

Judgment Formation in AI-Mediated Inquiry: A Preliminary Generative Relational Account

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

This paper develops a preliminary Generative Relational (GR) account of judgment formation in AI-mediated inquiry. It examines the history-bearing development of capacities for framing, discrimination, confidence calibration, selective reliance, revision, responsibility, and learning when artificial intelligence participates in generating options, reasons, evaluations, recommendations, and polished answers. The central distinction separates improved task performance from the development of judgment. AI assistance can improve performance while leaving users poorly calibrated about the quality or provenance of resulting decisions, and algorithmic advice can influence both decisions and confidence in ways conditioned by metacognitive sensitivity. Judgment formation is therefore treated as a relational process whose relevant analytical unit may be a coupled Human–AI configuration as well as an individual participant. The paper develops this account through analyses of the AI-mediated judgment environment, selective-reliance calibration, problem salience, differentiated evaluative registers, temporal and mediating judgment, reflexive effects on future judgment conditions, criterion power, delegation and ownership, practices of judgment, cultivation and formative cognitive friction, developmental assessment, and epistemic and normative boundaries on the use of judgment evidence. Unaided human performance is not used as a universal benchmark, and answer accuracy is treated as insufficient evidence of judgment development. The purpose is to provide a revisable conceptual architecture for studying how judgment can remain calibrated, answerable, and developmentally generative as artificial assistance becomes more capable.

Keywords: Generative Relational Theory; judgment formation; AI-mediated inquiry; metacognition; selective reliance

Preliminary Research Note

This manuscript belongs to the broader Generative Relational (GR) research programme. Its immediate object is judgment formation in AI-mediated inquiry: the historically developing configuration through which judgments are framed, calibrated, revised, acted upon, delegated, and learned from when artificial systems participate in the production of options, reasons, evaluations, recommendations, and representations. The paper treats judgment as situated in relations among persons, artificial systems, institutions, evidence environments, language systems, histories, and consequences rather than as an isolated mental act.

The inquiry adopts ontological suspension. Relational, configurational, dynamical, and developmental language is used as analytical representation without establishing an ultimate ontology of persons, artificial intelligence, cognition, judgment, or responsibility. Human–AI configurations are admissible analytical units where the chosen boundary clarifies the dependencies relevant to a judgment process; the representation itself does not establish that the participating elements constitute one metaphysical subject.

The manuscript distinguishes task performance from judgment formation. A correct answer, efficient recommendation, or improved score can be practically valuable while providing incomplete evidence about whether the configuration has developed better calibration, more appropriate selective reliance, stronger capacity for revision, clearer responsibility, or greater ability to learn from subsequent consequences. Conversely, a judgment process can be developmentally valuable even where a particular decision proves mistaken, provided that its uncertainty, reasons, provenance, and consequences remain available for later correction and learning.

Formal economy is a governing constraint. Specialized notation is introduced only when it clarifies a relational dependency, temporal transformation, or distinction that cannot be expressed adequately in prose. External literature is included only after source verification and recording in the manuscript's verification ledger.

Responsible Use Notice

The author encourages good-faith discussion, criticism, independent inquiry, and responsible use of the material in this work. Separately from the licence's terms, the author asks users to consider foreseeable harms when adapting or applying the proposed framework. This ethical request leaves the licence's permissions and legally authorized uses unchanged.

The author retains the rights preserved under CC BY-NC 4.0 and may pursue remedies to which the author is legally entitled for breach of the licence or violation of the author's independently applicable rights. Reuse remains independent from authorial endorsement. Third-party rights require authorization from their respective holders where applicable. Copyright exceptions and limitations remain fully available.

Notices

Release status. This release version contains the complete Sections 1–18 manuscript on judgment formation in AI-mediated inquiry. Sections 1–5 establish the analytical object, literature interfaces, Generative Relational foundations, the AI-mediated judgment environment, and Human–AI judgment configurations. Sections 6–12 develop selective-reliance calibration, salience judgment, differentiated evaluative registers, temporal and mediating judgment, reflexive future judgment conditions, criterion power, delegated-function ownership, and responsibility. Sections 13–17 develop judgment practice, cultivation, cognitive friction, developmental assessment, and boundaries on the institutional use of judgment evidence. Section 18 consolidates the relational developmental account and its research agenda. The manuscript has undergone manuscript-wide notation, cross-reference, citation-ledger, structural, and copyediting review. The theoretical framework remains preliminary and revisable; stronger descriptive, educational, professional, or normative conclusions require domain-specific empirical investigation.

Licence. Except where otherwise indicated, copyright 2026 Wanhong Huang. This work is made available under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), subject to the licence’s terms.

Statement on the use of language models. The exploratory discussions and preparation of this paper involved OpenAI’s ChatGPT. ChatGPT supported exploratory dialogue, source-oriented literature checking, formal reconstruction, argumentative criticism, drafting in $\text{\LaTeX}$, and technical checks. The author selected the research question, supplied the principal conceptual direction concerning judgment practice and cultivation in AI-mediated inquiry, directed and approved the theoretical commitments and epistemic status of the claims, and bears sole responsibility for the manuscript, including its definitions, formal constructions, arguments, conclusions, and errors. Authorship credit remains with the human author.

1 Introduction

This section establishes the analytical object, motivation, scope, and manuscript architecture of the proposed account of judgment formation in AI-mediated inquiry. Its role is to distinguish the quality of a particular answer or decision from the development of the capacities through which judgments are framed, calibrated, revised, delegated, acted upon, and learned from across time. The discussion proceeds through six steps. It first defines judgment formation as a developmental and relational object. It then characterizes AI mediation as a reorganization of the judgment environment, separates performance gains from judgment development, introduces a minimal Generative Relational orientation, specifies the status of coupled Human–AI configurations, and closes with the manuscript architecture. The method is conceptual and classificatory, with selective interfaces to empirical work on metacognition, reliance on AI advice, overreliance, and scientific understanding. The account is preliminary and conceptual; a universal psychology of judgment, a general decision theory, and a normative algorithm for correct choice remain outside its scope.

1.1 Judgment Formation as an Analytical Object

The increasing availability of artificial intelligence in inquiry makes it useful to distinguish a successful judgment episode from the formation of judgment itself. A person or team may reach a correct answer because an AI system supplied a strong recommendation, because independent reasoning converged with the recommendation, because the recommendation was accepted despite weak understanding, or because a coupled Human–AI process generated and tested alternatives more effectively than either component would have done alone. These cases can produce similar visible outcomes while leaving different developmental consequences for later inquiry. The immediate decision therefore provides only partial evidence about the capacities that generated, monitored, accepted, rejected, revised, or learned from it.

The present paper uses *judgment formation* to name this broader developmental object. Judgment formation concerns changes in the history-bearing configuration through which an inquirer becomes able to frame a situation, identify what deserves attention, compare alternatives, assess reasons and evidence, estimate uncertainty, decide when to rely on another source, revise a conclusion, connect a judgment to responsibility, and learn from later consequences. The term *formation* emphasizes that these capacities can be cultivated, reorganized, weakened, bypassed, distributed, or partially delegated over time. Judgment formation is therefore analytically distinct from both judgment performance in a single task and the possession of a fixed stock of decision rules.

This distinction also prevents the manuscript from treating judgment as an isolated mental act. Inquiry is conducted through relations among persons, texts, instruments, institutions, models, archives, interfaces, norms, and prior histories. The quality of judgment can depend on which alternatives are made visible, how evidence is represented, what confidence signals are available, which errors are recoverable, what costs attach to delay, who carries responsibility, and whether later outcomes remain available for revision. Judgment is consequently situated in an environment that can support or distort calibration even when the reasoning ability of the focal person remains unchanged.

The paper’s central proposal is that judgment formation can be investigated as a *history-bearing relational development of evaluative and action-guiding capacities*. The phrase *history-bearing* indicates that present judgment depends on accumulated cases, errors, corrections, exemplars, habits of attention, institutional experiences, and prior delegations. *Relational development* indicates that these capacities arise and operate through relations among agents, tools, evidence environments, language systems, and normative structures. *Evaluative and action-guiding capacities* indicates that judgment concerns more than proposition acceptance: it can include problem salience, aesthetic discrimination, normative assessment, trajectory-sensitive interpretation, practical timing, reliance, and responsibility. This proposal functions as a conceptual representation and leaves open whether judgment has any invariant essence across disciplines and cultures.

A further implication is that confidence alone provides insufficient evidence of judgment formation. Confidence can be informative when it tracks accuracy and uncertainty, yet it can also be inflated, sup-

pressed, or disconnected from the quality of the underlying judgment. The relevant object is calibration: whether confidence, reliance, revision, and action are appropriately related to the quality and limits of the available reasons. The paper treats calibration as one interface within a broader account of judgment formation rather than as a complete substitute for judgment.

1.2 AI Mediation and the Judgment Environment

AI-mediated inquiry changes the environment in which judgment is exercised because artificial systems can increasingly participate in stages that previously required substantial human search, formulation, comparison, or synthesis. An AI system can supply candidate explanations, produce arguments for multiple positions, translate evidence into a requested register, suggest actions, generate confidence-like language, rank options, or synthesize large bodies of material. These functions can be valuable. They can also alter which cognitive and practical burdens remain with the user. The relevant transformation extends beyond the provision of more information. AI can change the *structure of the judgment situation*: the number and presentation of options, the apparent completeness of reasons, the speed at which closure becomes available, the ease of obtaining a second opinion, and the relation between visible fluency and underlying warrant.

The scientific context illustrates why this change matters. Messeri and Crockett argue that AI tools can increase productivity while also creating conditions for illusions of understanding, in which users or communities take an AI-supported output to indicate more understanding than the underlying process warrants [1]. Their analysis concerns scientific research and should not be generalized mechanically to all judgment domains. It nevertheless identifies a structural possibility central to the present paper: improved production of answers, models, or summaries can coexist with weakened access to the limits of one's own understanding. When this occurs, the judgment environment changes even if final task performance improves.

The same environment can also alter how individuals relate their own judgment to algorithmic advice. Soleimanof and Neufeld report that confidence estimates are strongly associated with responses to AI advice, while lower metacognitive sensitivity is associated with less consistent use of confidence when deciding whether to shift toward or away from that advice [3]. This result does not establish a universal rule for Human–AI collaboration. It supports the narrower claim that the relation between internal confidence and external advice is itself a variable in AI-mediated judgment and that metacognitive sensitivity can influence the resulting reliance pattern.

These developments motivate a shift from a binary trust question to a relational one: under which conditions should an inquirer rely, verify, contest, defer, suspend, or revise? The answer can vary by task, consequence, reversibility, domain expertise, evidence quality, system reliability, and the availability of independent checks. A judgment architecture oriented toward calibrated use should therefore support selective reliance rather than global acceptance or global rejection of artificial assistance.

1.3 Task Performance and Judgment Development

The distinction between performance and formation is especially important because AI can improve the former without reliably improving the latter. Fernandes and colleagues report two studies of AI-assisted logical reasoning in which participants improved task performance relative to a norm population while substantially overestimating their own performance; the authors interpret the results as evidence of a disconnect between AI-supported performance and metacognitive monitoring [2]. The studies concern a bounded reasoning task rather than long-horizon professional judgment, and the comparison design should not be treated as a direct measure of durable judgment development. The finding nevertheless demonstrates the conceptual possibility required here: a user can perform better with AI while remaining poorly calibrated about the quality of that performance.

This distinction can be stated more generally. A judgment environment may improve immediate output by supplying missing expertise, reducing search costs, or filtering alternatives, while simultaneously reducing opportunities for the user to practice error detection, generate independent hypotheses,

experience uncertainty, or observe the consequences of their own initial judgment. Conversely, deliberately preserving some independent reasoning or review can impose short-term cost while contributing to later calibration. The developmental value of an inquiry episode therefore cannot be inferred solely from whether the final answer is correct or efficient.

Research on overreliance provides a second interface. Bućinca, Malaya, and Gajos found that cognitive forcing interventions reduced overreliance on incorrect AI recommendations relative to simpler explainable-AI conditions, while also producing less favorable subjective ratings and benefiting participants unevenly depending on need for cognition [4]. The result illustrates two points relevant to judgment formation. First, interaction design can affect whether users actively engage their own evaluative processes rather than merely consume a recommendation. Second, reducing overreliance can carry costs in effort, usability, or satisfaction. The practical objective is calibrated cognitive effort: conditions under which effort, verification, and reliance are matched appropriately to the stakes and uncertainty of the task.

The manuscript consequently treats *judgment development* as a change in later capability rather than as the immediate success of a single assisted episode. Relevant evidence may include improved calibration, better discrimination between cases that warrant reliance and cases that warrant verification, stronger ability to articulate uncertainty, greater sensitivity to counterevidence, more appropriate revision, or better transfer of judgment across related situations. These indicators can diverge. A system can improve one while weakening another. The paper therefore resists reducing judgment development to a single scalar score.

Claim 1.1 (Performance–formation distinction). Within AI-mediated inquiry, improved task performance is insufficient by itself to establish improved judgment formation. Evidence of formation requires changes in the capacities through which later judgments are framed, calibrated, revised, relied upon, owned, or learned from.

Claim 1.1 is conceptual rather than empirical in its full scope. Particular indicators of judgment formation require domain-specific validation. Its purpose is to prevent the practical success of an AI-assisted episode from being treated automatically as evidence that the participating person or configuration has become better at judging.

1.4 A Minimal Generative Relational Orientation

The GR perspective used in this paper begins from a history-bearing configuration rather than an isolated decision maker. A minimal representation is given in Equation 1.

$$\Sigma_t^J = (H_t, \mathcal{R}_t, \mathcal{E}_t, \mathcal{L}_t, \mathcal{N}_t, \mathcal{C}_t). \quad (1)$$

In Equation 1, H_t denotes the relevant accumulated history of prior judgments, consequences, feedback, and revisions; $\mathcal{R}_t$ denotes the relational structure linking persons, artificial systems, institutions, and other supports; $\mathcal{E}_t$ denotes the evidence and representation environment available to the inquiry; $\mathcal{L}_t$ denotes the operative language and conceptual resources through which the situation is framed; $\mathcal{N}_t$ denotes relevant normative and institutional conditions; and $\mathcal{C}_t$ denotes the operative capacities available for framing, comparison, calibration, reliance, revision, and action. The tuple is analytical and deliberately incomplete. Later sections may refine selected components where doing so clarifies a specific problem.

Judgment formation is represented minimally as a history-sensitive update of this configuration. The developmental relation is stated in Equation 2.

$$\Sigma_{t+1}^J = \mathcal{F}_t^J (\Sigma_t^J, \Delta_t^J), \quad (2)$$

where Δ_t^J denotes the task-relevant episode or intervention through which judgment is exercised and potentially changed. The update can include direct experience, AI advice, independent analysis,

criticism, outcome feedback, institutional review, or combinations of these elements. Equation 2 does not assume that formation is monotonic improvement. An episode can strengthen calibration, produce dependence, narrow attention, increase confidence without warrant, redistribute responsibility, or leave little durable change.

The representation also makes reflexivity visible. A judgment does not merely select an action under fixed conditions. The action can change the evidence environment, institutional relations, option space, habits of reliance, and later criteria of evaluation. Some of the strongest GR questions therefore concern loops in which present judgment helps generate the conditions of future judgment. Those loops are developed later in the manuscript rather than formalized fully in the introduction.

1.5 Coupled Human–AI Configurations and Scope

The paper does not assume that judgment formation belongs exclusively to an isolated human subject. In many inquiry settings, relevant functions are distributed across a person, an AI system, colleagues, databases, institutional procedures, instruments, and archived prior decisions. A narrow analysis may attribute a judgment to one person for legal or practical reasons, while a broader analysis may need to represent the coupled configuration that produced the available evidence and recommendation. The appropriate boundary depends on the claim under investigation.

This position avoids two symmetrical simplifications. The first is human exceptionalism, understood here as the assumption that calibration, revision, problem framing, or evaluative discrimination are inaccessible to artificial systems by definition. The present framework leaves that metaphysical question open. The second is performance equivalence, the inference that producing a similar answer establishes equivalent judgment formation. Similar outputs can arise from different histories, dependencies, capacities for revision, and responsibility structures. The relevant comparison must therefore be typed and task-relative.

Human–AI coupling also raises the question of judgment ownership. Delegating computation does not necessarily delegate evaluation; delegating recommendation does not necessarily delegate decision authority; and distributing causal contribution does not automatically distribute responsibility in the same way. These distinctions are reserved for later analysis, but they shape the scope of the present paper from the outset. The relevant question concerns how the judgment process was organized: which capacities were exercised or bypassed, who could contest or revise the result, and who remained answerable for acting on it.

The manuscript’s normative ambition is correspondingly limited. It does not prescribe a universal level of human involvement or claim that independent unaided reasoning is always preferable. Some tasks benefit from extensive automation; some require substantial verification; some may be safer when judgment is distributed across several human and artificial components. The aim is to make these differences available for principled analysis through separate treatment of performance, reliance, calibration, responsibility, and formation.

1.6 Manuscript Architecture

The manuscript is organized around four analytical movements. The first establishes the literature and formal foundations: metacognition and Human–AI decision support, the GR representation of judgment formation, the transformation of the judgment environment under AI mediation, and coupled Human–AI configurations. The second examines central operations of judgment: calibration and selective reliance, problem salience, aesthetic, normative, practical, temporal, and mediating judgment. The third examines reflexive and institutional consequences: how present judgment changes future judgment conditions, how power shapes criteria and option visibility, and how delegation interacts with responsibility and judgment ownership. The fourth turns to practice and cultivation: recurrent practices for AI-rich inquiry, judgment cultivation, formative and misallocated cognitive friction, assessment of judgment development, and the framework’s epistemic and normative boundaries.

The architecture preserves several distinctions developed in the preceding GR research programme while giving judgment formation an independent analytical object. Problem formation and contingency sensitivity matter because judgment begins partly with what becomes salient enough to evaluate. Aesthetic and normative judgment matter because intellectual attraction, social priority, and practical action do not collapse into one criterion. Human–AI configurations matter because AI can change both the manifestations of judgment and the conditions under which judgment capacities are practiced. The present paper extends these interfaces by focusing on how judgment is exercised across time: how confidence becomes calibrated or miscalibrated, how reliance is selected, how responsibility remains attached or becomes diffuse, how present decisions reshape later options, and how educational practices can cultivate revisable judgment rather than merely improve immediate answer production.

The resulting research question is therefore bounded but consequential: how should judgment formation be understood, practiced, and cultivated when AI becomes a recurrent participant in inquiry? The proposed GR account treats that question as developmental, relational, and reflexive. It asks not only whether a judgment is correct, but also what kind of configuration produced it, what that episode changes for the next judgment, and which capacities remain available for learning when the consequences arrive.

2 Metacognitive Interfaces in Human–AI Decision Support

This section positions the proposed account of judgment formation relative to five established interfaces: metacognition, automation use, trust and reliance, responses to algorithmic advice, and appropriate reliance on AI advice. Its role is diagnostic rather than synthetic: the cited traditions investigate related objects with different measures and evidential assumptions, and they should not be collapsed into a single theory of judgment. The discussion proceeds from metacognitive monitoring to automation reliance, then to algorithm-specific responses and contemporary Human–AI advice-taking. It closes by identifying the developmental remainder that motivates the Generative Relational (GR) account developed in the rest of the paper. The central boundary is that successful reliance in one episode, even when measured against known advice correctness, provides only partial evidence about how an inquirer’s later judgment capacities have changed.

Table 1 summarizes these interfaces. The table does not rank the traditions. Each isolates a different part of the judgment environment, and their methodological differences matter for the present paper’s scope.

2.1 Metacognitive Monitoring and Judgment Quality

Metacognition provides the most direct established interface for the paper’s concern with judgment about one’s own judgment. Flavell’s developmental formulation distinguished metacognitive knowledge and metacognitive experiences from the first-order cognitive activity being monitored, emphasizing knowledge about cognitive agents, tasks, strategies, and the monitoring of ongoing cognitive enterprises [5]. This tradition makes it possible to distinguish performing a task from representing the state and limits of one’s own performance while the task is being performed.

For the present paper, the importance of this distinction is structural. AI-mediated inquiry can alter first-order performance and second-order monitoring differently. An AI system may improve the quality of an answer while also changing the cues through which the user estimates whether the answer is reliable. Fluent explanation, rapid completion, polished language, or apparent comprehensiveness can become metacognitive cues even when they are only imperfectly related to the quality of the underlying reasoning. Conversely, an unfamiliar or visibly uncertain answer can elicit low confidence despite being well supported. Judgment formation therefore requires attention to the relation between first-order task success and second-order monitoring, rather than treating either as a sufficient proxy for the other.

Fleming and Lau provide a useful measurement-oriented refinement by distinguishing metacognitive bias, metacognitive sensitivity, and metacognitive efficiency [6]. Metacognitive bias concerns the overall level of expressed confidence; metacognitive sensitivity concerns the extent to which confidence

Table 1: Literature interfaces for judgment formation in AI-mediated inquiry.

Interface	Primary analytical object	Typical evidential focus	Boundary relative to this paper
Metacognition	Monitoring and evaluation of one’s own cognition	Confidence, accuracy, discrimination, calibration, monitoring	Does not by itself specify how AI-mediated episodes reorganize later judgment practices
Automation use and trust	Reliance on imperfect automation in context	Use, misuse, disuse, trust, reliability, monitoring	Often evaluates reliance on a system rather than formation of broader evaluative capacities
Algorithm aversion/appreciation	Comparative response to algorithmic versus human advice	Advice weighting, source preference, response to observed error	Shows context-sensitive source response without supplying a developmental account of judgment
Appropriate reliance on AI advice	Case-sensitive acceptance or rejection of AI advice	Reliance on correct advice and self-reliance when advice is incorrect	Requires a criterion of advice correctness and primarily evaluates episode-level reliance
Learning in Human–AI decision support	Change in the human component during repeated advice use	Learning, mental models, reliance, complementary performance	Moves toward formation, while leaving broader judgment criteria, responsibility, and long-horizon revision open

discriminates correct from incorrect judgments; and metacognitive efficiency asks how much metacognitive sensitivity is achieved relative to first-order task performance. These distinctions support a core premise of the present manuscript: high confidence, good performance, and good monitoring are analytically separable.

This separability becomes especially important under artificial assistance. A user can improve first-order performance because the AI supplies missing information while retaining weak sensitivity to when the AI is wrong. A different user can remain modestly accurate while possessing strong sensitivity to uncertainty and therefore knowing when further verification is needed. The latter configuration may be developmentally better positioned for later unfamiliar tasks even when its immediate score is lower. The present paper consequently treats metacognitive measures as evidence about important components of judgment formation without reducing judgment formation to metacognition alone.

Claim 2.1 (Confidence is an object of calibration). In AI-mediated inquiry, confidence is best treated as one variable to be calibrated against evidence, performance, uncertainty, and consequence rather than as a direct measure of judgment quality. The same confidence level can accompany well-grounded judgment, poorly monitored error, strategic deference, or unsupported fluency.

Claim 2.1 also limits the interpretation of self-reported certainty in later sections. Calibration matters because judgment often has to regulate action before correctness is fully known, yet the quality of that regulation depends on how confidence interacts with the available evidence and with opportunities for independent checking.

2.2 Automation Use, Misuse, Disuse, and Trust

Research on automation predates contemporary generative AI and provides a second interface for understanding reliance. Parasuraman and Riley distinguished use, misuse, disuse, and abuse of automation,

treating overreliance and underuse as different failure modes shaped by factors such as trust, workload, reliability, and system design [7]. The distinction remains conceptually useful because it prevents a simple equation of more automation use with better judgment. Underuse can be costly when a reliable system is ignored, while overuse can be costly when monitoring and independent discrimination deteriorate.

Lee and See developed the relation between trust and reliance further, emphasizing that trust becomes especially relevant when system complexity makes complete understanding impractical and that appropriate reliance depends on both automation characteristics and context [8]. Their account is particularly important for AI-mediated inquiry because contemporary systems can be too complex for users to inspect in full. The practical problem is therefore often not whether a user possesses complete causal understanding of a model, but whether available evidence about capability, limits, context, and uncertainty supports a particular degree of reliance for the present task.

The automation literature also supplies a caution for the present manuscript. A reliance pattern can be locally sensible without producing durable judgment formation. A professional may learn an effective rule for when to consult a particular system while remaining unable to transfer that discrimination to a different system, domain, or failure mode. Conversely, a temporary period of low reliance can be developmentally useful when it creates opportunities to understand recurrent system errors. Reliance behavior therefore has to be interpreted together with the history in which it was formed and the range of situations to which it can transfer.

2.3 Algorithm Aversion and Algorithm Appreciation

Studies of algorithm aversion and algorithm appreciation show why a stable psychological disposition toward artificial advice should not be assumed. Dietvorst, Simmons, and Massey found that participants could become especially reluctant to use algorithmic forecasts after observing algorithmic errors, even when the algorithm outperformed a human forecaster [9]. In a different set of tasks, Logg, Minson, and Moore found that participants often weighted algorithmic advice more heavily than comparable human advice, a pattern they termed algorithm appreciation [10]. These findings arise from different experimental settings and should not be read as a contradiction requiring one universal resolution. Together they demonstrate that source preference is sensitive to experience, framing, task structure, and comparison conditions.

For judgment formation, the important implication is that source preference and judgment quality are distinct. A person can display algorithm aversion for poorly grounded reasons, or algorithm appreciation without being able to distinguish strong from weak advice. The same outward decision, whether acceptance or rejection of an algorithmic recommendation, can therefore result from very different internal and relational configurations. This is one reason the present manuscript avoids treating trust, distrust, acceptance, rejection, or reliance frequency as direct indicators of judgment quality or judgment formation.

The aversion/appreciation literature also reveals a design-sensitive dimension of judgment. Small changes in how agency, error, or control are presented can alter whether people use algorithmic advice. Such effects matter for formation because repeated interface structures can become part of the history through which future reliance is learned. An inquiry environment that consistently encourages acceptance after polished explanation may cultivate a different reliance repertoire from one that requires explicit comparison with an independent first-pass judgment.

2.4 Appropriate Reliance on AI Advice

Contemporary Human–AI decision-support research increasingly replaces global trust with the more discriminating concept of *appropriate reliance*. Schemmer and colleagues conceptualize appropriate reliance as the ability to rely on AI advice when it is correct while withholding reliance when it is incorrect, and they propose a two-dimensional measurement framework for analyzing this discrimination [11]. This move is closely aligned with the present paper’s preference for selective reliance over global acceptance or rejection.

Appropriate-reliance research is analytically valuable because it makes correctness-sensitive discrimination observable at the level of individual advice episodes. It also allows intervention research to ask whether explanations, interface features, or training improve the user’s ability to distinguish when the AI should be followed. Bućinca and colleagues, for example, found that cognitive forcing interventions could reduce overreliance on AI advice, while also producing usability and preference trade-offs [4]. Soleimanof and Neufeld further connect reliance behavior to metacognition by showing that confidence estimates and metacognitive sensitivity are associated with how people respond to AI advice [3].

The present paper adopts the discrimination objective while preserving two boundaries. First, appropriate reliance usually requires a criterion by which advice can be classified as correct or incorrect. Many inquiry settings provide such a criterion only later, imperfectly, or contestably. Research prioritization, diagnosis under incomplete evidence, scientific interpretation, institutional forecasting, and emerging-risk assessment can involve delayed outcomes, plural values, shifting states, or disagreement over what counts as success. Second, correct acceptance and correct rejection in one task do not establish that the inquirer has developed a transferable capacity for later cases.

Caution 2.2 (Episode-level appropriateness and developmental inference). An appropriate reliance outcome can support a claim about the quality of a particular Human–AI decision episode when advice correctness is known. It does not by itself establish durable improvement in judgment formation. Stronger developmental inference requires evidence about transfer, calibration across time, changing error patterns, revision after feedback, and performance under altered conditions.

Caution 2.2 is not a criticism of appropriate-reliance metrics. It states the difference between their primary analytical target and the broader developmental target of this paper.

2.5 Learning within Human–AI Decision Support

Recent work begins to move from episode-level reliance toward learning. Schemmer and colleagues explicitly study human learning as a mediator of appropriate reliance and complementary Human–AI performance, reporting an experimental relationship between learning and appropriate reliance in a decision-support task [12]. This line of work is especially relevant because it treats the human component of the system as capable of changing through interaction rather than as a fixed decision-maker receiving repeated recommendations.

The present manuscript extends the developmental object in three directions. First, learning can concern more than a mental model of system accuracy. Repeated AI-mediated inquiry can change problem framing, standards of evidence, confidence habits, attention to alternatives, willingness to verify, tolerance for unresolved uncertainty, and assumptions about who remains responsible for closure. Second, these changes can be beneficial, harmful, mixed, or domain-specific. A person may learn the error profile of one model while becoming less practiced at generating independent alternatives. Third, the relevant history can belong to a coupled configuration rather than to one individual: teams, institutions, procedures, archives, and recurrent AI systems can jointly preserve and transform judgment routines.

This broader developmental scope also changes how interventions should be evaluated. An interface that maximizes immediate correct reliance may still have weak long-horizon value if it prevents users from learning which cues matter. Conversely, an intervention that initially slows performance may be developmentally valuable if it improves discrimination, transfer, or retrospective error analysis. The distinction parallels the performance/formation separation established in Section 1, but the decision-support literature now provides a concrete empirical interface for testing it.

2.6 Analytical Remainder for a Generative Relational Account

The literatures reviewed in this section establish substantial foundations for the present account. Metacognition distinguishes first-order performance from monitoring of that performance. Automation research shows that both overuse and underuse can be failures. Trust research emphasizes context-sensitive reliance under complexity. Algorithm-aversion and appreciation studies show that source response is not

a stable binary disposition. Appropriate-reliance research supplies episode-level measures of discriminating correct from incorrect advice. Learning-oriented Human–AI research begins to model change in the human component across interaction.

The remaining GR problem concerns how these processes become historically organized into later judgment capacity. The paper therefore asks what repeated episodes leave behind. An inquirer may acquire better calibration, a more accurate model of system limits, stronger habits of verification, broader problem sensitivity, or greater tolerance for uncertainty. The same sequence can also produce dependency, narrowing of alternatives, confidence inflation, erosion of independent discrimination, diffusion of responsibility, or procedural routines that remain effective only while the surrounding system is stable. These outcomes are developmental and relational because they depend on the prior configuration, the structure of interaction, the feedback available after action, and the institutions that preserve or suppress learning from error.

Claim 2.3 (Reliance is evidence about judgment, not a substitute for judgment formation). Reliance behavior can provide evidence about calibration, discrimination, trust, and source use. It is not sufficient evidence of judgment formation unless the analysis also establishes how the episode changes capacities available for subsequent inquiry.

Claim 2.3 marks the boundary between the literature interface and the formal work that follows. The GR task is not to replace established measures of metacognition or reliance. It is to locate those measures within a history-bearing configuration whose later capacities can be changed by the very support systems used to improve present decisions.

3 Generative Relational Foundations of Judgment Formation

This section develops the minimum formal architecture needed to treat judgment formation as a history-bearing relational process. Its role is to refine the configuration and update relations introduced in Section 1 without converting the paper into a general decision calculus. The objective is to distinguish four analytical objects that are easily conflated in AI-mediated inquiry: the situated judgment episode, the criteria and monitoring through which that episode is organized, the developmental residue left by the episode, and the later availability of any resulting change. The discussion proceeds by deriving each object from the existing judgment configuration Σ_t^J and update relation $\mathcal{F}_t^J$. The formalism is representational. It does not assert that judgment has a unique mathematical ontology, that all relevant dimensions are measurable on common scales, or that one formal decomposition applies unchanged across scholarly, educational, professional, and practical domains.

3.1 Judgment Configuration, Episode, and Analytical Projection

Equation 1 represents the history-bearing configuration within which judgment becomes possible. That configuration should be distinguished from a particular judgment episode. The configuration includes accumulated history, relational supports, evidence environments, conceptual resources, normative conditions, and operative capacities. A judgment episode is a task- and context-indexed realization of selected parts of that configuration under a particular demand for interpretation, comparison, reliance, closure, or action.

Let q denote the inquiry task and let ξ denote the local context relevant to that task. The present paper represents a situated judgment record by Equation 3.

$$\mathfrak{J}_t^{(q,\xi)} = \mathcal{P}_{q,\xi}^J (\Sigma_t^J, \Delta_t^J) = (\mathcal{S}_t, \mathcal{O}_t, \mathcal{W}_t, \mathcal{V}_t, \mathcal{M}_t, a_t)^{(q,\xi)}. \quad (3)$$

In Equation 3, $\mathcal{P}_{q,\xi}^J$ is an analytical projection rather than a psychological mechanism. It selects the aspects of the configuration and episode that are relevant to the inquiry being studied. The component $\mathcal{S}_t$ records what has become salient enough to frame the judgment; $\mathcal{O}_t$ records the options, interpretations,

hypotheses, or response possibilities under consideration; $\mathcal{W}_t$ records the available warrant structure, including evidence, reasons, counterevidence, and provenance where relevant; $\mathcal{V}_t$ records the criteria or evaluative considerations applied to those possibilities; $\mathcal{M}_t$ records monitoring variables such as confidence, uncertainty, reliance, and perceived need for verification; and a_t records the resulting response. The response may be an action, recommendation, commitment, revision, request for further evidence, explicit deferral, or preservation of unresolved alternatives.

The record in Equation 3 is intentionally broader than a choice among fixed alternatives. Judgment can begin before the option set is stable. It can change what is treated as salient, reject the framing supplied by an AI system, generate an additional alternative, alter the criteria of comparison, or decide that closure is premature. This is especially important in inquiry settings where the visible option set is itself produced by search, recommendation, institutional filtering, or model generation.

Definition 3.1 (Situating judgment episode). A *situated judgment episode* is a task- and context-indexed organization of salience, candidate possibilities, warrants, evaluative criteria, monitoring, and response within a history-bearing judgment configuration. The representation permits action, revision, deferral, and maintained plurality as distinct responses rather than treating immediate closure as the defining feature of judgment.

Definition 3.1 establishes the local object whose developmental consequences can later be studied. Similar responses can arise from substantially different judgment records. Two people may accept the same AI recommendation while differing in what alternatives they considered, what evidence they checked, how well calibrated their confidence was, and whether they could later reconstruct or revise the reasoning. Outcome equivalence therefore remains weaker than equivalence of judgment process or formation.

3.2 Salience and the Construction of the Option Space

The first two components of Equation 3 are upstream of many familiar evaluation problems. $\mathcal{S}_t$ concerns what becomes visible as requiring judgment; $\mathcal{O}_t$ concerns what becomes available as a possible interpretation or response. Both can be altered by AI mediation before explicit evaluation begins. Search ranking can affect which evidence is noticed, summarization can compress disagreement, recommendation systems can privilege some actions over others, and generative systems can enlarge or narrow the candidate set depending on how the interaction is structured.

A judgment architecture therefore cannot assume that the option set is externally given and neutral. For analytical purposes, the paper treats $\mathcal{O}_t$ as configuration-dependent. The relation is stated schematically in Equation 4.

$$\mathcal{O}_t^{(q,\xi)} = \mathcal{G}_{q,\xi}^O(\mathcal{S}_t, H_t, \mathcal{R}_t, \mathcal{E}_t, \mathcal{L}_t, \mathcal{C}_t), \quad (4)$$

where $\mathcal{G}_{q,\xi}^O$ denotes a context-specific option-construction relation. Equation 4 does not imply a deterministic generator. It records that the available option space can depend on history, relations, evidence presentation, conceptual language, and operative capacities. The same external situation can therefore yield different candidate sets for differently formed configurations.

This dependence matters for judgment formation because repeated use of a support system can change not only which option is selected but also which kinds of options become routinely imaginable. A system that repeatedly supplies plausible candidate explanations can enlarge a novice’s repertoire, yet habitual dependence on externally generated alternatives can also reduce practice in independent option generation. The developmental question is therefore not whether a large option set is intrinsically superior. It is whether the configuration develops appropriate sensitivity to missing alternatives, framing effects, and the conditions under which externally generated options should be extended or contested.

Claim 3.2 (Option-space dependence). In AI-mediated inquiry, the quality of judgment can depend on the construction of the candidate space before explicit comparison begins. Judgment formation therefore

includes development of sensitivity to omitted, generated, framed, or prematurely excluded possibilities, not only improved selection among a fixed set of options.

Claim 3.2 prepares the later analysis of problem salience and power over judgment criteria. It also limits a narrow interpretation of appropriate reliance: a person can rely appropriately among the recommendations shown while the more consequential failure lies in what the system never made available for consideration.

3.3 Criteria, Monitoring, and the Separation of Evaluation Layers

The components $\mathcal{V}_t$ and $\mathcal{M}_t$ serve different analytical roles. $\mathcal{V}_t$ concerns the criteria by which alternatives are assessed. Depending on the domain, these may include evidential strength, explanatory adequacy, expected consequences, reversibility, legal constraints, aesthetic fertility, fairness, timeliness, cost, or other task-relevant considerations. $\mathcal{M}_t$ concerns the configuration’s monitoring of its own judgment conditions: confidence, perceived uncertainty, source reliability, dependence on external advice, unresolved disagreement, and the need for further checking.

Keeping these layers distinct prevents a common collapse between evaluating an object and evaluating one’s own basis for evaluation. A person may judge one hypothesis superior to another while remaining uncertain about whether the available evidence is adequate. A clinician may prefer one diagnosis while recognizing low confidence because a decisive test is unavailable. A researcher may find an AI-generated explanation elegant while judging its provenance or evidential support too weak for adoption. The second-order monitoring state can therefore regulate reliance and action without simply repeating the first-order evaluation.

The relation can be represented without reducing either layer to a scalar. Equation 5 defines a paired profile.

$$\mathfrak{E}_t^{(q,\xi)} = \left(\mathcal{V}_t^{(q,\xi)}, \mathcal{M}_t^{(q,\xi)} \right). \quad (5)$$

The first component of Equation 5 concerns what reasons or values bear on the object of judgment; the second concerns how the configuration assesses the adequacy and limits of its own evaluative position. Later sections refine these layers into calibration, selective reliance, aesthetic judgment, normative judgment, practical judgment, and temporal and mediating judgment. At the foundational level, the important point is that a strong evaluative preference and a strong warrant for acting on that preference are analytically separable.

Caution 3.3 (Criteria and monitoring are historically situated). The sets $\mathcal{V}_t$ and $\mathcal{M}_t$ should not be interpreted as universal inventories of correct criteria or metacognitive variables. Which criteria are legitimate, which uncertainty signals are available, and which monitoring practices are institutionally recognized can vary across domains, histories, cultures, and technologies. The formalism records these layers without settling their substantive content.

Caution 3.3 is particularly important in AI-mediated inquiry because interface design can make some monitoring signals salient while hiding others. A calibrated probability, a fluent explanation, a citation, a warning banner, or the absence of a confidence estimate can each alter how a recommendation is interpreted. These presentation effects belong partly to $\mathcal{E}_t$ and partly to the monitoring practices available in $\mathcal{C}_t$; the distinction is analytical rather than ontological.

3.4 Closure, Deferral, and Action

Judgment is often represented as a transition from uncertainty to a selected option. That representation is useful in bounded decision tasks but too narrow for many inquiry settings. A contextually appropriate response can consist in acting, requesting more evidence, preserving several hypotheses, postponing commitment, changing the problem frame, or deciding that the present evidence does not support closure. The response variable a_t in Equation 3 therefore includes forms of controlled non-closure.

This broader representation is especially important when AI increases the apparent availability of answers. Fluent completion can create pressure to treat an inquiry as resolved even when the underlying evidence remains incomplete or the decision is highly consequential. The ability to suspend closure can therefore be part of judgment rather than a failure to judge. Conversely, indefinite suspension can itself be irresponsible when delay carries substantial cost. Whether closure, deferral, or continued plurality is appropriate depends on consequence, reversibility, time pressure, evidence quality, and the costs of further inquiry.

Definition 3.4 (Judgment closure). *Judgment closure* is the context-relative stabilization of a response sufficient for the next relevant action, communication, or institutional step. Closure need not imply certainty, final truth, or irreversible commitment. It can be provisional and explicitly revisable.

Definition 3.4 allows later sections to distinguish local closure for action from epistemic openness to revision. It also prevents the paper from treating maximal decisiveness as an intrinsic virtue of judgment.

3.5 Developmental Residue of a Judgment Episode

The central object of the paper appears after the local episode. Equation 2 states that Σ_t^J can change through Δ_t^J . The relevant question is what difference the episode leaves in the configuration that will participate in later judgments. Because the components of Σ_t^J are heterogeneous, this difference should not be represented as ordinary vector subtraction. The paper instead introduces a typed developmental residue in Equation 6.

$$\mathfrak{D}_t^J = \text{Chg}(\Sigma_t^J, \Sigma_{t+1}^J \mid q, \xi), \quad (6)$$

where Chg denotes a task-sensitive description of changes relevant to later judgment. Such changes can occur in the accumulated history H , relational dependencies $\mathcal{R}$, evidence environment $\mathcal{E}$, conceptual resources $\mathcal{L}$, normative or institutional conditions $\mathcal{N}$, or operative capacities $\mathcal{C}$. The residue can be beneficial, harmful, mixed, or negligible. It can include improved discrimination, new error-recognition habits, increased confidence without calibration, dependence on a particular interface, a changed standard of evidence, a preserved audit trail, or redistribution of who is expected to verify and own the decision.

This representation clarifies why immediate success does not settle formation. A correct answer can leave little useful residue when the user cannot reconstruct why it was accepted. An initially slower episode can leave substantial developmental residue when explicit comparison, feedback, and revision improve later discrimination. The developmental value of a judgment episode is therefore temporally extended and cannot be read directly from the final response variable a_t .

Claim 3.5 (Developmental residue principle). Evidence of judgment formation concerns changes in a history-bearing configuration that remain relevant to subsequent judgment. Episode-level accuracy, confidence, reliance, or satisfaction can contribute evidence, but none is sufficient by itself to identify the developmental residue represented in Equation 6.

Claim 3.5 generalizes the performance–formation distinction from Section 1 and the reliance boundary from Claim 2.3. It also identifies the empirical burden for later assessment: studies of judgment cultivation need designs capable of observing change across time, transfer, altered conditions, or later reconstruction rather than measuring only immediate assisted performance.

3.6 Assimilation, Availability, and Transfer

Exposure to a reason, recommendation, correction, or AI-generated explanation does not by itself establish that the material has been incorporated into later judgment capacity. The present paper uses *assimilation* in a limited functional sense: an episode has been assimilated to the extent that it changes how the configuration can later frame, discriminate, monitor, revise, or act in a relevant context. This usage does not imply a particular cognitive mechanism and does not require that the change be beneficial.

The distinction can be stated through a later-availability predicate. Let d denote a change identified within $\mathcal{D}_t^J$. Equation 7 records whether that change remains operative for a later task q' and context ξ' .

$$\text{Avail}_{t+k}^J(d \mid q', \xi') = 1. \quad (7)$$

In Equation 7, the value 1 indicates that d remains available to influence the later judgment episode. The expression is a predicate, not a claim that availability is intrinsically binary. Empirical work can replace it with graded, probabilistic, behavioral, or domain-specific measures. Its analytical purpose is to distinguish a transient effect from a change that remains available for later use.

Transfer imposes a stronger condition. If the later task reproduces the same cues, interface, and recommendation structure, successful reuse may reflect local procedural learning. Evidence of broader formation becomes stronger when a relevant change remains useful under altered examples, different surface features, changed models, new domains, or different distributions of error. The degree of transfer required depends on the claim being made; the framework does not assume that judgment formation must generalize without limit.

Definition 3.6 (Judgment formation). Within the present framework, *judgment formation* is a history-bearing change in a judgment configuration whose developmental residue remains available to influence later framing, evaluation, monitoring, reliance, revision, closure, or action in a task-relevant context. Stronger claims of formation require correspondingly stronger evidence concerning persistence, transfer, and revisability.

Definition 3.6 refines the conceptual definition introduced in Section 1. It also leaves room for negative formation. A recurrent AI-mediated practice can form judgment poorly by stabilizing overconfidence, dependency, premature closure, or narrow option generation. Formation names durable organization, while the normative evaluation of that organization remains a separate question.

3.7 Formation without a Scalar Ideal

The formal objects above do not imply that judgment formation can be represented by one developmental score. Calibration can improve while independent option generation weakens. Reliance can become more discriminating while responsibility becomes less clear. Faster closure can improve performance in one environment while making the configuration brittle under distribution shift. The relevant developmental profile is therefore multidimensional and task-sensitive.

For the same reason, the paper does not define desirable judgment formation through maximal independence from AI. Some configurations can become better formed by learning when to rely on specialized artificial support, how to interrogate it, and how to preserve provenance and revision pathways. Other configurations may require more independent practice because excessive delegation would remove the experiences through which discrimination is learned. The appropriate balance depends on what capacities the task requires and how errors become visible.

Caution 3.7 (Formation and normative evaluation). Durable formation is not equivalent to desirable formation. A stable habit of premature deference, unjustified confidence, narrow framing, or responsibility diffusion can be genuinely formed while remaining epistemically or normatively poor. The descriptive question of what has become durable should therefore remain distinct from the evaluative question of what ought to be cultivated.

Caution 3.7 establishes a boundary that becomes important throughout the manuscript. The GR framework tracks how judgment capacities are generated and reorganized; later sections evaluate calibration, reliance, mediation among judgment forms, responsibility, educational practice, and the conditions under which particular forms of development are worth preserving.

3.8 Formal Status and Scope

The constructions in this section are analytical representations. Σ_t^J is not proposed as the hidden substance of a judging mind, $\mathfrak{J}_t^{(q,\xi)}$ is not a complete causal model of a decision episode, and $\mathfrak{D}_t^J$ does not establish a measurable metric of development. Their role is to prevent several recurrent category errors: inferring formation from performance, inferring judgment quality from confidence, treating the option set as exogenously fixed, treating immediate closure as the only valid response, and treating exposure to AI advice as equivalent to durable learning.

The framework also preserves methodological pluralism. Experimental decision tasks can estimate calibration and reliance under controlled conditions; longitudinal studies can examine persistence and transfer; qualitative work can reconstruct how practitioners interpret responsibility and uncertainty; historical inquiry can study how judgment criteria become institutionalized; and computational models can test specific update mechanisms where variables can be operationalized responsibly. These modes of inquiry address partially different aspects of the architecture and should not be treated as interchangeable.

The resulting foundation can be summarized as a sequence of analytically distinct relations. A history-bearing configuration participates in a situated judgment episode; the episode organizes salience, options, warrants, criteria, monitoring, and response; the episode can change the configuration; and some changes remain available for later judgment. AI mediation can enter at each stage. The remainder of the paper examines how that mediation transforms the judgment environment, redistributes judgment across Human–AI configurations, and changes the practical and educational conditions under which judgment can remain calibrated, revisable, and responsibility-bearing.

4 AI-Mediated Transformation of the Judgment Environment

This section examines how artificial intelligence can change the structure of a judgment episode before any final decision or recommendation is evaluated. Its role is to move from the formal decomposition developed in Section 3 to the environment in which that decomposition is realized. The objective is to identify recurrent forms of mediation through which AI can alter salience, candidate options, warrants, metacognitive cues, temporal conditions, verification burdens, and responsibility relations. The discussion is conceptual and literature-informed rather than a claim that one transformation occurs in every AI-assisted task. The central proposition is that AI mediation can reorganize the conditions of judgment itself, so the relevant comparison is often broader than assisted versus unassisted accuracy.

Table 2 summarizes the dimensions developed below. The entries identify possible transformations rather than fixed effects. The same AI system can broaden one user’s option space while narrowing another user’s attention, and the developmental consequences depend on task structure, prior expertise, interface design, feedback, and the practices through which outputs are checked.

4.1 AI Mediation as Structural Transformation

Section 3 represented a situated judgment episode as $\mathfrak{J}_t^{(q,\xi)}$, with components for salience, options, warrants, evaluative criteria, monitoring, and response. AI mediation can enter through several of these components at once. A language model can propose how the problem should be stated, generate candidate alternatives, summarize reasons, evaluate those alternatives under requested criteria, express uncertainty, and recommend a response. The judgment episode is therefore sometimes reorganized before the user reaches the point conventionally described as “making a decision.”

The present paper represents that possibility through the analytical mediation map in Equation 8.

$$\tilde{\mathfrak{J}}_t^{(q,\xi)} = \mathcal{A}_t^{(q,\xi)} \left(\mathfrak{J}_t^{(q,\xi)}, \Sigma_t^J \right), \quad (8)$$

where $\mathcal{A}_t^{(q,\xi)}$ denotes AI mediation relevant to task q and context ξ , and $\tilde{\mathfrak{J}}_t^{(q,\xi)}$ denotes the resulting mediated judgment episode. Equation 8 is not a causal model of a particular AI system. It records

Table 2: Dimensions through which AI mediation can transform a judgment environment.

Dimension	Possible transformation	Judgment consequence	Developmental concern
Salience and framing	Highlights, summarizes, labels, or proposes problem representations	Changes what enters attention and what appears decision-relevant	Recurrent framing can stabilize selective attention or blind spots
Option construction	Generates, ranks, expands, or compresses alternatives	Changes the effective comparison set before evaluation	Breadth can increase while independent option generation weakens
Warrants and reasons	Produces explanations, evidence summaries, counterarguments, or citations	Changes the apparent support available for alternatives	Verification and provenance may become the limiting work
Metacognitive cues	Alters fluency, coherence, confidence cues, and presentation quality	Changes how users monitor answer quality and uncertainty	Performance and confidence can improve at different rates
Temporal structure	Reduces search, drafting, and comparison time; accelerates recommendation cycles	Changes the cost of delaying, checking, or reopening judgment	Fast closure can become habitual even when later tasks require patience
Verification burden	Produces more candidate claims than a user can independently inspect	Redistributes effort from generation toward checking and triage	Verification practices can strengthen or atrophy depending on use
Responsibility relations	Distributes contribution across user, model, data, interface, and institution	Changes who can inspect, contest, revise, and own the outcome	Repeated delegation can alter responsibility expectations and skills

that mediation can alter how an existing history-bearing configuration participates in the local episode. Different systems, prompts, interfaces, retrieval layers, recommendation procedures, and institutional policies instantiate different forms of $\mathcal{A}$.

Definition 4.1 (AI-mediated judgment environment). *An AI-mediated judgment environment is a task- and context-specific arrangement in which artificial systems materially affect one or more conditions through which a judgment episode is framed, populated with alternatives, supported by warrants, monitored, verified, closed, contested, or assigned responsibility.*

Definition 4.1 deliberately includes more than explicit recommendation. An AI system can transform judgment by controlling retrieval order, compressing a large evidence base into a summary, suggesting categories, ranking alternatives, or determining which uncertainty signals are visible. These forms of mediation can matter even when the final decision remains formally assigned to a human user.

4.2 Salience and Problem Representation

Judgment begins partly with what becomes salient enough to enter the episode. AI systems can influence salience by selecting sources, summarizing long materials, proposing diagnostic categories, suggesting questions, and foregrounding one interpretation of an ambiguous situation. Such assistance can reduce search costs and reveal considerations that a user would otherwise miss. It can also make one framing unusually available and thereby reduce attention to alternatives that were never generated or retrieved.

This effect matters because salience is upstream of evaluation. A well-calibrated comparison among the options in $\mathcal{O}_t$ cannot recover an important alternative that never entered the effective option space. The relevant judgment question therefore includes whether the framing and candidate set deserve acceptance, not only whether the selected item is preferable once the set has been fixed.

Messeri and Crockett’s analysis of AI-supported science provides an adjacent caution: tools that promise broader access to information and analytical capacity can also produce illusions of explanatory or exploratory reach when users mistake tool-supported output for a fuller understanding of the relevant space [1]. The present framework extends that concern to judgment formation. Recurrent dependence on AI-supplied framing can change which differences a user learns to notice independently and which forms of uncertainty remain visible.

Claim 4.2 (Upstream mediation principle). AI can influence judgment before explicit evaluation by changing what is represented as the problem and which considerations become salient. Assessment of judgment quality should therefore examine framing and option construction as well as the final choice or recommendation.

Claim 4.2 does not imply that human framing is neutral. Human inquiry is always selective. The point is comparative and developmental: AI introduces additional mechanisms of selection, compression, ranking, and suggestion whose effects can become recurrent parts of later judgment practice.

4.3 Expansion and Compression of Option Space

Generative systems can make alternative production unusually cheap. A user can request ten hypotheses, five legal interpretations, several policy scenarios, multiple counterarguments, or a set of possible diagnoses in seconds. This capacity can enlarge the effective option space and reduce premature fixation on the first plausible answer. It can also create a false appearance of comprehensiveness when the generated alternatives share hidden assumptions, training-data regularities, or a common framing supplied by the prompt.

AI mediation can also compress the option space. Ranking, summarization, and single-answer interfaces can reduce a complex field of possibilities to a small number of apparently dominant alternatives. Expansion and compression are therefore complementary transformations. A system may generate many textual options internally while presenting only one recommendation to the user, or it may present many surface variants that occupy a narrow conceptual region.

For judgment formation, the relevant capacity is not maximal option count. Mature judgment requires sensitivity to omitted alternatives, redundancy among generated candidates, the cost of further search, and the point at which additional options cease to improve discrimination. Repeated AI use can strengthen this capacity when users compare candidate sets and learn to identify missing dimensions. It can weaken it when option generation is consistently outsourced and users stop practicing independent problem representation.

Caution 4.3 (Candidate abundance and option adequacy). A large set of AI-generated alternatives does not establish that the relevant option space has been adequately represented. Surface diversity can coexist with shared assumptions, omitted categories, or repeated variants of one underlying approach.

Caution 4.3 becomes important later when the manuscript examines calibration and selective reliance. A user can be well calibrated about the relative quality of presented options while remaining unaware that the decisive alternative was excluded upstream.

4.4 Warrant Abundance and Provenance Burden

AI mediation also changes the production of reasons. Systems can generate explanations, supporting considerations, objections, analogies, citations, and plausible causal narratives at a speed that exceeds unaided human production. This abundance can improve judgment when it exposes hidden assumptions or supplies counterarguments. It can also shift the bottleneck from reason generation to warrant evaluation.

The distinction between a reason’s availability and its warrant becomes especially important in generative systems. Fluent explanation can make a claim easier to understand without establishing that the

explanation is evidentially adequate. A citation-like reference can increase apparent specificity while still requiring source verification. A balanced list of pros and cons can improve deliberative structure while omitting the dimensions that matter most for the actual decision. These possibilities are consistent with the broader concern that AI-supported inquiry can create an illusion of understanding when polished outputs are treated as substitutes for reconstructed epistemic support [1].

The present framework therefore treats provenance as part of the judgment environment. Users need different verification practices when a claim can be traced to inspectable primary evidence than when it is generated through an opaque synthesis whose supporting path is unavailable. The developmental question is whether repeated AI use strengthens habits of provenance checking, triangulation, and source-sensitive confidence, or whether polished warrant production gradually lowers the threshold for closure.

Claim 4.4 (Verification redistribution). When AI lowers the cost of generating reasons, alternatives, and explanations, the scarce work of judgment can shift toward verification, provenance assessment, comparison, and triage. The developmental effect depends on whether those activities remain practiced within the inquiry configuration.

Claim 4.4 does not assume that verification must always be performed manually. Verification itself can be distributed across tools and people. The relevant issue is whether the configuration retains reliable mechanisms for distinguishing supported from unsupported claims and for escalating uncertainty when those mechanisms are weak.

4.5 Metacognitive Cue Reconfiguration

AI systems change not only the informational content of judgment but also the cues from which users infer quality. Response fluency, grammatical polish, speed, length, structured formatting, apparent balance, and confident wording can all affect how an answer feels to evaluate. These cues can sometimes correlate with useful qualities such as coherence or completeness, yet they are imperfect indicators of correctness and warrant.

Recent evidence makes this separation concrete. Fernandes and colleagues report AI-assisted performance gains alongside overestimation of performance and imperfect metacognitive accuracy in their study settings [2]. Soleimanof and Neufeld likewise identify metacognitive sensitivity as relevant to how users respond to AI advice [3]. These findings do not establish one universal effect of AI on confidence. They support the narrower point that AI mediation can change first-order performance and second-order monitoring through partially distinct pathways.

For formation, repeated exposure to polished outputs can recalibrate users in different directions. Some users may learn that fluency deserves little evidential weight and become more source-sensitive. Others may internalize stylistic confidence as a proxy for quality. The resulting developmental residue belongs in $\mathcal{D}_t^J$ because it changes how later judgments are monitored even when the AI system is absent or different.

4.6 Temporal Compression, Verification Load, and Closure

Generative AI reduces the time required for many operations that previously slowed inquiry: initial search, drafting, comparison, reformulation, translation, and candidate generation. Lower friction can be valuable, especially when delay has real costs. It also changes the temporal structure of judgment by making rapid iteration and rapid closure technically easy.

The effect is not simply “faster decisions.” When generation becomes cheap, users can face more material to verify within the same available time. The cost of producing one additional explanation can approach zero while the cost of checking that explanation remains substantial. A judgment environment can therefore combine temporal compression at the generation stage with increased verification load at the evaluation stage.

The cognitive-forcing evidence reviewed in Section 2 illustrates the corresponding design tension: added friction can reduce overreliance while imposing usability costs [4]. Section 15 treats that tension as a developmental problem rather than a general argument for more effort.

The present section adds a temporal concern. Recurrent access to immediate recommendations can change the perceived legitimacy of delay. Requests for more evidence, independent reconstruction, or preservation of unresolved alternatives may begin to feel inefficient even when they are appropriate responses. Judgment formation therefore concerns the learned timing of closure as well as the correctness of the eventual choice.

Definition 4.5 (Closure pressure). *Closure pressure* denotes features of a judgment environment that increase the practical, social, interface-level, or cognitive tendency to treat a judgment as sufficiently settled before additional verification, alternative generation, or reconsideration occurs.

Definition 4.5 is intentionally descriptive. Closure pressure can be beneficial when delay is costly and evidence is sufficient, and harmful when uncertainty remains material. The relevant capacity is calibrated closure rather than maximal delay.

4.7 Responsibility, Contestability, and Judgment Ownership

AI mediation can distribute causal contribution across a wider configuration while leaving formal authority concentrated in one person or institution. A model may generate the recommendation, an interface may rank it, a database may determine which evidence was retrieved, organizational policy may require the tool’s use, and a human professional may remain legally or institutionally responsible for the final action. The resulting responsibility structure cannot be inferred from causal contribution alone.

Automation research has long emphasized that reliance depends on system reliability, trust, monitoring conditions, and task context [7, 8]. Contemporary generative systems add a further problem of contestability. A user can sometimes inspect and challenge an explicit recommendation while lacking access to the internal process that generated it. Conversely, a system can make contestation easier by exposing alternatives, uncertainty, sources, or counterarguments. The relevant judgment environment therefore includes the pathways through which outputs can be questioned, revised, escalated, or rejected.

For formation, responsibility is partly learned through practice. Repeated workflows can teach users that AI output is advisory and must be owned through independent review, or they can normalize a pattern in which the human supplies nominal approval after substantial evaluative work has already been delegated. The visible final decision may look identical in both cases while the developmental residue differs substantially.

Claim 4.6 (Responsibility-structure dependence). The developmental consequences of AI-assisted judgment depend partly on who is expected and able to inspect, contest, revise, and answer for the outcome. Similar recommendations can form different judgment practices under different responsibility and contestability structures.

Claim 4.6 prepares the later analysis of delegation and judgment ownership. At this stage the point is narrower: responsibility relations are part of the environment in which judgment is practiced and therefore part of what repeated AI mediation can reorganize.

4.8 Recurrent Mediation and Developmental Residue

The transformations above become a problem of judgment formation when they recur. A single mediated episode can change the local outcome without leaving a durable trace. Repeated exposure to the same forms of salience, option generation, warrant production, confidence cues, closure timing, and responsibility allocation can become part of the history H_t and operative capacities C_t of later judgment configurations.

The relation to Section 3 can be stated by distinguishing the local mediation map in Equation 8 from the longer-run judgment update. AI mediation enters the episode, the episode produces a developmental

residue, and some part of that residue may remain available later. The same technology can therefore support very different developmental trajectories depending on how it is embedded. A workflow that requires reconstruction, source checking, explicit confidence revision, and post-outcome review can leave different residue from a workflow that rewards immediate acceptance of fluent recommendations.

Claim 4.7 (Environment-formation coupling). AI-mediated judgment environments can contribute to judgment formation when recurrent features of mediation become assimilated into later framing, option construction, warrant assessment, monitoring, closure, verification, or responsibility practices. The direction of that formation remains an empirical and normative question.

Claim 4.7 sets the boundary for the next stage of the manuscript. The present section identifies transformations in the judgment environment. Subsequent analysis can then ask how functions are distributed across Human–AI configurations and how reliance should be calibrated within those configurations without treating environmental transformation itself as evidence of improvement or decline.

5 Human–AI Judgment Configurations

This section develops the configuration-level unit of analysis required once AI participates in more than one local step of inquiry. Its role is to distinguish a judgment configuration from a simple pairing between one human user and one model output. The objective is to identify where judgment-relevant functions are located, how they are coupled, how analytical boundaries should be drawn, and why high performance of a Human–AI arrangement does not by itself identify where judgment capacity resides. The discussion is functional and relational. It does not treat a Human–AI configuration as one metaphysical subject, assign moral responsibility by formal notation, or assume that greater distribution of functions is intrinsically desirable.

5.1 Judgment Functions and Configuration Boundaries

Human–automation research has long emphasized that automation can be introduced at different stages of information processing rather than only at final decision selection. Parasuraman, Sheridan, and Wickens distinguish automation of information acquisition, information analysis, decision and action selection, and action implementation [13]. The present framework adopts the underlying function-allocation insight while expanding the analytical vocabulary to match the broader judgment record in Equation 3.

Let $\mathcal{U}_t^{(q,\xi)}$ denote the set of components included within the analytical boundary of a Human–AI judgment configuration for task q and context ξ . Depending on the question, those components can include one or more human inquirers, AI models, retrieval systems, databases, sensors, interfaces, institutional procedures, and other recurrent supports. Let Φ^J denote the family of judgment-relevant function types used in the analysis. A provisional allocation relation is given in Equation 9.

$$\Lambda_t^{(q,\xi)} \subseteq \Phi^J \times \mathcal{U}_t^{(q,\xi)}. \quad (9)$$

Equation 9 is a relation rather than a function because the same judgment function can be jointly realized by several components, and one component can participate in several functions. The relevant function family can include problem framing and salience selection, evidence acquisition, option construction, warrant synthesis, criterion application, monitoring and calibration, closure or action selection, implementation, and post-outcome review. Table 3 summarizes these functions and the formation questions they create.

The analytical boundary in $\mathcal{U}_t^{(q,\xi)}$ should be chosen relative to the claim being investigated. A narrow boundary can be appropriate when studying whether a user detects a model error. A broader boundary may be necessary when retrieval infrastructure, institutional policy, or a downstream reviewer materially determines which evidence is visible or whether an AI-supported recommendation can be contested. The purpose of the boundary is therefore explanatory rather than ontological.

Table 3: Judgment-relevant functions in a Human–AI configuration

Function	Possible Human–AI distribution	Formation concern
Problem framing and salience	Human question, model reframing, search/ranking cues	What remains independently noticeable or contestable?
Evidence acquisition	Human observation, automated retrieval, databases, sensors	Can provenance and evidence limits still be reconstructed?
Option construction	Human generation, model generation, hybrid iteration	Does the configuration retain sensitivity to omitted alternatives?
Warrant synthesis	Human reasoning, model explanation, tool-supported comparison	Are reasons checked, compared, and source-sensitive?
Criterion application	Human standards, institutional rules, model scoring or ranking	Which criteria remain visible and revisable?
Monitoring and calibration	Human confidence, model uncertainty cues, external checks	Can the configuration detect when further verification is needed?
Closure and action selection	Human decision, model recommendation, policy constraint	Who can delay, veto, revise, or preserve plurality?
Review and learning	Human reflection, automated logging, outcome analytics	What residue remains available for later judgment?

Caution 5.1 (Boundary selection). A Human–AI configuration should not be expanded until every remote contributor or infrastructure provider is treated as part of one judging subject. The boundary should include components whose recurrent relations materially affect the judgment functions relevant to the claim under study, while preserving external dependencies as external when they need not be modeled internally.

Caution 5.1 prevents two opposite errors: reducing a complex workflow to a single user–model dyad when important functions are distributed elsewhere, and defining the configuration so broadly that attribution of capacity becomes analytically meaningless.

5.2 Coupling, Interdependence, and Joint Realization

Function allocation does not imply that components operate independently and then contribute separable outputs. In recurrent inquiry, a model response can change the human’s framing; the human’s reformulation can change the model’s search space; retrieved sources can alter both confidence and option construction; institutional rules can constrain which model outputs may be acted upon. The relevant object is therefore a coupled configuration in which local states can alter the later contribution of other components.

For analytical purposes, let $\mathcal{R}_t^{HA,(q,\xi)}$ denote the task-relevant relations among components in $\mathcal{U}_t^{(q,\xi)}$. The configuration-level representation in Equation 10 records both component membership and coupling.

$$\Omega_t^{HA,(q,\xi)} = \left(\mathcal{U}_t^{(q,\xi)}, \mathcal{R}_t^{HA,(q,\xi)}, \Lambda_t^{(q,\xi)} \right). \quad (10)$$

Equation 10 does not replace the history-bearing configuration Σ_t^J . Σ_t^J represents the capacities and conditions relevant to formation across time, while $\Omega_t^{HA,(q,\xi)}$ is a local structural representation of how judgment functions are distributed and coupled for a selected task. The two objects answer different questions.

Definition 5.2 (Coupled Human–AI judgment configuration). A *coupled Human–AI judgment configuration* is a task- and context-indexed arrangement in which judgment-relevant functions are distributed across human, artificial, and supporting components whose contributions are materially interdependent for the episode under study. The definition concerns analytical organization and does not establish collective personhood, agency, or moral status.

Definition 5.2 allows the paper to distinguish recurrent configuration-level judgment from incidental tool use. A calculator used once can influence an answer without materially reorganizing the judgment architecture. By contrast, a workflow in which AI repeatedly frames cases, generates alternatives, supplies reasons, shapes confidence, and participates in review can become part of the recurrent environment through which judgment is practiced and formed.

5.3 Functional Contribution, Authority, and Responsibility

A central analytical risk is to infer authority or responsibility directly from functional contribution. A model can generate most of the candidate options without possessing institutional authority to decide. A human can retain formal approval authority while contributing little to option generation or warrant assessment. An institution can determine the admissible criteria while neither the frontline human nor the model can revise those criteria. These layers should therefore be kept separate.

The present section uses $\Lambda_t^{(q,\xi)}$ only for functional participation. Let $\text{Auth}_t^{(q,\xi)}$ denote the local distribution of authority to close, veto, revise, or escalate a judgment episode. Responsibility is left as a distinct normative and institutional relation for later analysis rather than inferred from either Λ or Auth . This separation matters because a nominal “human in the loop” can retain formal authority while lacking time, evidence access, expertise, or practical permission to contest the system meaningfully.

Claim 5.3 (Functional-authority separation). The distribution of judgment-relevant work, the distribution of decision authority, and the distribution of responsibility can diverge within the same Human–AI configuration. Analysis of judgment formation should therefore avoid treating visible final approval as sufficient evidence that the approving component exercised the judgment capacities attributed to it.

Claim 5.3 prepares the later treatment of delegation and judgment ownership. Its immediate implication is methodological: studies of Human–AI judgment need to observe what each component actually contributes and can contest, rather than classify systems only by who performs the final click or signature.

5.4 Complementarity, Redundancy, and Correlated Failure

Distributed configurations can support judgment through complementarity. A model can search a larger candidate space while a human supplies contextual constraints; one model can generate alternatives while another tool verifies citations; a specialist can interpret domain-specific evidence while institutional procedures preserve auditability. Distribution can also create redundancy when several components independently inspect the same high-risk claim.

The same distribution can fail in a correlated way. Human and model may rely on the same database, inherit the same framing, or share the same omitted variable. A human reviewer can appear independent while evaluating only the reasons already selected by the model. Multiple AI systems can provide apparent plurality while drawing on strongly overlapping training distributions or retrieval sources. Configuration-level robustness therefore depends on the independence and contestability of relevant checks, not on the visible number of components.

This point extends established automation research, which emphasizes that automation can alter human monitoring, workload, and coordination demands rather than merely replacing manual operations [13, 7]. In generative settings, the coordination problem also includes epistemic diversity: whether different parts of the configuration can expose distinct assumptions, sources, alternatives, and failure modes.

Claim 5.4 (Configuration-level performance). Strong performance by a Human–AI configuration can result from complementarity, redundancy, substitution, or compensatory support. Configuration-level success therefore does not by itself establish that each component possesses the capacities required to reproduce, explain, or revise that success independently.

Claim 5.4 is the configuration-level analogue of the performance–formation distinction introduced in Section 1. It is especially important for educational and professional evaluation, where assisted success

can be mistaken for evidence that a particular human or model component has developed the corresponding judgment capacity.

5.5 Configuration Performance and Component Formation

Repeated episodes make the separation between configuration performance and component formation especially important. Suppose option generation is reliably supplied by AI while a human remains responsible for verification and closure. The coupled configuration may improve as the human becomes better at evaluating generated alternatives. At the same time, the human can receive less practice in generating alternatives independently. Formation can therefore be asymmetric across functions.

The relation can be expressed by projecting developmental residue onto a selected component or function family. Let $u \in \mathcal{U}_t^{(q,\xi)}$ and $f \in \Phi^J$. Equation 11 denotes the part of the developmental residue relevant to later exercise of function f by component u .

$$\mathfrak{D}_t^J[u; f] = \Pi_{u,f}(\mathfrak{D}_t^J), \quad (11)$$

where $\Pi_{u,f}$ is an analytical projection rather than an assumption that developmental changes can always be uniquely decomposed. Equation 11 is useful precisely because one episode can strengthen monitoring while weakening independent option construction, or strengthen source verification while increasing dependence on externally supplied criteria.

This functional asymmetry also clarifies why “keeping the human involved” is an incomplete developmental prescription. Human participation can occur in a function that provides little practice for the capacity one intends to preserve. If a training goal concerns independent problem framing, requiring a human to approve AI-generated options does not necessarily exercise that capacity. Conversely, if the target capacity is selective reliance, exposure to fallible AI advice with feedback can create useful practice even when the human does not generate every candidate independently.

Claim 5.5 (Formation follows practiced functions). Within a coupled judgment configuration, developmental residue is partly conditioned by which judgment functions a component repeatedly performs, monitors, contests, or reconstructs. Nominal participation in the final outcome is therefore weaker evidence of formation than repeated practice in the function whose development is being claimed.

Claim 5.5 will matter later when the paper considers cultivation, formative cognitive friction, and assessment. It also gives a configuration-level interpretation to the finding that appropriate reliance can involve learning across repeated Human–AI interactions [12].

5.6 Boundary Transparency and Hidden Supports

Judgment configurations can appear simpler than they are. A user may interact with one visible model while the result depends on retrieval infrastructure, hidden system instructions, external evaluators, organizational defaults, or previously curated data. Conversely, an analysis that includes every infrastructural dependency can make it impossible to identify which relations matter for the local formation claim.

The framework therefore requires *boundary transparency*: an account of why particular components are inside the analytical configuration and which consequential dependencies remain outside it. This requirement is especially important when comparing “human,” “AI,” and “Human–AI” performance. A comparison can become misleading when one condition includes extensive hidden support and another is treated as an isolated individual.

Boundary transparency also protects against a common attribution error in AI-mediated inquiry. A polished model output can embody substantial human labor in data creation, evaluation, prompt design, tool construction, or institutional filtering. Recognizing those relations does not require expanding the judging subject indefinitely. It requires making the support structure visible enough that claims about capacity, independence, and formation are scoped correctly.

Caution 5.6 (Capacity attribution across boundaries). A capacity attributed to a Human–AI configuration should not automatically be attributed to a human component, an AI component, or the configuration under a different boundary. Attribution should track the functions, supports, and couplings required for the observed performance and the persistence of those relations across the tasks to which the claim is extended.

Caution 5.6 preserves the revisability of configuration analysis as AI systems and institutional work-flows change. A function that currently requires recurrent human correction can later become more reliably automated; a capability that appears model-internal can turn out to depend strongly on retrieval or human curation; a coupled configuration can become more or less robust as its interfaces change.

5.7 Developmental Implications of Distributed Judgment

The principal contribution of configuration analysis to this paper is developmental. Judgment formation is not determined only by whether AI is present. It depends on which functions are distributed, how strongly components depend on one another, where errors become visible, who can contest closure, and which capacities remain practiced across time.

This perspective also avoids treating independence as the universal educational goal. Mature judgment can include knowing when to rely on specialized systems, when to seek redundancy, how to preserve provenance, how to escalate uncertainty, and how to recognize that a configuration is operating outside the conditions under which its distributed competence was formed. The relevant developmental object is the capacity to participate responsibly in a configuration while retaining enough reconstructive and monitoring ability to detect when the configuration itself should be questioned.

Claim 5.7 (Distributed-formation principle). Judgment formation in Human–AI inquiry can occur at multiple analytical levels: within a human component, within recurrent coordination practices, and within the configuration’s retained relations, records, and correction pathways. Claims about development should specify the level at which the relevant residue is expected to persist.

Claim 5.7 establishes the boundary for the next section. Configuration analysis identifies where judgment-relevant functions and dependencies are located. The further question is how a configuration should calibrate reliance across those functions when human and artificial components vary in competence, uncertainty, transparency, and consequence.

6 Calibration of Selective Reliance

This section develops calibration and selective reliance as judgment capacities within recurrent Human–AI inquiry. Its role is to move from the configuration-level allocation of judgment functions in Section 5 to the local question of how an inquirer or coupled configuration should relate to artificial assistance under uncertainty. The objective is to distinguish confidence calibration, reliance calibration, verification intensity, and closure timing without reducing them to a single scalar trust score. The discussion is conceptual, with interfaces to metacognition, automation reliance, algorithm aversion and appreciation, cognitive forcing, and appropriate-reliance research. The framework does not prescribe one universal reliance policy. It identifies dimensions that make selective reliance inspectable, revisable, and developmentally consequential.

6.1 Calibration across Confidence, Warrant, and Consequence

Calibration is often discussed through the relation between confidence and accuracy. That relation is indispensable but incomplete for AI-mediated inquiry. A user can be well calibrated about personal confidence while being poorly calibrated about the reliability of a model on the local task, the adequacy of the evidence base, or the consequences of acting too quickly. Conversely, a user can have low confidence

Table 4: Selective-reliance modes in AI-mediated inquiry

Mode	Local relation to AI contribution	Principal judgment burden
Adopt	Use the contribution substantially as supplied	Establish that local reliability and consequence justify low additional checking
Verify	Use after independent or partially independent checking	Select checks with enough independence and diagnostic value
Contest	Challenge framing, evidence, inference, or recommendation	Identify a specific basis for disagreement and preserve alternatives
Defer or escalate	Seek another person, system, source, or procedure	Recognize limits in the present configuration and choose an appropriate escalation path
Suspend	Preserve unresolved alternatives and postpone closure	Judge that current uncertainty or consequence warrants delay
Proceed independently	Continue without the AI contribution for the relevant function	Preserve or exercise an independent capacity where external assistance would add insufficient value or obscure evaluation

while making an appropriate decision to rely on a system whose local reliability is substantially higher and independently supported.

The present framework therefore distinguishes three related calibration problems. *Metacognitive calibration* concerns how internal confidence tracks the quality and uncertainty of one’s own judgment. *Reliance calibration* concerns whether the degree and form of dependence on an external component are appropriate to its task-local contribution and limitations. *Closure calibration* concerns whether the inquiry is treated as sufficiently settled at an appropriate point relative to uncertainty, stakes, reversibility, and the availability of further checks. These forms of calibration can diverge within the same episode.

The metacognitive distinctions reviewed in Section 2 separate confidence level, confidence discrimination, and first-order performance; recent AI-advice evidence likewise links confidence and metacognitive sensitivity to reliance behavior [6, 3]. The narrower proposition used here is that internal confidence is one input into reliance, while selective reliance also depends on task-local warrant, consequence, verification, and available alternatives.

Claim 6.1 (Multidimensional calibration). Calibration in AI-mediated inquiry is multidimensional. Appropriate confidence does not by itself establish appropriate reliance or appropriately timed closure, and successful reliance in one episode does not by itself establish durable improvement in metacognitive calibration.

Claim 6.1 prevents a common collapse between feeling appropriately uncertain, using AI appropriately, and forming better judgment. Each can support the others, yet each requires distinct evidence.

6.2 Selective Reliance as a Mode Choice

A binary trust question obscures the range of responses available to an inquirer. AI-mediated inquiry often permits several distinct reliance modes. Table 4 summarizes six recurrent modes and the judgment burden associated with each.

The set of modes in Table 4 is not exhaustive and the categories can be combined. A researcher can adopt a low-stakes formatting suggestion while independently verifying a factual claim in the same response. A clinician can use AI-generated differential possibilities while escalating final interpretation to another specialist. A student can request hints while deliberately preserving independent problem construction. Selective reliance is therefore function-specific and task-relative.

Definition 6.2 (Selective reliance). Selective reliance is the capacity of an inquirer or recurrent judgment configuration to choose, justify, and revise among different modes of dependence on an external component according to task-local competence, uncertainty, provenance, stakes, reversibility, available verification, and the capacities whose continued exercise matters for later judgment.

Definition 6.2 extends the appropriate-reliance literature without replacing its empirical measures. Schemmer and colleagues treat appropriate reliance through the pattern of correct acceptance and correct rejection of AI advice [11]. The present definition adds a developmental dimension: a reliance pattern can be locally appropriate while still reorganizing later capacities in ways that matter for formation.

6.3 Task-Local Reliability and the Limits of Global Trust

Trust in automation is shaped by system properties, task characteristics, user factors, and context [8, 7]. In generative AI, this contextuality becomes especially important because the same system can vary substantially across functions within one inquiry. It can be strong at language reformulation, useful for candidate generation, weak at provenance-sensitive factual reconstruction, or difficult to evaluate in a specialized domain. A global disposition such as “I trust this model” or “I do not trust AI” is therefore too coarse for many judgment tasks.

Selective reliance requires localizing the claim. The relevant question concerns whether the contribution associated with a specified function, domain, representation, and consequence is sufficiently reliable for the intended mode of use. This local requirement also applies to humans. Expertise can be narrow, institutional authority can exceed local competence, and confidence can fail to track performance. The framework therefore treats comparative competence as a relation among components and functions rather than as a permanent property of one class of agent.

Algorithm aversion and algorithm appreciation demonstrate why global reliance tendencies can move in opposite directions under different conditions. Dietvorst and colleagues show that observing an algorithm err can produce avoidance even when the algorithm remains comparatively useful [9]; Logg and colleagues show conditions under which people give greater weight to algorithmic than human advice [10]. The present account uses these results to motivate local calibration rather than to infer a universally optimal level of algorithmic reliance.

Caution 6.3 (Global trust compression). Global labels such as “trusting AI” and “distrusting AI” can compress function-specific differences in competence, provenance, uncertainty, and consequence. They should not substitute for task-local analysis when those differences materially affect judgment.

Caution 6.3 motivates the task-local analysis that follows: verification requirements should respond to the reliability and consequence structure of the particular judgment relation rather than to a global attitude toward AI.

6.4 Verification Intensity, Stakes, and Reversibility

Reliance also depends on the cost of being wrong and the cost of checking. Verification can consume time, money, attention, specialist access, and opportunity. Requiring maximal verification for every AI contribution would defeat many legitimate benefits of assistance, while requiring almost none can allow fluent but weakly warranted outputs to accelerate closure. The practical problem is therefore one of calibrated verification intensity.

The relevant dimensions include at least uncertainty, consequence magnitude, reversibility, error detectability, availability of independent evidence, and the degree to which later correction remains possible. A low-stakes reversible formatting decision can justify immediate adoption under conditions where a legal, medical, engineering, or high-consequence research claim would warrant independent verification or escalation. The same factual uncertainty can rationally support different verification burdens when the consequences differ.

Cognitive-forcing evidence shows why verification effort should be calibrated rather than maximized: added engagement can reduce overreliance while increasing effort and lowering subjective preference [4]. The relevant burden should therefore track the epistemic and practical structure of the case.

Claim 6.4 (Calibrated verification). Verification intensity is appropriately analyzed relative to uncertainty, consequence, reversibility, available independence of checks, and verification cost. A judgment architecture can therefore be poorly calibrated both by accepting too easily and by imposing costly verification where little is gained.

Claim 6.4 will later support the paper’s distinction between formative cognitive friction and misallocated difficulty.

6.5 Reliance Policies as Developmental Objects

Selective reliance becomes a problem of judgment formation when repeated episodes generate a relatively stable policy about when and how to use artificial assistance. Such a policy need not be explicit. A user can gradually become more likely to accept fluent recommendations, more likely to verify citations, less likely to formulate independent alternatives, or more skilled at recognizing domains in which escalation is required. These changes can enter the history H_t and operative capacities $\mathcal{C}_t$ of the judgment configuration.

Let $\varrho_t^{(q,\xi)}$ denote a task- and context-indexed reliance policy: a dispositional mapping from features of a judgment situation to one or more reliance modes of the kind listed in Table 4. Equation 12 represents its possible revision after a mediated episode.

$$\varrho_{t+1}^{(q,\xi)} = \mathcal{T}_t^{\text{rel}} \left(\varrho_t^{(q,\xi)}, \mathcal{D}_t^J, \mathcal{I}_t^{\text{fb}} \right), \quad (12)$$

where $\mathcal{D}_t^J$ is the developmental residue defined earlier and $\mathcal{I}_t^{\text{fb}}$ denotes outcome or feedback information available after the episode. The transition operator $\mathcal{T}_t^{\text{rel}}$ is a formal representation, not a specified learning algorithm. Equation 12 makes one developmental dependency explicit: reliance can improve only when relevant feedback is available, retained, and interpreted well enough to revise future use.

Experimental work linking human learning and appropriate reliance provides an empirical interface for this update problem [12]. The developmental object here is broader, including verification, escalation, reconstruction, closure, and responsibility practices.

6.6 Overreliance, Underreliance, and Pseudo-Calibration

Two familiar failure modes are overreliance and underreliance. Overreliance occurs when external advice receives more weight than its local warrant supports; underreliance occurs when useful assistance is rejected or discounted despite relevant evidence of comparative reliability. Automation research has long treated misuse and disuse as distinct problems [7]. AI-mediated inquiry adds a third failure mode that is especially important for judgment formation: *pseudo-calibration*.

Pseudo-calibration occurs when a configuration appears selective because reliance varies across cases, yet the variation tracks weak cues rather than epistemically relevant differences. A user might verify outputs that sound uncertain while accepting fluent outputs without checking, rely more heavily on answers accompanied by lengthy explanations regardless of source quality, or distrust a model after one visible error while accepting human authority without comparable scrutiny. The resulting behavior is selective in form but poorly calibrated in basis.

Definition 6.5 (Pseudo-calibration). Pseudo-calibration is a pattern in which reliance varies across cases but the variation is driven primarily by cues that do not adequately track the relevant competence, uncertainty, provenance, consequence, or verification structure of those cases.

Definition 6.5 is deliberately broader than one interface effect. It identifies a developmental risk: repeated exposure to persuasive but weak cues can form a stable reliance policy that looks discriminating while remaining epistemically fragile.

6.7 Selective Reliance as a Judgment Capacity

Selective reliance is therefore more than a local behavioral outcome. As a judgment capacity, it involves recognizing the relevant function, locating uncertainty, comparing available competence, choosing an appropriate reliance mode, calibrating verification effort, preserving contestability, and updating future reliance from outcome feedback. These operations can be distributed across a Human–AI configuration, but their developmental residue should remain attributable at the analytical level at which later judgment is expected to improve.

The capacity also has an important independence condition. Mature reliance does not require that a human reproduce every contribution independently. It requires enough reconstructive, monitoring, or escalation capacity to detect when the configuration is operating outside the conditions under which reliance is warranted. This preserves the possibility of genuine specialization and distributed competence while avoiding a purely deferential model of judgment.

Claim 6.6 (Selective-reliance formation). Selective reliance is formed when repeated mediated episodes leave a revisable, task-sensitive capacity to discriminate among reliance modes using epistemically and practically relevant features, and when that capacity remains available under changed cases rather than being tied only to one familiar interface or model.

Claim 6.6 closes the section’s developmental argument. Calibration determines how the configuration relates confidence, warrant, dependence, verification, and closure. The next analytical problem concerns what enters judgment before reliance is even considered: which phenomena, questions, and alternatives become salient enough to be judged at all.

7 Salience Judgment in Problem Formation

This section develops problem formation and the judgment of salience as an upstream component of judgment practice. Its role is to refine the salience term $\mathcal{S}_t$ introduced in Section 3 and to distinguish the generation or recognition of a possible problem from the later judgment that the problem deserves preservation, monitoring, or further inquiry. The objective is narrower than a theory of creativity and broader than a model of visual or attentional capture. The discussion relates the framework to problem-recognition research, distinguishes experienced salience from endorsed salience, examines AI-mediated changes in visibility and problem supply, and identifies developmental evidence for salience judgment. The central claim is that judgment formation begins partly before explicit option evaluation: it includes learning what differences, anomalies, absences, recurrences, and problem framings deserve to remain available for later judgment.

7.1 Problem Recognition before Problem Evaluation

Inquiry does not always begin with a well-formed problem. Pretz, Naples, and Sternberg distinguish recognizing, defining, and representing a problem from the later procedures used to solve it, and emphasize that many problems must first be identified in the environment before they can be addressed [15]. Csikszentmihalyi and Getzels provide an earlier empirical example in which discovery-oriented behavior during problem formulation was positively related to the originality of the resulting creative products in their art-school study [14]. These literatures provide an important boundary for the present section: problem formation is analytically distinguishable from solving or evaluating a problem that has already been supplied.

The present paper adds a judgment-specific question. A phenomenon can be noticed and a problem can be articulated without yet being judged worthy of continued attention. Conversely, a configuration can preserve a weak or poorly understood signal for later inquiry even when it has not yet generated a well-formed question. The relevant judgment is therefore neither full problem solution nor final research priority. It concerns whether something should remain inside the configuration’s active or recoverable field of inquiry long enough to receive stronger evaluation.

The distinction is especially important in AI-mediated inquiry. Generative systems can produce large numbers of candidate questions, anomalies, summaries, and apparent tensions. Search and ranking systems can make some observations highly visible while rendering others peripheral. The judgment problem is consequently not exhausted by generating more candidates. It includes deciding which candidates or encounters deserve further cognitive, evidential, or institutional attention.

Claim 7.1 (Problem availability precedes stronger evaluation). Aesthetic, normative, and practical judgment can operate only over problems, phenomena, or candidate framings that remain sufficiently available to the judging configuration. Judgment formation therefore includes upstream capacities for recognizing, preserving, and revising what enters the field of later evaluation.

Claim 7.1 does not imply that one stable salience filter should govern every domain. It states only that later evaluation is conditioned by what has survived earlier processes of noticing, framing, retention, and exclusion.

7.2 Experienced Salience and Endorsed Salience

The word *salience* can refer to at least two different relations. *Experienced salience* concerns what attracts or occupies attention. It can arise through surprise, vividness, novelty, repetition, interface position, emotional charge, prestige, or algorithmic ranking. *Endorsed salience* concerns a revisable judgment that a phenomenon, discrepancy, question, or framing deserves preservation, monitoring, or further inquiry. The second relation can agree with the first, but it need not.

For analytical purposes, let $\mathcal{X}_t^{(q,\xi)}$ denote the task- and context-relative field of encounters, candidate questions, observations, and framings accessible to the judgment configuration. The salience component of the situated judgment episode can then be refined as Equation 13.

$$\mathcal{S}_t^{(q,\xi)} = \left(\mathcal{S}_{t,\text{att}}^{(q,\xi)}, \mathcal{S}_{t,\text{end}}^{(q,\xi)} \right), \quad (13)$$

where $\mathcal{S}_{t,\text{att}}^{(q,\xi)}$ records what becomes attentionally prominent and $\mathcal{S}_{t,\text{end}}^{(q,\xi)}$ records what the configuration endorses for continued inquiry. Equation 13 is an analytical decomposition rather than a claim that attention and endorsement are psychologically independent modules.

Definition 7.2 (Judgment of salience). A *judgment of salience* is a revisable determination that an accessible phenomenon, discrepancy, question, absence, recurrence, or candidate framing merits preservation, monitoring, or further inquiry before stronger aesthetic, normative, or practical commitment is made.

Definition 7.2 explains why immediate prominence can be epistemically weak. A fluent AI answer, dramatic anomaly, highly cited claim, or visually emphasized recommendation can dominate $\mathcal{S}_{t,\text{att}}$ while receiving little justified support in $\mathcal{S}_{t,\text{end}}$. The reverse relation also matters. A small residual, ordinary recurrence, missing datum, or unresolved inconsistency can be deliberately retained even when it produces little immediate attentional pull.

The distinction creates a practical form of judgment restraint. Mature inquiry can resist the demand to convert every noticed irregularity into a major problem while also resisting the opposite tendency to discard weak signals merely because they are not already represented by dominant categories. Salience judgment therefore includes both admission and restraint.

7.3 Salience Cues and Their Failure Modes

A judgment of salience requires reasons, but those reasons need not form a universal ranking metric. Different domains can treat different cues as relevant, and the same cue can support attention under one context while misleading under another. Table 5 summarizes recurrent cues that can justify further examination and the failure mode associated with treating each cue as self-validating.

Table 5: Illustrative cues in judgments of salience and their principal failure modes

Cue	Possible reason for continued inquiry	Principal failure mode
Discrepancy	An observation conflicts with an expectation, model, baseline, or prior judgment.	Noise or measurement error can be mistaken for theoretically important anomaly.
Recurrence	A pattern persists across occasions, cases, or representations.	Repetition can reflect shared bias, duplicated data, or stable but trivial structure.
Unresolvedness	Available reasons leave a material gap, contradiction, or ambiguity.	Lack of closure can arise from missing background rather than deep difficulty.
Relational connectivity	A phenomenon appears linked to several problems, domains, scales, or consequences.	Dense association can be mistaken for causal or structural importance.
Neglect or exclusion	A signal, case, or framing is repeatedly omitted by dominant classifications or workflows.	Marginality alone does not establish epistemic value.
Representational sensitivity	The apparent problem changes materially when language, scale, aggregation, or framing changes.	Framing dependence can reflect unstable description rather than a substantive phenomenon.

The cues in Table 5 are diagnostic prompts rather than criteria of final worth. An item can deserve further inquiry because one or several cues are present, yet later aesthetic judgment may find the problem shallow, normative judgment may assign it low priority, or practical judgment may conclude that available resources should be directed elsewhere. Salience judgment therefore preserves a problem for stronger evaluation without pre-empting those later judgments.

This sequencing also makes omission inspectable. If a later failure can be traced to a phenomenon that was visible but repeatedly excluded from $\mathcal{S}_{t,\text{end}}$, the relevant developmental question concerns the salience policy that filtered it out. Conversely, if a configuration repeatedly escalates vivid but uninformative anomalies, the formation problem concerns over-admission. Both error types can persist even when downstream reasoning is technically competent.

7.4 AI-Mediated Salience and the Problem Supply

AI systems alter the supply and visibility of possible problems. Search systems rank; summarizers compress; language models reformulate; recommendation systems foreground; anomaly detectors flag; and generative systems can produce candidate questions at a scale that exceeds unaided human review. These functions can expand inquiry, yet they can also produce a mediated salience structure in which what appears important partly reflects model priors, training distributions, interface defaults, retrieval coverage, prompt framing, or institutional optimization.

The relevant distinction is between *problem supply* and *salience judgment*. A large supply of candidate questions can be useful while leaving the difficult judgment untouched: which candidates should be preserved, investigated, combined, postponed, or discarded? Similarly, an AI system can surface a statistically unusual pattern without establishing that the pattern deserves conceptual or practical escalation.

This creates a new form of scarcity. When candidate problems become inexpensive to generate, attention, verification, historical reconstruction, and discrimination can become relatively more important. The configuration must decide not only whether a candidate is plausible but whether it deserves a place in the limited continuation space of inquiry. This is one reason that AI-mediated problem generation does not eliminate the need for judgment of salience.

Caution 7.3 (Mediated visibility and epistemic importance). Visibility produced by ranking, generation, summarization, repetition, or fluent presentation should not be treated as direct evidence of epistemic

importance. AI mediation can legitimately reveal neglected problems, but it can also amplify what is easy to generate, retrieve, benchmark, or narrate.

Caution 7.3 applies symmetrically. Human attention is also shaped by prestige, disciplinary habit, memorable cases, institutional incentives, and perceptual limits. The distinction makes the production of visibility part of the judgment architecture while avoiding any presumption of unbiased human attention or uniformly biased AI mediation.

7.5 Formation of Salience Judgment

Salience judgment becomes a formation problem when repeated inquiry episodes alter what a person or configuration independently notices, preserves, questions, or dismisses later. A learner may become better at recognizing missing controls, unexamined assumptions, unstable categories, or weak signals. The same learner can also form brittle habits: treating novelty as importance, chasing anomalies without triage, accepting whatever an AI system foregrounds, or ignoring phenomena that do not fit familiar prompt templates.

The developmental residue introduced in Equation 6 can therefore include changes in salience discrimination. Let d_{sal} denote a change in the configuration's later capacity to recognize and endorse potentially inquiry-relevant phenomena. Evidence of stronger formation becomes more credible when that change remains available under altered cases rather than only when the same interface or cue recurs.

Several forms of evidence would be relevant. An inquirer may independently notice a consequential omission after AI assistance is removed; reconstruct why a previously ignored signal mattered; resist a highly visible but poorly warranted candidate; preserve an ambiguous observation without premature escalation; or revise a salience judgment after subsequent evidence shows that an earlier cue was misleading. These outcomes concern discrimination and revision rather than raw question count.

Claim 7.4 (Salience-judgment formation). Judgment of salience is formed when repeated inquiry leaves a revisable, transferable capacity to discriminate which phenomena or candidate problems merit continued attention, and when that discrimination remains responsive to provenance, counterevidence, changed framing, and later outcomes rather than being tied primarily to vividness or interface prominence.

Claim 7.4 also explains why problem-finding measures and discovery-oriented behavior are relevant but not sufficient for the present object. They establish that early problem formulation can be studied as a distinct process [15, 14]. The present paper additionally asks how the selection and retention of problem candidates are calibrated and how that selection changes through AI-mediated practice.

7.6 Boundary between Salience and Stronger Forms of Judgment

A judgment of salience is intentionally weak. It licenses continued attention, not final commitment. A problem can be salient yet aesthetically unpromising, normatively low priority, practically infeasible, or insufficiently evidenced. Conversely, a problem can inherit high normative importance from external conditions even when it is not generated through unusual discovery-oriented attention. Keeping these layers separate prevents the manuscript from treating every admitted problem as intellectually or socially valuable.

This boundary is especially important under AI mediation because the system can make later evaluative language available too early. A generated answer may immediately label a question “important,” “promising,” “high impact,” or “urgent” before the configuration has reconstructed why the question entered the field of inquiry or which standards support those labels. Judgment formation requires preserving the possibility that salience, aesthetic worth, normative importance, and practical priority diverge.

The section therefore leaves the configuration with a deliberately incomplete result: a set of problems, phenomena, or framings retained for stronger evaluation. The following analytical layer separates aesthetic, normative, and practical judgment while examining how those evaluations can diverge and require mediation.

Table 6: Three evaluative registers in AI-mediated inquiry

Register	Primary evaluative object	Illustrative considerations	Characteristic collapse risk
Aesthetic	Intellectual form, problem structure, theory, method, or explanatory organization	Elegance, depth, compression, structural fertility, unexpected connection, generative promise	Difficulty, novelty, prestige, or elegance is treated as sufficient reason for priority or action.
Normative	Pursuit, allocation, treatment of persons or groups, institutional policy, or foreseeable consequence	Harm, fairness, legitimacy, responsibility, distribution, rights, social priority, ecological consequence	Moral or political importance is treated as sufficient evidence of feasibility or as a substitute for epistemic quality.
Practical	Situated action, timing, sequencing, commitment, verification, escalation, or deferral	Feasibility, resources, reversibility, urgency, dependencies, authority, uncertainty, available alternatives	Expedience or current feasibility is treated as sufficient justification for what ought to be pursued or believed.

8 Evaluative Registers of Judgment

This section develops three evaluative registers that can operate over the problems, options, interpretations, and actions preserved by the preceding stages of judgment. Its role is to refine the evaluative component $\mathcal{V}_t$ of the situated judgment episode in Section 3. The objective is to distinguish aesthetic judgment concerning intellectual worth and structural promise, normative judgment concerning value, obligation, legitimacy, and priority, and practical judgment concerning situated action under present constraints and uncertainty. The discussion uses philosophy of science and reflective-practice literature as interfaces while keeping the three registers analytically non-equivalent. The central claim is that judgment formation requires learning both how to exercise each register and how to preserve their differences when they support divergent conclusions.

8.1 Three Evaluative Registers

The evaluative component $\mathcal{V}_t$ introduced in Equation 3 can contain heterogeneous reasons. A research problem can be judged elegant or fertile, socially important or ethically troubling, and feasible or infeasible under current conditions. Treating these reasons as one undifferentiated score hides the source of agreement and conflict. For the present analysis, Equation 14 refines $\mathcal{V}_t$ into three task- and context-indexed registers.

$$\mathcal{V}_t^{(q,\xi)} = \left(\mathcal{V}_{t,\text{aest}}^{(q,\xi)}, \mathcal{V}_{t,\text{norm}}^{(q,\xi)}, \mathcal{V}_{t,\text{prac}}^{(q,\xi)} \right). \quad (14)$$

Equation 14 is a decomposition of evaluative considerations, not a claim that all judgment belongs exclusively to one of three psychological modules. The registers can overlap in practice, operate on different objects, and become mutually conditioning over time. Their analytical separation is useful because one register can change while another remains comparatively stable. New evidence about environmental harm can alter normative judgment without changing the perceived elegance of a theory; new resource constraints can alter practical judgment without settling whether a project remains intellectually fertile; historical learning can change aesthetic standards without immediately determining what action is appropriate.

Table 6 summarizes the principal analytical roles of the three registers and the failure mode that follows when each is treated as a substitute for the others.

The distinctions in Table 6 do not assign permanent boundaries to disciplines. Aesthetic considerations can matter in mathematics, experimental science, design, and policy reasoning; normative considerations can enter scientific agenda setting as well as explicit ethical deliberation; practical judgment can be required wherever action must proceed before uncertainty is fully resolved. The classification concerns the role a consideration plays in the local judgment record.

8.2 Aesthetic Judgment and Intellectual Worth

Aesthetic judgment in inquiry concerns the perceived intellectual form and promise of a problem, explanation, theory, proof, method, or research direction. McAllister analyzes aesthetic evaluations in scientific theory assessment and emphasizes that scientific aesthetic preferences have histories rather than constituting timeless or purely decorative tastes [16]. The present framework uses this literature as an interface rather than as a complete theory of intellectual beauty.

Within judgment formation, aesthetic judgment matters because an AI-expanded inquiry environment can make difficult, novel-sounding, or formally elaborate candidates abundant. Those surface properties do not determine which candidates are intellectually deep or structurally generative. Aesthetic judgment can therefore concern relations such as explanatory compression, unexpected unification, conceptual economy, depth, tractability without triviality, or the ability of a problem to reorganize a wider inquiry space. These considerations can remain partly tacit and historically formed.

Definition 8.1 (Aesthetic judgment in inquiry). An *aesthetic judgment in inquiry* is a revisable evaluation of the intellectual form, depth, coherence, elegance, or generative promise of a problem, representation, explanation, theory, method, or research direction. It concerns intellectual worth without by itself establishing normative priority or present practical commitment.

Definition 8.1 preserves the distinction in Section 7. Salience judgment determines whether something deserves to remain available for stronger consideration. Aesthetic judgment then discriminates among available candidates according to intellectual form and promise. A candidate can be worth preserving because it is unresolved while later receiving a weak aesthetic evaluation, and a candidate can appear aesthetically powerful while still receiving low normative priority.

Aesthetic judgment is also prospective and fallible. Structural fertility is often inferred before the full continuation of an inquiry is known. An elegant hypothesis can fail empirically, a seemingly narrow problem can later open a major field, and aesthetic canons can change historically. Judgment formation therefore involves not only developing aesthetic discrimination but also retaining evidence of past aesthetic successes and failures so that the standards themselves remain revisable.

8.3 Normative Judgment and Priority

Normative judgment concerns what ought to be pursued, prioritized, permitted, constrained, funded, communicated, or acted upon given relevant values, obligations, rights, distributions, and foreseeable consequences. Philosophy of science provides several interfaces for this problem. Kitcher's analysis of scientific significance and well-ordered science places the organization of inquiry in relation to social priorities and democratic deliberation [17]. Douglas argues against a simple value-free ideal and emphasizes the role of values and moral responsibility in scientific reasoning, especially where uncertainty and policy consequence interact [18]. These accounts differ in their normative programmes; the present section uses them to establish that research and expert judgment cannot always be evaluated solely through internal epistemic criteria.

Normative judgment can operate at several scales. An individual researcher can judge whether a line of inquiry is responsible to pursue; an institution can allocate limited resources; a regulator can determine acceptable risk; a professional community can establish obligations concerning disclosure, consent, or downstream harm. The relevant normative considerations therefore depend on the role and authority of the judging configuration.

Definition 8.2 (Normative judgment in inquiry). A *normative judgment in inquiry* is a revisable evaluation of the legitimacy, responsibility, priority, permissibility, or distributive consequence of pursuing, supporting, communicating, or acting upon an inquiry option under a specified relational and institutional context.

Definition 8.2 does not imply that one universal normative metric governs every inquiry domain. It identifies a register in which reasons concern what should be done or prioritized, rather than only what is intellectually attractive or presently feasible. Normative disagreement can persist even when parties agree on evidence and predicted consequences because they weight duties, rights, distributions, or forms of harm differently.

The distinction from aesthetic judgment is especially important under generative abundance. An AI system can help formulate intellectually striking research questions whose intensive pursuit would consume substantial resources, create unacceptable risk, or reinforce a harmful institutional objective. Conversely, a normatively urgent problem can be conceptually inelegant, methodologically messy, or aesthetically unattractive while still deserving priority. Mature judgment formation therefore includes the capacity to keep intellectual admiration from silently determining normative commitment.

8.4 Practical Judgment under Situated Constraints

Practical judgment concerns what response is appropriate here and now given the available evidence, time, resources, authority, uncertainty, reversibility, and dependency structure. It is narrower than a complete ethics of action and broader than technical feasibility. Schön’s analysis of professional reflection-in-action provides an adjacent account of practitioners responding to indeterminate situations through a reflective conversation with the situation rather than by applying technical rules mechanically [19]. The present framework uses practical judgment in a similarly situated sense while preserving its own formal vocabulary.

Practical judgment can recommend commitment, staged experimentation, verification, escalation, deferral, preservation of alternatives, or withdrawal. It is therefore closely related to the closure and selective-reliance problems developed earlier. A normatively attractive action can be practically premature when evidence is weak and reversal is costly. A high-value research programme can be practically deferred because a required instrument or dataset does not yet exist. A feasible action can remain normatively unacceptable. Practical judgment coordinates current conditions without converting current conditions into final values.

Definition 8.3 (Practical judgment in inquiry). A *practical judgment in inquiry* is a revisable determination of an appropriate response, timing, sequence, verification demand, or degree of commitment under the present constraints, uncertainties, dependencies, and action possibilities of a situated judgment episode.

The word *appropriate* in Definition 8.3 does not imply normative correctness by definition. It refers to the action-guiding role of the register. Whether the practical judgment is epistemically or normatively defensible remains open to evaluation. This separation is necessary because efficient execution can coexist with poor ends, and morally defensible ends can coexist with poorly chosen timing or implementation.

Practical judgment is also especially sensitive to changes in context. A decision to verify rather than adopt can become inappropriate when delay creates greater harm; an earlier decision to defer can become obsolete once new evidence or resources arrive. Formation therefore requires responsiveness to changing constraints rather than memorization of fixed action rules.

8.5 Coordination without a Common Scalar

The three registers can converge, but their analytical differentiation becomes especially visible when they diverge. A problem can be aesthetically fertile, normatively low priority, and practically infeasible. A public-health intervention can be normatively urgent, practically implementable, and intellectually

ordinary. A technically elegant model can be practically useful while embedding criteria that receive serious normative objection. These combinations should remain visible rather than being compressed prematurely into one overall score.

For a candidate x within the current option or problem space, the paper therefore represents a multi-register evaluation profile schematically in Equation 15.

$$\mathfrak{V}_t^{(q,\xi)}(x) = \left(\mathbf{a}_t^{(q,\xi)}(x), \mathbf{n}_t^{(q,\xi)}(x), \mathbf{p}_t^{(q,\xi)}(x) \right), \quad (15)$$

where $\mathbf{a}_t$, $\mathbf{n}_t$, and $\mathbf{p}_t$ denote structured aesthetic, normative, and practical evaluation profiles. The bold notation indicates that each profile can contain several considerations rather than one scalar value. Equation 15 does not prescribe a conversion rule among them.

Caution 8.4 (Cross-register commensuration). The coexistence of aesthetic, normative, and practical considerations does not establish that they share one cardinal scale or that disagreement can be resolved by assigning numerical weights. Where a common metric is introduced for a particular institutional purpose, the construction and authority of that metric require separate justification.

Caution 8.4 is important for AI-mediated decision support because ranking systems often require compression. A model can return a single recommendation even when the underlying reasons belong to different registers and remain in tension. The appearance of one ordered output can therefore conceal unresolved evaluative plurality.

8.6 AI-Mediated Criterion Compression

AI systems can support each evaluative register. They can identify stylistic or structural regularities associated with elegant solutions, retrieve arguments concerning harms and obligations, estimate feasibility, compare resource requirements, and simulate consequences under specified assumptions. These capabilities can expand the evidential and comparative basis for judgment. They can also obscure which register a recommendation is expressing.

A generated answer may describe one project as “better” because it is more novel, more socially beneficial, easier to implement, more publishable, or better aligned with an institutional objective. If these reasons are compressed into one recommendation, the user can accept the ranking without recognizing that several evaluative criteria have been combined. The present paper calls this *criterion compression*.

Definition 8.5 (Criterion compression). *Criterion compression* occurs when heterogeneous evaluative considerations are combined into a recommendation, ranking, or summary in a way that makes the distinct aesthetic, normative, practical, or other grounds of evaluation insufficiently visible for independent inspection or revision.

Definition 8.5 identifies a representational failure that can arise from AI generation, interface design, institutional scoring systems, or ordinary human communication. AI mediation increases its importance because fluent synthesis can make heterogeneous considerations appear naturally unified. The relevant practice is to preserve enough *register transparency* during synthesis that the user can identify which kind of reason is doing what evaluative work.

Claim 8.6 (Register transparency). Where a consequential AI-mediated recommendation depends on heterogeneous evaluative considerations, judgment is more inspectable when the configuration can distinguish the relevant registers, identify the principal criteria within each, and revise one register without silently rewriting the others.

Claim 8.6 does not require every recommendation to expose an exhaustive value model. The required transparency depends on stakes, uncertainty, reversibility, and the degree to which different judgments would materially change the response.

8.7 Formation of Differentiated Judgment

The developmental problem is not simply learning three vocabularies. Differentiated judgment is formed when an inquirer becomes better able to identify which register is active, preserve conflict among registers, revise the appropriate register when new information arrives, and avoid transferring success in one register into unwarranted authority in another. A researcher can learn that a theory remains aesthetically compelling despite a new normative concern; a policymaker can learn that a normatively preferred intervention requires practical staging; a student can learn that an AI-generated answer is elegant but unsupported.

Evidence of formation can therefore include several patterns: more explicit separation of intellectual and normative reasons; improved sensitivity to changing constraints without corresponding drift in underlying values; stronger recognition that feasibility does not imply legitimacy; preservation of normatively important options that receive weak aesthetic appeal; and the ability to explain why a final response reflects coordination among several registers rather than one dominant criterion.

Claim 8.7 (Differentiated-judgment formation). Aesthetic, normative, and practical judgment are more plausibly formed when repeated inquiry leaves a transferable and revisable capacity to distinguish their grounds, detect cross-register conflict, and coordinate action without treating agreement in one register as sufficient authority over the others.

Claim 8.7 preserves the developmental focus of the present paper. The relevant achievement is not the permanent separation of intellectual, normative, and practical considerations. It is the capacity to relate them without losing track of their different objects, warrants, and consequences. Their interaction creates the further problem of judgment across time: present action can depend on an assessment of how a situation is unfolding, and a present judgment can alter the conditions under which later judgment will be made.

9 Mediating Judgment across Temporal Horizons

This section develops the temporal dimension of judgment and the coordination problem that arises when a present inquiry must be evaluated in relation to changing conditions. Its role is to extend the multi-register architecture in Section 8 without turning trajectory sensitivity into a separate universal faculty. The objective is to distinguish present-state evaluation from revisable interpretation of incipient change, and then to specify how aesthetic, normative, practical, and trajectory-sensitive considerations can be coordinated when they support different conclusions. The discussion uses naturalistic decision-making research and selected passages from the *Zhouyi* as comparative interfaces. The classical material is used to illuminate attention to subtle change and timely response; it is not treated as empirical validation of divinatory or cosmological prediction.

9.1 Temporal Extension of the Judgment Record

The situated judgment record in Equation 3 is indexed by a present task and context, yet many judgments concern processes whose significance cannot be evaluated from the present state alone. A small institutional change can matter because it may initiate a durable shift; a weak anomaly can matter because it may precede a regime change; a research direction can remain aesthetically attractive while becoming practically ill-timed; and an apparently minor dependency can become consequential if it narrows later options. Temporal judgment therefore requires the configuration to consider possible continuations of the present situation without converting those continuations into certain forecasts.

For a task q , context ξ , and decision horizon h , let the family of currently admissible trajectory hypotheses be stated in Equation 16.

$$\hat{\Gamma}_t^{(q,\xi;h)} = \left\{ \hat{\gamma}_{k,t}^{(q,\xi;h)} \right\}_{k \in K_t}. \quad (16)$$

Each $\hat{\gamma}_{k,t}^{(q,\xi;h)}$ is a revisable representation of one possible continuation judged sufficiently relevant to preserve over horizon h . Equation 16 is deliberately weaker than a probabilistic forecast. The family can contain qualitatively different branches, can leave probabilities unspecified, and can change when new evidence alters the relation among current signals, historical analogues, and future possibilities.

Definition 9.1 (Temporal judgment). A *temporal judgment* is a revisable interpretation of present evidence, relations, and weak changes in relation to one or more possible future trajectories over a specified decision horizon, together with an assessment of what those possible continuations imply for monitoring, timing, preservation of options, or action.

Definition 9.1 keeps temporal judgment distinct from prediction. A configuration can judge that a trajectory deserves monitoring without assigning it high probability; it can prepare for a difficult-to-reverse branch without asserting that the branch will occur; and it can defer action because the available temporal evidence remains too weak. The relevant object is the action-guiding significance of possible change under uncertainty.

9.2 Incipient Change as a Comparative Resource

The classical vocabulary of *ji* (幾) provides a useful comparative resource for this temporal problem. In the *Xici II*, the *Zhouyi* characterizes *ji* as “幾者，動之微” and links recognition of such incipience with timely response: “君子見幾而作，不俟終日” [22]. The present paper uses these lines as an historically specific articulation of sensitivity to movement before it becomes fully manifest.

The initial line of *Kun*, “履霜，堅冰至，” similarly presents a small current manifestation in relation to a later condition [23]. The analytical value of the image lies in its trajectory structure: the present state is interpreted as part of a developing process. The inference need not be deterministic. Real scientific, institutional, political, and technological processes can branch, reverse, accelerate, or dissolve. The comparison therefore concerns the form of attention to incipient change rather than the predictive validity of any classical cosmology.

The *Zhouyi* also contains the formulation “觀乎天文，以察時變；觀乎人文，以化成天下” in the *Bi* commentary [24]. Tze-Ki Hon’s account of the *Yijing* emphasizes human finitude, continuous change, sign-mediated understanding, and decision under uncertainty while also preserving the text’s layered historical uses [25]. These materials make temporal sensitivity visible as a practical problem: the observer must act without complete knowledge while remaining attentive to changing relations.

Naturalistic decision-making research provides a modern neighboring interface. Work in that tradition studies experienced decision makers operating under time pressure, shifting conditions, incomplete information, and consequential responsibility, often emphasizing recognition of patterns and situation assessment grounded in prior experience [20, 21]. The present paper does not equate naturalistic expertise with the classical sage ideal. The comparison is narrower: both traditions make visible a form of judgment in which present cues are interpreted through retained histories of how situations develop.

Caution 9.2 (Incipient-change sensitivity and predictive validity). Sensitivity to subtle or early change does not by itself establish the validity of the interpretive system used to explain or forecast that change. The preservation of weak signals, the construction of trajectory hypotheses, and the warrant for acting on those hypotheses should remain analytically separable.

Caution 9.2 applies to contemporary AI-assisted trend analysis as much as to classical omen interpretation. A fluent narrative can retrospectively make an early signal appear diagnostic even when many alternative trajectories were prospectively plausible. Temporal judgment therefore requires explicit attention to alternatives, reversals, missing variables, and the possibility that the apparent signal is noise.

9.3 Trajectory Hypotheses and Decision Horizons

Trajectory hypotheses become practically relevant only relative to a decision horizon. The same present signal can support different judgments at different temporal scales. A minor software dependency may

Table 7: Illustrative dimensions of temporal judgment under uncertainty

Dimension	Analytical question	Characteristic failure mode
Signal provenance	Where did the apparent early signal come from, and how was it selected or generated?	Interface prominence or retrospective storytelling is mistaken for independent evidence.
Decision horizon	Over what timescale would the possible trajectory matter for present judgment?	Short-horizon urgency or long-horizon speculation dominates without an explicit temporal frame.
Alternative trajectories	Which materially different continuations remain plausible enough to preserve?	One coherent narrative displaces branching, reversal, or no-change possibilities.
Path dependence	Which present commitments could alter what remains possible later?	Current convenience is evaluated without its effects on future option spaces.
Reversibility	How difficult would correction become if the trajectory judgment proves wrong?	High-confidence intervention is used where evidence is weak and correction is costly.
Monitoring trigger	What later evidence would strengthen, weaken, or retire the trajectory hypothesis?	The hypothesis becomes self-sealing because no prospective revision condition was specified.

be irrelevant over a one-day task but consequential over a five-year infrastructure horizon. A research direction can be practically infeasible this year while worth preserving for later development. A weak institutional change may warrant observation without immediate intervention because its longer-term significance remains uncertain.

Table 7 summarizes dimensions that can make temporal judgment inspectable without reducing it to a single forecast score.

The dimensions in Table 7 support temporal calibration. A temporal judgment can remain action-guiding while explicitly preserving uncertainty. The relevant action can be monitoring, reversible preparation, information acquisition, option preservation, delayed commitment, or escalation. Immediate intervention is only one possible response.

This structure also clarifies why historical experience can support temporal judgment without guaranteeing it. Retained histories provide comparison classes, yet analogy across trajectories is always partial. Experienced practitioners can recognize recurrent patterns, but familiar histories can also encourage over-projection onto genuinely novel situations. Historical depth contributes to temporal judgment when it expands discrimination among possible continuations and remains open to revision.

9.4 Mediating Judgment across Evaluative Registers

Temporal judgment complicates the three evaluative registers developed in Section 8. A problem can be aesthetically fertile while the surrounding trajectory makes immediate investment normatively questionable. A project can be normatively urgent while current conditions make direct execution practically irresponsible. A weak but potentially irreversible trend can justify preserving options even when neither its probability nor its intellectual attractiveness is high.

The present paper uses *mediating judgment* for the coordination required when these registers and temporal interpretations diverge.

Definition 9.3 (Mediating judgment). *Mediating judgment* is the revisable coordination of aesthetic, normative, practical, and temporal considerations when an inquiry configuration must determine what deserves commitment, delay, monitoring, preservation, verification, or action without collapsing the distinct evaluative grounds into a single score.

Definition 9.3 does not describe compromise by averaging. Some conflicts should remain unresolved.

Aesthetic value may remain high even when normative reasons block implementation. Practical urgency may justify temporary action while leaving normative disagreement intact. A temporal hypothesis can alter the appropriate timing of action without changing the underlying normative standard. Mediation therefore preserves register transparency while connecting those registers to closure and action.

The multi-register profile in Section 8 can be extended by keeping the trajectory family $\hat{\Gamma}_t^{(q,\xi;h)}$ jointly available during closure. The relevant practice makes visible which register, trajectory assumption, and horizon supports a particular response without reducing them to one total value. This is especially important when AI systems produce recommendations whose fluent summaries can compress heterogeneous grounds into one apparently unified conclusion.

Claim 9.4 (Mediation without scalar reduction). A stronger form of mediating judgment preserves the separability of evaluative registers and trajectory hypotheses long enough to identify which conflict, assumption, or uncertainty is responsible for a proposed action, and remains able to revise that action when one component changes without silently rewriting the others.

Claim 9.4 connects this section to criterion compression in Section 8. The danger is not only that AI combines several criteria. The deeper problem arises when the configuration can no longer reconstruct which criterion, trajectory assumption, or temporal horizon drove the recommendation.

9.5 AI-Mediated Temporal Judgment

AI systems can assist temporal judgment by retrieving historical analogues, detecting weak patterns, maintaining scenario libraries, comparing signals across scales, generating alternative continuations, and monitoring changes over time. These functions can expand the set of trajectories that a configuration can consider. They can also create distinctive failure modes.

One risk is *narrative overcompletion*. A generative system can turn sparse evidence into a coherent story whose internal fluency exceeds the evidential warrant for the trajectory. A second risk is *scenario proliferation*: generating many plausible futures can create the appearance of breadth while providing little discrimination about which differences should change present action. A third risk is *temporal anchoring*: a model, benchmark, or retrieval window can privilege recent patterns or familiar historical analogues while suppressing structural breaks. A fourth risk is *opaque horizon selection*: the recommendation can depend strongly on an implicit timescale that the user never inspects.

These risks motivate a stronger practice of AI-assisted temporal judgment. The configuration should preserve provenance for key signals, request materially different trajectory hypotheses rather than stylistic variants, state the relevant decision horizon, identify what evidence would reverse the current interpretation, and distinguish monitoring recommendations from intervention recommendations. AI support is most developmentally useful when it makes temporal assumptions more inspectable rather than merely making forecasts more abundant.

The same principle applies to mediating judgment. An AI system can help articulate conflicts among aesthetic, normative, practical, and temporal considerations, yet the existence of a synthesized recommendation does not establish that the registers have been responsibly coordinated. A mature Human–AI configuration should remain able to reopen the recommendation, identify the trajectory assumption that changed the action, and revise one register without automatically changing the others.

9.6 Formation of Temporal and Mediating Judgment

Temporal judgment becomes a formation problem when repeated episodes change what the inquirer later recognizes as an early signal, which alternative trajectories remain available, how far ahead the configuration reasons, and when it chooses monitoring, preparation, commitment, or restraint. The developmental residue in Equation 6 can therefore include temporal discrimination and mediation capacities.

Evidence of stronger temporal formation would include prospective rather than merely retrospective success. An inquirer might record a trajectory hypothesis before the outcome is known, preserve alternatives, specify revision triggers, later compare the realized path with the original hypothesis family, and

update which cues deserve attention in future cases. Such practices reduce hindsight reconstruction and make calibration of temporal judgment possible.

Mediating judgment is formed when the configuration becomes better able to preserve disagreement among evaluative registers while still reaching proportionate action. Development can be visible when the inquirer can explain why an aesthetically attractive option is being delayed for normative reasons, why a normatively urgent intervention is being designed reversibly because the trajectory evidence remains weak, or why a practical constraint affects timing without being mistaken for a change in intellectual worth.

Claim 9.5 (Formation through prospective revision). Temporal and mediating judgment are more credibly formed when the configuration retains prospective records of trajectory hypotheses, evaluative conflicts, decision horizons, and revision conditions, and later uses observed outcomes to recalibrate what it notices, preserves, coordinates, and acts upon in subsequent cases.

Claim 9.5 keeps formation distinct from occasional foresight. A single successful prediction can reflect chance, selective recall, or a favorable environment. Durable judgment development requires a revisable history of how trajectories were interpreted, how conflicts were mediated, and which cues or commitments proved misleading.

9.7 Epistemic Boundary of Prognostic Reasoning

Temporal judgment is especially vulnerable to retrospective distortion. Once an outcome is known, earlier events can appear more diagnostic than they were at the time. Successful actors can remember the signals they followed and forget the many signals that led nowhere. AI-generated historical summaries can intensify this problem by producing coherent causal narratives from selectively retrieved evidence.

The framework therefore treats prognosis as a bounded component of judgment rather than privileged access to the future. A trajectory hypothesis should remain answerable to provenance, alternatives, prospective revision criteria, and later outcome comparison. The same discipline protects the classical comparison used in this section: sensitivity to *ji* is analytically useful as a model of attention to incipient change, while stronger claims about divinatory prediction remain outside the paper's evidential scope.

The result is a narrower and more practical account of temporal wisdom. Mature judgment can sometimes recognize that the conditions of inquiry are changing before those changes become fully explicit, yet such recognition remains fallible and responsibility-bearing. Its value lies less in claiming certainty about the future than in supporting calibrated monitoring, reversible preparation, preservation of options, and coordination among forms of judgment whose appropriate weight changes with time.

10 Reflexive Formation of Future Judgment Conditions

This section develops the reflexive dimension of judgment formation. Its role is to extend the history-bearing update architecture from Section 3 beyond the developmental residue carried by an inquirer and toward changes in the conditions under which later judgments become possible. The objective is to distinguish a judgment that concerns only the immediate response from a judgment that also attends to how present action can reorganize later evidence, option spaces, conceptual categories, reliance relations, institutional criteria, and capacities for correction. The discussion proceeds through a task-relative representation of future judgment conditions, an analysis of path dependence, a typology of reflexive effects in AI-mediated inquiry, and a final account of reversibility and responsibility. The argument is conceptual and developmental. It does not assume that second-order consequences can be predicted with precision or that preserving future revisability should dominate every other practical consideration.

10.1 Reflexive Scope of a Judgment Episode

A situated judgment episode can be evaluated at more than one temporal level. At the first level, the inquiry concerns the immediate response: whether an interpretation is accepted, whether an action is

taken, whether a recommendation is verified, or whether closure is deferred. At a second level, the inquiry concerns what the episode changes for later judgment. Section 3 introduced this developmental question through the residue $\mathfrak{D}_t^J$. Reflexivity adds a further emphasis: some consequences of present judgment operate by changing the environment and relational structure through which future judgments will be framed and evaluated.

The distinction is important because an action can be locally reasonable while generating conditions that later make good judgment more difficult. A workflow can improve present efficiency while reducing independent option generation. A classification can simplify present coordination while narrowing which future cases remain legible. A decision to discard apparently irrelevant observations can reduce storage costs while removing evidence needed for later reinterpretation. Conversely, a slower or more reversible response can preserve records, alternatives, or contestation pathways that improve later judgment even when those benefits are not visible in the immediate episode.

Definition 10.1 (Reflexive judgment). Within the present framework, *reflexive judgment* is judgment that includes, among its relevant considerations, the foreseeable ways in which a present response may alter the conditions under which later judgment will be framed, evidenced, monitored, contested, revised, or acted upon. Reflexivity concerns the relation between present judgment and future judgment conditions; it does not require certainty about those future conditions.

Definition 10.1 separates reflexivity from retrospective reflection. A later review can be valuable without having influenced the earlier decision. Reflexive judgment, by contrast, gives at least some present weight to the possibility that the current response will reorganize the next inquiry. This temporal extension is especially relevant where actions are path dependent, where AI systems become recurrent supports, or where institutional decisions alter the evidence and option environment available to later participants.

10.2 Task-Relative Future Judgment Conditions

The history-bearing configuration Σ_t^J already contains the components needed to represent future judgment conditions. A separate ontology is therefore unnecessary. What is needed is a task-relative projection identifying which components of a later configuration materially condition a selected future judgment. Let q' and ξ' denote a later task and context. Equation 17 defines a projection $\Pi_{\text{cond},q',\xi'}^J$ from the later configuration to the subset of conditions relevant to that judgment.

$$\Theta_{t+k}^{J,(q',\xi')} = \Pi_{\text{cond},q',\xi'}^J (\Sigma_{t+k}^J). \quad (17)$$

In Equation 17, $\Theta_{t+k}^{J,(q',\xi')}$ does not introduce a new kind of entity. It is an analytical view of the later judgment configuration. Depending on the future task, the projection can include the available evidence environment $\mathcal{E}$, conceptual and representational resources $\mathcal{L}$, normative and institutional conditions $\mathcal{N}$, relational supports $\mathcal{R}$, operative capacities $\mathcal{C}$, and parts of the retained history H . Different future tasks can therefore make different consequences of the same present judgment relevant.

The relation between a present response and later judgment conditions can be represented by refining the update relation from Equation 2. Let ω_{t+1} denote consequences or feedback produced after the present response a_t , including changes that may become visible only after the action enters a wider environment. Equation 18 states the future-condition projection after the configuration has been updated by an episode that includes both the response and its consequences.

$$\Theta_{t+1}^{J,(q',\xi')} = \Pi_{\text{cond},q',\xi'}^J [\mathcal{F}_t^J (\Sigma_t^J, \Delta_t^J(a_t, \omega_{t+1}))]. \quad (18)$$

Equation 18 is intentionally schematic. It does not imply that the inquirer can calculate ω_{t+1} in advance. Its analytical purpose is to prevent the response a_t from being treated as the terminal point of judgment. The response can alter later evidence, create or destroy options, redistribute authority, change reliance patterns, and produce new feedback. Some of those changes then become part of the configuration from which the next judgment is made.

Table 8: Channels through which present judgment can reshape future judgment conditions.

Channel	Present judgment effect	Possible future consequence
Evidence environment	Preserve, collect, discard, summarize, or privilege selected evidence	Later judgments inherit richer provenance, missing observations, compressed disagreement, or altered feedback visibility
Option environment	Commit resources, standardize workflows, set defaults, or close alternatives	Some later options become easier, costlier, institutionally unavailable, or difficult to imagine
Conceptual language	Adopt classifications, labels, metrics, or problem framings	Later inquiry treats particular distinctions as natural, salient, or institutionally required
Relational structure	Reallocate judgment functions among persons, AI systems, teams, and procedures	Contestability, expertise development, dependency, and review authority are redistributed
Normative and institutional conditions	Stabilize thresholds, incentives, policies, or accepted standards	Later judgments inherit changed criteria of adequacy, urgency, legitimacy, or acceptable risk
Capacity practice	Repeatedly exercise, delegate, or bypass selected judgment functions	Independent discrimination, verification habits, calibration, and reliance policies can strengthen, weaken, or specialize

Caution 10.2 (Reflexivity and prediction). Reflexive judgment should not be interpreted as a requirement to predict all downstream effects of an action. The relevant capacity is to identify materially plausible pathways by which present judgment can reshape later judgment conditions, preserve uncertainty about those pathways, and revise the assessment when consequences become observable.

Caution 10.2 limits a possible overextension of the framework. A demand for exhaustive second-order prediction would make action impossible and would reward speculative narratives. Reflexivity is better understood as structured attention to consequential dependencies, especially where effects are difficult to reverse or where the present action can suppress the evidence needed to recognize error later.

10.3 Channels of Reflexive Change

The components of Σ_t^J provide several recurrent channels through which present judgment can alter future judgment conditions. Table 8 summarizes six analytically distinct channels. The examples are illustrative rather than exhaustive, and a single intervention can operate through several channels at once.

Table 8 shows why reflexive effects cannot be reduced to learning inside an individual. Judgment can reshape the archive, the interface, the institution, the category system, the distribution of labor, and the practices through which later judgment is exercised. The relevant history may therefore reside partly in externalized records and institutional routines as well as in retained personal experience.

The evidence channel is particularly important for inquiry. Decisions about what to measure, archive, summarize, or discard affect which later hypotheses can be evaluated. A present decision to preserve uncertainty and provenance can make later reinterpretation possible; a present decision to compress conflicting observations into a single clean representation can make later correction more difficult. The issue is not that all evidence should be retained indefinitely. The reflexive question is whether the current evidential simplification removes distinctions whose future relevance remains materially plausible.

The option and conceptual-language channels operate similarly. Selecting one research programme, workflow, classification scheme, or institutional metric can create infrastructure that makes related future judgments easier while making alternatives less visible or more expensive. The resulting path dependence can be beneficial when it stabilizes productive coordination, yet it can also create lock-in when the conditions that justified the original judgment later change.

Claim 10.3 (Path dependence of judgment conditions). Present judgments can alter the evidence, options, language, relations, institutional criteria, and practiced capacities through which later judgments are made. Evaluation of a judgment can therefore require attention to both its immediate object and its contribution to the path dependence of future judgment conditions.

Claim 10.3 does not imply that path dependence is inherently undesirable. Accumulated expertise, shared terminology, standardized evidence practices, and stable institutions are themselves forms of path dependence. The analytical concern is whether the resulting structure remains revisable when evidence, purposes, or conditions change.

10.4 Reflexive Effects of Recurrent AI Mediation

AI-mediated inquiry makes reflexive effects especially visible because the same system can participate repeatedly in framing, retrieval, option generation, explanation, evaluation, monitoring, and review. Recurrent use can therefore change both the immediate judgment episode and the infrastructure of later judgment.

One reflexive pathway concerns *evidence ecology*. If AI becomes the recurrent interface through which evidence is retrieved and summarized, future inquiry can increasingly depend on what the system indexes, ranks, compresses, cites, or omits. The user may become more efficient at accessing an evidence environment while becoming less exposed to the structure of disagreement or to sources that the interface does not surface readily. The relevant developmental question concerns both user capacity and the mediated evidence ecology that the user comes to inhabit.

A second pathway concerns *option ecology*. Generative systems can broaden the supply of hypotheses and alternatives, yet recurrent dependence on generated options can also shape expectations about what a plausible option looks like. Over time, an inquirer may learn productive patterns from generated alternatives, or may become less practiced at detecting assumptions shared across them. Configuration-level breadth and component-level independent generation can therefore move in different directions.

A third pathway concerns *criteria and monitoring*. AI-generated explanations, confidence language, rankings, and critique can affect which criteria become routinely foregrounded and which metacognitive cues users learn to trust. Repeated exposure can improve calibration if feedback connects those cues to later outcomes. It can also stabilize pseudo-calibration if fluency, citation density, or stylistic markers become substitutes for competence or provenance.

A fourth pathway concerns *responsibility structure*. Once an AI-mediated workflow becomes routine, participants can begin to treat generated recommendations as procedural defaults. Formal decision authority may remain human while the practical burden of contesting the system rises. The resulting configuration can make later responsibility more diffuse even when the nominal allocation of authority has not changed. This issue becomes central when the paper turns to delegation and judgment ownership.

The reflexive effect of AI mediation is therefore neither simple deskilling nor simple augmentation. The same workflow can strengthen some capacities while weakening others, improve local performance while increasing dependence, or broaden configuration-level option generation while narrowing component-level practice. Evaluation requires specifying the level and function at which change is being claimed.

10.5 Reflexive Residue and Future Revisability

The developmental residue $\mathfrak{D}_t^J$ from Section 3 can be projected onto those changes that materially affect later judgment conditions. Equation 19 defines a reflexive-residue projection without adding a second update process.

$$\mathfrak{D}_{t,\text{ref}}^{J,(q',\xi')} = \Pi_{\text{cond},q',\xi'}^J (\mathfrak{D}_t^J). \quad (19)$$

In Equation 19, $\mathfrak{D}_{t,\text{ref}}^{J,(q',\xi')}$ denotes the part of the developmental residue relevant to the selected later judgment conditions. The expression can include changes in archives, criteria, institutional routines,

conceptual distinctions, reliance policies, and practiced capacities. It is therefore broader than personal learning while remaining derived from the same history-bearing update framework.

A central concern is *future revisability*: whether later participants retain practical means to inspect, contest, reinterpret, or reverse a judgment when new evidence or changed conditions warrant revision. Revisability is partly epistemic and partly structural. Records can preserve provenance; institutions can preserve appeal; workflows can preserve independent checks; technical systems can preserve version history; and decision design can preserve options whose destruction would otherwise make later correction impossible.

Revisability does not require maximal reversibility. Some decisions are inherently irreversible, and delay can itself create irreversible losses. A responsible architecture therefore treats reversibility as one consideration among practical, normative, and temporal registers. The more difficult an intervention is to reverse, however, the stronger the case for making uncertainty, provenance, alternatives, and revision conditions explicit before closure when circumstances permit.

Claim 10.4 (Revisability-sensitive judgment). When present action can materially narrow the evidence, options, contestation pathways, or capacities available to later judgment, a stronger form of judgment evaluates not only immediate performance but also whether proportionate opportunities for future correction remain available.

Claim 10.4 is not a universal priority rule. Preservation of future correction can conflict with urgency, safety, confidentiality, cost, and other obligations. Its role is to make the loss of revisability visible as a judgment-relevant consequence rather than allowing that loss to disappear inside an immediate performance metric.

10.6 Reflexive Responsibility and Developmental Evidence

Reflexivity also extends the temporal object of responsibility. Responsibility for a judgment need not concern only the immediate action or outcome. Where the downstream effects are reasonably foreseeable, a decision can also be evaluated by how it shapes later judgment conditions. An institution that adopts an automated assessment procedure, for example, can influence what evidence is collected, how cases are classified, what reviewers practice, which exceptions remain visible, and how future decisions are audited. The same principle applies at smaller scales when a recurrent personal workflow determines which capacities continue to be exercised and which are routinely bypassed.

This extension should remain proportionate. Agents cannot be responsible for unforeseeable consequences merely because those consequences later become important. The relevant question is whether the present configuration had reasonable grounds to recognize a pathway through which the judgment could alter later evidence, options, criteria, capacities, or responsibility structures, and whether proportionate safeguards or revision pathways were available.

Evidence of reflexive judgment formation is therefore longitudinal. A configuration becomes more reflexively capable when it learns to identify such pathways prospectively, records the assumptions that made them relevant, monitors whether the anticipated effects occur, and updates later practice accordingly. Prospective documentation is important because retrospective narratives can make downstream consequences appear obvious after the fact.

Schön's account of reflection-in-action provides an adjacent professional interface: practitioners can engage in a reflective conversation with a changing situation rather than apply static rules mechanically [19]. The present framework extends that situated insight temporally. Reflection concerns not only how the present situation responds to intervention, but also how an intervention can reorganize the environment in which later reflection and judgment will occur.

Definition 10.5 (Reflexive judgment formation). *Reflexive judgment formation* is the durable development of a revisable capacity to recognize, monitor, and learn from the ways present judgments alter future judgment conditions, including evidence environments, option spaces, conceptual categories, relational dependencies, institutional criteria, practiced capacities, and opportunities for correction.

Definition 10.5 makes the developmental standard stronger than awareness of second-order effects in one episode. Formation requires later availability and revision. A person or institution can acknowledge reflexivity rhetorically while repeatedly failing to change practice after consequences become visible. Conversely, a configuration can become more reflexively capable by retaining records of assumptions, comparing anticipated and realized effects, and changing future procedures when the original judgment proves brittle.

10.7 Scope and Interface with Judgment Power

The account in this section remains preliminary in three respects. First, it does not provide a general causal model of how actions transform institutions or environments. The update relations identify analytical dependencies whose empirical mechanisms vary by domain. Second, it does not supply a universal rule for balancing immediate performance against future revisability. That balance depends on urgency, consequence, reversibility, authority, and normative commitments. Third, reflexive awareness can itself become overextended: concern with every possible second-order effect can generate paralysis, speculative overfitting, or indefinite postponement.

The section nevertheless establishes an important structural result. Judgment is reflexive when present responses participate in producing the conditions of later judgment, and judgment formation is reflexive when configurations become better able to recognize and revise those effects. Once future judgment conditions are acknowledged as partly generated, a further question follows: who has the capacity to determine which criteria, classifications, defaults, and options become stabilized for others? That is a problem of power over judgment criteria rather than reflexivity alone.

11 Power over Judgment Criteria

This section develops the power dimension of judgment formation. Its role is to make explicit that criteria are not merely private standards applied by an isolated evaluator. In many inquiry settings, the admissibility of reasons, the weighting of considerations, the thresholds for action, the defaults presented by an interface, and the conditions under which a judgment is treated as closed are shaped by institutions, professional communities, technical infrastructures, platforms, and recurrent Human–AI configurations. The objective is to distinguish ordinary influence over a judgment from power over the criteria through which judgments are subsequently formed. The discussion proceeds by locating criteria within the existing GR configuration, identifying recurrent channels of criterion power, representing asymmetric criterion modification, examining AI-mediated criterion opacity, and specifying forms of contestability and developmental evidence. The analysis is conceptual. It does not assume that every shared standard is oppressive, that institutional authority is illegitimate, or that criterion pluralism eliminates the need for closure.

11.1 Criterion Conditions in Situated Judgment

Section 3 represented evaluative considerations through $\mathcal{V}_t$ and normative or institutional conditions through $\mathcal{N}_t$. Sections 8 and 9 then showed that the evaluative layer can contain several non-equivalent registers whose relative relevance depends on the task, context, time horizon, and action conditions. A further issue now becomes unavoidable: who or what determines which criteria enter $\mathcal{V}_t$, which evidence is admissible to them, how conflicts among criteria are resolved, and when their satisfaction is treated as sufficient for closure?

The present framework uses the term *criterion regime* for the task-relative judgment conditions that organize those questions. A criterion regime can include the set of admissible criteria, classifications that determine which cases are comparable, relative weighting or ordering among considerations, thresholds for adequacy or action, defaults that apply when active choice is absent, and closure or review rules. The

regime need not be codified in one document. It can be distributed across professional norms, organizational routines, interface design, database fields, model objectives, legal rules, inherited vocabulary, and tacit expectations.

Because these conditions are already contained within the history-bearing judgment configuration, a separate ontology is unnecessary. Equation 20 defines the criterion regime as a task-relative projection of Σ_t^J .

$$\mathcal{K}_t^{J,(q,\xi)} = \Pi_{\text{crit},q,\xi}^J(\Sigma_t^J). \quad (20)$$

In Equation 20, $\mathcal{K}_t^{J,(q,\xi)}$ denotes those elements of the current configuration that materially determine how the selected judgment is evaluated. Depending on the domain, the projection can draw from conceptual language $\mathcal{L}_t$, institutional and normative conditions $\mathcal{N}_t$, relational authority structures $\mathcal{R}_t$, operative capacities $\mathcal{C}_t$, and retained historical practices in H_t . The projection can also condition how the episode-level warrant structure $\mathcal{W}_t$, evaluative layer $\mathcal{V}_t$, and monitoring layer $\mathcal{M}_t$ are populated and interpreted.

Definition 11.1 (Criterion regime). Within this paper, a *criterion regime* is the task- and context-relative structure through which a judgment configuration determines which evaluative considerations and reasons are admissible, how they are differentiated or combined, which thresholds and defaults organize action, and under what conditions judgment is treated as sufficiently settled, reviewable, or reopenable.

Definition 11.1 is deliberately broader than a scoring rule. A research committee can possess a criterion regime even when members deliberate qualitatively. A clinician can operate under one without assigning numerical weights. A scholar can inherit one through disciplinary training. An AI interface can participate in one through ranking, filtering, confidence displays, or default recommendations even when the underlying model never contains an explicit list of criteria.

Classification and standardization literatures provide important interfaces for this claim. Bowker and Star show how categories and standards become part of information infrastructure and can shape visibility, interaction, and consequence [27]. Bourdieu’s account of symbolic power likewise provides an adjacent vocabulary for understanding why some authorized forms of expression and classification acquire greater social force than others [26]. The present paper does not import either framework wholesale. Their relevance is narrower: criterion regimes are socially and technically situated, and the practical force of a criterion can depend on more than its abstract epistemic merits.

11.2 Channels of Criterion Power

Power over judgment criteria can operate through several analytically distinct channels. Table 9 summarizes six recurrent forms. A single institution or AI-mediated workflow can operate through several at once, and the relevant power relation may be distributed across designers, administrators, experts, models, interfaces, and inherited institutional routines.

Table 9 makes two distinctions visible. First, criterion power can operate before explicit evaluation. Classification and admissibility determine which objects and reasons become available to evaluation at all. Second, criterion power can operate after evaluation through thresholds, defaults, and review rules that determine when a judgment becomes action-guiding or institutionally durable.

Commensuration deserves particular attention because it can make unlike objects comparable by converting them into a common metric. Espeland and Stevens define commensuration as comparison according to a common metric and analyze it as a mode of power with cognitive and political stakes [28]. In the present framework, the relevant point is not that numerical comparison is inherently suspect. Metrics can improve consistency, transparency, and coordination. The judgment problem arises when the metric’s selection, scope, or weighting becomes practically authoritative while the dimensions excluded by the common scale become difficult to reintroduce.

The same issue appears in classification. A category can facilitate coordination and preserve important distinctions, yet categories also condition what evidence is collected, which cases are treated

Table 9: Recurrent channels through which criterion power can shape judgment.

Channel	Criterion-level effect	Judgment consequence
Admissibility	Determines which reasons, evidence, variables, or perspectives can enter evaluation	Relevant considerations can become institutionally visible, marginal, or excluded before comparison begins
Classification	Defines categories, case types, labels, and comparison classes	The same phenomenon can become legible under one category while losing distinctions important under another
Commensuration	Places heterogeneous objects or values on a shared metric or ordered scale	Comparison becomes easier while non-measured differences can lose practical force
Thresholding	Sets evidential, risk, performance, or confidence levels required for action	Similar evidence can produce different actions because closure thresholds differ
Defaulting	Determines what occurs when no active override is made	Interface or institutional defaults can become practical judgments without explicit endorsement of their criteria
Closure and review	Determines when judgment is treated as settled and what conditions permit reopening or appeal	Criteria can become self-stabilizing when contestation, revision, or counterevidence is procedurally costly

as similar, and which anomalies remain visible. Bowker and Star’s analysis of classification infrastructures is therefore especially relevant to the reflexive framework of Section 10: once classifications enter forms, databases, workflows, and reporting requirements, later judgments inherit them as part of the environment rather than encountering them as fresh theoretical proposals [27].

11.3 Asymmetric Criterion Modification

Influence over criteria becomes a problem of power when the capacity to alter another configuration’s criterion regime is materially asymmetric and consequential. Ordinary disagreement does not suffice. A colleague can argue for a new standard without possessing power over the listener. A journal, platform, regulator, funding body, employer, examination system, or recurrent AI interface can have a different position when its classifications, metrics, defaults, or thresholds materially determine which judgments become actionable, visible, rewarded, or reviewable.

Let A and B denote components or configurations within the relevant relational field. A criterion intervention by A on the regime governing judgments by or about B can be represented schematically by Equation 21.

$$\mathcal{K}_{B,t+1}^{J,(q,\xi)} = \mathcal{H}_{A \rightsquigarrow B,t}^{\text{crit}} \left(\mathcal{K}_{B,t}^{J,(q,\xi)}, \zeta_t \right), \quad (21)$$

where ζ_t denotes the policy, classification, metric, interface rule, model-mediated representation, or other intervention through which the criterion regime is modified. Equation 21 is not a causal model and does not imply unilateral control. Its purpose is to identify the object being changed: the conditions under which subsequent judgments are evaluated.

Definition 11.2 (Criterion power). Within the present framework, *criterion power* is a relationally asymmetric capacity to stabilize, modify, or constrain another judgment configuration’s admissible criteria, classifications, weightings, thresholds, defaults, or closure conditions in ways that materially affect the judgments or opportunities available to that configuration.

Definition 11.2 makes criterion power narrower than influence and broader than intentional coercion. Intent is not required. A technical standard can acquire criterion power through institutional uptake even when no designer sought to dominate later users. A benchmark can become practically authoritative

because funding, publication, or deployment decisions coordinate around it. A language model can re-currently foreground particular notions of relevance because of retrieval, ranking, or generation practices without itself holding institutional authority. The relation of power depends on the larger configuration through which those outputs become consequential.

Caution 11.3 (Power without malicious intent). Criterion power should not be inferred from disagreement, influence, or technical mediation alone, and it should not be restricted to cases of malicious or strategic intent. The relevant evidence concerns asymmetric capacity to shape criterion conditions together with material consequences for judgment, contestation, or future opportunity.

Caution 11.3 matters for AI-mediated inquiry because models, interface designers, organizations, and users can contribute jointly to a criterion regime. Attributing the resulting power to a single visible component can obscure the infrastructure that stabilizes the criterion. Conversely, describing the regime as merely “technical” can hide the fact that choices about categories, metrics, thresholds, and defaults redistribute practical authority.

11.4 AI-Mediated Criterion Opacity

AI mediation can make criterion power difficult to inspect because systems often present an integrated output rather than the heterogeneous evaluative decisions that produced it. Section 8 introduced *criterion compression*: several evaluative registers can be blended into one fluent recommendation. Criterion power adds a different question. Even when users know that several criteria are involved, they may lack visibility into who selected them, how they were operationalized, what proxy variables stand in for them, or which threshold converts evaluation into recommendation.

For example, a system that ranks research proposals can appear to provide an informational summary while implicitly establishing a criterion order among novelty, feasibility, expected impact, methodological conformity, cost, and institutional priorities. A retrieval system can appear to answer a factual question while its ranking architecture determines which evidence receives early visibility. A writing assistant can appear to improve clarity while repeatedly normalizing one disciplinary register or argumentative form. In each case, the output can influence judgment partly because criterion-setting has been embedded upstream in the technical or institutional process.

The sociotechnical-abstraction literature provides an important interface here. Selbst and colleagues argue that technical abstractions can become misleading when relevant social context falls outside the system boundary, and they recommend treating decision systems as sociotechnical processes rather than isolated technical artifacts [29]. For the present analysis, this supports a specific methodological rule: the criterion regime of an AI-mediated judgment should be traced across model objectives, data and retrieval infrastructure, interface defaults, organizational policies, and human interpretation when those components materially affect the evaluation.

This broader boundary also clarifies why explanation alone may be insufficient. An explanation can disclose why a system produced a recommendation under its operative criteria while leaving the selection of those criteria unquestioned. A highly interpretable score can remain criterion-powerful if users cannot contest the metric, alter the threshold, introduce excluded considerations, or appeal the classification. Transparency of application and contestability of criterion selection are therefore different properties.

Claim 11.4 (Criterion transparency and criterion contestability). A judgment configuration can make the application of criteria transparent while leaving the selection, weighting, thresholding, or revision of those criteria practically uncontestable. Evaluation of AI-mediated judgment should therefore distinguish transparency about criterion application from contestability of the criterion regime itself.

Claim 11.4 does not imply that every user should be able to rewrite every institutional standard. Stable standards can be necessary for coordination, comparability, safety, and legal consistency. The relevant question is whether the degree of contestability, appeal, and revision is proportionate to the consequences, uncertainty, heterogeneity, and potential for criterion error in the domain.

11.5 Contestability through Counter-Criteria

Criterion regimes become especially difficult to evaluate when the only available criticism must be expressed in the regime's own terms. A metric may be challenged only by presenting a better score; a classification only by requesting reassignment within the same category system; a benchmark only by improving performance on the benchmark. Such procedures can be useful for local correction, yet they provide weak protection against the possibility that the underlying criterion system is itself incomplete or misaligned with the inquiry.

The present framework therefore distinguishes *within-regime correction* from *counter-criterion contestation*. Within-regime correction asks whether a case was evaluated correctly under existing standards. Counter-criterion contestation asks whether relevant reasons, distinctions, thresholds, or alternatives have been excluded or mis-specified by those standards. The distinction parallels the difference between error correction inside a model and revision of the model's representation of the problem.

A counter-criterion need not replace the existing regime. It can function as a diagnostic remainder: a reason or distinction that cannot yet be incorporated without loss. Maintaining such remainders can be especially important when commensuration is useful for coordination but weakly justified as a complete representation of value. The objective is not maximal pluralism. An inquiry that never stabilizes criteria cannot act. The objective is to preserve enough external standpoint that criteria remain revisable when persistent anomalies, excluded harms, changed purposes, or new evidence reveal a mismatch.

Claim 11.5 (Counter-criterion review). A more power-sensitive judgment configuration preserves some practical capacity to distinguish errors made under an accepted criterion regime from reasons for revising the regime itself. This capacity can include access to alternative classifications, uncompressed evidence, appeal routes, independent review, or the ability to articulate relevant considerations that the dominant metric does not represent.

Claim 11.5 connects criterion power to the future-revisability analysis of Section 10. If a criterion regime progressively determines what evidence is collected and what alternatives remain institutionally available, the capacity to formulate counter-criteria may itself erode over time. Power over criteria can therefore become reflexive: the regime shapes the later conditions under which its own adequacy is judged.

11.6 Developmental Evidence within a Bounded Account

Criterion-power awareness becomes a problem of judgment formation when an inquirer develops a transferable capacity to identify who or what set the operative criteria, distinguish application from selection, detect criterion compression and excluded dimensions, estimate the consequences of thresholds and defaults, and recognize when counter-criterion review is warranted. The developmental object is not habitual suspicion of institutions or AI systems. It is improved discrimination about where criterion-setting occurs and when that setting materially affects the validity, legitimacy, or revisability of judgment.

Evidence for such formation can include the ability to reconstruct the criterion regime of a new domain, compare how alternative thresholds change action, identify when a metric is being used outside the context that justified it, preserve non-commensurable considerations when a common scale is only partially adequate, and revise one's own preferred criteria after target-side or outcome feedback. It can also include recognizing legitimate reasons for stable standards and resisting the opposite failure mode in which every criterion is treated as arbitrary simply because it is historically or institutionally situated.

The account remains bounded in four respects. First, it does not provide a general theory of social power. The concept of criterion power is restricted to relational capacities that shape the conditions of judgment. Second, the analysis does not establish which substantive criteria are correct in any particular domain. Third, asymmetric criterion-setting can be legitimate when supported by expertise, procedural authority, safety obligations, or coordination needs; the relevant issue is how that authority is justified, limited, reviewed, and revised. Fourth, criteria are sometimes delegated deliberately. The existence of criterion power therefore does not by itself settle who should bear responsibility for a resulting judgment.

These boundaries preserve the interface between power and delegation. Once criteria can be selected, stabilized, or operationalized by components other than the final decision-maker, responsibility depends partly on which judgment functions were delegated, which criteria were inherited, who retained authority to contest them, and who remained capable of reconstructing the basis of closure. Criterion power therefore prepares the ground for a more precise analysis of delegation and judgment ownership without reducing either concept to formal decision authority alone.

12 Judgment Ownership in Delegated Configurations

This section examines what remains of judgment when substantial parts of inquiry are delegated to artificial or institutional components. Its role is to distinguish delegation of judgment-relevant functions from delegation of authority, responsibility, and ownership of a judgment. The objective is to identify conditions under which an individual or institution can meaningfully stand behind a judgment even when evidence retrieval, option generation, warrant synthesis, criterion application, or recommendation has been partly delegated. The discussion proceeds by representing delegation within the existing function-allocation framework, distinguishing levels of delegation, developing a bounded account of judgment ownership, relating ownership to responsibility gaps and moral crumple zones, and specifying developmental consequences for recurrent Human–AI inquiry. The analysis is conceptual. It does not provide a complete theory of moral or legal responsibility, and it does not assume that responsible judgment requires unaided human performance.

12.1 Delegation within a Distributed Judgment Configuration

Section 5 represented judgment-relevant work through the allocation relation $\Lambda_t^{(q,\xi)}$ and kept functional contribution separate from local authority $\text{Auth}_t^{(q,\xi)}$. Delegation introduces a directional relation within that architecture. A component can remain institutionally associated with a task while arranging for another component to perform a substantial part of the relevant judgment function.

Let $u, v \in \mathcal{U}_t^{(q,\xi)}$ denote components within the selected configuration and let $f \in \Phi^J$ denote a judgment-relevant function. Equation 22 represents task-relative delegation.

$$\text{Del}_t^{(q,\xi)} \subseteq \mathcal{U}_t^{(q,\xi)} \times \Phi^J \times \mathcal{U}_t^{(q,\xi)}. \quad (22)$$

A tuple $(u, f, v) \in \text{Del}_t^{(q,\xi)}$ indicates that, within the selected task and configuration, component u relies on component v to perform a substantial share of function f . Equation 22 is descriptive. It does not by itself establish whether u had legitimate authority to delegate, whether v is competent, whether the delegation is reversible, or where responsibility should lie.

Delegation therefore differs from simple participation. A retrieval system can participate in evidence acquisition without anyone having deliberately delegated the whole evidential function to it. Conversely, an organization can formally delegate a screening function to an AI system even when humans still perform occasional review. For judgment formation, the important issue is how recurrent function allocation changes what different components continue to practice, inspect, contest, and reconstruct, irrespective of the workflow label.

Definition 12.1 (Judgment-function delegation). Within this paper, *judgment-function delegation* is a task- and context-relative arrangement in which a component that remains connected to a judgment episode relies on another component to perform a substantial part of a judgment-relevant function. Delegation can concern framing, evidence acquisition, option construction, warrant synthesis, criterion application, monitoring, closure, implementation, or review, and it need not transfer authority or responsibility in the same proportion.

Definition 12.1 makes it possible to distinguish beneficial specialization from unexamined substitution. Delegation can improve judgment when it places a function with a more competent component, in-

Table 10: Recurrent forms of delegation in AI-mediated judgment.

Delegated function	Typical AI-mediated form	Judgment-formation question
Computation or transformation	Calculation, sorting, summarization, formatting, translation	Does the user still understand the assumptions and limits relevant to later use?
Evidence acquisition	Search, retrieval, extraction, source triage	Can omitted, low-ranked, or provenance-poor evidence still be detected and contested?
Option generation	Candidate hypotheses, interpretations, scenarios, plans	Does independent option construction remain practiced or reconstructable?
Warrant synthesis	Explanations, argument assembly, comparisons, confidence cues	Can the receiving configuration distinguish supplied reasons from reasons it can independently defend?
Criterion application	Scoring, ranking, eligibility or risk classification	Are the operative criteria, thresholds, and exclusions visible and contestable?
Closure or action recommendation	Recommended answer, action, escalation, or deferral	Does formal human approval correspond to meaningful ability to revise or refuse closure?
Review and learning	Error diagnosis, feedback, post-hoc explanation, performance summary	Can the configuration learn from outcomes rather than merely consume a system-generated account of them?

creases redundancy, reduces avoidable workload, or preserves scarce attention for tasks requiring contextual interpretation. The same arrangement can weaken formation when delegation removes the practice through which a human or institution previously learned to generate alternatives, recognize anomalies, test reasons, or revise criteria.

12.2 Forms and Depths of Delegation

Delegation is not one-dimensional. A system can perform calculations while leaving interpretation untouched, produce candidate reasons while leaving criterion selection to the user, or generate a recommendation that effectively determines closure unless a human actively overrides it. Table 10 distinguishes several recurrent forms.

Table 10 shows why the amount of delegated work is weaker information than the *location* of delegation within the judgment architecture. Delegating arithmetic can reduce unnecessary cognitive burden while leaving problem framing and evaluative judgment intact. Delegating criterion application or closure can have much stronger consequences because those functions determine which reasons acquire practical force and when inquiry is treated as sufficiently settled.

The relation is also path dependent. A person who first learns to perform a function and later delegates it can retain a reconstructive model that supports oversight. A novice who encounters the function only through AI outputs can display the same immediate performance while lacking comparable independent discrimination. The difference need not justify a universal preference for manual performance. It does show why the developmental significance of delegation depends partly on prior formation, continued practice, feedback, and the possibility of reacquiring the function when conditions change.

12.3 Judgment Ownership beyond Final Approval

The phrase *judgment ownership* can be misleading if it is treated as property, authorship, or exclusive causal production. The present paper uses it more narrowly. A component owns a judgment to the extent that it can meaningfully stand behind the judgment as its answerable position while understanding enough of the basis, criteria, limits, and revision conditions to contest or revise it when warranted.

A provisional ownership profile for component u is given in Equation 23.

$$O_t^{J,(q,\xi)}(u) = (\text{Rec}_u, \text{Crit}_u, \text{Rev}_u, \text{Ans}_u)_t^{(q,\xi)}. \quad (23)$$

In Equation 23, Rec_u denotes reconstructive access to the judgment’s relevant basis and limitations; Crit_u denotes practical ability to understand and, where appropriate, contest the operative criteria; Rev_u denotes authority and capacity to revise, reopen, veto, or escalate the judgment; and Ans_u denotes an answerability relation through which the component recognizes that the judgment can properly require explanation, justification, or response from it. The profile is not a scalar and is not proposed as a metric of blameworthiness.

Definition 12.2 (Judgment ownership). Within this paper, *judgment ownership* is a task-relative relation in which a person or institution can reasonably present a judgment as its answerable position because it retains sufficient reconstructive access, criterion awareness, revision capacity, and responsibility awareness for the consequences and limits relevant to that judgment. Ownership can coexist with extensive delegation and need not imply that the owner performed each contributing operation personally.

Definition 12.2 rules out two shortcuts. First, final approval does not by itself establish ownership. A person who clicks “accept” while lacking access, time, competence, or practical permission to contest the recommendation can retain formal authority without meaningful ownership. Second, substantial AI contribution does not automatically eliminate ownership. A clinician, researcher, analyst, or institution can rely heavily on specialized systems while remaining able to reconstruct the relevant basis, understand the limitations, contest the criteria, and reopen the decision.

The ownership relation is therefore closer to *answerable endorsement* than to causal authorship. This distinction matters in collaborative inquiry as well as AI use. Many judgments have always depended on instruments, laboratories, statistical packages, colleagues, databases, and institutional standards. AI increases the depth and opacity of some dependencies, but distributed production alone is insufficient to show that no one can own the resulting judgment.

12.4 Responsibility Gaps and Misallocated Responsibility

The ethics of autonomous and learning systems provides an important neighboring literature. Matthias introduced the term *responsibility gap* to describe cases in which increasingly autonomous learning systems complicate traditional responsibility attribution because relevant behavior may not be reasonably predictable or directly controlled by the manufacturer or operator [30]. The present paper does not assume that every generative-AI judgment creates such a gap. The relevance is narrower: as judgment functions become distributed, responsibility cannot safely be inferred from visibility, formal approval, or proximity to the final action.

A different failure occurs when responsibility is assigned too easily to the most visible human component. Elish’s analysis of *moral crumple zones* shows how humans in complex automated systems can absorb blame for system failures despite having limited practical control over the behavior that produced the outcome [31]. This is especially relevant to nominal human oversight. A workflow can be designed so that a human is formally required to approve an AI-supported judgment while the surrounding time pressure, information asymmetry, organizational expectations, and interface design make meaningful contestation difficult. In such a case, “human in the loop” language can preserve responsibility attribution without preserving the conditions required for responsible judgment.

These literatures motivate a separation among at least four questions: who contributed which judgment functions, who possessed practical authority, who could reconstruct and contest the relevant basis, and who was institutionally or morally answerable for the outcome. The answers can overlap, but they need not coincide.

Caution 12.3 (Ownership without scapegoating). A requirement that a human or institution own an AI-mediated judgment should not be implemented by assigning responsibility to a visible approver while

withholding the information, authority, time, competence, or contestation pathways required for meaningful ownership. Responsibility structures should be assessed together with the judgment functions and practical control available to the purported owner.

Caution 12.3 also applies in the opposite direction. Distributed technical complexity should not become a general device for dissolving responsibility. An organization that deliberately adopts an opaque workflow, removes appeal routes, or benefits from AI-mediated decisions can remain relevant to responsibility even when no individual can predict each model output. The present paper leaves the full normative allocation of responsibility open, but it requires the institutional design of delegation to remain visible in the analysis.

12.5 Meaningful Control and Answerable Delegation

Work on meaningful human control provides a useful interface because it rejects the idea that merely placing a human somewhere in the causal chain is sufficient for control or responsibility. Mecacci and Santoni de Sio develop meaningful human control in terms of reason-responsiveness and discuss two conditions inherited from the broader framework: tracking relevant human reasons and tracing system behavior to human agents who understand the system’s capabilities and appreciate their responsibility [32]. The present paper does not import this framework as a complete theory of inquiry. It uses it to sharpen the distinction between nominal authority and judgment ownership.

For AI-mediated inquiry, meaningful ownership need not require continuous manual intervention. It can be supported when delegated systems remain sufficiently connected to the reasons relevant to the task, when responsibility-bearing humans or institutions understand the system’s role and limits, and when practical mechanisms exist for contestation, escalation, and revision. The key condition is not maximal control over each operation. It is a sufficiently robust relation between the reasons that justify the judgment, the components that perform delegated functions, and the agents who remain answerable for closure or action.

This gives a more precise interpretation of *answerable delegation*. Delegation is answerable when the delegating configuration can specify why a function was delegated, under what competence assumptions, what evidence would trigger verification or override, who can reopen the judgment, and where responsibility remains located when the delegated component fails or operates outside its domain.

Claim 12.4 (Answerable delegation). Delegation is more consistent with judgment ownership when the delegating configuration retains task-relevant reconstructability, criterion awareness, revision capacity, and explicit responsibility relations, and when the delegated component’s competence conditions and failure modes remain sufficiently inspectable to support selective reliance and contestation.

Claim 12.4 does not require complete technical explainability. In many complex domains, no single user can reconstruct the internal operation of every instrument or model. Reconstructability is task-relative: the owner should understand enough of the evidential basis, competence boundary, uncertainty, criteria, and revision conditions to know what is being endorsed and what would justify reopening it.

12.6 Delegation as a Developmental Intervention

Delegation also feeds back into judgment formation. Section 5 showed that formation follows the functions a component repeatedly performs, monitors, contests, or reconstructs. The present section adds that recurrent delegation can change those practice conditions even when immediate configuration-level performance improves.

Several trajectories are possible. Delegating routine computation can free attention for higher-level framing and evaluation. Delegating option generation can expose a learner to a wider candidate space and improve comparative judgment. Delegating verification can weaken source-checking habits if the user loses contact with provenance. Delegating closure can gradually turn review into ritual approval. Delegating post-hoc explanation can make error learning dependent on the same system whose behavior

is being evaluated. The developmental effect therefore depends on which functions are delegated, what complementary practice remains, and whether the recipient can later reconstruct or resume the delegated function when needed.

The relevant educational objective is *delegation literacy*: the capacity to decide what can be delegated, what should remain practiced, what requires redundancy, and what must remain contestable because losing the function would compromise later judgment. Universal independence from AI is unnecessary for this objective. Delegation literacy is closely related to selective reliance but adds a longitudinal question: what pattern of delegation is the current workflow training the configuration to become?

Claim 12.5 (Delegation and formation). The developmental effect of AI delegation depends partly on whether the delegated function remains reconstructable, monitorable, contestable, and recoverable within the later judgment configuration. Immediate performance gains therefore provide incomplete evidence about the long-term effect of delegation on judgment capacity.

Claim 12.5 prepares the practical turn of the paper. Once ownership and responsibility are separated from raw functional contribution, judgment practice can be designed around explicit delegation boundaries, preserved reconstruction, calibrated verification, counter-criteria, and post-decision learning rather than around a binary choice between unaided judgment and automation.

13 Judgment Practice in AI-Rich Inquiry

This section translates the analytical architecture developed in Sections 3–12 into recurrent practices of inquiry. Its role is practical and architecture-oriented: the objective is to identify what kinds of activity preserve opportunities for judgment formation when AI participates in problem framing, option generation, evidence organization, evaluation, verification, or closure. The section therefore organizes practices around the temporal structure of a judgment episode rather than offering a universal checklist. It distinguishes preparatory structuring, active inquiry, closure or deferral, and post-judgment reconstruction; relates each phase to salience, calibration, criterion contestability, delegation, and ownership; and identifies the traces through which repeated episodes can contribute to later judgment capacity. The discussion remains task-relative. Appropriate practice depends on stakes, reversibility, available expertise, institutional constraints, and the competence profile of the Human–AI configuration.

13.1 Judgment Practice as Recurrent Configuration

A situated judgment episode, as defined in Section 3, records how salience, options, warrants, evaluative criteria, monitoring, and response are organized for a particular task. A *judgment practice* concerns the recurrent organization through which such episodes are prepared, conducted, closed, reconstructed, and linked to later inquiry. This distinction is important because a single well-reasoned answer provides limited evidence about the configuration that produced it. Formation depends on patterns that recur across episodes: which functions are habitually delegated, which criteria are made explicit, what triggers verification, whether counterevidence remains accessible, who can reopen a decision, and whether errors are converted into later changes in practice.

Definition 13.1 (Judgment practice). Within this paper, a *judgment practice* is a recurrent, task-sensitive organization of judgment functions, reliance relations, criteria, monitoring, closure conditions, responsibility relations, and reconstructive activity through which a history-bearing configuration conducts inquiry and modifies the conditions of later judgment.

Definition 13.1 gives practice a broader scope than repeated task execution. A practice can include deliberate pauses, preserved disagreement, source checking, independent comparison, escalation to another person, or explicit deferral. It can also include decisions about where AI participation is appropriate and where independent human or institutional judgment should remain available. The relevant unit is therefore a recurrent configuration of inquiry rather than a fixed sequence of manual steps.

Table 11: Analytical phases of judgment practice in AI-rich inquiry

Practice phase	Primary judgment work	AI-mediated vulnerability	Developmentally useful trace
Preparatory structuring	Problem boundary, salient unknowns, criteria, delegation, verification conditions	Premature framing, inherited defaults, opaque criterion selection	Initial framing, declared criteria, delegation rationale, expected failure conditions
Active inquiry	Option generation, warrant assessment, comparison, counter-criteria, uncertainty monitoring	Fluent convergence, provenance loss, correlated generation, pseudo-calibration	Alternatives considered, provenance checks, disagreements, criterion changes, verification triggers
Closure or deferral	Commitment, action, postponement, escalation, retained plurality	Automation bias, ritual approval, premature closure, responsibility diffusion	Closure rationale, unresolved uncertainty, ownership location, reopening conditions
Reconstruction	Error analysis, surprise, outcome interpretation, responsibility review, practice revision	Post-hoc rationalization, AI-generated retrospective coherence, loss of independent diagnosis	Revised reliance policy, failure-mode record, preserved counterexample, modified practice condition

This account is compatible with traditions of reflective professional practice without identifying judgment formation with reflection alone. Schön’s account of reflection-in-action is useful because it treats competent practice as involving situated interpretation and revision during indeterminate activity rather than only retrospective application of explicit rules [19]. The present framework adds a specific AI-mediated concern: reflection can itself be distributed across human and artificial components, and the developmental significance of that distribution depends on whether relevant functions remain inspectable, contestable, and recoverable.

Table 11 summarizes the practice architecture used in this section. The table is analytical rather than prescriptive. Different domains can compress, repeat, reorder, or omit phases when justified by the task.

13.2 Preparatory Structuring of Inquiry

Judgment practice begins before an AI system produces an answer. Preparatory structuring makes explicit enough of the inquiry configuration to prevent the first generated response from silently becoming the problem definition, option boundary, and criterion regime at once. The relevant activities include identifying the provisional object of inquiry, preserving salient unknowns, distinguishing externally imposed constraints from revisable criteria, allocating functions across human and artificial components, and deciding what kinds of uncertainty or failure will trigger additional verification.

This preparation need not require a complete specification of the problem. Section 7 emphasized that problem formation itself is a judgment process. A useful practice therefore preserves a difference between a provisional framing and an endorsed framing. In exploratory research, for example, a user can begin with a tentative question while recording which assumptions came from the user, which emerged through AI interaction, and which remain disputed. In professional settings, an institution can distinguish mandatory legal or procedural constraints from model-suggested priorities. The objective is not to freeze the inquiry before it begins. It is to keep the genealogy of the framing sufficiently visible that later revision remains possible.

Preparatory structuring also provides a natural location for *delegation declarations*. Section 12 treated delegation as a directional relation among judgment functions. In practice, the declaration can be modest: which component is expected to generate alternatives, retrieve evidence, summarize disagreement, estimate uncertainty, check sources, or recommend closure; which functions require independent

confirmation; and who retains authority to alter the criterion regime. Such declarations are especially useful when the same AI system can move fluidly from retrieval to interpretation to recommendation, because interface continuity can otherwise conceal functional transitions.

The preparation phase should also establish verification conditions at a level proportionate to the task. Section 6 argued that verification intensity should respond to task-local reliability, uncertainty, stakes, and the consequences of error. A practice can operationalize that principle by identifying events that trigger stronger checking: unfamiliar sources, large model disagreement, high-stakes claims, surprising numerical results, missing provenance, changes in criterion definition, or recommendations that substantially alter the future option space. These triggers are not universal. Their role is to convert calibration from a retrospective judgment into an actionable part of inquiry design.

Claim 13.2 (Preparatory visibility). AI-mediated judgment is more reconstructable when provisional framing, criterion selection, functional delegation, and verification conditions are made sufficiently visible before or during the early stages of inquiry to support later comparison and revision.

Claim 13.2 does not imply that every inquiry requires formal documentation. The required visibility is task-relative. A low-stakes exploratory conversation may need little more than awareness of provisional assumptions, whereas a consequential research, clinical, legal, or institutional judgment may require durable records and explicit review conditions.

13.3 Active Inquiry through Generation, Verification, and Contestation

During active inquiry, AI can expand the number and speed of candidate explanations, arguments, examples, interpretations, and plans available to the configuration. The formative value of this abundance depends on how generation is coupled to discrimination. Repeated exposure to plausible outputs can increase option supply while reducing practice in recognizing missing alternatives, evaluating provenance, or distinguishing independent support from stylistic confidence. Judgment practice therefore benefits from organizing generation and evaluation as related but separable activities.

One recurrent pattern is *comparative generation*. The configuration preserves more than one plausible candidate long enough for reasons, evidence, and failure conditions to become visible. Candidates can come from the human participant, the AI system, external sources, or independent model runs. The practice matters because the first coherent completion can otherwise acquire an anchoring role disproportionate to its evidential status. Comparative generation is especially useful for problem framing, hypothesis formation, interpretive disputes, and tasks in which omitted alternatives carry substantial cost.

A second pattern is *provenance-sensitive verification*. Verification concerns more than asking the same system whether its previous answer is correct. When evidence claims materially affect the judgment, practice should preserve a route to the underlying source, dataset, calculation, institutional rule, or independent expert assessment. The required route can vary by domain, but verification becomes weaker when generation, critique, and confirmation all depend on the same unexamined informational pathway. This is consistent with the selective-reliance framework in Section 6: appropriate reliance depends on discrimination among conditions of competence, not on a stable attitude of trust or distrust toward AI.

A third pattern is *criterion contestation*. Section 11 showed that transparent criteria can remain difficult to challenge. Active inquiry therefore requires some practical possibility of asking whether the operative criterion regime is adequate to the case. Counter-criteria can be introduced when a ranking, recommendation, or synthesis appears stable only because one evaluative register has displaced another. A research recommendation optimized for tractability may deserve reconsideration under scientific significance; an institutional recommendation optimized for efficiency may deserve review under fairness or reversibility; an apparently elegant theoretical account may require separate evidential assessment. Counter-criterion practice keeps disagreement about the rules of evaluation distinguishable from disagreement within those rules.

A fourth pattern is *metacognitive interruption*. Human–AI studies of cognitive forcing indicate that interventions which slow or restructure reliance can reduce overreliance under some experimental conditions, although such interventions can impose usability costs [4]. The present framework generalizes

the underlying design intuition cautiously: an interruption is useful when it creates an occasion for task-relevant monitoring that would otherwise be bypassed. The interruption can take the form of an independent estimate before viewing an AI recommendation, a requirement to state a confidence range, a prompt to identify disconfirming evidence, or a source check triggered by uncertainty. Its value depends on whether it improves discrimination rather than merely adding procedural burden.

Claim 13.3 (Generation–discrimination coupling). The developmental value of expanded AI-mediated generation depends partly on whether the inquiry configuration preserves recurrent opportunities to discriminate among alternatives, warrants, criteria, competence conditions, and failure modes. Greater candidate abundance alone therefore provides limited evidence of stronger judgment formation.

Claim 13.3 therefore makes comparative and reconstructive capacity, rather than mandatory unaided generation, the relevant safeguard against generated convergence.

13.4 Closure, Deferral, and Answerable Commitment

Closure is itself a judgment practice. Section 3 defined closure broadly enough to include action, commitment, revision, explicit deferral, or preservation of unresolved alternatives. AI-rich inquiry can make closure easier because a fluent answer can supply the appearance of completion even when uncertainty, criterion conflict, or responsibility remains unresolved. A mature practice therefore treats the conditions of closure as part of the object of judgment.

Three distinctions are useful. First, *epistemic closure* concerns whether the configuration has enough warrant for the claim being made. Second, *practical closure* concerns whether action is justified given time, stakes, reversibility, and available alternatives. Third, *institutional closure* concerns whether a procedure has reached the point at which an organization records, authorizes, or implements a decision. These forms can coincide, but they often diverge. An institution may need to act while epistemic uncertainty remains substantial; a researcher may have enough evidence for a provisional interpretation while preserving competing explanations; a student may need to submit work even though additional inquiry would still be valuable.

Deferral is therefore a substantive response rather than an absence of judgment. It can preserve optionality when the expected value of additional evidence is high, when responsibility is unclear, when criterion conflict remains unresolved, or when the available system is operating outside a competence boundary. Conversely, repeated deferral can itself become a failure when action is time-sensitive. Temporal judgment from Section 9 is therefore inseparable from closure practice: the question is partly whether the current uncertainty justifies acting now, waiting, escalating, or preserving alternatives.

Answerable closure also requires an identifiable location of ownership. The person or institution that closes the judgment should retain enough reconstructive access, criterion awareness, revision capacity, and responsibility awareness to explain what is being endorsed and under what limitations. A final AI-generated summary can support this process, but it should not substitute automatically for the configuration’s own ability to identify unresolved assumptions, the basis of reliance, and the conditions under which the judgment would be reopened.

Definition 13.4 (Answerable closure). *Answerable closure* is a task-relative completion, commitment, action, or explicit deferral for which the responsible configuration can state the operative basis of the response, the material uncertainties and criteria relevant to it, the location of judgment ownership, and the conditions that would justify revision or reopening where such conditions are identifiable.

Definition 13.4 permits bounded uncertainty. It does not require exhaustive reasons, perfect prediction, or complete explanation of every model operation. Its purpose is to distinguish owned commitment from procedural finality produced by workflow momentum.

13.5 Reconstruction after Judgment

Judgment formation becomes longitudinal when an episode leaves a usable trace. Section 3 represented developmental residue as the change in later-accessible capacities, criteria, monitoring dispositions, and

relational conditions produced by an episode. Post-judgment reconstruction is one practice through which that residue can become inspectable and revisable. Its object extends beyond answer correctness to how the judgment was formed, where uncertainty entered, which delegated functions mattered, what evidence was missing, whether a criterion regime distorted the comparison, and what should change in the next configuration.

Outcome feedback is informative but incomplete. A correct result can follow from weak judgment, fortuitous model behavior, or a fragile shortcut; an incorrect result can follow from a reasonable process under genuine uncertainty. Reconstruction therefore benefits from preserving a distinction between *outcome error* and *judgment-process error*. The latter can include premature closure, unrecognized domain shift, poor criterion choice, inadequate source verification, miscalibrated confidence, unjustified delegation, or failure to preserve a relevant alternative. This distinction allows learning to target the mechanism that should change rather than treating every bad outcome as evidence that the opposite decision rule should be adopted.

The reconstructive phase also requires caution about retrospective AI assistance. A system can generate coherent post-hoc explanations that were not part of the original judgment process. Such explanations can be useful for hypothesis generation, but they should not be confused with historical reconstruction unless supported by preserved records. A minimal judgment trace can therefore have developmental value: the provisional framing, material alternatives, important warrants, confidence or uncertainty, delegation points, criterion changes, and closure rationale. The trace need not record every conversational turn. Its role is to preserve enough structure to distinguish what actually influenced the judgment from a later rationalization.

Research on appropriate reliance also examines learning across Human–AI interaction rather than treating reliance only as an isolated episode. Schemmer and colleagues, for example, investigate human learning in relation to appropriate reliance in Human–AI decision-making [12]. The present framework interprets such learning broadly: repeated interaction can modify not only reliance on an AI source but also the user’s option-generation habits, criterion sensitivity, verification patterns, closure thresholds, and expectations about which judgment functions remain personally or institutionally available.

Claim 13.5 (Reconstructive learning). Post-judgment reconstruction contributes to judgment formation when it can distinguish outcome feedback from process diagnosis and convert identified failures, surprises, or successful discriminations into revisable changes in later reliance, criteria, verification, delegation, or closure practice.

Claim 13.5 treats a judgment trace as developmentally useful only when it supports later discrimination and revision.

13.6 Practice Variation across Tasks and Configurations

A practice architecture should not be mistaken for a single optimal workflow. Human–AI configurations differ in expertise, time, stakes, institutional authority, reversibility, model capability, evidential access, and the maturity of existing procedures. The same practice can therefore be formative in one setting and wasteful or distorting in another. An experienced researcher may use AI for rapid alternative generation while independently controlling evidence and theoretical closure. A novice may benefit from more explicit provenance work and independent problem formulation. A multidisciplinary team may distribute criterion contestation across members rather than requiring every participant to reconstruct every technical component. A high-stakes institution may require durable audit and appeal pathways that would be disproportionate for exploratory private inquiry.

Naturalistic decision-making research is relevant here because expert professional judgment often develops under time pressure, uncertainty, shifting conditions, and domain-specific experience rather than through exhaustive comparison of formally enumerated alternatives [21, 20]. AI-rich practice should therefore preserve room for domain-sensitive pattern recognition while also recognizing that model mediation can alter the cues from which such recognition develops. The aim is not to replace situated ex-

pertise with a universal deliberative protocol. It is to keep the conditions of expertise formation visible when AI changes what is encountered, noticed, practiced, and reconstructed.

One implication is that independent performance should be preserved selectively rather than ritualistically. Requiring unaided work on every component can waste resources and suppress the benefits of specialized tools. Eliminating independent practice across a function can also make later monitoring or recovery difficult. The relevant design question is which functions require continuing first-hand practice, which can be delegated while remaining reconstructable, which require periodic independent calibration, and which can be transferred more fully because the configuration retains adequate alternatives and oversight elsewhere.

A second implication is that redundancy should be functional rather than merely numerical. Asking several similar models the same question can reveal instability, yet correlated systems can reproduce the same omission or criterion bias. More useful redundancy can involve different evidence sources, different disciplinary perspectives, a human counter-analysis, or a deliberately altered criterion regime. Redundancy contributes to judgment when it creates genuinely different opportunities for error detection and contestation.

Claim 13.6 (Configuration-sensitive practice). Judgment practice in AI-rich inquiry should be evaluated relative to the task, stakes, reversibility, competence distribution, and available contestation pathways of the configuration. Uniform requirements for either maximal AI use or maximal independent human performance are therefore poorly aligned with the task-relative account developed in this paper.

Claim 13.6 therefore makes configuration description part of practice evaluation because identical routines can have different developmental significance under different competence, stakes, reversibility, or contestation conditions.

13.7 Practice Architecture and the Cultivation Problem

The practices developed in this section identify recurrent opportunities through which judgment can remain active within AI-mediated inquiry: provisional framing, explicit delegation, calibrated verification, comparative generation, counter-criterion review, answerable closure, and post-judgment reconstruction. Their common function is to preserve discriminations that would otherwise become difficult to observe once generation, evaluation, and closure are compressed into a fluent interaction.

These practices do not by themselves explain how a person or institution acquires the capacity to perform them well. A learner can follow a verification routine mechanically without developing sensitivity to when verification matters. A researcher can record alternative hypotheses without learning to recognize a genuinely missing alternative. An institution can require human approval while weakening the authority and time needed for meaningful contestation. Practice architecture therefore specifies occasions for judgment but does not guarantee formation.

This limitation defines the next analytical step. Section 13 has identified what recurrent inquiry activity can preserve; the following section turns to cultivation: how judgment capacities can develop through experience, guided comparison, apprenticeship, reflection, feedback, and appropriately structured Human–AI participation. The later analysis of cognitive friction then asks which difficulties are formative and which merely consume attention or exclude participation.

14 Judgment Cultivation

This section examines how the capacities presupposed by the practice architecture in Section 13 can develop over time. Its role is developmental: the objective is to identify recurrent conditions through which a history-bearing configuration can become more capable of framing problems, discriminating among alternatives, calibrating reliance, contesting criteria, mediating among judgment registers, and reconstructing its own decisions. The discussion proceeds from the developmental-residue account in Section 3. It first distinguishes cultivation from exposure and immediate performance, then considers experience

across varied cases, guided comparison and articulation, reconstructive feedback, staged Human–AI participation, and institutional conditions that sustain formation. The section does not assume a single pedagogy of judgment. It identifies a family of developmental relations whose usefulness depends on domain, prior expertise, stakes, available feedback, and the functions that a Human–AI configuration is expected to retain.

14.1 Cultivation as History-Bearing Reconfiguration

Section 3.5 represented judgment formation through changes in a history-bearing configuration that remain available to later inquiry. *Cultivation* concerns the deliberate or institutionally supported organization of experiences through which such changes become more likely, more discriminating, more revisable, or more transferable. This makes cultivation broader than instruction. A lecture can contribute to cultivation, yet so can comparison among cases, supervised participation, retrospective reconstruction, exposure to disagreement, revision after error, or repeated movement between independent and AI-supported performance.

Definition 14.1 (Judgment cultivation). Within this paper, *judgment cultivation* is the organization of experiences, relations, feedback, roles, and reflective opportunities intended to produce developmental residues that remain available for later task-relevant judgment and that can themselves be inspected, revised, and recontextualized.

Definition 14.1 places the developmental target in the configuration rather than in a single observable behavior. Repeatedly obtaining correct answers with AI assistance can improve task performance while leaving little independent sensitivity to framing, provenance, missing alternatives, or model failure. Conversely, a sequence containing disagreement, failed hypotheses, or delayed closure can contribute to formation when it changes later discrimination and monitoring. The relevant evidence therefore includes how a learner’s later option construction, confidence, verification behavior, criterion use, delegation choices, and reconstruction change across episodes.

The distinction also clarifies why cultivation cannot be reduced to information transfer. Judgment often concerns relations among heterogeneous considerations whose relevance becomes visible only within concrete situations. Metacognitive knowledge and monitoring, for example, concern what a person knows about cognition and how ongoing cognition is assessed rather than merely the possession of task facts [5]. In the present framework, such monitoring forms one component of a wider configuration that also includes salience, evaluative criteria, temporal sensitivity, reliance relations, and responsibility. Cultivation therefore includes learning what to notice, what to compare, when to doubt, what to preserve, and what requires another form of judgment.

Claim 14.2 (Cultivation–performance distinction). Immediate assisted performance provides incomplete evidence of judgment cultivation. Stronger evidence concerns later changes in framing, discrimination, monitoring, criterion use, reliance, reconstruction, and transfer under conditions that differ meaningfully from the original episode.

Claim 14.2 follows from the developmental-residue principle in Claim 3.5. It also establishes the main constraint on the mechanisms considered below: a practice is developmentally relevant when it changes what the configuration can later do; immediate ease of completion alone is insufficient.

14.2 Experience across Variation and Contrast

Judgment becomes discriminating through encounters with differences that matter. Repetition of nearly identical cases can improve fluency while leaving boundary conditions obscure. Cultivation therefore benefits from structured variation: cases in which the correct reliance relation changes, similar surface cues conceal different underlying structures, one evaluative criterion becomes decisive while another recedes, or a previously successful heuristic encounters a domain shift. The purpose is not maximal novelty. It is exposure to contrasts that make the basis of a judgment more visible.

This point is especially important for AI-mediated inquiry because an artificial system can smooth variation by supplying similarly structured explanations, standardizing terminology, or filling gaps before the learner experiences the uncertainty that distinguishes one case from another. The same system can also be used constructively to generate contrastive cases, alternative framings, counterexamples, or competing interpretations. The developmental consequence depends on whether the variation becomes an occasion for human discrimination or is merely consumed as additional generated content.

Naturalistic decision-making research provides a neighboring perspective in which expertise develops through experience with recurring patterns, anomalies, time pressure, and changing situational cues [21, 20]. The present account adds explicit criterion comparison, calibration, and reconstruction, while retaining the shared implication that encountered contrasts shape later recognition.

A useful cultivation design therefore asks what dimensions of variation the learner has actually encountered. For calibration, the relevant contrast may involve when an AI source is reliable and when it fails. For problem formation, it may involve cases with similar data but different questions worth asking. For normative judgment, it may involve similar outcomes produced under different rights, obligations, or distributions of burden. For temporal judgment, it may involve situations in which early signals resemble noise until later trajectory information makes their relevance visible. Cultivation deepens when the configuration learns to discriminate among these relations rather than memorizing a fixed response associated with a surface category.

Claim 14.3 (Contrastive cultivation). Judgment cultivation is strengthened when experience varies along dimensions that are relevant to later discrimination and when the learner must identify which differences change the appropriate framing, criterion, reliance relation, or response.

Claim 14.3 is deliberately task-relative. Random variation can distract rather than cultivate. The developmental value of contrast depends on whether the difference is connected to a judgment function the learner may later need to exercise.

14.3 Guided Comparison and Articulated Judgment

Experience becomes more informative when comparison is guided without converting judgment into rule following alone. A teacher, supervisor, peer, or AI system can direct attention to differences in evidence quality, missing alternatives, criterion conflicts, or confidence without dictating the final conclusion. The learner can then articulate why two superficially similar cases require different treatment or why two competing judgments remain reasonable under different assumptions. Such articulation makes parts of the operative judgment structure available for correction and later reconstruction.

The value of articulation should remain qualified. Skilled judgment is not exhausted by what can be verbalized. Schön's account of professional practice emphasizes reflection within indeterminate situations in which competent action cannot be represented adequately as the mechanical application of explicit technical rules [19]. Naturalistic accounts likewise emphasize experience-based recognition and situated adaptation [21, 20]. The present framework therefore treats articulation as one cultivation device among others. Its role is to externalize enough of a judgment to support comparison, feedback, criterion inspection, and later revision while preserving the possibility that some domain-sensitive perception remains only partially explicit.

Guided comparison can be organized around several recurrent contrasts. A learner may compare an initial unaided judgment with an AI-supported judgment; two AI recommendations produced under different prompts or evidence conditions; an expert's judgment with the learner's own; or a pre-outcome rationale with a post-outcome reconstruction. The educational value lies in the relation among these artifacts. A difference in answer matters less than identifying which evidence, criterion, cue, or assumption generated the divergence.

This also changes the role of explanation in AI-supported learning. An explanation can function as content to accept, as a hypothesis to test, or as a prompt for comparison. The third role is often more compatible with judgment cultivation because the learner must determine which parts of the explanation

are relevant, sufficient, misleading, or missing. AI-generated rationales can thereby become objects of judgment rather than substitutes for judgment.

Claim 14.4 (Articulated comparison). Articulation contributes to judgment cultivation when it supports comparison and revision of the reasons, criteria, uncertainties, and reliance relations that shaped a judgment. Verbal production alone does not establish that the relevant judgment capacity has been formed.

Claim 14.4 therefore treats articulation as an evidence-producing device without requiring expert judgment to become fully explicit.

14.4 Feedback, Reconstruction, and Revision

Feedback contributes to cultivation when it identifies what should change in later judgment. Outcome feedback can reveal whether an action succeeded, yet the developmental target may lie elsewhere: a missed cue, an unjustified confidence estimate, a narrow problem frame, an inappropriate criterion, or a delegation decision that concealed a source of error. Section 13.5 therefore distinguished outcome error from judgment-process error. Cultivation extends that distinction by asking whether feedback changes the configuration that produced the process.

A reconstructive learning episode can compare at least three records: what was believed or expected before closure, what occurred, and how the earlier judgment should now be interpreted. This temporal separation matters because retrospective coherence can hide the uncertainty that existed at the time. Preserved judgment traces make it possible to identify genuine surprise, overconfidence, ignored counterevidence, or criteria that shifted after the fact. AI can assist by organizing records, surfacing inconsistencies, or generating candidate explanations, but the system's retrospective narrative should remain distinguishable from the historical record of the original judgment.

Work linking human learning and appropriate reliance provides a direct empirical interface for this developmental problem [12]. Reconstructive feedback can alter a wider set of functions: what is noticed before consulting AI, when confidence is withheld, which sources are checked, when alternative criteria are introduced, and whether closure is delayed or escalated.

Correction should also remain connected to the reason for the error. If a learner accepts poor AI advice because of domain ignorance, the developmental response may require substantive knowledge. If the problem is confidence miscalibration, additional domain facts alone may be insufficient. If the error arose from institutional pressure to follow the model, individual metacognitive training may miss the operative relation. Cultivation is therefore more precise when feedback is typed to the function or relation that should change.

Claim 14.5 (Function-sensitive feedback). Feedback contributes more directly to judgment cultivation when it identifies the judgment function, relation, or criterion implicated in an episode and supports a corresponding change that remains available to later inquiry.

Claim 14.5 prevents an incorrect outcome from being treated as a unitary developmental signal: different error pathways require different revisions.

14.5 Staged Human–AI Participation

Human–AI participation can itself be organized developmentally. Section 5 showed that formation follows the functions a component repeatedly performs, monitors, contests, or reconstructs. Section 12.6 then treated delegation as a developmental intervention. The cultivation problem is therefore partly architectural: which judgment functions should be performed first-hand, jointly, under supervision, or through delegated support at different stages of development?

A fixed answer is unlikely to fit all domains. In some tasks, early independent problem formulation may be valuable because it reveals what the learner notices before an AI system establishes the frame. In other tasks, immediate AI-supported comparison may expose possibilities the learner could not yet

Table 12: Cultivation mechanisms and Human–AI participation

Cultivation mechanism	Developmental object	Constructive AI participation	Developmental vulnerability
Structured variation	Sensitivity to boundary conditions and task-relevant differences	Generate contrasting cases, counterexamples, or altered conditions	Surface variety without learner discrimination
Guided comparison	Reasons, criteria, alternatives, and uncertainty	Provide competing analyses or organize comparison	Model explanation becomes the default standard of evaluation
Reconstructive feedback	Error diagnosis, calibration, reliance, and revision	Organize traces, compare predictions with outcomes, surface inconsistencies	Retrospective coherence replaces historical reconstruction
Role variation	Delegation literacy and recoverability of judgment functions	Alternate generation, critique, verification, and synthesis roles	Persistent delegation removes first-hand practice of a needed function
Longitudinal revisit	Persistence, transfer, and revisability	Reintroduce prior cases under changed evidence or models	Repetition tests memory for the prior interaction rather than later judgment capacity

generate. A novice can benefit from model-generated alternatives while still being required to classify their evidential basis, identify missing assumptions, or rank which claims require verification. An experienced practitioner may delegate routine information synthesis while retaining direct engagement with anomalous evidence, criterion conflict, or irreversible closure.

The general design principle is *staged participation*: the distribution of functions changes as competence, task demands, and developmental objectives change. Support can be increased where the learner lacks access to relevant examples or expert feedback, and reduced where continued delegation would remove the practice needed to maintain a judgment function. The sequence need not move monotonically toward less AI use. Mature judgment can involve more sophisticated delegation because the practitioner has developed better calibration, verification, and reconstruction capacities.

Table 12 summarizes several cultivation mechanisms and their possible Human–AI arrangements. The table describes developmental possibilities rather than a fixed progression.

Staged participation also permits role reversal. The learner can ask AI to propose an answer and then critique it, generate the initial analysis and ask AI for counterarguments, inspect AI-selected evidence against independently gathered sources, or compare a model’s confidence cues with later outcomes. Such role variation prevents the Human–AI relation from becoming synonymous with one habitual division of labor. It can also reveal which functions have become recoverable and which remain dependent on a particular interface.

Claim 14.6 (Developmental role allocation). The developmental value of Human–AI participation depends partly on the distribution and variation of judgment functions across episodes. A configuration can use AI extensively while still cultivating judgment when relevant functions remain actively discriminated, monitored, contested, reconstructed, and recoverable where the task requires them.

Claim 14.6 makes the division of labor itself a developmental variable.

14.6 Cultivating Multi-Register and Temporal Judgment

Judgment cultivation also concerns relations among forms of judgment. Section 8 distinguished aesthetic, normative, and practical registers, while Section 9 examined temporal and mediating judgment.

Development can therefore involve more than becoming better within one register. It can involve learning when a register is being compressed into another, when several criteria should remain jointly visible, and when timing changes the appropriate relation among them.

Comparative cases can make these distinctions practicable. A technically elegant research direction may have weak social significance; a normatively attractive intervention may be impracticable under present institutional constraints; a practically effective shortcut may erode a future capability that the configuration later needs. Cultivation occurs when the learner can recognize these as different evaluative relations and can explain why the resulting tension should be mediated rather than collapsed into one score.

Temporal cultivation adds another layer. The same judgment can look different when evaluated at immediate, intermediate, and long-term horizons. Learners can therefore revisit earlier decisions after additional evidence arrives, compare expected with realized trajectories, and identify early changes that were initially dismissed. Such practices can develop sensitivity to timing and incipient change without assuming that every weak signal is meaningful. The developmental objective is discrimination among trajectories, including learning when waiting, acting, preserving optionality, or reopening a judgment becomes warranted.

AI can broaden the available comparison space by simulating scenarios, retrieving historical analogues, or making alternative criteria explicit. The risk is compression: generated summaries can make distinct registers appear commensurable merely because they are expressed in a common prose format. Cultivation requires the learner to preserve where translation remains partial and where a practical decision is necessary despite unresolved evaluative plurality.

14.7 Institutional Conditions of Cultivation

Judgment is cultivated within relations of authority, time, assessment, access, and responsibility. Individual exercises therefore operate inside an institutional environment that can support or undermine the capacities they are intended to develop. An organization can formally encourage critical review while rewarding speed and conformity; an educational programme can require independent reasoning while assessing only the final answer; a workplace can assign responsibility to a human reviewer while providing insufficient time or authority to contest an AI recommendation. Such conditions become part of the developmental environment because they determine which judgment functions are repeatedly practiced and which are discouraged.

The criterion-power analysis in Section 11 is directly relevant. Cultivation requires some opportunity to encounter, articulate, and sometimes contest the criteria by which judgments are evaluated. If assessment rewards only agreement with a model answer, learners can become skilled at reproducing an authorized criterion regime without developing sensitivity to its boundaries. Conversely, unconstrained criterion choice can make feedback incoherent. Institutions therefore need both relatively stable standards and pathways through which the grounds, scope, and revisability of those standards can become visible.

Responsibility structures matter in parallel. Learners and practitioners are unlikely to cultivate ownership when they are asked to approve judgments they cannot reconstruct or revise. Answerable delegation, developed in Section 12, therefore has an educational dimension. A training environment can make delegation rationales explicit, expose model limitations, require reopening under selected conditions, and allow the learner to experience consequences of both reliance and override. Such arrangements cultivate responsibility through participation rather than through abstract exhortation alone.

Institutional cultivation also requires preserving a diversity of competent relations. Peers, mentors, domain experts, affected stakeholders, documentary sources, instruments, and AI systems can each reveal different aspects of a judgment. Exclusive dependence on one channel can narrow the error-correction environment even when that channel is individually strong. A relational account therefore treats access to disagreement, alternative expertise, and appeal pathways as part of the conditions under which judgment can continue to develop.

Claim 14.7 (Cultivation environment). Judgment cultivation is partly an institutional property because time, authority, assessment, access to alternatives, responsibility, and criterion contestability shape which judgment functions are repeatedly exercised and which developmental residues remain available.

Claim 14.7 locates part of the developmental process in the institutional allocation of time, authority, feedback, and alternative expertise.

14.8 Cultivation without a Fixed Developmental Endpoint

The mechanisms considered in this section do not define a single mature state of judgment. Domains change, AI systems change, institutional criteria change, and the configurations in which judgment is exercised acquire new histories. Cultivation is therefore better represented as continued capacity for discriminating, revising, and reorganizing judgment relations than as movement toward a final level that eliminates dependence, error, or uncertainty.

This open developmental view also preserves the distinction between formation and normative evaluation established in Section 3.7. A person can become highly effective at navigating an institution's criteria while the criteria themselves remain questionable. A configuration can become excellently calibrated to one AI system and brittle when the system changes. A learner can become fluent in model-assisted reasoning while losing independent access to a function needed under failure conditions. Cultivation therefore includes the capacity to revisit what has been cultivated and to recognize when a previously adaptive pattern has become constraining.

The analysis also leaves one important question unresolved: what kinds of difficulty should a cultivation environment preserve? Structured variation, independent attempts, reconstruction, and contestation can all demand effort. Some effort may make relevant differences perceptible or keep a judgment function active; other effort may consume attention, exclude participants, or preserve obsolete procedures. The developmental value of difficulty cannot be inferred from effort alone. Section 15 therefore turns to cognitive friction in AI-mediated learning and inquiry, distinguishing friction that can contribute to formation from friction that is unnecessary, misallocated, or developmentally counterproductive.

15 Cognitive Friction in Judgment Cultivation

Section 14 identified an unresolved design problem for judgment cultivation: some forms of effort keep relevant judgment functions active, while other forms consume attention without producing useful developmental residue. This section develops *cognitive friction* as a relational category for analyzing that difference. Its role is discriminative rather than celebratory. The objective is to identify when resistance, delay, uncertainty, generation effort, verification burden, or reconstructive work can contribute to judgment formation, when such demands become misallocated or exclusionary, and how AI redistributes friction across a Human–AI configuration. The discussion proceeds by treating friction as task-, function-, and configuration-relative; distinguishing function-relevant formative friction from burdens that mainly consume capacity; examining AI-mediated friction redistribution; and specifying conditions for selective friction design. The section does not treat difficulty as intrinsically educational, nor does it assume that reducing effort necessarily weakens judgment.

15.1 Friction as a Relational Property

The ordinary language of “making a task easier” can obscure where effort has been removed and what role that effort previously played. A user may experience less difficulty because an AI system retrieves sources, proposes candidate explanations, formats a calculation, generates alternatives, evaluates those alternatives, or supplies a recommendation. These interventions reduce different burdens. Their developmental consequences can therefore differ even when the user's experienced effort decreases by a similar amount.

The present paper uses *cognitive friction* to name resistance encountered while a judgment-relevant function is being performed, coordinated, monitored, reconstructed, or transferred across components of a configuration. The resistance can arise from the task, the representation, the interface, the learner's prior formation, the need to coordinate heterogeneous criteria, the requirement to verify a source, the absence of an obvious option, or an institutional procedure. Friction is therefore neither a fixed property of an activity nor a synonym for cognitive load. The same requirement can be formative for one configuration, negligible for another, and obstructive for a third.

Definition 15.1 (Cognitive friction). Within this paper, *cognitive friction* is task- and configuration-relative resistance that requires the exercise, coordination, monitoring, reconstruction, or reallocation of one or more judgment-relevant functions before inquiry can proceed or close.

Definition 15.1 makes the relevant unit of analysis relational. Reading an untranslated archival source can create productive linguistic and interpretive engagement for an expert whose inquiry concerns historical wording, while imposing largely extraneous burden on a novice whose immediate objective is to understand a well-established conceptual distinction. Requiring independent generation before revealing AI suggestions can exercise option construction in one learning design; imposing the same requirement during an urgent high-stakes operational episode can delay access to needed support. Friction becomes developmentally interpretable only after the affected function, configuration, and task horizon are identified.

This relational treatment is compatible with several neighboring learning traditions while remaining distinct from them. Bjork and Bjork describe *desirable difficulties* as conditions that can make acquisition performance appear worse while supporting later retention or transfer [33]. Kapur's productive-failure research shows that initial struggle with complex problems can, under specified instructional sequences, support later learning and transfer [35]. Cognitive-load research provides the complementary warning that demanding problem-solving activity can consume processing capacity in ways that hinder schema acquisition [34]. In Human-AI decision support, cognitive-forcing interventions show that deliberately introduced effort can reduce overreliance, while also producing usability and motivational trade-offs [4]. These literatures support a scoped conclusion: effort can have developmental value under some arrangements, yet the value depends on what the effort makes the learner process, discriminate, retain, or revise.

15.2 Function-Relevant Formative Friction

Friction becomes potentially formative when it keeps a judgment function active in a way that produces diagnostically useful experience. An independent attempt before AI assistance can reveal what options the learner can generate without prompting. Delayed access to a recommendation can require the learner to commit to an initial estimate that later supports calibration. Requiring provenance inspection can make evidential differences visible. Asking for a reasoned comparison between two plausible outputs can exercise criterion articulation rather than mere answer selection. Reopening an earlier decision after outcomes are observed can make temporal and reconstructive judgment available for revision.

The common feature is functional relevance. The effort is connected to a capacity whose later availability matters. The resulting difficulty also needs to expose a difference that the configuration can learn from. Repeatedly struggling with an opaque interface may be effortful without sharpening any task-relevant discrimination. By contrast, difficulty generated by two closely matched cases that require different reliance policies can reveal a boundary in the learner's current judgment.

Definition 15.2 (Formative friction). *Formative friction* is cognitive friction for which the exercised function, the experienced contrast or resistance, and the available feedback or reconstruction pathway jointly create a plausible route to developmental residue relevant to later judgment.

Definition 15.2 makes developmental consequence, rather than subjective difficulty alone, the criterion for treating friction as potentially formative.

The definition deliberately uses a conditional formulation. Formative friction is not guaranteed to produce beneficial formation. A learner can misunderstand the contrast, receive misleading feedback, or retain an unhelpful rule. The definition identifies why a friction-bearing episode is developmentally interpretable: it connects present effort to a later judgment function through some pathway of discrimination, feedback, reconstruction, or transfer.

Several recurrent forms can be distinguished without treating them as an exhaustive taxonomy. *Generative friction* requires production of an initial framing, hypothesis, interpretation, or option before external completion. *Comparative friction* requires discrimination among close alternatives or conflicting reasons. *Verification friction* requires independent checking of evidence, provenance, or computation. *Criterion friction* requires articulation or contestation of the standards used for evaluation. *Temporal friction* requires delay, revisit, or preservation of unresolved alternatives when immediate closure would be easier. *Reconstructive friction* requires a judgment to be explained, traced, or reopened after action or feedback. These forms map onto capacities developed earlier in the paper rather than adding a new set of independent faculties.

Claim 15.3 (Function-relevance condition). Difficulty contributes evidence about judgment cultivation only when the analysis specifies which judgment-relevant function is exercised, what discriminative information the friction makes available, and how that information can remain available to later judgment.

Claim 15.3 prevents a simple equation between effort and formation. It also explains why removing one form of difficulty can improve cultivation. If a procedural burden consumes attention that could otherwise be used for comparison, verification, or criterion articulation, automation can reduce total effort while increasing the share of effort devoted to judgment-relevant functions.

15.3 Load, Accessibility, and Friction Misallocation

A cultivation environment can therefore contain friction whose location is poorly matched to its developmental objective. *Misallocated friction* occurs when resistance is concentrated in a function whose exercise is not the intended developmental target, or when the burden prevents access to the function that is. The distinction is important because the same visible struggle can arise from very different sources.

Cognitive-load research supplies one reason for caution. Sweller's analysis of means–ends problem solving argues that demanding search can occupy processing capacity that might otherwise support schema acquisition [34]. The present framework uses this cognitive-load result only as a boundary condition rather than as a general theory of judgment: some difficulty competes with the processing needed for formation. A judgment-cultivation design should therefore ask whether the burden is attached to the desired discriminative work or merely co-occurs with it.

Accessibility makes this distinction ethically and practically consequential. Friction produced by inaccessible formatting, unnecessary linguistic complexity, poor interface design, motor demands, sensory overload, or avoidable memory requirements can differentially exclude participants while contributing little to the target judgment capacity. Removing such friction can enlarge participation and preserve cognitive resources for the actual object of cultivation. The same principle applies to institutional burden. Requiring a practitioner to navigate duplicated forms, fragmented databases, or repeated low-value documentation does not become educational merely because it demands attention.

At the same time, accessibility should not be reduced to effortless completion. A supportive environment can remove barriers while preserving meaningful participation in the judgment function. A screen reader, translation aid, worked example, calculation tool, or AI-based drafting aid may change how a person accesses an inquiry without eliminating the need to discriminate evidence, compare criteria, calibrate reliance, or own a consequential decision. The developmental question concerns which function remains available for participation after the accommodation or automation is introduced.

Table 13 summarizes this diagnostic distinction.

The table also shows why friction is not intrinsically located in the human component. A verification burden can be distributed across tools, records, peers, and AI systems. What matters is whether

Table 13: Diagnostic distinctions for cognitive friction in judgment cultivation

Friction locus	Potential developmental contribution	Characteristic misallocation risk
Option generation	Makes independent framing and missing alternatives observable	Repetitive generation after the relevant option space is already mastered
Comparison and evaluation	Sharpens discrimination among reasons, criteria, and uncertainty	Artificially multiplying alternatives until attention is dispersed
Verification and provenance	Develops source sensitivity, checking, and escalation judgment	Requiring redundant checks where error cost and uncertainty are already low
Representation and procedure	Can support fluency when the representation itself is part of expertise	Interface, notation, or clerical burden displaces attention from the target judgment
Temporal delay and revisit	Preserves optionality, calibration evidence, and later reconstruction	Delay persists after further information has little expected value
Responsibility and explanation	Supports ownership, reconstructability, and criterion awareness	Ritual justification after a decision whose rationale or authority cannot actually be changed

the configuration retains a meaningful path for detecting, interpreting, and responding to error. Reducing manual work can strengthen that path when automation makes provenance more visible; increasing manual work can weaken it when procedural overload encourages superficial compliance.

15.4 AI-Mediated Redistribution of Friction

AI changes friction less by eliminating it globally than by redistributing it. Generative systems can remove search, drafting, translation, formatting, recall, and calculation burdens while creating new demands for prompt formulation, source checking, model comparison, uncertainty interpretation, hallucination detection, privacy judgment, and responsibility allocation. A configuration can therefore become easier to operate at one interface while becoming harder to evaluate at another.

This redistribution has direct developmental significance. When AI supplies a plausible answer before the learner has framed the problem, generative friction is reduced at the option-construction stage. If the saved effort is redirected toward comparative evaluation, provenance checking, or counterexample generation, the configuration may still cultivate substantial judgment. If the interaction closes immediately after answer acceptance, the same reduction can remove practice from several upstream functions at once. The developmental consequence depends on what replaces the removed friction.

The cognitive-forcing result already introduced above illustrates the inverse intervention: adding engagement can improve one reliance outcome while imposing usability and heterogeneous participation costs [4]. A formation-oriented design therefore needs evidence about the later capacity supported by added effort as well as the behavior observed during one experiment.

AI can also make friction adaptive. Assistance can be withheld, expanded, or redirected according to demonstrated competence, uncertainty, task stakes, and the developmental target. A novice may benefit from a worked example before independent comparison, while an advanced learner may benefit from delayed hints. A low-stakes reversible task may justify immediate automation, while a rare high-consequence task may justify preserving reconstructive practice even when automation is highly accurate. A user who already demonstrates stable provenance judgment may gain little from repeated mandatory source checks on trivial claims, yet still need explicit friction around unfamiliar or consequential evidence.

Claim 15.4 (Friction-redistribution principle). AI mediation should be analyzed as a redistribution of cognitive friction across judgment functions and configuration components. Developmental evaluation therefore concerns which burdens are removed, which new burdens appear, what functions receive the released capacity, and what residues remain available after repeated use.

Claim 15.4 also guards against a nostalgia for unaided difficulty. Manual search, arithmetic, drafting, or memory work can be valuable in some formation contexts, but their historical presence does not by itself establish that they should be preserved. A function warrants preserved practice when its continued availability matters for later judgment, error recovery, understanding, transfer, responsibility, or participation in a broader configuration.

15.5 Configuration-Sensitive Friction Design

The design problem is therefore selective: locate friction where it contributes to the intended formation while reducing burdens that displace or prevent that formation. At least five variables affect that judgment.

First, *prior expertise* changes what counts as productive resistance. Difficulty that reveals a novice's conceptual boundary may become clerical repetition for an expert. Conversely, advanced expertise can create new needs for friction around anomalous evidence, criterion revision, or domain transfer because familiar patterns otherwise encourage premature closure.

Second, *task stakes and reversibility* change the acceptable relation between cultivation and performance. A training environment can preserve an independent attempt even when it slows completion. A consequential operational environment may appropriately prioritize timely access to strong assistance, while using simulation, retrospective reconstruction, or lower-stakes rehearsal to preserve developmental practice elsewhere.

Third, *error visibility and feedback delay* affect whether struggle can generate useful residue. Findings on productive failure depend on instructional sequencing and subsequent opportunities for consolidation rather than on failure as an isolated event [35]. In the present framework, difficulty without interpretable feedback or reconstruction can leave uncertainty without improving later discrimination.

Fourth, *accessibility and participation* constrain legitimate friction. A design that exercises one judgment function by unnecessarily burdening another can narrow who is able to participate. Equivalent developmental objectives may therefore require different interface supports, modalities, pacing, or divisions of labor across learners and contexts.

Fifth, *function allocation* determines whether friction is experienced by the human, the AI, or the coupled configuration. A system can automate retrieval while requiring human source adjudication; automate calculation while preserving model selection; generate alternatives while requiring human criterion construction; or automate routine screening while preserving escalation judgment. The relevant question is whether the allocation retains sufficient practice, visibility, and recoverability for the capacities that the configuration is expected to exercise later.

These variables suggest a design discipline based on targeted rather than generalized difficulty. A friction-bearing intervention should identify its developmental target, remain proportionate to stakes and prior formation, provide a feedback or reconstruction route, allow accessibility-preserving alternatives where possible, and be revisable when evidence shows that the burden has ceased to serve its intended function.

Claim 15.5 (Selective friction design). A judgment-cultivation environment should preserve or introduce friction only with a specified function, developmental rationale, feedback pathway, and exit condition. Friction that lacks these relations should not receive a presumption of educational value merely because it demands effort.

Claim 15.5 makes friction a revisable intervention whose continuation depends on evidence linking the burden to the intended judgment function.

The exit condition is especially important in AI-mediated environments. A training scaffold can become a ritual long after a capacity has stabilized. Mandatory independent generation can cease to reveal anything new; repeated explanation can become performative; persistent verification of low-risk routine outputs can consume resources needed for harder cases. Cultivation therefore includes learning when friction itself should be reduced, relocated, or retired.

15.6 Friction as an Object of Developmental Assessment

Cognitive friction becomes analytically useful only if its developmental consequences can be examined across episodes. Immediate task time, subjective effort, or user preference cannot by themselves establish whether a friction-bearing design is formative. A slower episode may support later transfer; a preferred interface may encourage overreliance; an effortful procedure may teach little; and an accessible automation may free capacity for higher-order comparison. Assessment therefore needs to connect the friction profile of an episode to later changes in judgment.

The developmental-residue framework in Section 3.5 provides the relevant bridge. Evidence can include whether a learner later generates alternatives without prompting, detects a shifted reliability boundary, calibrates confidence more accurately, verifies selectively rather than ritualistically, identifies criterion conflict, reconstructs a prior decision, transfers a judgment function to a changed domain, or recognizes when AI assistance should be escalated, constrained, or bypassed. Such outcomes are temporally extended and function-specific.

This perspective also changes how efficiency should be interpreted. Efficiency is valuable when the configuration achieves its task while preserving the capacities, records, correction pathways, and participation conditions needed for later judgment. The paper therefore has no general preference for friction or fluency. It treats both as design variables whose consequences depend on the relations they reorganize.

Section 16 develops the assessment problem directly. The next question is how judgment development can be evidenced without collapsing formation into answer accuracy, speed, confidence, or one composite score. Cognitive friction supplies one important test case: an adequate assessment architecture should be able to distinguish temporary performance cost from later developmental gain, and genuine formative effort from burdens that merely make inquiry harder.

16 Judgment-Development Assessment

Sections 14 and 15 developed judgment cultivation and cognitive friction as temporally extended properties of recurrent Human–AI inquiry. This section specifies what would count as evidence that judgment has actually developed. Its role is methodological and diagnostic. The objective is to distinguish episode-level performance from developmental evidence, identify assessment dimensions that correspond to the typed developmental residue introduced in Section 3, and clarify how persistence, transfer, reconstruction, calibration, criterion sensitivity, role recoverability, and configuration-level change can be studied without collapsing them into one score. The discussion proceeds from the assessment object to longitudinal evidence, altered-condition transfer, reconstructive evidence, calibration and reliance, criterion-sensitive assessment, configuration-level analysis, and validity conditions. The resulting architecture is intentionally plural: different methods can provide evidence about different developmental relations, and the strength of a formation claim should track the scope of the evidence available.

16.1 Developmental Change as the Assessment Object

The central measurement problem follows directly from Claim 3.5. Immediate success can be informative about a judgment episode while remaining weak evidence about the developmental residue represented by Equation 6. A person can reach a correct answer because an AI system supplied the decisive comparison, because an interface highlighted the relevant source, because a supervisor corrected an error, or because the person independently recognized the relevant distinction. These pathways can yield similar episode-level outcomes while leaving different changes in later capacity.

Assessment of judgment development therefore begins by specifying the developmental object. The object can be improved discrimination among cases, more appropriate reliance, stronger provenance sensitivity, better recognition of missing alternatives, more calibrated confidence, increased ability to articulate or contest criteria, improved timing of closure, stronger reconstructive capacity, or greater recoverability of a function after delegation. The object can also be a harmful or constraining residue, such

Table 14: Evidence layers for assessing judgment development

Evidence layer	Developmental question	Recurrent false positive
Repeated performance	Does a relevant discrimination recur across episodes?	Repetition of the same cues or memorized response
Delayed persistence	Does a change remain available after time and intervening activity?	Short-term retention interpreted as durable formation
Transfer	Does the change remain useful when examples, interfaces, models, or conditions change?	Surface similarity mistaken for broader transfer
Reconstruction	Can the basis, uncertainty, alternatives, and revision path be recovered?	Fluent post-hoc explanation generated after the fact
Calibration and reliance	Do confidence, verification, reliance, and closure respond appropriately to local conditions?	One aggregate trust or accuracy measure masking distinct errors
Criterion sensitivity	Can the configuration identify, apply, compare, and where appropriate contest evaluative criteria?	Agreement with an authorized criterion regime treated as judgment maturity
Role recoverability	Can a delegated judgment function be resumed, checked, or reassigned when conditions change?	High coupled performance masking dependency on one component
Configuration development	Has the distribution of competence, authority, monitoring, and responsibility improved for the task?	Attributing a distributed outcome to an isolated individual or tool

as habitual premature closure, dependency on a particular interface, stable overconfidence, or narrowing of the option space. Assessment is descriptive before it becomes evaluative: it asks what has become more available, less available, more stable, or differently distributed across the configuration.

Definition 16.1 (Developmental assessment). Within the present framework, *developmental assessment* is the collection and interpretation of evidence about changes in a judgment configuration that persist, recur, transfer, become reconstructible, or otherwise remain available to influence later judgment under a specified range of tasks and conditions.

Definition 16.1 preserves the later-availability requirement in Equation 7. It also prevents an assessment instrument from becoming the definition of the capacity it measures. A confidence calibration task, for example, can provide evidence about one monitoring relation without exhausting judgment development. A reliance experiment can reveal discrimination between correct and incorrect AI advice without establishing that the participant can reconstruct why reliance was warranted, recognize a missing option, or transfer the policy to another model. The assessment claim should therefore remain matched to the relation actually observed.

Table 14 summarizes the principal evidence layers developed in this section and recurrent false positives that can arise when each layer is read too broadly.

The rows in Table 14 are complementary rather than cumulative stages. A study can legitimately focus on one layer when its claim is correspondingly narrow. Stronger claims about judgment formation require several layers because different developmental patterns can produce the same observable result in one task.

16.2 Longitudinal Traces and Persistence

The minimum temporal extension beyond an episode is persistence. A developmental residue becomes empirically consequential when some change observed or inferred after one episode remains available during later judgment. The relevant interval depends on the claim. A training study can examine persistence over days or weeks; professional formation can require observation over months or years; in-

stitutional judgment can change across personnel, procedures, and technological generations. No single delay establishes maturity across these domains.

Longitudinal assessment benefits from repeated traces rather than a single before–after comparison. Such traces can include judgments made before AI assistance, judgments made after assistance, confidence reports, source selections, verification actions, criterion statements, deferral decisions, delegation records, later revisions, and outcomes that become visible only after time has passed. The trace need not be exhaustive. Its purpose is to make enough of the developmental trajectory observable that the analyst can distinguish transient facilitation from a change that participates in later episodes.

The learning interfaces reviewed in Section 15 provide a methodological precedent for delayed observation: immediate acquisition performance and later retention or transfer can diverge under specified conditions [33, 35]. Judgment formation remains broader because the residue under study can concern criterion sensitivity, reliance, responsibility, or reconstruction rather than retained task knowledge alone.

Repeated traces also make reversals visible. A learner can initially improve at identifying weak AI advice, then become less discriminating as familiarity with a system produces habitual acceptance. A practitioner can become more cautious after a visible model failure, then overgeneralize that caution to contexts where the system is reliable. A configuration can preserve independent verification during training and abandon it under workload pressure. Developmental assessment should therefore be capable of representing non-monotonic trajectories rather than assuming continuous improvement.

Claim 16.2 (Persistence requirement). Evidence of durable judgment formation requires observation beyond the episode in which assistance, feedback, friction, or correction occurred. The appropriate delay and recurrence pattern depend on the developmental claim, but immediate assisted performance alone cannot establish persistence.

Claim 16.2 does not require every study to become long-term ethnography. It requires precision in inference. An experiment that measures immediate reliance can support an episode-level claim. A study that measures delayed performance can support a persistence claim within its tested interval. Claims about professional or institutional formation require evidence with correspondingly broader temporal scope.

16.3 Transfer under Altered Conditions

Persistence under identical conditions remains compatible with narrow procedural learning. Transfer provides stronger evidence by perturbing the conditions under which a developmental residue is expected to remain available. The relevant perturbation can concern content, surface representation, interface, AI model, error distribution, stakes, time pressure, source quality, or the allocation of judgment functions across people and systems.

The transfer requirement should remain proportional to the claim. A domain-specific judgment skill need not generalize to unrelated domains in order to count as genuine formation. An experienced radiologist’s calibration with diagnostic support, for example, can be meaningful even if it does not transfer to financial forecasting. The stronger question is whether the learned relation survives changes that are relevant to the competence being claimed. If a reliance policy disappears whenever the interface changes color or the recommendation appears in a different location, the residue may be tied to superficial cues. If it survives a change of model while remaining sensitive to model-specific reliability, the evidence supports a more robust form of selective reliance.

Several perturbations are especially informative in AI-mediated inquiry. *Model substitution* tests whether the user has learned a relation to local competence rather than a habit attached to one branded system. *Assistance withdrawal* tests whether a function remains recoverable when external support is absent. *Error-profile shift* tests whether reliance adapts when the distribution or type of model error changes. *Criterion shift* tests whether the configuration can recognize that a previously adequate standard no longer fits the task. *Stake shift* tests whether verification and closure policies respond to consequence and reversibility. *Role reassignment* tests whether a configuration can redistribute generation, critique, verification, or synthesis without losing answerability.

Transfer tests should also preserve diagnosticity. Removing AI entirely can reveal independent capacity, yet such a test may be irrelevant when the target competence is judgment within a coupled Human–AI configuration. Conversely, evaluating only the coupled configuration can hide loss of a human capacity that matters during system failure, contestation, or escalation. Assessment therefore needs to specify which mode of later functioning matters: independent performance, coupled performance, switching between modes, or recoverability under disruption.

Claim 16.3 (Perturbation-sensitive transfer). Transfer evidence becomes informative when the altered condition targets a relation that the developmental claim says should remain available. The strength of a transfer claim depends on the relevance of the perturbation, rather than on maximal distance between training and assessment tasks.

Claim 16.3 makes transfer diagnostic when the perturbation tests a relation that the developmental claim predicts should persist or remain recoverable.

This formulation avoids equating “far transfer” with superior development in every domain. It treats transfer as a test of relational stability under theoretically relevant change.

16.4 Reconstruction and the Fidelity of Judgment Traces

Sections 13 and 14 treated reconstruction as a practice through which judgment can become inspectable and revisable. Assessment can use the same practice as evidence, provided that reconstruction is distinguished from retrospective fluency. A person who can produce a persuasive explanation after seeing the correct answer may still be unable to recover the considerations that actually guided the earlier judgment.

Reconstructive evidence is strongest when contemporaneous traces preserve part of the historical record. Useful traces can include an initial framing, confidence estimate, option set, reasons for accepting or rejecting AI advice, sources consulted, unresolved alternatives, criterion choices, delegation rationale, and conditions under which the judgment was intended to be reopened. Later reconstruction can then be compared with the earlier record rather than evaluated only for present coherence.

This comparison permits several developmental observations. A learner can become better at identifying which evidence actually changed a judgment. A practitioner can become more precise about uncertainty while retaining the same final decision. A team can become better at documenting who held authority for verification. A researcher can recognize that an apparently principled rejection of AI advice was partly driven by an earlier negative experience. Reconstruction can therefore reveal changes in metacognitive access, causal attribution, provenance sensitivity, and responsibility even when outcome accuracy remains stable.

AI introduces a special validity risk because fluent retrospective narratives can be generated cheaply. An assessment that asks the participant to “explain your reasoning” after unrestricted model assistance can measure the quality of a jointly produced rationale rather than the participant’s reconstructive access to the earlier judgment. This does not make AI-assisted reconstruction useless. It means that the assessment object should be specified: unaided reconstruction, assisted reconstruction, comparison between the two, or configuration-level production of an auditable account.

Caution 16.4 (Retrospective coherence). A coherent later explanation should not be treated automatically as evidence that the same reasons organized the original judgment. Developmental assessment should preserve enough temporal trace to distinguish historical reconstruction from plausible post-hoc narration when that distinction matters to the claim.

Caution 16.4 is particularly important when responsibility is attached to reasons. A configuration can produce an excellent ex-post justification while remaining unable to identify who verified the decisive premise, what uncertainty was recognized at the time, or why an alternative was excluded.

16.5 Calibration and Selective Reliance as Partial Indicators

Calibration and selective reliance are among the most measurable components of the framework, but their interpretation should remain function-specific. The metacognitive and appropriate-reliance measures

reviewed in Sections 2 and 6 provide strong evidence about particular monitoring and reliance relations without exhausting judgment development.

A developmental assessment can extend those measures longitudinally. It can ask whether confidence sensitivity improves after repeated AI-mediated inquiry, whether reliance becomes more responsive to task-local reliability, whether verification intensity changes with stakes, whether the user learns to suspend closure when evidence is weak, and whether those patterns persist when the model or interface changes. Such designs make the transition from episode-level calibration to calibration development empirically visible.

Several asymmetries should remain visible. Reduced overreliance can coexist with increased underreliance. Better confidence calibration can coexist with poorer source verification. A user can learn when an AI answer is likely wrong while becoming more dependent on AI for generating alternatives. A configuration can improve overall accuracy while responsibility for checking gradually becomes opaque. A single accuracy or trust measure can therefore conceal offsetting developmental changes.

Selective-reliance assessment should also include action modes beyond adoption and rejection where the domain permits them. Verification, escalation, suspension, independent continuation, and partial use can each be appropriate responses. Assessment can examine whether the configuration selects among these modes in ways that track reliability, uncertainty, consequence, reversibility, and available alternatives. This preserves the richer reliance architecture developed in Section 6.

Claim 16.5 (Partial-indicator principle). Calibration, reliance, accuracy, confidence, verification behavior, and closure timing are informative developmental indicators when linked to the judgment functions they represent. Improvement in one indicator should not be generalized to judgment development as a whole without evidence concerning the other relations implicated by the claim.

Claim 16.5 therefore limits inference to the function represented by an indicator unless additional evidence supports a broader developmental claim.

16.6 Criterion Sensitivity and Multi-Register Assessment

Judgment development also concerns what counts as a reason, which criteria are treated as relevant, and whether the configuration can recognize conflicts among evaluative registers. These capacities are harder to assess than answer accuracy because the criterion regime may itself be part of the object under judgment. Section 11 showed that institutional and technical systems can shape which criteria become visible, legitimate, and actionable. Assessment can therefore reproduce criterion power if it treats conformity to one authorized rubric as the complete measure of judgment development.

Criterion-sensitive assessment can examine several relations separately. First, can the person identify which criteria are being used? Second, can the person apply them consistently where they are appropriate? Third, can the person recognize when a criterion changes in relevance because the task, stakeholders, evidence, or time horizon changes? Fourth, can the person identify tensions among aesthetic, normative, practical, and temporal considerations without immediately compressing them into one score? Fifth, where the setting permits contestation, can the person articulate a reasoned challenge to the criterion regime and identify what alternative consideration is being excluded?

These questions do not imply that every assessment should reward disagreement. Stable standards are necessary in many domains, including safety-critical and legally regulated practice. The developmental object is sensitivity to the scope, grounds, and consequences of criteria. Agreement can be evidence of good judgment when the criteria are appropriate and well understood; disagreement can be evidence of good judgment when the criteria fail to capture a material feature of the case. The assessor therefore needs an explicit account of which relation is being evaluated.

Multi-register assessment also benefits from case comparison. Two cases can share a technically attractive solution while differing normatively because affected parties or distributional consequences differ. Two options can be normatively similar while differing practically because reversibility, authority, or resource constraints differ. Repeated comparison allows assessment of whether the learner detects these relational changes and revises the weighting or mediation of criteria accordingly.

Caution 16.6 (Criterion conformity). Agreement with an assessment rubric can demonstrate competent use of a specified criterion regime, but it does not by itself establish sensitivity to the regime’s scope, exclusions, or revisability. Claims about broader judgment development require evidence that preserves this distinction.

Caution 16.6 preserves the distinction between competent use of a criterion regime and sensitivity to its scope or revisability.

16.7 Configuration-Level Development and Role Recoverability

AI-mediated judgment is frequently distributed. Assessment therefore needs to decide whether its unit is the human participant, the AI component, a team, or the recurrent configuration through which a task is performed. Each unit answers a different question. Human-only assessment can reveal what capacities remain independently available. Configuration-level assessment can reveal whether the coupled system reaches better decisions, preserves provenance, escalates appropriately, and allocates responsibility coherently. Component-level comparison can reveal where the configuration is strong or brittle.

The configuration perspective also makes *role recoverability* assessable. A judgment function can be delegated during ordinary operation yet remain available for human resumption when the system fails, when the task becomes novel, or when responsibility requires independent review. Recoverability can be tested by selectively removing or degrading a component, changing the model, introducing disagreement, or asking another member of the configuration to reconstruct the delegated function. The aim is not maximal redundancy. It is to identify whether the configuration retains enough alternative capacity to remain answerable under relevant disruption.

Assessment can also examine coordination quality. A configuration may contain individually competent components yet perform poorly because uncertainty is not communicated, verification responsibility is ambiguous, or one component’s output becomes an unquestioned input to another. Conversely, a configuration can outperform each isolated component because disagreement is exposed, complementary expertise is available, and escalation paths are clear. These are developmental properties of relations and allocation that cannot be reduced to properties of individuals.

The distinction matters for educational evaluation. A student who performs well with AI may have developed effective collaborative judgment, or the system may be carrying most relevant functions. Removing AI can reveal independent capacity, but only configuration variation can reveal whether the student knows when to rely, how to verify, when to contest, and how to resume a function when support becomes unreliable. A formation-oriented assessment therefore benefits from comparing several allocations rather than treating one mode as the universal benchmark.

Claim 16.7 (Configuration-sensitive assessment). When judgment functions are distributed across human and artificial components, developmental assessment should specify the unit of competence and include evidence about allocation, coordination, monitoring, and recoverability whenever those relations are material to the claim.

Claim 16.7 therefore limits attribution: configuration performance, human-alone performance, and component competence answer different developmental questions.

16.8 Triangulation, Validity, and Developmental Profiles

The preceding dimensions suggest an assessment architecture based on triangulation. A developmental profile can combine repeated performance, delayed probes, altered-condition transfer, contemporaneous traces, later reconstruction, calibration measures, reliance behavior, criterion-sensitive comparison, and configuration perturbation. These observations need not share one scale. Their value lies in revealing convergences and tensions among different parts of the judgment configuration.

A non-scalar profile is especially important when development is mixed. Consider a learner whose accuracy and confidence calibration improve while independent option generation declines; a professional whose selective reliance improves while criterion contestation narrows; or a team whose coupled

performance improves while responsibility becomes harder to locate. Compressing these trajectories into one number can conceal the very trade-offs that formation-oriented assessment is intended to expose. A profile can instead state where change is supported, where evidence is absent, and where developmental directions conflict.

Assessment validity also depends on the relation between the assessment environment and the capacity being claimed. Repeated exposure to the same item types can produce cue learning. Model-generated rationales can inflate apparent reconstructive ability. Hidden access to external tools can blur what unit of competence is being measured. Strict tool prohibition can make an assessment ecologically irrelevant when real practice is necessarily collaborative. Uniform time limits can confound judgment with processing speed. Inaccessible interfaces can transform an assessment of criterion sensitivity into an assessment of sensory, linguistic, or motor burden. These threats require design choices matched to the developmental object rather than one universal testing protocol.

Measurement can itself alter formation. Once a rubric becomes consequential, learners and institutions can reorganize practice around what is scored. An assessment that rewards rapid closure can cultivate rapid closure; one that requires source traces can cultivate documentation practices; one that treats disagreement as error can reduce criterion contestation. Assessment is therefore part of the judgment environment analyzed in Section 4, not an external observation device standing outside it.

Claim 16.8 (Evidence triangulation). Claims about judgment development become stronger when evidence converges across theoretically relevant time intervals, task variations, reconstructive traces, monitoring measures, criterion-sensitive comparisons, and configuration changes. Divergence among these sources is itself developmentally informative and should remain visible rather than being forced into a single composite score.

Claim 16.8 is the methodological counterpart to the paper's non-scalar account of formation. Triangulation narrows uncertainty without eliminating underdetermination about latent capacity, future transfer, or the desirability of the criteria being measured; the traces that strengthen assessment can also create privacy, surveillance, and power concerns. Section 17 therefore examines the boundaries of inference and institutional use.

17 Boundaries of Judgment-Development Evidence

Section 16 developed a plural architecture for assessing judgment formation through persistence, transfer, reconstruction, calibration, criterion sensitivity, role recoverability, and configuration-level change. The availability of such evidence creates a second-order problem. Observability does not by itself establish what may be inferred from a trace, who may collect or interpret it, which consequences may legitimately follow, or whose criteria may govern the interpretation. This section specifies the epistemic and normative boundaries of the framework. Its objective is to distinguish data availability, epistemic warrant, and normative authorization; identify recurrent inference errors in the interpretation of judgment traces; analyze privacy, surveillance, accessibility, and criterion authority; and formulate conditions for institutionally answerable use of judgment-development evidence. The discussion proceeds from limits of inference to boundaries of data use, contestability, accessibility, temporal validity, and institutional governance. The resulting account treats boundary setting as part of judgment practice itself rather than as an external compliance layer.

17.1 Three Boundary Layers of Judgment Evidence

AI-mediated inquiry can generate unusually dense traces. Prompts, model outputs, revisions, source selections, confidence reports, response times, verification behavior, delegation choices, and later corrections can make parts of a judgment process visible at a level that ordinary unaided inquiry rarely records. Such visibility can improve reconstruction and developmental assessment. It can also tempt an assessor to treat what is recorded as equivalent to what is known about the person or configuration.

The present framework requires three distinct questions. First, *observability* asks whether a feature of an inquiry episode can be recorded or reconstructed. Second, *epistemic warrant* asks what claim the available evidence can support and with what scope. Third, *normative authorization* asks whether the evidence may legitimately be collected, retained, combined, interpreted, or used for a specified purpose by a specified actor. These questions can yield different answers for the same trace.

A complete interaction log, for example, can make repeated deference to AI observable. The log can support a limited inference about behavior under the recorded tasks. It may remain insufficient to establish a stable disposition toward deference, independent incapacity, or a general lack of judgment. Even where the behavioral inference is strong, an institution can still lack adequate justification for indefinite retention or employment-related use of the trace. Conversely, an institution can possess authorization to administer an assessment while the resulting data remain too weak to support the broad developmental claim it wishes to make.

Definition 17.1 (Boundary distinction). For judgment-development evidence, *observability* concerns whether a relation leaves an inspectable trace; *epistemic warrant* concerns the scope of inference supported by that trace; and *normative authorization* concerns the legitimacy of collecting, retaining, interpreting, transferring, or acting on the evidence for a specified purpose and institutional relation.

Definition 17.1 prevents technical availability from silently becoming epistemic entitlement or institutional permission. The distinction is particularly important when AI systems make formerly ephemeral acts durable. A hesitation that once disappeared after conversation can become a timestamped record; exploratory prompts can be reinterpreted as commitments; provisional language can be extracted from the context in which it was produced. Greater trace density can therefore increase both analytical opportunity and the consequences of misinterpretation.

17.2 Inference Limits in Capacity Attribution

Developmental assessment necessarily infers beyond directly observed acts. The problem is not that inference is avoidable. The problem is that several latent histories or capacities can produce similar observable traces. A person can verify extensively because the task is high stakes, because confidence is low, because organizational policy requires verification, because the interface is unfamiliar, or because prior experience has taught that the model fails in a particular domain. The same behavior can therefore support different interpretations depending on task conditions, institutional rules, prior history, and the function being assessed.

The reverse problem also occurs. Similar capacities can produce different observable behavior under different constraints. A learner can know how to evaluate a source while failing to do so under severe time pressure. A multilingual researcher can display slower written reconstruction in one language while retaining strong criterion sensitivity. A professional can accept an AI recommendation because an organizational mandate removes practical authority to contest it. In such cases, the recorded act remains real, yet attribution to an internal capacity can exceed the evidence.

This underdetermination is intensified by AI mediation because the system can contribute to both the judgment and the trace through which that judgment is assessed. Generated rationales can improve the apparent coherence of a reconstruction. Interface defaults can shape which sources are inspected. Suggested prompts can change the questions asked. Automated confidence cues can affect reported confidence. An assessor therefore needs to distinguish evidence about the human participant, evidence about the artificial component, and evidence about the coupled configuration.

Claim 17.2 (Inference proportionality). An inference about judgment development should remain proportional to the observed task conditions, temporal range, configuration, and measurement relation that produced the evidence. Broader claims require additional evidence that addresses plausible alternative explanations rather than treating a dense trace as self-interpreting.

Claim 17.2 is compatible with strong conclusions when the design supports them. Repeated observation across altered conditions can substantially narrow alternative explanations. Longitudinal evidence

can support persistence claims. Configuration perturbations can identify where a function is located or recoverable. The boundary concerns the scope of the inference, not a presumption that developmental evidence is inherently inconclusive.

The framework also distinguishes an absence of evidence from evidence of absence. A trace that contains no explicit counter-criterion does not establish that the participant lacked one. A participant may have recognized a concern without recording it, or the interface may have provided no practical place to express it. Conversely, repeated opportunities to articulate a criterion combined with convergent evidence across interviews, behavior, and altered tasks can support a stronger conclusion. Triangulation in Section 16 is therefore partly a method for managing underdetermination.

17.3 Data Governance of Judgment Traces

Formation-oriented assessment creates a tension between developmental visibility and informational restraint. Longitudinal evidence becomes more informative when inquiry traces can be linked across time, tasks, models, and institutional contexts. The same linkage can produce an extensive record of uncertainty, failed attempts, changing beliefs, interpersonal dependencies, and exploratory questions. Those records can reveal more than the developmental relation for which they were initially collected.

The learning-analytics literature provides a direct adjacent warning. Slade and Prinsloo identify informed consent, privacy, deidentification, classification, data management, surveillance, transparency, and power among the ethical issues created by educational data analytics [36]. The present framework extends that concern to AI-mediated inquiry more broadly. Judgment traces are especially sensitive because they can contain unfinished reasoning and developmental vulnerability. Their value for learning can depend on a protected space in which provisional positions, uncertainty, and error can be expressed without every trace becoming a durable evaluation record.

A privacy boundary therefore concerns more than secrecy. It concerns the relation among the purpose of collection, the actors who gain access, the duration of retention, the contexts in which data can be recombined, and the consequences that follow from interpretation. A trace collected for formative feedback can acquire a different meaning when reused for hiring, ranking, disciplinary review, insurance, or eligibility. The underlying bytes can remain unchanged while the institutional relation changes substantially.

Surveillance can also alter the phenomenon being measured. If a learner knows that every exploratory prompt contributes to a permanent competence profile, the learner can reduce experimentation, avoid recording uncertainty, or optimize behavior for the visible metric. The resulting cost reaches beyond privacy because surveillance can change the judgment environment and thereby change the developmental process itself. Measurement pressure can therefore contaminate both the ethical legitimacy and epistemic validity of the evidence.

Caution 17.3 (Trace abundance). The availability of detailed Human–AI interaction traces should not be treated as a general authorization for comprehensive retention or secondary use. Developmental value depends on preserving enough evidence for the stated purpose while limiting uses that change the institutional meaning of inquiry, discourage exploratory judgment, or exceed the warranted inference.

Caution 17.3 supports proportional data practice. A short-lived local trace can be sufficient for immediate reconstruction; a longitudinal study can require persistent identifiers; an institutional safety review can require auditable records. The appropriate design depends on purpose, stakes, affected parties, and the availability of less intrusive evidence. The paper does not propose one universal retention rule.

17.4 Criterion Authority in Consequential Interpretation

Section 11 showed that criteria structure admissibility, classification, commensuration, thresholds, defaults, and closure. Developmental assessment adds a second layer of criterion power: institutions can

define what counts as evidence of judgment development and then use that definition to allocate opportunities. The result can make an assessment regime consequential even where its formal object appears narrow.

Categories embedded in standards and databases can become durable institutional infrastructure [27]. Authorized classifications can also acquire differential institutional force [26]. An assessment criterion therefore becomes a relation of authority when one actor can make its interpretation consequential for another actor's opportunities, reputation, or access to resources.

Criterion authority does not by itself invalidate assessment. Professional practice frequently requires shared standards. Educational institutions need criteria for progression. Safety-critical organizations need thresholds for independent review. The boundary concerns the relation between the consequence imposed and the contestability of the criterion and its application. A person can competently satisfy a criterion while reasonably disputing whether that criterion captures the relevant judgment capacity. Section 16 therefore treated criterion conformity as narrower than judgment maturity.

Contestability has at least four objects. A participant can contest the factual record, the inference drawn from it, the criterion used to evaluate it, or the consequence attached to the evaluation. These forms should remain distinguishable. Correcting a mislabeled trace does not necessarily resolve disagreement about the criterion. Allowing an appeal against a final score does not necessarily reveal how the score was generated. Publishing a rubric does not necessarily create practical power to challenge it.

Claim 17.4 (Interpretive contestability). When judgment-development evidence is used to impose material consequences, the affected person or configuration should, to a degree proportionate to the stakes, have access to mechanisms for correcting the record, understanding the material basis of the inference, contesting criterion application, and seeking review of consequential interpretations.

Claim 17.4 is a normative proposal within the present framework rather than a universal legal rule. Its rationale is relational. Developmental assessment already recognizes that evidence is partial, history-sensitive, and configuration-dependent. Consequential use becomes more answerable when those subject to interpretation can expose missing context, measurement error, criterion mismatch, or changes that occurred after the evidence was collected.

17.5 Accessibility-Sensitive Measurement Conditions

Section 15 distinguished formative friction from burdens that are functionally irrelevant to the developmental target. The same distinction is necessary in assessment. A task can impose linguistic, sensory, motor, attentional, temporal, or interface burdens that materially affect observed performance while remaining only weakly related to the judgment function being assessed. Treating those burdens as neutral background can transform an assessment of judgment into an assessment of access conditions.

Accommodation therefore need not be understood as a relaxation of the relevant standard. It can be part of construct preservation. If the developmental object is provenance sensitivity, providing an accessible interface can make the intended judgment relation more rather than less visible. If the object is rapid emergency triage, time pressure can be constitutive of the task and cannot simply be removed. The proper decision depends on whether a constraint belongs to the target configuration or enters as an incidental barrier.

AI creates additional complexity because assistive use can simultaneously reduce irrelevant friction and perform part of the target function. Translation support, text-to-speech, structured note generation, summarization, retrieval, and drafting assistance can each alter the allocation of judgment functions differently. A rule that classifies all AI assistance as equivalent obscures these differences. Assessment design should specify which functions may be supported, which are under evaluation, and whether the claim concerns independent capacity or capacity within an accessible Human–AI configuration.

This issue also limits cross-person comparison. Equivalent observed speed can reflect very different effort and access conditions; different observed outputs can arise from comparable judgment under different expressive constraints. The framework therefore favors task-relative profiles and within-person

developmental evidence where appropriate, while leaving open the use of shared standards when the construct and consequences justify them.

Claim 17.5 (Accessibility-sensitive validity). An assessment of judgment development is stronger when the conditions of access and expression are examined for their relation to the target capacity. Constraints that materially affect performance should be treated as part of the assessment configuration rather than presumed to be transparent background conditions.

Claim 17.5 also applies to institutional resources. Access to paid models, specialized databases, expert supervision, quiet time, computational infrastructure, and permission to challenge automated recommendations can shape the traces from which developmental conclusions are drawn. These relations can be analytically relevant even when the assessment is not designed as a study of inequality.

17.6 Temporal Validity of Developmental Evidence

Judgment-development evidence is history-bearing. The same property that makes longitudinal assessment valuable also means that old evidence can lose relevance as the person, task, institution, or AI configuration changes. A weak performance observed before training can remain historically informative without remaining a valid estimate of current capacity. A reliance pattern learned with one model generation can change after repeated exposure to a later system. An institutional criterion can persist after the practice it was designed to evaluate has changed.

Evidence therefore has a temporal domain of applicability. That domain need not be expressed through a fixed expiration date. It can be defined by the stability of the relevant relations: similarity of task conditions, continuity of the criterion regime, persistence of the measured function, and absence of material reconfiguration. A major change in any of these can justify re-observation before a consequential inference is carried forward.

This temporal boundary has a normative counterpart. Permanent labels can freeze a developmental state into an identity. Terms such as “over-reliant,” “poorly calibrated,” or “weak independent judgment” can begin as local descriptions of behavior and later function as durable classifications. The historical fact should not be erased when it matters, yet later development should have a practical route to modify the institutional interpretation.

Claim 17.6 (Revisable developmental record). Consequential interpretations of judgment-development evidence should remain revisable when later evidence materially changes the supported inference or when the task, criterion regime, or Human–AI configuration has changed enough to alter the evidential relation.

Claim 17.6 follows the broader reflexive account in Section 10. An assessment can shape future opportunities and thereby participate in the trajectory it later appears merely to describe. Revisability helps prevent a historical trace from becoming an unnecessarily self-stabilizing constraint on future judgment conditions.

17.7 Institutional Boundary Conditions

The preceding analysis can be summarized as a set of boundary tests for institutional use. Table 15 organizes the tests by the relation that needs to remain inspectable. The table is not a compliance checklist. Different domains can require different thresholds, and some high-stakes settings can justify denser records or stronger standardization. Its purpose is to prevent several analytically distinct questions from being compressed into the fact that data exist.

The tests in Table 15 also reveal that epistemic and normative quality can reinforce one another. Purpose limitation can reduce incentives to overinterpret exploratory traces. Contestability can expose missing context and improve validity. Accessibility analysis can distinguish target competence from

Table 15: Boundary tests for the institutional use of judgment-development evidence

Boundary	Principal question	Recurrent failure
Observational	What relation was actually recorded, and under which task and configuration?	Treating trace density as direct access to latent capacity
Inferential	What claim does the evidence support, over what population, interval, and conditions?	Generalizing from local behavior to stable identity or broad competence
Purpose and data use	Why is the trace collected, who can access it, and how long is it retained?	Reusing formative evidence for materially different consequential purposes
Criterion authority	Who defines maturity, normality, acceptable reliance, or adequate reconstruction?	Treating institutional authorization as epistemic self-validation
Contestability	Can the record, inference, criterion application, and consequence be challenged?	Offering formal transparency without practical review or correction
Accessibility	Which burdens belong to the target capacity and which arise from access conditions?	Interpreting incidental barriers as deficits in judgment
Temporal validity	When does earlier evidence cease to support a present consequential inference?	Converting a local developmental state into a persistent identity label
Configuration attribution	Is the claim about a person, an AI component, a team, or a coupled system?	Assigning distributed performance or failure to the wrong unit

measurement burden. Temporal revision can prevent stale evidence from being treated as current. Privacy protections can preserve a space in which uncertainty is expressed more candidly, improving the evidence available for formative purposes.

The relations can also conflict. Stronger auditability can require more persistent traces. Privacy can limit retrospective reconstruction. Standardization can improve comparability while reducing sensitivity to context. Individual contestability can impose institutional cost and delay closure. The framework does not resolve these tensions through a single priority rule. It requires the governing institution to state which values, risks, and responsibilities justify its design and to preserve the possibility of revision when the balance proves inadequate.

This position also limits the ambition of the paper. The framework can identify where boundary questions arise and can specify relations that should remain visible. It does not supply a complete theory of privacy, distributive justice, disability, professional regulation, or legal due process. Those domains contain independent normative and institutional structures that cannot be derived from a theory of judgment formation alone. The appropriate interface is therefore one of disciplined connection rather than absorption.

Claim 17.7 (Boundary-sensitive governance). Institutional use of judgment-development evidence is more answerable when the institution specifies the observational basis, scope of inference, purpose of data use, criterion authority, contestation pathway, access conditions, temporal validity, and unit of attribution relevant to the consequential judgment.

Claim 17.7 completes the movement from formation to institutional use. Greater observability can support revision while also enabling overmeasurement, over-inference, and criterion entrenchment; epistemic and normative restraint therefore become linked design requirements. Section 18 draws together these results.

18 Conclusion

This paper has developed a preliminary Generative Relational account of judgment formation in AI-mediated inquiry. Its central analytical move has been to relocate attention from the visible quality of

a particular answer toward the history-bearing configuration through which judgments become possible, are exercised, and leave consequences for later inquiry. The paper has therefore treated judgment formation as a developmental object rather than a synonym for decision accuracy, confidence, independent reasoning, or resistance to artificial assistance. Its objective has also remained narrower than a general theory of judgment: to provide a revisable conceptual architecture for studying how framing, salience, calibration, reliance, evaluation, temporal interpretation, delegation, responsibility, cultivation, and assessment are reorganized when AI becomes a recurrent participant in inquiry. This concluding section consolidates the analytical structure, identifies its developmental implications, states the principal research agenda, and clarifies the open status of the proposed account.

18.1 Relational Developmental Synthesis of Judgment

The first result concerns the object of analysis. Section 1 separated immediate performance from judgment formation, and Section 3 represented formation as a history-bearing change whose developmental residue remains available for later judgment. This distinction changes what counts as evidence. A correct AI-assisted answer can be educationally thin when the participating person or configuration cannot reconstruct the grounds of acceptance, recognize uncertainty, detect a changed error regime, or transfer the relevant discrimination to a later task. A slower or initially less accurate episode can have greater developmental significance when comparison, feedback, revision, and later availability strengthen relevant capacities. The framework therefore directs empirical inquiry toward persistence, transfer, revisability, and the organization of later judgment rather than treating task-level success as a sufficient developmental endpoint.

The second result concerns the unit of judgment. Sections 4 and 5 showed that AI can alter the environment within which judgment occurs and can participate in different judgment functions without thereby becoming identical to the full judgment configuration. Generation, retrieval, comparison, evaluation, confidence estimation, verification, recommendation, closure, and record keeping can be distributed differently across persons, models, institutions, and other supports. Functional contribution, authority, responsibility, and developmental formation consequently require separate analysis. The relevant unit can be an individual for one question, a coupled Human–AI configuration for another, or a wider institutional arrangement when hidden supports and criterion-setting structures materially shape the episode. Boundary selection is therefore part of the analysis rather than a neutral preliminary choice.

The third result concerns the internal differentiation of judgment. Section 6 treated calibration and selective reliance as multidimensional and task-relative. Section 7 placed problem recognition and salience judgment upstream of later evaluation. Sections 8 and 9 distinguished aesthetic, normative, practical, temporal, and mediating judgment while resisting their reduction to a common scalar. These distinctions matter because an AI system can improve one part of an inquiry while compressing another. Greater option abundance can coexist with weaker discrimination; stronger practical convenience can coexist with poorly examined normative criteria; fluent temporal narratives can coexist with weak evidence for trajectory claims. The quality of AI-mediated inquiry therefore depends partly on preserving enough differentiation for these registers and functions to remain inspectable and revisable.

The fourth result concerns reflexivity and power. Section 10 showed that present judgments can alter the evidence environments, conceptual resources, dependencies, institutional rules, and capacities through which later judgments will be made. Section 11 extended this reflexive analysis to criterion regimes, emphasizing that classification, commensuration, thresholds, defaults, admissibility rules, and closure conditions can shape what counts as a good judgment before a particular person begins evaluating options. Recurrent AI mediation can therefore influence judgment through more than recommendation quality. It can participate in the generation and stabilization of the criteria, defaults, and representations through which later inquiry becomes intelligible.

The fifth result concerns delegation and ownership. Section 12 separated delegation of a judgment function from transfer of authority or responsibility. The capacity to approve a final output is a weak basis for judgment ownership when the approving actor cannot reconstruct relevant grounds, identify govern-

ing criteria, contest the recommendation, or revise the process under changed conditions. Answerable delegation instead requires a relation among functional allocation, inspectability, contestability, revision capacity, and responsibility. This relation is developmental as well as institutional because repeated delegation changes which functions remain practiced and which capacities remain recoverable.

Taken together, these results support a relational developmental interpretation of AI-mediated judgment. The same artificial capability can have different consequences because task structure, stakes, reversibility, expertise, criterion regimes, feedback, role allocation, and accumulated history determine how judgment functions are exercised, delegated, contested, learned, and retained.

18.2 Developmental Implications for AI-Mediated Inquiry

Sections 13 through 16 translated the analytical architecture into a developmental problem. The resulting implication is neither a general prescription for more independent reasoning nor a general prescription for deeper AI integration. Judgment practice is configuration-sensitive. Productive arrangements preserve or create opportunities to exercise the functions that matter for later judgment while using artificial assistance where it improves inquiry without making those functions opaque, irrecoverable, or institutionally unanswerable.

The practice architecture in Section 13 therefore emphasized preparatory structuring, comparative generation, provenance-sensitive verification, criterion review, controlled closure or deferral, and post-judgment reconstruction. These practices are analytically useful because they make relations among generation, discrimination, monitoring, commitment, and learning more visible. Their value remains task-relative. A low-stakes reversible inquiry can warrant extensive automation and light reconstruction, while a high-stakes or poorly reversible judgment can warrant independent checks, stronger provenance requirements, explicit criterion review, and preservation of a trace that makes later challenge possible.

Section 14 extended this analysis from recurrent practice to cultivation. The central developmental mechanism is structured reconfiguration across histories of experience rather than mere exposure to more tasks or more AI outputs. Variation across cases, guided comparison, articulated reasons, function-sensitive feedback, changing Human–AI role allocation, and later reconstruction can contribute different kinds of developmental residue. Cultivation remains open-ended because the target environment is itself changing. A reliance policy that is well calibrated to one model generation can become poorly calibrated when system capabilities, interfaces, institutional incentives, or failure patterns change. Mature judgment in this setting therefore includes the capacity to revise its own reliance relations and evaluative routines.

Section 15 further qualified the developmental picture by treating cognitive friction as function-relative. Difficulty has no intrinsic educational virtue. Friction becomes potentially formative when it is attached to a judgment function whose exercise contributes to later capacity, when feedback or reconstruction can convert effort into a usable developmental residue, and when the burden is proportionate to the learner and task. AI mediation can remove obsolete procedural burdens while preserving or relocating more diagnostically useful effort toward comparison, verification, criterion articulation, uncertainty monitoring, or reconstruction. The relevant design problem is therefore selective redistribution of friction rather than preservation of hardship for its own sake.

Assessment must follow the same developmental logic. Section 16 rejected a single judgment score in favor of triangulated evidence concerning persistence, transfer, reconstruction, calibration, criterion sensitivity, role recoverability, and configuration-level change. This approach is especially important in AI-rich environments because immediate performance can become increasingly decoupled from the capacities that remain available when assistance changes, fails, or disappears. Perturbations such as model substitution, altered error profiles, role reassignment, changed stakes, and assistance withdrawal can therefore provide evidence that ordinary performance tests miss. Divergence among indicators is itself informative: improved accuracy alongside declining recoverability describes a different developmental profile from improved accuracy accompanied by stronger calibration and transfer.

The developmental implications are consequently institutional as well as individual. Educational programmes, research groups, professions, and organizations shape which functions participants practice,

which errors receive feedback, what traces remain available for reconstruction, which criteria become authoritative, and whether questioning an AI-mediated judgment carries social or procedural cost. Judgment cultivation therefore depends partly on the design of participation, assessment, review, and responsibility structures. The framework does not determine one optimal institutional arrangement. It provides a vocabulary for asking what an arrangement is forming over time, which capacities it is preserving or weakening, and whether its apparent efficiencies are accompanied by losses in later discrimination, contestability, or recoverability.

18.3 Research Agenda for Situated Judgment Formation

The present account generates an empirical and interpretive research programme rather than a closed explanatory model. Eight priorities follow from the architecture developed here.

First, developmental residue requires longitudinal designs that compare immediate performance with later reconstruction, calibration, source discrimination, option generation, criterion articulation, and revision under delayed or altered conditions. Second, comparative Human–AI studies should vary the organization of assistance itself: when AI enters, which functions it performs, whether independent commitments precede advice, how provenance is displayed, how verification is allocated, and who can revise criteria. These two directions can test whether similar final performance conceals different developmental residues or hidden dependence.

Third, research on salience and problem formation should examine whether recurrent AI use changes which questions, anomalies, literatures, hypotheses, and counterexamples become available for inquiry. Fourth, studies of differentiated and mediating judgment should test when AI helps articulate aesthetic, normative, practical, and temporal registers and when interface design compresses them into a single recommendation or score. The relevant outcomes include later reconstruction, contestability, and the ability to mediate partially incommensurable considerations under real constraints.

Fifth, reflexive effects and criterion power require longitudinal organizational and historical study because recurrent AI infrastructures can change reporting categories, evidential standards, workflow expectations, professional roles, thresholds, defaults, and review pathways. Sixth, delegation research should examine responsibility allocation, documentation, substantive rather than merely formal oversight, and recoverability of delegated functions before as well as after failure.

Seventh, cultivation research can compare sequences of Human–AI participation, including independent attempt before assistance, AI-supported exploration followed by reconstruction, model comparison, staged support withdrawal, and deliberate role reassignment. Outcomes should include transfer, calibration, problem formation, criterion sensitivity, and accessibility-sensitive participation. Eighth, governance research should examine data minimization, retention, interpretive contestability, temporal expiry, accessibility-sensitive validity, and the institutional consequences of inferred developmental profiles, preserving the distinction among observability, epistemic warrant, and normative authorization developed in Section 17.

These priorities require methodological pluralism. Controlled experiments can isolate reliance and calibration mechanisms; longitudinal studies can test persistence and transfer; qualitative reconstruction can examine uncertainty, responsibility, and criteria; historical analysis can trace institutional changes; and computational models can formalize selected update processes under explicit assumptions. Their claims should remain commensurate with what each method can support.

18.4 Closing Orientation for Judgment in AI-Rich Inquiry

The accelerating capability of artificial systems can make judgment appear less important precisely because high-quality outputs become easier to obtain. The argument of this paper points in a different analytical direction. As generation, synthesis, recommendation, and evaluation become increasingly available through artificial assistance, the organization of judgment shifts toward questions of salience, discrimination, calibration, criterion authority, selective reliance, temporal interpretation, responsibility, reconstruction, and learning across changing configurations. Some of these functions can themselves be

supported or partially realized by AI. Their distribution therefore remains an empirical and institutional question rather than a fixed boundary between human and machine.

What remains central is the developmental relation. Recurrent arrangements of inquiry can leave a residue: it can strengthen a capacity, redistribute it, make it dependent on a particular support, render a criterion less visible, preserve a reconstruction pathway, or narrow the set of questions that later become available. The relevant evaluation of AI-mediated inquiry consequently extends beyond the present answer toward the conditions it generates for the next judgment. This temporal extension also keeps the framework open to revision. The capabilities of AI systems, the institutions that deploy them, and the practices through which people learn with them will continue to change; judgments about appropriate reliance and cultivation must therefore remain revisable as well.

The proposed Generative Relational account should be read in that spirit. It offers a set of distinctions and representations for investigating judgment formation without treating them as an ultimate ontology, a universal educational prescription, or a complete theory of responsible AI use. Its principal contribution is to make several relations jointly visible: performance and formation; individual and configuration; generation and discrimination; reliance and calibration; present judgment and future judgment conditions; criterion power and contestability; delegation and ownership; practice and developmental residue; assessment and its epistemic boundaries. These relations provide a basis for more precise empirical, educational, professional, and institutional inquiry.

Judgment in an AI-rich environment is therefore best approached as a continuing history of formation. The practical question is not whether inquiry should remain purely human or become maximally automated. It is how a changing configuration can use artificial capability while preserving or developing the capacities needed to notice relevant problems, discriminate among generated possibilities, understand the criteria that organize evaluation, calibrate reliance, remain answerable for consequential commitments, reconstruct prior reasoning, revise under changed conditions, and continue learning from what follows. The present paper leaves that question deliberately open to domain-specific investigation. Its aim has been to provide an architecture through which the question can be posed with greater analytical precision and developmental responsibility.

References

- [1] Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58, 2024. doi:10.1038/s41586-024-07146-0.
- [2] Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175:108779, 2026. doi:10.1016/j.chb.2025.108779.
- [3] Leily Soleimanof and Derrick Neufeld. AI advice and human metacognition. *Decision Support Systems*, 210:114757, 2026. doi:10.1016/j.dss.2026.114757.
- [4] Zana Buçinca, 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. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):Article 188, 1–21, 2021. doi:10.1145/3449287.
- [5] John H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906–911, 1979. doi:10.1037/0003-066X.34.10.906.
- [6] Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014. doi:10.3389/fnhum.2014.00443.

- [7] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997. doi:10.1518/001872097778543886.
- [8] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi:10.1518/hfes.46.1.50_30392.
- [9] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi:10.1037/xge0000033.
- [10] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019. doi:10.1016/j.obhdp.2018.12.005.
- [11] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023. doi:10.1145/3581641.3584066.
- [12] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. Towards effective Human–AI decision-making: The role of human learning in appropriate reliance on AI advice. In *Proceedings of the 44th International Conference on Information Systems (ICIS 2023)*, 2023.
- [13] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, 2000. doi:10.1109/3468.844354.
- [14] Mihaly Csikszentmihalyi and Jacob W. Getzels. Discovery-oriented behavior and the originality of creative products: A study with artists. *Journal of Personality and Social Psychology*, 19(1):47–52, 1971. doi:10.1037/h0031106.
- [15] Jean E. Pretz, Adam J. Naples, and Robert J. Sternberg. Recognizing, defining, and representing problems. In Janet E. Davidson and Robert J. Sternberg, editors, *The Psychology of Problem Solving*, pages 3–30. Cambridge University Press, 2003. doi:10.1017/CBO9780511615771.002.
- [16] James W. McAllister. *Beauty and Revolution in Science*. Cornell University Press, 1996.
- [17] Philip Kitcher. *Science, Truth, and Democracy*. Oxford University Press, 2001. doi:10.1093/0195145836.001.0001.
- [18] Heather E. Douglas. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.
- [19] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York, 1983.
- [20] Kathleen Mosier, Ute Fischer, Robert R. Hoffman, and Gary Klein. Expert professional judgments and “Naturalistic Decision Making”. In K. Anders Ericsson, Robert R. Hoffman, Aaron Kozbelt, and A. Mark Williams, editors, *The Cambridge Handbook of Expertise and Expert Performance*, second edition, pages 453–475. Cambridge University Press, 2018. doi:10.1017/9781316480748.025.
- [21] Gary Klein. *Sources of Power: How People Make Decisions*. 20th Anniversary Edition. The MIT Press, 2017.

- [22] *Zhouyi (Book of Changes), Xici II (Appended Statements II)*. Chinese Text Project. Classical Chinese text; accessed 2026-10-03.
- [23] *Zhouyi (Book of Changes), Kun Hexagram and Xiang Commentary*. Chinese Text Project. Classical Chinese text; accessed 2026-10-03.
- [24] *Zhouyi (Book of Changes), Bi Hexagram, Tuan Commentary*. Chinese Text Project. Classical Chinese text; accessed 2026-10-03.
- [25] Tze-Ki Hon. Chinese philosophy of change (Yijing). In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Summer 2024 edition. Metaphysics Research Lab, Stanford University, 2024.
- [26] Pierre Bourdieu. *Language and Symbolic Power*. Edited by John B. Thompson; translated by Gino Raymond and Matthew Adamson. Harvard University Press, Cambridge, MA, 1991.
- [27] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, 1999. doi:10.7551/mitpress/6352.001.0001.
- [28] Wendy Nelson Espeland and Mitchell L. Stevens. Commensuration as a social process. *Annual Review of Sociology*, 24:313–343, 1998. doi:10.1146/annurev.soc.24.1.313.
- [29] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet A. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 59–68. Association for Computing Machinery, 2019. doi:10.1145/3287560.3287598.
- [30] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004. doi:10.1007/s10676-004-3422-1.
- [31] Madeleine Clare Elish. Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019. doi:10.17351/ests2019.260.
- [32] Giulio Mecacci and Filippo Santoni de Sio. Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, 22:103–115, 2020. doi:10.1007/s10676-019-09519-w.
- [33] Elizabeth L. Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Morton A. Gernsbacher, Richard W. Pew, Leaetta M. Hough, and James R. Pomerantz, editors, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth Publishers, New York, 2011.
- [34] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. doi:10.1207/s15516709cog1202_4.
- [35] Manu Kapur. Productive failure. *Cognition and Instruction*, 26(3):379–424, 2008. doi:10.1080/07370000802212669.
- [36] Sharon Slade and Paul Prinsloo. Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10):1510–1529, 2013. doi:10.1177/0002764213479366.