work receives opportunities for use and return. The interventions may interact, and a single institution may display all four mechanisms. Still, an explanation of one contrast does not substitute for evidence about the others. Paper III contributes observer-level constructs and testable distinctions to that larger programme, while leaving the political legitimacy of decision rules, the ownership of infrastructures, and the circulation of benefits for the companion inquiries.

The proposed contribution is accordingly neither a new discovery of selective attention, reconstructive memory, Bayesian assessment, epistemic vigilance, or trust in automation, nor a replacement for the substantial literatures studying these processes. It is a scoped attempt to articulate their relations in a research problem involving evolving, representation-sensitive histories of contribution. Its strongest theoretical commitments are distinctions that can be stated precisely and subjected to failure: availability versus encounter, content memory versus provenance reconstruction, surprise versus diagnostic evidence, confidence versus warranted assessment, and judgment versus reliance. The conditional processing architecture is valuable only insofar as it enables empirical comparisons more informative than an undifferentiated measure of trust or recognition. Future research should determine which stages require separate modeling, which interactions defeat a convenient staged representation, and which proposed mechanisms meaningfully improve the explanation or prediction of human judgments in actual evidentiary environments.

Human recognition of relational histories is therefore a historically situated accomplishment requiring both evidentiary opportunities and cognitive work. Its quality depends on the fit among the documentary record, the question under examination, the observer's resources, the provenance and independence of sources, and the grounds for subsequent belief or action. An encounter with an unfamiliar person, archive, or representation may open a route to previously overlooked evidence; what follows can be recognition of an earlier contribution, warranted skepticism, or an explicit preservation of uncertainty. Understanding these possibilities requires tracing how evidence is encountered and evaluated rather than treating favorable recognition as the necessary result of greater novelty or connectivity. The present paper offers a conditional and testable vocabulary for that tracing, while leaving its empirical adequacy and broader institutional implications open to further investigation.

References

- [1] Amin Abolghasemi et al. “Evaluation of Attribution Bias in Generator-Aware Retrieval-Augmented Large Language Models”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025, pp. 21105–21124. DOI: 10.18653/v1/2025.findings-acl.1087.
- [2] Liz Allen et al. “Publishing: Credit where credit is due”. In: *Nature* 508.7496 (2014), pp. 312–313. DOI: 10.1038/508312a.
- [3] Tatsuya Amano, Juan P. González-Varo, and William J. Sutherland. “Languages Are Still a Major Barrier to Global Science”. In: *PLOS Biology* 14.12, e2000933 (2016). DOI: 10.1371/journal.pbio.2000933.
- [4] Peter M. Aronow and Cyrus Samii. “Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment”. In: *The Annals of Applied Statistics* 11.4 (2017), pp. 1912–1947. DOI: 10.1214/16-AOAS1005.
- [5] Alan D. Baddeley and Graham J. Hitch. “Working Memory”. In: *Psychology of Learning and Motivation*. Ed. by Gordon H. Bower. Vol. 8. Academic Press, 1974, pp. 47–89. DOI: 10.1016/S0079-7421(08)60452-1.
- [6] Gagan Bansal et al. “Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16. DOI: 10.1145/3411764.3445717.
- [7] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922.
- [8] José M. Bernardo. “Expected Information as Expected Utility”. In: *The Annals of Statistics* 7.3 (1979), pp. 686–690. DOI: 10.1214/aos/1176344689.
- [9] Tom Bielik and Moritz Krell. “Developing and Evaluating the Extended Epistemic Vigilance Framework”. In: *Journal of Research in Science Teaching* 62.3 (2025). First published online 14 September 2024, pp. 869–895. DOI: 10.1002/tea.21983.
- [10] Pedro Bordalo et al. “Stereotypes”. In: *The Quarterly Journal of Economics* 131.4 (2016), pp. 1753–1794. DOI: 10.1093/qje/qjw029.
- [11] Glenn W. Brier. “Verification of Forecasts Expressed in Terms of Probability”. In: *Monthly Weather Review* 78.1 (1950). Publisher DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, pp. 1–3. URL: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.
- [12] Richard W. Brislin. “Back-Translation for Cross-Cultural Research”. In: *Journal of Cross-Cultural Psychology* 1.3 (1970), pp. 185–216. DOI: 10.1177/135910457000100301.
- [13] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, 188 (2021), pp. 1–21. DOI: 10.1145/3449287.
- [14] Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. “Categorization and Representation of Physics Problems by Experts and Novices”. In: *Cognitive Science* 5.2 (1981), pp. 121–152. DOI: 10.1207/s15516709cog0502_2.
- [15] John M. Darley and Paget H. Gross. “A Hypothesis-Confirming Bias in Labeling Effects”. In: *Journal of Personality and Social Psychology* 44.1 (1983), pp. 20–33. DOI: 10.1037/0022-3514.44.1.20.
- [16] A. Philip Dawid. “Conditional Independence in Statistical Theory”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979), pp. 1–15. DOI: 10.1111/j.2517-6161.1979.tb01052.x.
- [17] Robert Desimone and John Duncan. “Neural Mechanisms of Selective Visual Attention”. In: *Annual Review of Neuroscience* 18 (1995), pp. 193–222. DOI: 10.1146/annurev.ne.18.030195.001205.
- [18] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err”. In: *Journal of Experimental Psychology: General* 144.1 (2015), pp. 114–126. DOI: 10.1037/xge0000033.
- [19] Yifan Ding et al. “Citations and Trust in LLM Generated Responses”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 22. 2025, pp. 23787–23795. DOI: 10.1609/aaai.v39i22.34550.

- 4543 [20] Ullrich K. H. Ecker et al. “The Psychological Drivers of Misinformation Belief and Its Resistance to Cor-
4544 rection”. In: *Nature Reviews Psychology* 1 (2022), pp. 13–29. DOI: 10.1038/s44159-021-00006-y.
- 4545 [21] Lisa K. Fazio et al. “Knowledge Does Not Protect Against Illusory Truth”. In: *Journal of Experimental*
4546 *Psychology: General* 144.5 (2015), pp. 993–1002. DOI: 10.1037/xge0000098.
- 4547 [22] Klaus Fiedler. “Beware of Samples! A Cognitive-Ecological Sampling Approach to Judgment Biases”. In:
4548 *Psychological Review* 107.4 (2000), pp. 659–676. DOI: 10.1037/0033-295X.107.4.659.
- 4549 [23] Jessica K. Flake, Jolynn Pek, and Eric Hehman. “Construct Validation in Social and Personality Research:
4550 Current Practice and Recommendations”. In: *Social Psychological and Personality Science* 8.4 (2017),
4551 pp. 370–378. DOI: 10.1177/1948550617693063.
- 4552 [24] Andrew Gelman and Cosma Rohilla Shalizi. “Philosophy and the Practice of Bayesian Statistics”. In:
4553 *British Journal of Mathematical and Statistical Psychology* 66.1 (2013), pp. 8–38. DOI: 10.1111/j.2044-
4554 8317.2011.02037.x.
- 4555 [25] Tilmann Gneiting and Adrian E. Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation”.
4556 In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378. DOI: 10.1198/
4557 016214506000001437.
- 4558 [26] Alvin I. Goldman. “Experts: Which Ones Should You Trust?” In: *Philosophy and Phenomenological Re-*
4559 *search* 63.1 (2001), pp. 85–110. DOI: 10.1111/j.1933-1592.2001.tb00093.x.
- 4560 [27] Jacqueline Gottlieb and Pierre-Yves Oudeyer. “Towards a Neuroscience of Active Sampling and Curiosity”.
4561 In: *Nature Reviews Neuroscience* 19 (2018), pp. 758–770. DOI: 10.1038/s41583-018-0078-0.
- 4562 [28] Dale Griffin and Amos Tversky. “The Weighing of Evidence and the Determinants of Confidence”. In:
4563 *Cognitive Psychology* 24.3 (1992), pp. 411–435. DOI: 10.1016/0010-0285(92)90013-R.
- 4564 [29] Chengcheng Han et al. “Phantom Citations: An Empirical Study of Non-existent and Unverifiable Refer-
4565 ences in Scholarly Literature”. In: *Journal of the Association for Information Science and Technology* 77.10
4566 (2026), pp. 1287–1316. DOI: 10.1002/asi.70104.
- 4567 [30] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Freely available, periodically updated
4568 online book; 2020 cited as bibliographic edition. Boca Raton: Chapman and Hall/CRC, 2020. URL: <https://miguelhernan.org/whatifbook> (visited on 10/27/2026).
4569
- 4570 [31] Kevin Anthony Hoff and Masooda Bashir. “Trust in Automation: Integrating Empirical Evidence on Factors
4571 That Influence Trust”. In: *Human Factors* 57.3 (2015), pp. 407–434. DOI: 10.1177/0018720814547570.
- 4572 [32] Robin M. Hogarth and Hillel J. Einhorn. “Order Effects in Belief Updating: The Belief-Adjustment Model”.
4573 In: *Cognitive Psychology* 24.1 (1992), pp. 1–55. DOI: 10.1016/0010-0285(92)90002-J.
- 4574 [33] Almut Hupbach, Rebecca Gomez, and Lynn Nadel. “Episodic Memory Reconsolidation: Updating or
4575 Source Confusion?” In: *Memory* 17.5 (2009), pp. 502–510. DOI: 10.1080/09658210902882399.
- 4576 [34] Almut Hupbach et al. “Reconsolidation of Episodic Memories: A Subtle Reminder Triggers Integration of
4577 New Information”. In: *Learning & Memory* 14.1-2 (2007), pp. 47–53. DOI: 10.1101/lm.365707.
- 4578 [35] Kosuke Imai, Luke Keele, and Dustin Tingley. “A General Approach to Causal Mediation Analysis”. In:
4579 *Psychological Methods* 15.4 (2010), pp. 309–334. DOI: 10.1037/a0020761.
- 4580 [36] Laurent Itti, Christof Koch, and Ernst Niebur. “A Model of Saliency-Based Visual Attention for Rapid Scene
4581 Analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254–
4582 1259. DOI: 10.1109/34.730558.
- 4583 [37] Larry L. Jacoby. “A Process Dissociation Framework: Separating Automatic from Intentional Uses of Mem-
4584 ory”. In: *Journal of Memory and Language* 30.5 (1991), pp. 513–541. DOI: 10.1016/0749-596X(91)
4585 90025-F.
- 4586 [38] Ziwei Ji et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys*
4587 55.12, 248 (2023), pp. 1–38. DOI: 10.1145/3571730.
- 4588 [39] Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. “Source Monitoring”. In: *Psychological*
4589 *Bulletin* 114.1 (1993), pp. 3–28. DOI: 10.1037/0033-2909.114.1.3.
- 4590 [40] Daniel Kahneman. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

- [41] Carolin Kaiser et al. “A New Era of Online Search? A Large-Scale Study of User Behavior and Personal Preferences during Practical Search Tasks with Generative AI versus Traditional Search Engines”. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Extended abstract, Article 46. Association for Computing Machinery, 2025, pp. 1–7. DOI: 10.1145/3706599.3720123.
- [42] Olav Kallenberg. *Foundations of Modern Probability*. 3rd ed. Cham: Springer, 2021. DOI: 10.1007/978-3-030-61871-1.
- [43] Jeffrey D. Karpicke and III Roediger Henry L. “The Critical Importance of Retrieval for Learning”. In: *Science* 319.5865 (2008), pp. 966–968. DOI: 10.1126/science.1152408.
- [44] Celeste Kidd and Benjamin Y. Hayden. “The Psychology and Neuroscience of Curiosity”. In: *Neuron* 88.3 (2015), pp. 449–460. DOI: 10.1016/j.neuron.2015.09.010.
- [45] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. “Trust and Reliance on AI: An Experimental Study on the Extent and Costs of Overreliance on AI”. In: *Computers in Human Behavior* 160, 108352 (2024). DOI: 10.1016/j.chb.2024.108352.
- [46] Asher Koriati. “Monitoring One’s Own Knowledge During Study: A Cue-Utilization Approach to Judgments of Learning”. In: *Journal of Experimental Psychology: General* 126.4 (1997), pp. 349–370. DOI: 10.1037/0096-3445.126.4.349.
- [47] Solomon Kullback and Richard A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.
- [48] Ziva Kunda. “The Case for Motivated Reasoning”. In: *Psychological Bulletin* 108.3 (1990), pp. 480–498. DOI: 10.1037/0033-2909.108.3.480.
- [49] Dani’el Lakens. “Sample Size Justification”. In: *Collabra: Psychology* 8.1, 33267 (2022). DOI: 10.1525/collabra.33267.
- [50] Nilli Lavie. “Distracted and Confused?: Selective Attention under Load”. In: *Trends in Cognitive Sciences* 9.2 (2005), pp. 75–82. DOI: 10.1016/j.tics.2004.12.004.
- [51] John D. Lee and Katrina A. See. “Trust in Automation: Designing for Appropriate Reliance”. In: *Human Factors* 46.1 (2004), pp. 50–80. DOI: 10.1518/hfes.46.1.50_30392.
- [52] Falk Lieder and Thomas L. Griffiths. “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources”. In: *Behavioral and Brain Sciences* 43, e1 (2020). Published online 4 February 2019. DOI: 10.1017/S0140525X1900061X.
- [53] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. “Algorithm Appreciation: People Prefer Algorithmic to Human Judgment”. In: *Organizational Behavior and Human Decision Processes* 151 (2019), pp. 90–103. DOI: 10.1016/j.obhdp.2018.12.005.
- [54] Charles G. Lord, Lee Ross, and Mark R. Lepper. “Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence”. In: *Journal of Personality and Social Psychology* 37.11 (1979), pp. 2098–2109. DOI: 10.1037/0022-3514.37.11.2098.
- [55] Roger C. Mayer, James H. Davis, and F. David Schoorman. “An Integrative Model of Organizational Trust”. In: *Academy of Management Review* 20.3 (1995), pp. 709–734. DOI: 10.5465/amr.1995.9508080335.
- [56] Robert K. Merton. “The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered”. In: *Science* 159.3810 (1968), pp. 56–63. DOI: 10.1126/science.159.3810.56.
- [57] National Information Standards Organization. *Contributor Role Taxonomy (CRediT)*. ANSI/NISO Z39.104-2022; online resource continuously maintained. 2022. URL: <https://credit.niso.org/> (visited on 10/27/2026).
- [58] Raymond S. Nickerson. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”. In: *Review of General Psychology* 2.2 (1998), pp. 175–220. DOI: 10.1037/1089-2680.2.2.175.
- [59] Brian A. Nosek et al. “The Preregistration Revolution”. In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2600–2606. DOI: 10.1073/pnas.1708274114.
- [60] Raja Parasuraman and Victor Riley. “Humans and Automation: Use, Misuse, Disuse, Abuse”. In: *Human Factors* 39.2 (1997), pp. 230–253. DOI: 10.1518/001872097778543886.
- [61] Luis Pérez-González. “Context in Translation and Interpreting Studies”. In: *The Cambridge Handbook of Language in Context*. Ed. by Jesús Romero-Trillo. Cambridge University Press, 2023, pp. 371–392. DOI: 10.1017/9781108989275.018.

- 4643 [62] Lawrence D. Phillips and Ward Edwards. “Conservatism in a Simple Probability Inference Task”. In: *Journal of Experimental Psychology* 72.3 (1966), pp. 346–354. DOI: 10.1037/h0023653.
- 4644
- 4645 [63] Diane L. Putnick and Marc H. Bornstein. “Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research”. In: *Developmental Review* 41 (2016), pp. 71–90. DOI: 10.1016/j.dr.2016.06.004.
- 4646
- 4647
- 4648 [64] Charan Ranganath and Maureen Ritchey. “Two Cortical Systems for Memory-Guided Behaviour”. In: *Nature Reviews Neuroscience* 13 (2012), pp. 713–726. DOI: 10.1038/nrn3338.
- 4649
- 4650 [65] Evan F. Risko and Sam J. Gilbert. “Cognitive Offloading”. In: *Trends in Cognitive Sciences* 20.9 (2016), pp. 676–688. DOI: 10.1016/j.tics.2016.07.002.
- 4651
- 4652 [66] III Roediger Henry L. and Kathleen B. McDermott. “Creating False Memories: Remembering Words Not Presented in Lists”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21.4 (1995), pp. 803–814. DOI: 10.1037/0278-7393.21.4.803.
- 4653
- 4654
- 4655 [67] Donald B. Rubin. “Inference and Missing Data”. In: *Biometrika* 63.3 (1976), pp. 581–592. DOI: 10.1093/biomet/63.3.581.
- 4656
- 4657 [68] Daniel L. Schacter. “The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience”. In: *American Psychologist* 54.3 (1999), pp. 182–203. DOI: 10.1037/0003-066X.54.3.182.
- 4658
- 4659 [69] Daniel L. Schacter, Scott A. Guerin, and Peggy L. St Jacques. “Memory Distortion: An Adaptive Perspective”. In: *Trends in Cognitive Sciences* 15.10 (2011), pp. 467–474. DOI: 10.1016/j.tics.2011.08.004.
- 4660
- 4661 [70] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin, 2002. ISBN: 0395615569.
- 4662
- 4663 [71] Claude E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
- 4664
- 4665 [72] Herbert A. Simon. “A Behavioral Model of Rational Choice”. In: *The Quarterly Journal of Economics* 69.1 (1955), pp. 99–118. DOI: 10.2307/1884852.
- 4666
- 4667 [73] Dan Sperber et al. “Epistemic Vigilance”. In: *Mind & Language* 25.4 (2010), pp. 359–393. DOI: 10.1111/j.1468-0017.2010.01394.x.
- 4668
- 4669 [74] Sara Steegen et al. “Increasing Transparency Through a Multiverse Analysis”. In: *Perspectives on Psychological Science* 11.5 (2016), pp. 702–712. DOI: 10.1177/1745691616658637.
- 4670
- 4671 [75] Emma Steigerwald et al. “Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future”. In: *BioScience* 72.10 (2022), pp. 988–998. DOI: 10.1093/biosci/biac062.
- 4672
- 4673
- 4674 [76] Joshua B. Tenenbaum et al. “How to Grow a Mind: Statistics, Structure, and Abstraction”. In: *Science* 331.6022 (2011), pp. 1279–1285. DOI: 10.1126/science.1192788.
- 4675
- 4676 [77] Elizabeth Tipton. “How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations”. In: *Journal of Educational and Behavioral Statistics* 39.6 (2014), pp. 478–501. DOI: 10.3102/1076998614558486.
- 4677
- 4678
- 4679 [78] Andrew Tomkins, Min Zhang, and William D. Heavlin. “Reviewer Bias in Single- versus Double-Blind Peer Review”. In: *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12708–12713. DOI: 10.1073/pnas.1707323114.
- 4680
- 4681
- 4682 [79] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science* 185.4157 (1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.
- 4683
- 4684 [80] Melina R. Uncapher and Michael D. Rugg. “Selecting for Memory? The Influence of Selective Attention on the Mnemonic Binding of Contextual Information”. In: *Journal of Neuroscience* 29.25 (2009), pp. 8270–8279. DOI: 10.1523/JNEUROSCI.1043-09.2009.
- 4685
- 4686
- 4687 [81] Tyler J. VanderWeele and Peng Ding. “Sensitivity Analysis in Observational Research: Introducing the E-Value”. In: *Annals of Internal Medicine* 167.4 (2017), pp. 268–274. DOI: 10.7326/M16-2607.
- 4688
- 4689 [82] Tyler J. VanderWeele, Stijn Vansteelandt, and James M. Robins. “Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder”. In: *Epidemiology* 25.2 (2014), pp. 300–306. DOI: 10.1097/EDE.0000000000000034.
- 4690
- 4691

- [83] Robert C. Wilson et al. “Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma”. In: *Journal of Experimental Psychology: General* 143.6 (2014), pp. 2074–2081. DOI: 10.1037/a0038199.
- [84] Andrew P. Yonelinas. “The Nature of Recollection and Familiarity: A Review of 30 Years of Research”. In: *Journal of Memory and Language* 46.3 (2002), pp. 441–517. DOI: 10.1006/jmla.2002.2864.

Research Ethics and Source Provenance

Research status. Sections 1–10 have completed a section-by-section revision cycle in this release. Integrated cross-sectional audits and empirical validation remain pending. All formal constructs are exploratory models. No original participant experiment or validated causal estimates are reported.

Bibliographic policy. References in this document were checked against publisher, archival or trusted bibliographic records. Claim scopes and access limitations are recorded in the verification ledger.

Illustration policy. Illustrative histories and observers are hypothetical and non-identifiable. The proposed architecture is representation-sensitive and does not claim a fixed psychological ontology.