SR7D Framework

Decision Governance in AI-Abundant Environments

v1.3.0 · 2026-05-28

Lars Bracker

Contents

1 Decision Governance in AI-Abundant Environments

1.1 A Framework for Preserving Human Judgment

SR7D Framework — Full Specification

Draft 1.3.0 — 2026-05-28 (industrial-grade pass + Pass-6 imbad-skill validation + composite-self-skepticism integration; V1.3.0 = V1.2.1 content + Steerable-branded LaTeX template) Status: Final Draft for Review — V1.2.1 incorporates Pass-5 (lcrawfurd 5-framework) + Pass-6 (imbad 7-agent peer review v1.9.1, real Anthropic-marketplace skill) + composite-self-skepticism integration (§3.7c) + RQ8/RQ9 + Schema field_role discipline. V1.2 baseline incorporates 4-pass external review (LTF-R11 4-model triangulation + ce-doc-review parallel persona agents + Doppelkritik V2 + Paper-Review-Skills simulation), Sonar-verified BaFin/MiFID/AI-Act norm corrections, and a structural response to the introspection-illusion counter (§3.7a). See companion file `WHITEPAPER_V1.2_REVISION_TRAIL.md` *for the full audit log. V1.1 evidence: SIC-026 (stat-prior confidence), SIC-028 (Bayesian posterior layer), Appendix F field-validated practitioner objections.*

1.2 Abstract

In AI-abundant environments, the limiting factor shifts from analysis quality to governance quality. This paper proposes and instantiates Decision Governance as an architectural discipline — formalized through the SR7D framework (seven core patterns, three ethical guardrails) — for making consequential decisions visible, traceable, contestable, and improvable without removing the human from the loop. We ground the framework in decision science (Kahneman), forecasting research (Tetlock), information theory (Friston), and abductive logic (Peirce), and document a reference implementation in which the information-theoretic bridge is instantiated as a Bayesian posterior layer (Section 3.7, Appendix E.1) and Judgment Quality Metrics (JQM) is instantiated through a stat-prior confidence subcomponent (Section 4.4a), and each of the seven SR7D patterns carries an explicit reference-implementation status annotation (Chapter 2). Financial advisory under the EU AI Act, narrowed to the HNW-CFP segment for first validation (Section 6.0), serves as primary application domain. Practitioner objections collected during the framework's iterative development are catalogued with architectural counter-evidence in Appendix F.

1.3 Intended Audience and Reading Outcome

Added 2026-05-24 — primary-persona acceptance function for industrial-grade quality calibration. This section governs all subsequent revision passes: edits are accepted into the paper only when they

preserve or improve the primary reader's ability to perform the three acceptance actions below.

1.3.1 Primary Reader

The primary reader of this paper is a CFP-credentialed advisory practitioner serving high-net-worth households in the DACH region under MiFID II and BaFin MaComp regulation. Secondary readers — regulators, academic peer reviewers, technology evaluators inside advisory firms, investors — are explicitly out of scope for the primary acceptance function below. Their distinct needs are addressed in the appendices, in Chapter 6's regulatory mapping, and in separate companion documents.

1.3.2 Acceptance Function — "Industrial Grade" Operationalized

The paper is considered industrial-grade when, after thirty minutes of unassisted reading by the primary reader described above, that reader can perform three concrete actions:

1. **Identify** which of the seven SR7D patterns (Chapter 2) addresses the most pressing governance pain in their current advisory practice, naming a specific recent client situation as the test case.
2. **Articulate** in one sentence why a Decision Packet (Appendix C) would either have changed the outcome of that situation or would have preserved the reasoning that is today lost in CRM notes and post-meeting summaries.
3. **Articulate** in their own words which of the two operationally instantiated components — SIC-026 stat-prior confidence (Section 4.4a) or SIC-028 Bayesian posterior layer (Section 3.7 and Appendix E.1) — would be the more plausible integration candidate with their existing AI-tooling stack, *independent of whether any follow-up conversation is subsequently scheduled*. (Earlier drafts of this acceptance function graded the paper on the reader's decision to schedule a discovery call. That formulation was withdrawn in V1.2.1 because grading on a conversion action would have been a self-recursive violation of Ethics 1 (Non-Normativity). The §6.7 Discovery-Call Decision Page remains available as a practical exit-ramp for readers who choose to use it, but is no longer part of the acceptance function the paper grades itself against.)

A reader who cannot perform any of these three actions after thirty minutes of reading establishes a failure of the industrial-grade test for that reader. Patterns of such failures across readers are the input to the next revision pass.

1.3.3 Out-of-Scope Acceptance Tests

The following are explicitly *not* the industrial-grade test for this paper, even though the paper may incidentally support them:

Out-of-scope test	Why excluded	Where partially addressed
Peer-reviewed publication acceptance (academic journal)	Different reader, different conventions (IMRAD, formal proofs, methods replication)	Chapter 3 formal grounding; Appendix A pre-registration draft
Regulatory consultation reference (BaFin, ESMA, EU AI Office)	Different reader, different language register (normative-prescriptive, paragraph-citation density)	Chapter 6.1 regulatory mapping; Appendix C regulatory-correspondence table
Investor due-diligence material	Different reader, different selection function (commercial traction, market size, defensibility)	Separate investor materials, not in this paper
Internal team alignment document	Different reader (Steerable team), different scope (operational sequencing, prioritization)	Companion roadmap and persona-sequencing documents in the operations layer

Including these as primary acceptance tests would force trade-offs between mutually incompatible optimizations (e.g., academic-rigor density versus practitioner-readability concision). The CFP-practitioner test is the *primary* filter; the others are *secondary* sanity checks that may inform appendix-level revisions but do not gate main-chapter edits.

1.3.4 Implications for the V1.2 Sharpening Pass and Beyond

The V1.2 sharpening pass — planned with external-reviewer tooling (LLM-Triangulation across four models, parallel-persona document review, second-pass adversarial critique, and scientific-paper-review tooling) — applies all reviewer recommendations through the acceptance function above as a filter:

- Recommendations that improve all three primary-reader actions → accepted into the main chapters.
- Recommendations that improve some primary-reader actions and degrade others → require explicit trade-off documentation in this section, then a judgment call by the editor.
- Recommendations that improve secondary-test outcomes (academic, regulatory, investor) at the cost of primary-reader actions → deferred to the appropriate companion document, not incorporated into main chapters.
- Recommendations that improve neither primary nor secondary tests → rejected, with rejection rationale recorded in the revision trail.

This filter is itself an instance of the framework the paper describes: Non-Normativity (the editor is

not told what the right text is), Constructed Ambiguity (recommendations that genuinely conflict are surfaced as documented trade-offs rather than silently resolved), and Archaeological Governance (the revision trail preserves the decision logic for any future revision-pass reviewer).

1.4 Changelog

1.4.1 V1.2.1 — 2026-05-24 — Pass-6 Skill-Validation + Composite-Self-Skepticism

Incorporates findings from two additional review passes performed after V1.2 commit: - **Pass 5 (lcrawfurd/paper-review):** 5-framework review (Edmans, Nyhan, Humphreys, Blattman, Evans/Bellemare) — 6 NEW findings, Edmans-verdict Minor Revision. - **Pass 6 (imbad-academic-paper-reviewer v1.9.1, real Anthropic-marketplace skill):** 7-agent multi-perspective peer review with 0-100 rubrics. 14 NEW findings, 1 CRITICAL (composite self-skepticism). Aggregate 71/100. Major Revision under IRON RULE 4 (DA CRITICAL prevents Accept until composite addressed).

Substantive V1.2.1 changes:

- **§3.7c NEW Composite Self-Skepticism section (~600 words, addresses Pass-6 CRITICAL N-5-11):** Composes Confabulation Counter + Audit-Theater Risk + Methodology Self-Audit-Violation into a single composite-risk statement; specifies three empirical failure-conditions (confabulation dominance $\kappa<0.2$, ritualization signature in JQM variance, methodology-self-audit non-remediation) under which the framework should be considered to have failed at its own standard.
- **Chapter 7 (RQ8 + RQ9 added):** Open Research Question 8 (Decision-Packet reasoning-field audit, addressing §3.7a Response 1 elevation criterion); Open Research Question 9 (methodology-self-audit remediation pathway, addressing §3.7b Pattern-1+Pattern-6 violations).
- **§1.4b (Pass-6 N-5-5):** Constitutional AI literature reference added (Bai et al. 2022).
- **Pattern 7 (Pass-6 N-5-12):** Klein 1998 RPD cherry-picking caveat added with back-reference to §3.7a confabulation counter.
- **Acceptance Function Action 3 (lcrawfurd Finding 1):** Reformulated from "Decide whether to schedule discovery call" to "Articulate which component would be the more plausible integration candidate, independent of any follow-up conversation." Removes Non-Normativity-Self-Reference violation that graded paper on reader conversion-action.
- **Appendix C Schema (lcrawfurd Finding 3):** `field_role` enum added to Decision Packet schema (structural vs annotation). 3 annotation fields explicitly tagged (decision_reasoning, check_notes, override_notes) with `audit_caveat` referencing §3.7a Response 1 and RQ8. New §C.5 Field-Role Discipline section documents validator behavior for the two field-role classes; backwards-compatible schema migration v0.1 → v0.1.1.
- **Appendix C Lines 2093+2098 (Pass-6 HIGH N-5-2):** "Art.14.1d (logging)" → "Art. 12 (record-keeping)" — V1.2 Batch-A Norm-Korrektur did not propagate to Appendix C originally; now corrected to match §6.1 Sonar-verified text.
- **Chapter 5 §5.3 + Appendix B.4 headers (Pass-6 N-5-1):** "Merkle Chain" → "Tamper-Evident Hash Chain" in 2 remaining headers (V1.2 global-replace missed these).

- **Abstract (Pass-6 N-5-4):** "operationalized" → "instantiated (validation pending)" for consistency with post-Batch-B status language.

V1.2.1 is a Minor Revision under Edmans framework; remaining Pass-6 findings deferred to V1.3: - N-5-6 ESMA Guidelines reference in §6.1 narrative - N-5-7 Sonar-verify FinVermV §§12-22 references - N-5-8 Cross-domain (§6.6) domain-literature anchors - N-5-13 JQM weight specification in reference deployment - N-5-14 Quality-gate-check blocks at chapter ends (cosmetic) - Various MEDIUM/LOW items from prior passes (11 total)

No SR7D pattern definition has been altered between V1.2 and V1.2.1. V1.2.1 is a backwards-compatible patch release; all V1.2-shaped Decision Packets remain valid against schema v0.1.1.

1.4.2 V1.2 — 2026-05-24 — Industrial-Grade Pass

Incorporates findings from 4-pass external review (LTF-R11 with Gemini-2.5-Flash/GPT-4o-mini/Mistral-Large/Perplexity-Sonar-Pro; ce-doc-review with 4 custom personas; Doppelkritik V2; Paper-Review-Skills simulation). Sonar-verified BaFin/MiFID/AI-Act norm corrections. Substantive changes:

- **Chapter 1 (§1.4b):** Unverified citations to Brimage 2026 + Agents of Chaos 2026 removed; Kozyrkov 2024 reframed as industry-blog reference.
- **Chapter 2 (§2.0a, §2.11):** New Pain-to-Pattern Map as reading entry-point; §2.11 Reflexive Governance by Design cut (internal-state leak).
- **Chapter 2 Patterns:** Reference-implementation status annotations changed from "operational" to "instantiated (validation pending)" — alignment with Appendix E.1.3.
- **Chapter 3 (§3.7, new §3.7a):** Practitioner-summary prepended to §3.7; new §3.7a Engaging the Confabulation Counter (~1300 words) — three architectural responses (constrain capture, couple-to-outcome, preserve contestability) + Open Research Question 8.
- **Chapter 4 (§4.2.3, §4.3, §4.4a):** Bradley-Terry transitivity caveat (Tversky 1969); Arrow's Impossibility caveat (Perspective Diversity as deliberate IIA-relaxation); JQM "Brier score adaptation" terminology dropped (category error); §4.4a practitioner-summary prepended.
- **Chapter 5 (throughout):** "Merkle chain" replaced with "tamper-evident hash chain" (8 occurrences) — Git analogy preserved.
- **Chapter 6 (§6.0, §6.1, §6.5, new §6.6a, new §6.7):** EU AI Act §6.1 REWRITTEN (Sonar-verified Annex III + Art. 12 logging); §6.5 Legal Forensics 10 row-level Norm-corrections (Art. 23 not 27, MaComp BT 5/BT 7 corrected); §34h GewO out-of-scope caveat; Power 1997 + Hadfield-Menell 2017 noted; new §6.6a Companion-Document Map; new §6.7 Discovery-Call Decision Page.
- **Chapter 7:** Open Research Question 8 added (Decision-Packet reasoning-field audit).
- **Appendices relabeled:** C/D/E/E-Annex/F/G → B/C/D/D-Annex/E/F (close missing-B gap).
- **Appendix B (was C) Validator Spec:** German forbidden-lexicon added.
- **Appendix E.1 (was F.1):** PMI base-choice rationale corrected.
- **Appendix F (was G) Objection 4:** Reframed as category-error + hedged ROI framing.
- **Companion file:** V1.2 Revision Trail moved out of front matter.

Deferred to V1.3: I1 Chapter 3 academic-depth subtraction, I3 Appendix A migration, I4 Appendix D (was E) migration, 11 MEDIUM items.

No SR7D pattern definition has been altered between V1.1 and V1.2.

1.4.3 V1.1 — 2026-05-24 — Sharpening Pass

Incorporates reference-implementation evidence accumulated since V1.0 (2026-03-01). Substantive additions:

- **Chapter 2:** Each SR7D pattern (1-7) now carries a reference-implementation status annotation indicating whether the pattern is enforced, operational, or prototype-stage in the reference architecture.
- **Chapter 3:** New Section 3.7 documents the Bayesian posterior layer that operationalizes the information-theoretic bridge of Sections 3.2-3.5. Implementation detail is in Appendix E.1.
- **Chapter 4:** New Section 4.4a documents the stat-prior confidence operational subcomponent of Assumption Coverage. New Section 4.1a is a reflexive note on the four-role review protocol used in the whitepaper's authoring (detail in Appendix E.2).
- **Chapter 5:** New Section 5.3a documents production hardening through frozen, signed governance-constraint sets.
- **Chapter 6:** New Section 6.0 Segment-Scope paragraph specifies HNW-CFP as the first-validation segment. New Section 6.2a references the practitioner-objection catalogue in Appendix F. RQ4-Extension added to Chapter 7 for objection-validation research.
- **Chapter 7:** RQ4 extended with practitioner-objection generalization sub-question.
- **Appendices F and G:** New. Appendix E documents reference-implementation detail; Appendix F catalogues field-collected practitioner objections with methodology disclosure.
- **Appendix D:** New Annex (D-Annex) documents reference compliance tooling as proposed extension to the 10-point manual checklist.

No SR7D pattern definition, no JQM dimension definition, and no Decision Packet schema field has been modified from V1.0; the V1.1 additions are documentation of operational instantiation, not framework redefinition.

1.5 V1.2 Revision Trail (Companion File)

The full audit log of external-reviewer findings (Pass 1-4) and the synthesis that produced V1.2 from V1.1 has been moved to a companion file, `WHITEPAPER_V1.2_REVISION_TRAIL.md`, to keep this paper's front matter focused on reader-facing content. The companion file preserves the verbatim trail and the V1.2 Apply-Summary. Cross-repo master plan in `Greenfield/docs/plans/wp01-sharpening/`; Sonar verification in `Greenfield/docs/plans/wp01-sharpening/sonar_verification_results.md`.

2 Chapter 1 — Noise, New Noise, and the Governance Gap

2.1 1.0 — The Problem That Abundance Creates

For most of the history of professional advisory work — financial planning, medical diagnosis, legal counsel, strategic consulting — the primary constraint was analytical capacity. A financial advisor in 1985 needed weeks to model multiple portfolio scenarios, stress-test assumptions, and compare tax-optimization strategies across different income trajectories. A physician evaluating a complex case could access perhaps a handful of relevant studies from memory or from a library that required physical presence. Analysis was expensive, slow, and bottlenecked by human cognitive and institutional capacity.

That constraint is structurally gone. A financial advisor in 2026 can generate a Monte Carlo simulation across 10,000 portfolio scenarios, a tax-loss harvesting analysis, an insurance gap assessment, and a retirement income projection in under 90 seconds. A physician can query a system trained on millions of clinical records and surface the five most statistically relevant diagnoses in moments. Analysis has become abundant.

Abundance, however, does not automatically produce better decisions. It produces more inputs. And more inputs, without governance, produce more noise.

This chapter argues that AI has solved one class of noise problem — analytical consistency — while simultaneously creating a new class: noise at the human-AI decision boundary. The gap between "AI generates output" and "human commits to decision" is structurally ungoverned. This governance gap is not a temporary oversight waiting to be corrected by better AI. It is a structural property of the transition between machine output and human judgment, and it requires infrastructure to manage.

2.2 1.1 — What Noise Is (and What It Isn't)

The most important conceptual clarification for this chapter comes from Kahneman, Sibony, and Sunstein's 2021 book *Noise: A Flaw in Human Judgment*. Noise is not bias. The distinction matters because the two problems have fundamentally different remedies.

Bias is systematic error — a consistent deviation in one direction. If a medical screening test consistently overestimates disease severity, it is biased. Bias can, in principle, be corrected once identified: adjust the calibration, control for the known confound, subtract the systematic offset.

Noise is random error — unwanted variance in judgments that should, by any reasonable standard,

be identical. If ten experienced radiologists evaluate the same scan and produce ten different severity assessments, the variance between them is noise. No single radiologist is consistently wrong in a directional sense — each is highly trained, each has good intentions — but the outputs vary. Patients with identical conditions receive different diagnoses depending on which radiologist they happen to see, on which day, in which city.

The empirical evidence for noise in professional judgment is extensive and deeply uncomfortable. Kahneman and colleagues documented inter-rater variance across domains that cannot be explained by legitimate differences in case complexity. Insurance underwriters producing widely divergent risk assessments for identical cases. Judges handing down sentences that vary by years for comparable offenses. Fingerprint analysts disagreeing on match determinations for the same prints examined twice.

This variance is not the result of incompetence. It is the result of the inherent variability of human judgment systems operating without external calibration structures. System 1 thinking — fast, intuitive, automatic — is exquisitely sensitive to context, mood, fatigue, and the order in which information is presented (Kahneman, 2011). The same case looks different on a Tuesday morning after a good night's sleep than on a Friday afternoon at the end of a difficult week. The noise is not in the professional's expertise — it is in the conditions under which that expertise is exercised.

Why does this matter for Decision Governance? Because noise is structurally different from incompetence, and the remedies are different. Incompetence calls for better training. Noise calls for better systems — structures that reduce the variance of equally capable professionals exercising their judgment under varied conditions. This is a governance problem, not a capability problem.

2.3 1.2 — How AI Solves Analytical Noise

The first-order effect of AI on professional judgment is genuinely positive. Models are not subject to the human noise sources described above. An AI system evaluating a portfolio will produce the same recommendation at 9 AM on a Monday as at 4 PM on a Friday. It is not affected by hunger, fatigue, the anchoring effect of the previous case, or the confirmation bias introduced by a client who happens to remind the advisor of a previous success.

For analytical tasks — processing structured data, identifying statistical patterns, running computational simulations — AI reduces inter-rater noise dramatically. Two advisors using the same AI system to generate a portfolio analysis will get similar starting points. The variance between their starting points shrinks. In this sense, AI functions exactly as noise-reduction infrastructure.

The capability has grown so quickly that it has outpaced the governance structures surrounding it. AI can now perform tasks that previously required substantial human judgment: generating coherent written analysis, producing narrative explanations of complex financial positions, synthesizing multi-source research into readable summaries. The boundary of what counts as "analytical" has expanded significantly.

This is the setup for the second-order problem — because the elimination of analytical noise at one layer does not eliminate noise at the decision layer. It relocates it.

2.4 1.3 — The New Noise Categories

When AI enters a professional workflow, it does not remove noise. It moves it upstream and changes its character. The following taxonomy of AI-generated noise categories represents the new frontier — largely undiscussed in the governance literature, which remains focused on AI bias rather than AI noise.

2.4.1 1.3.1 — Model Noise

The same prompt submitted to different AI models produces meaningfully different outputs. GPT-4, Claude, Gemini, and Llama — trained on different datasets, with different RLHF procedures, different system prompts — will generate different portfolio recommendations, different risk assessments, different narrative framings from identical input data. This is not a bug to be fixed in the next model version. It is a structural feature of probabilistic language models: the output distribution depends on model architecture, training data, and alignment procedures, none of which are fully transparent to the end user.

For a professional who uses one AI system, model noise is invisible — they never see the alternative outputs. For an organization with multiple AI tools, model noise produces variance between practitioners that looks exactly like the human inter-rater variance Kahneman documented, but with a new source: tool selection.

2.4.2 1.3.2 — Prompt Noise

The same model, given semantically similar prompts with different wording, will produce different outputs. "Analyze the risk profile of this portfolio" and "assess how conservative this portfolio is" may produce outputs that diverge significantly in their characterization of the same underlying data. Prompt sensitivity is a well-documented property of language models, and the gap between what a professional intends to ask and what the model responds to is a consistent source of variance.

Prompt noise is particularly insidious because it is attributable — incorrectly — to human judgment differences. If two advisors using the same AI system phrase their queries differently and get different outputs, the resulting advice variance looks like professional variance when it is actually prompt variance.

2.4.3 1.3.3 — Context Noise

AI systems are sensitive to context window composition — the documents, prior conversation turns, and system instructions that precede the query. An analysis run with a client's previous session in

context may differ from the same analysis run fresh. Context noise compounds across interactions: the AI "remembers" prior framings, earlier risk assessments, previous client goals. Each session's output is shaped by the accumulated context, which may or may not reflect the current situation accurately.

In financial advisory, context noise manifests as an AI system that weights outdated market assumptions or stale client goals because they are present in the context window. The professional may not be aware that the output is shaped by context they have not consciously chosen to include.

2.4.4 1.3.4 — Automation Bias

The most consequential of the new noise categories is not a property of the AI system — it is a property of the human-AI interface. Automation bias is the tendency of humans to over-rely on automated outputs, accepting them without critical evaluation, especially when the system presents its outputs in confident, authoritative language (Parasuraman & Manzey, 2010).

Automation bias is not a character flaw. It is a predictable response to cognitive load reduction. If the AI analysis is available, detailed, and authoritative-sounding, engaging System 2 to critically evaluate it requires effort that feels disproportionate to the apparent certainty of the output. System 1 accepts the confident output. The decision is made faster but with less deliberation.

The governance consequence of automation bias is severe: it collapses the decision process. The transition from "AI generates analysis" to "human commits to recommendation" becomes instantaneous rather than deliberate. The human becomes the implementation layer for an AI decision, not the decision-maker who uses AI as a tool. The governance gap is not just unfilled — it is unnoticed.

2.5 1.4 — The Governance Gap: A Formal Definition

The preceding analysis makes clear that a structural gap exists in AI-assisted professional workflows. This gap requires a formal definition that can anchor the architectural proposals in subsequent chapters.

> **The Governance Gap is the structural absence of infrastructure at the transition from AI-generated output to human-committed decision. It is the space where accountability disappears, variance compounds, and audit trails end.**

This definition has three components, each of which identifies a specific failure mode:

Accountability disappears because there is no record of how an AI output became a human decision. The AI log records the generation event; the CRM records the outcome; nothing records the reasoning that connected them. When a decision is later questioned — by a regulator, by a client, by the decision-maker themselves — the reasoning is unrecoverable. "I thought it seemed right given what the client said" is not an audit trail.

Variance compounds because the new noise categories documented in Section 1.3 are not self-correcting. Without governance infrastructure, model noise, prompt noise, context noise, and automation bias operate simultaneously and interact. A practitioner using a different AI tool (model noise) who phrases their query differently (prompt noise) in a context shaped by outdated assumptions (context noise) and accepts the result without deliberation (automation bias) has compounded four variance sources into a single recommendation. This is worse, structurally, than the pre-AI noise baseline.

Audit trails end at the point where AI output enters human judgment. Every step before is typically logged — the model used, the query submitted, the output generated. Every step after is typically logged — the recommendation made, the client action taken, the portfolio change executed. The gap between them is where the professional's reasoning lives. That reasoning is currently invisible to any record system.

This is not a technology problem. AI logging capabilities are more than sufficient to capture the decision process if the process were structured. The gap exists because no one has designed the infrastructure for it — because the transition from analysis to decision has been treated as instantaneous and unstructured.

2.6 1.4a — System Governance vs. Decision Governance

Understanding the governance gap requires distinguishing two layers that current AI governance discourse frequently conflates.

System Governance addresses what agents are permitted to do: permission boundaries, data access scope, identity attribution, monitoring infrastructure, resource limits, and policy enforcement. This layer is increasingly well-specified — through the EU AI Act, NIST AI RMF, and organizational AI policies — and addresses a genuine risk.

Decision Governance addresses a different layer: why a human converted agent output into real-world action. Specifically: which alternatives were considered and rejected; which agent assumptions were accepted versus independently verified; what authority grounded the decision; and what reasoning would, in principle, falsify the commitment post-action. System Governance specifies what the agent can do. Decision Governance specifies what the human did with what the agent produced.

The gap between these layers creates **accountability diffusion**: system logs capture what agents did, but not *why humans acted on agent outputs*. This is the specific gap that §1.4 defines — and the specific gap that neither monitoring infrastructure nor compliance tools fill.

Field observation from autonomous-agent deployments. Practitioner-facing reports from recent autonomous-agent deployments (publicly discussed in industry forums in 2025-2026; sources not yet peer-reviewed) describe the same pattern: even well-constrained agents create an account-

ability gap at the **output→action boundary**. When practitioners act on agent recommendations, structured analyses, or confidence-scored outputs, the reasoning bridge from output to action is ungoverned unless explicitly captured. The agent's behavior is logged. The human's reasoning is not.

Decision Governance infrastructure fills this gap. It does not replace system governance — both layers are necessary. It provides the artifact that system logs structurally cannot: the reconstruction of human judgment at the moment of commitment.

2.7 1.4b — The Adversarial Case: Why Decision Governance When the World Builds System Governance?

The dominant AI Governance discourse converges on a single architectural thesis: if we constrain the system sufficiently, the governance problem is solved. This manifests as guardrails for agentic systems, containment architectures for autonomous agents, output filtering, hallucination detection, bias mitigation, red-teaming, and constitutional AI (Bai et al. 2022, *Constitutional AI: Harmlessness from AI Feedback* — the canonical articulation of training-time alignment via principle-based feedback). The entire field is oriented toward making AI output trustworthy enough that human oversight becomes a formality.

The legitimate adversarial question follows directly: if the industry succeeds at making AI output reliably trustworthy, why would anyone need infrastructure that documents human judgment over that output?

Three independent arguments establish that system-side governance, even when fully successful, does not eliminate the need for decision-side governance — and that the gap between them widens rather than narrows as AI capabilities improve.

Argument 1: Regulatory Reality. Financial regulation does not ask whether a system's output was constrained. It asks whether the human professional can document why a specific recommendation was appropriate for a specific client in a specific situation. MiFID II Article 25, SEC Rule 17a-4, FINRA Suitability Rule 2111, and BaFin's MaComp share a common structural requirement: the advisor must demonstrate that a documented reasoning process connected the client's circumstances to the recommendation given. No guardrail architecture produces this documentation — because it originates in professional judgment, not in the system. System-side governance produces evidence of system behavior. Regulatory requirements demand evidence of human judgment. These are different evidentiary artifacts with different authors and different audit requirements.

Argument 2: The Inverse Governance Paradox. The relationship between AI capability and human governance risk is counterintuitive: better AI output increases, rather than decreases, the decision governance gap. When AI output is unreliable, professionals exercise genuine judgment — they question, verify, override. The friction generates natural documentation. When AI output is consistently excellent, the professional's role shifts from active judgment to passive ratification.

Outputs are signed without substantive evaluation — not from negligence, but because the system has earned trust through demonstrated reliability. This produces what we term **Ghost Ownership**: the professional bears formal responsibility for a decision they did not substantively make. The signature exists. The judgment does not. System-side governance cannot detect Ghost Ownership because the system performed correctly at every step. The governance failure occurs entirely on the human side of the interface. This pattern aligns with the "trust debt" framing in industry blog literature (Kozyrkov 2024, *data-based.com / Medium*; not peer-reviewed) — the gap between delegated authority and verified competence compounds over time when the verification step is unmonitored.

Argument 3: Complementarity, Not Competition. System-side governance and decision-side governance address different failure modes, different regulatory requirements, and different risk dynamics. Layer 1 (System Governance) asks: "Is the AI output trustworthy?" — implemented via guardrails, constitutional AI, red-teaming. Layer 2 (Decision Governance) asks: "Did a human demonstrably exercise judgment over this output?" — implemented via structured externalization, contestability mechanisms, provenance tracking, and auditable decision artifacts. These two layers operate on opposite sides of the same human-AI interface. Layer 1 governs what enters the interface. Layer 2 governs what exits it. Neither can substitute for the other because they produce different types of evidence about different actors. Adjacent industry frameworks have arrived at structurally similar conclusions from different starting points (e.g. references to "Bound Decision Artifact" or "Admissible-by-Construction" patterns in 2025-2026 practitioner literature). Where these references are not peer-reviewed publications, this paper treats them as convergent practitioner observation rather than independent academic triangulation.

The sharpest formulation: AI Governance without Decision Governance solves the easier half of the problem. The system side is an engineering challenge — constrain outputs, filter hallucinations, enforce guardrails. The human side is a governance challenge — document judgment, preserve agency, maintain contestability, create audit trails for decisions that exist in professional minds, not in system logs. Decision Governance addresses the governance problem that remains when the engineering problem is solved.

2.8 1.5 — Why Behavioral Economics Makes the Gap Visible

The governance gap would matter less if human decision-making under AI assistance were well-calibrated. If professionals reliably corrected for AI errors, balanced AI outputs against their own judgment, and maintained deliberate decision processes, the lack of formal infrastructure would be inconvenient but not catastrophic.

The behavioral economics literature indicates that this calibration is not reliably present — and provides a structural explanation for why.

Prospect Theory, developed by Kahneman and Tversky (1979) and published in *Econometrica*, estab-

lished that human decision-making under uncertainty is systematically non-linear in ways that no expected utility framework predicts. People weight losses approximately twice as heavily as equivalent gains. They are risk-averse in the domain of gains and risk-seeking in the domain of losses. Preferences are reference-dependent — the same objective outcome feels radically different depending on whether it is framed as a loss or a gain relative to a reference point.

The implication for AI-assisted advisory is precise: a client's stated risk tolerance — captured on a standard Likert-scale questionnaire — does not measure their actual loss aversion. A client who marks "moderate risk tolerance" on a five-point scale is not communicating how they will actually respond to a 20% portfolio drawdown. The scale elicits a self-concept, not a revealed preference.

This is why pairwise elicitation is structurally superior to Likert-based preference capture. A forced choice — "Which would you find harder to accept: losing €10,000 in a market downturn or missing a €10,000 gain when the market recovers?" — cannot be answered by marking a midpoint. It requires the respondent to actually resolve the tradeoff. The pairwise structure eliminates scale-use noise (different people use "moderate" differently) and mood noise (a person feeling cautious may mark "low risk" on a Tuesday morning and "moderate risk" on the same questionnaire given on a Thursday afternoon).

Prospect Theory provides the behavioral grounding for why the pairwise elicitation protocol embedded in the reference architecture's onboarding flow is not just a UX choice — it is a noise-reduction mechanism grounded in the empirical structure of human preference formation. Constructed Ambiguity (Pattern 4) and SPECTRADING (Pattern 2), in this context, are not design preferences. They are behavioral interventions.

2.9 1.6 — The Infrastructure Analogy

Why has the governance gap persisted? If the problem is structural and well-defined, why does no standard infrastructure exist for governing the transition from AI output to human decision?

The most clarifying analogy is the history of financial accounting before double-entry bookkeeping. Before Pacioli's system (1494), merchants tracked their financial state through various informal methods — single-entry records, memory, physical inventories. Each worked adequately for simple transactions. None provided the infrastructure for strangers to transact with strangers at scale, because none created a self-consistent record that could be audited by someone who was not present.

The governance gap is the epistemic equivalent of pre-bookkeeping financial recording. Each professional has their own informal method: a note here, a memory there, a PDF in a folder, a CRM entry that records what was done but not why. Each method works adequately for simple cases. None provides the infrastructure for regulatory accountability, institutional learning, or cross-professional consistency, because none creates a self-consistent decision record that can be evaluated by someone who was not in the room.

The parallel is precise. Pacioli's contribution was not a new way of doing business — it was a new way of making business visible. Decision Governance is not a new way of making decisions — it is a new way of making decisions visible.

Chapter 2 describes the architectural framework that fills this gap. Subsequent chapters ground it in information theory, formalize its measurement, describe its enforcement mechanisms, and demonstrate its application in the domain where the governance gap is most consequential and most immediately addressable: financial advisory under regulatory pressure.

2.10 1.7 — What Comes Next

The governance gap is not a software problem, an AI alignment problem, or a regulatory problem. It is an infrastructure problem — the absence of structured artifacts at the point where professional judgment is exercised. The argument in this chapter implies a specific architectural response:

1. The transition from AI output to human decision must be **structured** — not instantaneous, not informal, not optional.
2. The decision process must produce **records** — not impressions, not memories, not notes, but structured artifacts that capture reasoning, assumptions, and the specific human judgment that connected AI analysis to committed decision.
3. These records must be **immutable** — once created, they cannot be silently modified, because the value of the record depends on it reflecting what was actually decided, not what would have been decided in retrospect.
4. The records must be **measurable** — the quality of the decision process, not just the outcome, must be assessable, so that governance improves over time rather than remaining static.

These four requirements are the architectural demands that the SR7D framework is designed to meet. They arise directly from the structural analysis of the governance gap. They are not design preferences — they are the logical implication of taking the governance problem seriously.

The governance gap is real, it is measurable, and it is growing as AI capability increases faster than governance infrastructure. This paper proposes that filling it is the defining infrastructure challenge of the AI-abundant era — and that the architecture to fill it exists.

---

3 Chapter 2 — Decision Governance as Architectural Discipline: The SR7D Framework

3.1 2.0 — Why Architecture, Not Policy

Most responses to the governance challenge in AI-assisted environments are policy responses: guidelines, principles, ethical frameworks, responsible AI checklists. These are necessary but insufficient. A policy says "decisions should be transparent." An architecture *enforces* that decisions cannot be opaque.

The distinction matters because policy depends on compliance — someone choosing to follow the rule. Architecture depends on structure — the system making non-compliance structurally difficult or impossible. In safety-critical engineering (aviation, nuclear, medical devices), this distinction is well understood: you do not rely on the pilot remembering to check the altimeter. You design the system so that altitude information is unavoidable.

The distinction crystallizes in a single contrast: policy hopes for compliance; architecture compels it. A policy says "exercise judgment here." An architecture makes bypassing that judgment structurally difficult or impossible.

Decision Governance, as proposed in this paper, is an architectural discipline. It is realized through ten formal design constraints — seven core patterns and three ethical guardrails — collectively referred to as the SR7D framework. These constraints are not aspirational. They are implementable, testable, and in several cases machine-enforceable.

3.2 2.0a — Pain-to-Pattern Map (Reading-Path-by-Pain)

The seven SR7D patterns and three ethical guardrails are introduced below in their canonical architectural order. A CFP practitioner reading this paper for the first time may find it more useful to enter the patterns through the practical pain that is currently most pressing in their advisory practice. The table below maps recurring practitioner pains (collected during the framework's development; see Appendix F — Practitioner Objection Index) to the SR7D patterns that address them most directly. Use this as a non-linear reading path: locate the pain, jump to the pattern, return for the systematic treatment.

If your Tuesday-morning pain is…	Start with Pattern / Ethic	Read §	Why
"I can't reconstruct why I recommended X to the Müllers three years ago"	Archaeological Governance (Pattern 7) + Decision Packet	§2.1 Pattern 7 + §6.5 + Appendix C	Reconstruction is the regulatory bar; Decision Packet is the artifact that meets it
"BaFin audit asked for my reasoning chain and my CRM notes weren't enough"	Immutability First (Pattern 6) + Decision Packet	§2.1 Pattern 6 + §5.3 + Appendix C	Tamper-evident hash-chain on advisor-committed records
"My client acts on inherited intuitions they cannot articulate"	Constructed Ambiguity (Pattern 4) + pairwise elicitation	§2.1 Pattern 4 + §4.2.3	Bradley-Terry pairwise surfaces priorities Likert misses
"Two partners want different things and the AI averaged them"	Perspective Diversity (JQM Dimension 3) + SPECTRADING (Pattern 2)	§4.2.3 + §2.1 Pattern 2 + §6.2	Separate stakeholder profiles + documented divergence
"I suspect a client's preferences drifted but cannot prove it"	Drift Awareness (JQM Dimension 2) + Archaeological Governance	§4.2.2 + §2.1 Pattern 7	Temporal trajectory analysis over Decision Packet chain
"AI tools keep nudging me toward 'recommended' allocations I do not want to make"	Non-Normativity (Ethic 1) + machine-enforced lexicon	§5.4 + Appendix B.3	Pydantic-contract rejects normative language at architecture layer
"Smaller firms cannot afford complex tooling — and I am that smaller firm"	Digital Guardian (Pattern 1) + frozen constraint sets	§2.1 Pattern 1 + §5.3a	Architecture is implementable on commodity infrastructure; cost is editorial discipline
"Different AI tools give different answers and I don't know whom to trust"	Triangulated Truth (Pattern 3) + Bayesian posterior layer	§2.1 Pattern 3 + §3.7 + Appendix E.1	Multi-source convergence is the trust signal; single-source confidence is the warning

The table is illustrative; the actual mapping between a specific advisory pain and the most useful pattern may vary by practitioner context. Practitioners are encouraged to identify their own variant of each pain and to read the patterns through that lens.

3.3 2.1 — Seven Core Patterns

3.3.1 Pattern 1: Digital Guardian

Definition: The system protects the user — from external exploitation, from the system itself, and where possible from unreflected self-harm in the decision process. Data sovereignty is non-negotiable. No optimization runs against the user's interests, even implicitly.

Architectural implication: The system never monetizes user data, never optimizes for engagement, never introduces dark patterns that nudge toward a preferred outcome. Every data flow is auditable by the user. The user can export, inspect, and delete their complete decision history at any time.

Scientific grounding: Digital Guardian operationalizes what Zuboff (2019) describes as the antithesis of surveillance capitalism: a system whose economic model does not depend on behavioral prediction or modification. In the context of financial advisory, it directly supports the fiduciary duty — the legal and ethical obligation to act in the client's interest, not the platform's.

Relevance to Decision Governance: Without Digital Guardian, every other pattern is compromised. If the system has a hidden incentive to steer users toward certain decisions (ad revenue, product placement, engagement metrics), then Non-Normativity becomes theater and Agency Preservation becomes meaningless. Digital Guardian is the trust foundation on which the remaining patterns stand.

Reference implementation: enforced architecturally through data-flow constraints; no telemetry, no third-party data export, user-controlled export and deletion.

3.3.2 Pattern 2: SPECTRADING

Definition: The system operates on a configurable spectrum between autonomy and guidance. At one end, the system acts purely as a mirror — reflecting the user's mental model without interpretation. At the other end, the system actively structures, challenges, and extends the user's reasoning. The position on this spectrum is never hardcoded. It is context-dependent, user-configurable, and transparent.

Architectural implication: Low-stakes interactions (formatting, data presentation, navigation) permit higher system agency. High-stakes interactions (consequential decisions, value judgments, irreversible commitments) require maximum user agency. The system makes its current position on the

spectrum visible and adjustable. The user always knows how much the system is "leading" versus "following."

Scientific grounding: SPECTRADING addresses a core finding in human-AI interaction research: the automation bias problem (Parasuraman & Manzey, 2010). When systems present outputs with high confidence and low friction, humans default to accepting them — even when their own judgment would have been more appropriate. Conversely, systems that offer no structure at all impose unsustainable cognitive load. The optimal operating point is neither full automation nor full manual control, but a calibrated middle that shifts with context and stakes.

Relevance to Decision Governance: SPECTRADING prevents the two failure modes that plague AI-assisted decision tools: over-automation (the system decides, the human rubber-stamps) and under-support (the system dumps data, the human drowns). By making the autonomy-guidance spectrum explicit and configurable, it preserves the human's role as decision-maker while acknowledging that different moments in a decision process require different levels of system involvement.

Reference implementation: prototype-stage. The autonomy-guidance spectrum is implemented as a per-component configuration; UI surfacing of the current spectrum position is partial and remains an open design question.

A concrete way to picture SPECTRADING: imagine having a team of highly capable analysts who have done all the preparatory work — processed the data, evaluated the scenarios, stress-tested the assumptions — and laid everything out clearly on your desk. Their work is excellent. But none of them can put the signature on the contract. That final step — the actual commitment — belongs solely to you. SPECTRADING is the architectural specification of exactly this division of labor: AI does the preparation, the human makes the decision.

3.3.3 Pattern 3: Triangulated Truth

Definition: The system never relies on a single source of truth. Every significant claim, score, classification, or recommendation is derived from converging evidence across multiple independent sources. Where convergence is absent, the system reports divergence — it does not silently select one source over another.

Architectural implication: Pillar weights are not derived from a single questionnaire or a single conversation. They emerge from multiple intake modalities (narrative, quantitative, behavioral) processed through independent analytical paths. When these paths converge, confidence is high. When they diverge, the system surfaces the divergence as information, not as error.

Scientific grounding: Triangulated Truth is the epistemological principle that underlies meta-analysis, multi-method research design, and the Delphi method. In forecasting science, Tetlock (2015) demonstrated that aggregating independent judgments consistently outperforms individual expert judgment — not because any single source is more accurate, but because convergence across independent sources cancels out idiosyncratic noise. In information theory, multiple inde-

pendent signals reduce entropy faster than repeated measurements from a single source (Shannon, 1948).

Relevance to Decision Governance: In AI-abundant environments, the danger is not lack of information but lack of triangulation. Five AI models can produce five different portfolio recommendations from the same input data. Without Triangulated Truth, the user receives whichever output happens to appear first. With Triangulated Truth, the user sees where the models agree (high-confidence zone) and where they disagree (requiring human judgment). This transforms the abundance problem from noise into signal.

The analogy from medical practice is instructive. A practitioner who immediately delivers a confident single diagnosis for a complex symptom picture sounds authoritative — but is actually being epistemically dishonest. The practitioner worth trusting is the one who says: the evidence points in three directions, here are the competing explanations, here is what each implies for treatment. Paradoxically, this acknowledgment of uncertainty generates deeper and more durable trust than false precision. Constructed Ambiguity requires the system to behave like the second practitioner: structurally compelled to surface divergence rather than average it away.

Reference implementation: instantiated (validation against decision-outcome data pending — see Appendix E.1.3 and §3.7a). Multi-modal intake feeds into the Bayesian posterior layer documented in Section 3.7; convergence/divergence across paths is recovered from posterior marginals rather than asserted by the application layer.

3.3.4 Pattern 4: Constructed Ambiguity

Definition: The system deliberately maintains and communicates uncertainty where uncertainty exists. It does not resolve ambiguity prematurely. It does not present confidence it does not have. Where the evidence is mixed, the system says so. Where the model is uncertain, the system quantifies and displays that uncertainty. Constructed Ambiguity is epistemic honesty as a design principle.

Architectural implication: Every system output carries an explicit confidence indicator. This is not a cosmetic probability bar — it is a structurally computed measure derived from evidence density, source convergence, and model calibration. Where confidence is low, the system expands the option space rather than narrowing it. Where multiple plausible interpretations exist, the system presents them as alternatives rather than selecting one.

Scientific grounding: Constructed Ambiguity addresses what Kahneman (2011) calls the "illusion of certainty" — the human tendency to prefer definitive answers over calibrated uncertainty. AI systems amplify this tendency by presenting outputs in confident, declarative language regardless of their actual reliability. In epistemology, the concept maps to what philosophers of science call "honest uncertainty quantification" — the obligation to communicate not just what is known, but the boundaries of what is known and the confidence attached to it. Friston's Free Energy Principle (2010) provides a formal framework: a system that accurately models its own uncertainty makes

better predictions than one that collapses uncertainty into false precision.

Relevance to Decision Governance: Constructed Ambiguity is the direct countermeasure to AI-generated false precision. When a financial planning tool says "your retirement probability is 73%," that number carries an illusion of computational authority. Constructed Ambiguity requires the system to show what assumptions produce that number, how sensitive it is to changes in those assumptions, and what the confidence interval actually looks like. This does not make the user less informed — it makes them honestly informed. And honestly informed decisions are structurally better decisions.

Reference implementation: instantiated (validation against decision-outcome data pending — see Appendix E.1.3 and §3.7a). The reference architecture exposes confidence as three calibrated categories derived from posterior max-marginal entropy (see Section 3.7 and Appendix E.1); see Appendix F, Objection 5 for the user-facing rendering.

3.3.5 Pattern 5: Principled Inefficiency

Definition: The system deliberately introduces friction at points where speed would compromise judgment quality. It does not auto-execute consequential decisions. It does not optimize for time-to-completion. Where slowing down improves decision quality, the system slows down — by design, not by accident.

Architectural implication: Consequential decisions require explicit human confirmation that cannot be bypassed through UI shortcuts. The system introduces mandatory reflection points — not as interruptions, but as structured moments where the user is asked to confirm, revise, or reconsider before a decision becomes final. The time between "AI generates recommendation" and "human commits to decision" is a governed interval, not an instantaneous handoff.

Scientific grounding: Principled Inefficiency draws on Kahneman's (2011) System 1 / System 2 framework: fast, automatic processing (System 1) is efficient but error-prone for complex decisions. Slow, deliberate processing (System 2) is costly but necessary for consequential choices. AI tools, by making analysis instantaneous, push the entire decision process into System 1 territory — the user receives the answer so quickly that there is no natural pause for System 2 engagement. Principled Inefficiency reintroduces that pause architecturally.

In safety engineering, this principle is well-established: nuclear reactor control systems have mandatory confirmation sequences. Aircraft autopilot systems require explicit pilot authorization for mode changes. The principle is not that slowness is good — it is that certain decisions benefit from a structurally enforced interval between information and action.

Relevance to Decision Governance: Principled Inefficiency is the pattern most likely to be cut by product managers, investors, and users who equate speed with value. That resistance is precisely why it must be architectural rather than optional. In AI-abundant environments, the default is instant output → instant acceptance → no governance. Principled Inefficiency breaks this chain at the critical

juncture: the moment between receiving an AI-generated analysis and committing to a decision based on it. It is not a brake on productivity. It is a brake on unconsidered action. Intentional, programmed slowness at high-stakes decision points is not a concession to human limitation — it is the mechanism by which human judgment is kept in the loop rather than gradually engineered out of it.

Reference implementation: enforced. Consequential-decision confirmation gates are non-bypassable in the reference implementation; the gate set is configured per decision class.

3.3.6 Pattern 6: Immutability First

Definition: Once a decision record is written, it is never overwritten, modified, or deleted. Corrections are appended as new records that reference the original. The complete history — including errors, revisions, and changes of mind — is preserved.

Architectural implication: Decision Packets are stored in an append-only chain with cryptographic linking (tamper-evident hash chain). Each packet contains a parent_hash referencing its predecessor. Modifying a historical record would break the hash chain, making tampering structurally detectable. This is not blockchain — it is the same integrity principle applied at the individual decision level.

Scientific grounding: Immutability First borrows from two established disciplines. In accounting, the principle that ledger entries are never erased — errors are corrected with contra-entries — is the foundation of auditability. Pacioli's system (1494) established this five centuries ago: the integrity of the record depends on nothing being silently removed. In software engineering, version control systems (Git) operate on the same principle: every state is preserved, every change is a diff against a known baseline, and any historical state can be reconstructed.

Relevance to Decision Governance: Immutability is the precondition for two things that matter in regulated environments: audit trails and learning. An auditor needs to see the original decision, not the cleaned-up version. A decision-maker who wants to improve needs to see their actual reasoning at the time of the decision, not their retroactive rationalization. Immutability First ensures that the record is honest — even when the decision-maker would prefer it not to be.

Reference implementation: enforced. Decision Packets are written to an append-only chain with SHA-256 hash linking (Appendix C); the chain-validation check is part of the reference implementation's CI.

3.3.7 Pattern 7: Archaeological Governance

Definition: Every decision is reconstructable — not just what was decided, but why, on what basis, with what assumptions, by whom, and at what point in time. The system creates artifacts that

allow any future observer — a regulator, a client, the decision-maker themselves years later — to reconstruct the full decision context without interviewing anyone who was present.

Architectural implication: Decision Packets are self-contained records that include: the AI-generated analysis (model, version, input data, output, timestamp), the human decision (chosen option, rejected alternatives, reasoning, contextual factors), the assumption set (explicit and surfaced assumptions with confidence indicators), and the relational context (links to prior decisions, stakeholder profiles, temporal position in the decision trajectory). The record stands independent of its creator — it can be read, understood, and evaluated by someone who was not in the room.

Decision Packets are **reconstruction infrastructure**, not compliance logging. The distinction matters: compliance logging records what happened; reconstruction infrastructure preserves why the human decided to act on what happened. System logs already capture the former. Decision Packets fill the latter.

System Logs Show	Decision Packets Add
Agent invoked at timestamp T	Why human acted on that specific output
Agent produced output O	Which alternatives were rejected, with reasons
Agent confidence score 0.85	Human's confidence in accepting that score
Agent used data source S	Whether human verified vs. trusted that source

Scientific grounding: The archaeological metaphor is precise. An archaeologist reconstructs past civilizations not by interviewing the inhabitants, but by reading the artifacts they left behind. The quality of the reconstruction depends entirely on the quality of the record. In decision science, this maps to what Klein (1998) calls "recognition-primed decision making" — expert decisions are often made rapidly and intuitively, but their quality can only be evaluated retroactively if the decision context is preserved. *Caveat: Klein's finding cuts both ways for this pattern — see §3.7a (Engaging the Confabulation Counter), which acknowledges that expert decisions may be non-articulable by construction and that what Decision Packets preserve is a contestable record of the practitioner's stated reasoning, not a faithful reconstruction of the underlying cognitive process.* Without Archaeological Governance, expert intuition is invisible and unchallengeable. With it, intuition becomes an auditable artifact.

The temporal dimension of Archaeological Governance is critical. A single decision snapshot captures a moment. A sequence of decision snapshots — the same person, the same domain, over months or years — captures a trajectory. Trajectories reveal patterns that no single snapshot can: gradual drift in risk tolerance, slow erosion of confidence in a previously held position, systematic blind spots that only become visible over time. This temporal reconstructability transforms the decision history from a static archive into an analytical resource.

Relevance to Decision Governance: Archaeological Governance is what Pacioli's double-entry bookkeeping did for financial records: it makes the decision state legible to someone who was not

there. In AI-abundant environments, where decisions are made faster and with more AI involvement than ever, the ability to reconstruct why a decision was made — years later, by someone who was not present — is not a luxury. It is the infrastructure that allows trust to scale beyond personal relationships.

Reference implementation: instantiated (validation against decision-outcome data pending — see Appendix E.1.3 and §3.7a). Decision Packet schema v0 (Appendix C) is frozen; the reference implementation produces compliant Packets and supports their re-loading and re-rendering from disk.

3.4 2.2 — Three Ethical Guardrails

The seven core patterns define *how* the system is built. The three ethical guardrails define *what the system must never do* and *what the user must always retain*. They are constraints on the system's behavior, not features of it.

3.4.1 Ethics 1: Non-Normativity

Definition: The system never tells the user what to decide. It surfaces information, structures options, quantifies uncertainty, and makes assumptions explicit — but the evaluative judgment remains entirely with the human. The system describes; it does not prescribe.

Enforcement: Non-Normativity is machine-enforceable. A validation layer (implemented as a Pydantic contract in the reference architecture) scans every system output for normative language — "should," "must," "recommend," "the best option is." Outputs containing normative language are rejected before reaching the user. This is not a guideline that depends on developer discipline. It is a structural invariant that the system cannot violate without triggering an error.

Why it is an ethics, not a pattern: Non-Normativity is not a design choice among alternatives. It is the ethical boundary that separates a decision governance system from a recommendation engine. The moment the system recommends, it has substituted its judgment for the user's. In a world where AI-generated recommendations are cheap and abundant, the scarce resource is the human's own evaluative capacity. Non-Normativity protects that capacity by refusing to replace it.

3.4.2 Ethics 2: Agency Preservation

Definition: The system makes the human more capable of deciding — not less likely to engage with the decision. Every feature, every output, every interaction is evaluated against the question: does this increase or decrease the probability that the human will exercise their own judgment?

Enforcement: Agency Preservation is harder to enforce mechanically than Non-Normativity, but it has testable proxies. Does the system present options or a single answer? Does the user need to

actively confirm or can they passively accept? Does the system surface assumptions the user can challenge, or does it present conclusions the user can only take or leave? These are design tests that can be applied at every feature decision.

Why it is an ethics, not a pattern: Agency Preservation is the moral commitment that motivates the entire framework. AI systems, by default, reduce human agency — they make choices easier by making them automatic. For low-stakes decisions (choosing a restaurant, formatting a document), this is benign. For consequential decisions (financial strategy, medical treatment, business pivot, policy direction), reduced agency means reduced responsibility, reduced learning, and reduced capacity to course-correct when circumstances change. Agency Preservation is the assertion that for consequential decisions, the human must remain the agent — not the spectator.

3.4.3 Ethics 3: Contestability

Definition: Every system output is challengeable. The user can trace any score, classification, recommendation, or insight back to its sources, assumptions, and computational path. If the user disagrees, they can override the system with documented reasoning — and that override becomes part of the record.

Enforcement: Contestability requires provenance tracking. Every output carries metadata: which model produced it, what data went in, what assumptions were applied, what parameters were used. The user can inspect this metadata at any level of detail. Overrides are first-class objects in the system — they are not exceptions or workarounds, but expected and documented events.

Each Decision Packet captures not just *what* was decided, but *on whose authority*. The `actor_id` field grounds accountability; the `provenance` entries trace each assumption to its source (user-confirmed, AI-generated, inferred). This structure prevents the "the AI said so" defense — post-hoc inspection reveals which agent claims were accepted, which were verified, and what reasoning justified the conversion from output to action.

Why it is an ethics, not a pattern: Contestability is the democratic principle of the framework. It asserts that no output is above challenge — not because the system is unreliable, but because the user's right to question is more important than the system's authority. In a regulated environment (financial advisory, healthcare, legal), contestability is not optional — it is the mechanism through which human oversight becomes meaningful rather than ceremonial. A system that cannot be contested cannot be governed.

3.5 2.3 — The Relationship Between Patterns and Ethics

The seven patterns and three ethics are not two separate lists. They are interdependent.

Digital Guardian enables Non-Normativity — if the system has hidden incentives, its non-normative outputs are not trustworthy. Immutability First enables Contestability — you can only challenge an output if the original record is preserved. Archaeological Governance enables Agency Preservation — the user can only learn from their own decisions if the full context is reconstructable. Principled Inefficiency supports all three ethics — by creating a governed interval between information and action, it gives the user time to exercise judgment, contest outputs, and maintain agency.

Constructed Ambiguity has a dual role: it is a core pattern (how the system represents uncertainty) and a direct enabler of Non-Normativity (a system that honestly communicates its uncertainty cannot simultaneously tell the user what to decide). SPECTRADING operationalizes Agency Preservation — by making the autonomy-guidance spectrum explicit, it gives the user control over how much agency they retain in each interaction. Triangulated Truth supports Contestability — when the user can see that an output is based on converging evidence from multiple sources, they can contest the output by challenging any individual source.

The full SR7D framework, therefore, is not a menu from which individual principles can be selected. It is an integrated architecture where each element reinforces the others. Removing any single element weakens the entire structure — not catastrophically, but measurably.

3.6 2.4 — Why Ten Constraints, Not Four

The Substack articles accompanying this paper present four simplified principles: Non-Normativity, Contestability, Archaeological Governance, and Agency Preservation. This simplification is appropriate for a general audience and a 2,000-word format.

The full SR7D framework, however, requires all ten constraints for architectural completeness. The structure is a card house: remove any single element and the architecture collapses, because the constraints are mutually load-bearing. The simplification collapses important distinctions:

- Archaeological Governance subsumes the temporal dimension — trajectory analysis, drift detection, belief updating over time — but without Immutability First, the records that Archaeological Governance reconstructs cannot be trusted.
- Agency Preservation states the goal, but without Principled Inefficiency (the mechanism that creates space for agency) and SPECTRADING (the mechanism that calibrates how much support the system provides), the goal is aspirational rather than structural.
- Non-Normativity prevents the system from prescribing, but without Constructed Ambiguity (the system communicating its own uncertainty) and Triangulated Truth (the system showing evidence convergence rather than single-source conclusions), the user lacks the information needed to form their own judgment.
- Contestability requires provenance, but without Digital Guardian (ensuring the system has no incentive to obscure provenance), the provenance itself cannot be trusted.

Four principles describe *what* Decision Governance achieves. Ten constraints describe *how* it is enforced. The whitepaper requires the latter.

---

4 Chapter 3 — The Information-Theoretic Foundation

4.1 3.0 — From Design Philosophy to Formal Epistemics

Chapter 2 established the SR7D framework as an architectural specification for Decision Governance: ten constraints — seven core patterns and three ethical guardrails — that make decisions visible, traceable, contestable, and improvable. The justification offered there was primarily structural: these patterns, taken together, form an integrated architecture that resists the failure modes of AI-assisted decision-making.

This chapter offers a different kind of justification. Decision Governance is not only good design. It has formal roots in information theory and logic — roots that allow two of its central mechanisms to be understood not as design choices but as formal epistemic procedures with mathematically characterizable properties.

The two bridges developed in this chapter are:

1. **Active Inference / Entropy Reduction (Friston, Shannon):** The externalization protocol — pairwise querying of value priorities — is formally an entropy-reduction process. Each question asked reduces uncertainty in the user's value hierarchy in a way that is measurable using Shannon's information measure. The isomorphism to Friston's Free Energy Principle provides a theoretical framework for understanding why this process works as well as it empirically does.
2. **Abductive Decision-Making (Peirce):** Aspiration-Backward reasoning — the process of reasoning from a desired future state to the most plausible description of the present conditions needed to reach it — is formally abductive inference. The logical structure established by Charles Sanders Peirce in the nineteenth century maps precisely onto the reverse simulation methodology embedded in the reference architecture's goal-setting protocol.

These are not metaphors or analogies deployed for rhetorical effect. They are claims about the formal structure of the mechanisms in question. As required by the epistemic standards of this paper, both claims are flagged at their appropriate confidence level: the Friston bridge is proposed formalization, not proven theorem; the Peirce bridge is established logical structure applied to a new domain, not a new logical result.

4.2 3.1 — The Entropy Problem in Value Elicitation

Picture the cockpit of a commercial airliner where every instrument — altimeter, navigation display, fuel gauge — shows a blank screen. The plane is ready to fly, but the system has no idea where it is starting from, and the pilot has only a vague sense of where they want to land. This is precisely the epistemic situation of any decision-support system confronting a new client: technically capable, but operating in near-total ignorance of what the person actually values.

Before a pairwise querying sequence begins, an advisor or governance system confronting a new client has very high uncertainty about that client's value hierarchy. The client may have stated broad goals — "I want to retire comfortably" — but the specific ordering of competing values is unknown. Which matters more: housing security or income stability? Education funding for children or early retirement? Maintaining lifestyle in a downturn or maximizing long-term wealth?

Each of these tradeoffs is a dimension along which the client's preferences are, initially, nearly opaque. In Shannon's framework (1948), this initial state can be understood as high entropy: the space of possible preference orderings is large, and no individual ordering is substantially more probable than alternatives.

Shannon entropy is defined as:

$$H(X) = -\sum p(x) \log_2 p(x)$$

where $p(x)$ is the probability assigned to each possible state x in the space of interest. At the beginning of a value elicitation process, if there are n pillars — housing, retirement, income, education, legacy, liquidity — and their relative weights are unknown, the entropy over the space of possible orderings is maximal. Each pairwise question asked — "Which would you find harder to sacrifice in a difficult year: housing security or retirement savings?" — provides information that updates the probability distribution over possible preference orderings.

Claim level: The formal claim here is that pairwise querying is an entropy-reduction process in Shannon's sense, and that the sequence of questions can be ordered to maximize information gain per question. This is proposed formalization. The mathematical development would require specifying the probability distribution over pillar weights, the update rule for each pairwise comparison, and a proof that the process converges. The intuition is sound and the structure is well-defined; the full formal treatment is an open research question.

What can be stated with confidence is the empirical structure: pairwise forced choices produce more stable and consistent preference orderings than Likert-scale questionnaires, a result that follows from the elimination of scale-use noise and the forcing of genuine tradeoff resolution (see Chapter 4 for the Bradley-Terry/Elo formalization). The information-theoretic framing provides a principled explanation for why this is so. A useful way to picture the process: imagine playing the ultimate hyper-optimized game of twenty questions. If asked to guess an animal, the expert does not waste their first question asking "is it my neighbor's golden retriever?" — they ask "does it have fur?" because that question cuts the universe of possibilities as close to in half as possible. Each pairwise question in the value elicitation protocol is chosen by the same logic: the query that maximally

reduces entropy in the current model of the user's preferences, shrinking the space of plausible value orderings with every answer.

4.3 3.2 — The Free Energy Principle and Active Inference

Karl Friston's Free Energy Principle (2010) provides a unified theoretical account of how adaptive systems — biological and potentially artificial — interact with their environment in order to minimize prediction error, or equivalently, to minimize entropy in their model of the world.

The core claim of the Free Energy Principle is that biological agents act so as to minimize *free energy* — a formal quantity that upper-bounds the surprise (negative log probability) of their sensory observations given their model of the world. Minimizing free energy is equivalent to minimizing the divergence between the agent's generative model and the actual structure of the environment. An agent that does this successfully maintains accurate beliefs about its environment; an agent that fails accumulates surprise and degrades in its ability to behave adaptively.

In neuroscience, this provides a unified account of perception (updating the model to fit the data) and action (changing the data to fit the model). Both reduce free energy; the difference is the direction of the intervention.

The application to value elicitation in Decision Governance is as follows:

A governance system that conducts a pairwise querying protocol is implementing a structured form of active inference. At each step, the system: 1. Maintains a current model of the client's value hierarchy (the generative model) 2. Selects the query that maximally reduces uncertainty in that model (action chosen to minimize expected free energy) 3. Updates the model based on the client's response (perception: updating the model to fit the data)

This is the active inference loop: model → action → observation → model update → next action. The pairwise querying protocol is not a static questionnaire — it is an adaptive inference process that, at each step, chooses the question most likely to reduce entropy in the current model.

Claim level: The isomorphism between active inference and pairwise value elicitation is structurally well-defined. Whether the specific free energy functional used in Friston's neuroscientific framework applies unchanged to preference elicitation is an open question. The claim is that the structural parallel holds — both processes aim to minimize entropy in a model through structured inquiry — not that the mathematical details transfer without modification. This distinction matters for intellectual honesty.

Relevance to SR7D: Constructed Ambiguity (Pattern 4) is the architectural expression of accurate free energy minimization. A system that honestly represents its own uncertainty — rather than collapsing uncertainty into false precision — maintains an accurate generative model. Archaeological Governance (Pattern 7) provides the temporal record needed to track belief updating over time, which is the empirical evidence that the active inference loop is functioning rather than stagnating.

4.4 3.3 — Abductive Inference: Reasoning Backward from Goals

The second formal bridge requires a brief tour of the three modes of inference identified by Charles Sanders Peirce in his 1903 Harvard Lectures on Pragmatism.

Deduction proceeds from general rules to specific consequences: Given the rule "all stocks in this sector lose value in a rate-hiking cycle" and the fact "rates are rising," we deduce "this sector will lose value." Deduction is truth-preserving — if the premises are true, the conclusion is guaranteed. It is also, for this reason, limited: it can only produce conclusions already implicit in its premises.

Induction proceeds from specific observations to general rules: "This advisor's portfolios consistently outperformed in volatile markets. And this one. And this one. Therefore, this advisor has skill in volatile markets." Induction expands our knowledge beyond the premises but with a probabilistic guarantee rather than a logical one. It is the primary mode of empirical science.

Abduction is the third mode, and the least intuitively familiar. It proceeds from a goal or observed outcome to the most plausible hypothesis about its causes or preconditions: "I want to retire at 60 with €80,000 annual income, sustainable for 30 years, inflation-adjusted, with 95% probability. What must be true about my portfolio composition, savings rate, and risk exposure today?" Abduction does not prove the hypothesis it generates. It identifies the most plausible explanatory hypothesis given the evidence and the goal.

Peirce understood abduction as the engine of scientific discovery: hypotheses are not deduced from data, they are abduced — generated as the most plausible explanations of puzzling observations. The creative inference that produces a new scientific hypothesis is abductive.

4.5 3.4 — Aspiration-Backward Reasoning as Abduction

The goal-setting protocol in the reference architecture — sometimes described as "Aspiration-Backward reasoning" or "Reverse Simulation" — is formally abductive inference applied to personal financial goals.

The protocol works as follows: Rather than starting from a current portfolio and projecting forward to outcomes ("given your current savings rate and asset allocation, you have a 70% probability of reaching €80,000 annual retirement income"), the protocol starts from the goal state and asks what must be true today for that goal to be achievable. "You want €80,000 annual retirement income at 60. What savings rate, asset allocation, and risk tolerance would need to be true today for that goal to be reachable with high probability?"

This inversion is not merely a reframing. It is a different logical operation. Forward projection is deductive: given premises (current state), derive consequences (future states). Aspiration-Backward

reasoning is abductive: given the goal (desired future state), identify the most plausible current conditions that make the goal reachable.

The practical implication is significant. Forward projection anchors the client on their current situation and asks what it implies. This activates Kahneman's System 1: the current situation is salient, familiar, and psychologically weighted. Aspiration-Backward reasoning anchors the client on their desired future state and asks what changes in the present would make it reachable. This forces System 2 engagement: the client must evaluate the plausibility of different current conditions, not just project from a given starting point.

On the limits of AI abduction: It is important to be precise here about what AI can and cannot do in this framework. AI systems — language models and their underlying architectures — are, as of 2026, primarily deductive and inductive engines. They excel at pattern recognition across observed training data (induction) and at generating consistent outputs from given premises (deduction). They are structurally weak at abduction in Peirce's precise sense: generating genuinely novel explanatory hypotheses that are not implicit in their training distribution.

The governance implication of this limitation is precise: abductive goal-setting — defining what future state to aspire to, and evaluating the plausibility of different paths to that state — is exactly the cognitive task that should remain with the human. The reference architecture's Aspiration-Backward protocol does not claim to automate abductive reasoning. It structures the conversation in a way that invites the human to perform the abduction, while the AI performs the deductive computational work (projecting forward from specified conditions to probability distributions over outcomes).

The practical implication can be stated in a single contrast: the human sets the destination; the AI calculates the route. Abductive goal-setting — determining what future state is worth working toward — is not a computational task. It is a judgment about values, plausibility, and willingness to change. The AI's role is to make the consequences of that judgment legible: to show, with mathematical precision, what each proposed destination requires.

This is not a limitation to be apologized for. It is the correct division of cognitive labor: machine computation for the deductive paths, human judgment for the abductive goals.

On the limit of reverse inference: The abductive reasoning literature identifies a specific failure mode — what has been called "the limit of reverse inference." Recovering a unique causal structure from observed outcomes is inherently underdetermined: multiple different causal histories can produce identical observations. In neuroscience, this is the problem of inferring cognitive processes from brain activation patterns — many cognitive states produce similar activation signatures. In financial planning, this is the problem of inferring genuine risk preferences from observed behavior — many different preference structures can produce similar portfolios.

Decision Governance is designed to operate in precisely this gap. The framework does not claim to uniquely recover the client's "true" preferences from their responses. It provides structured instruments (pairwise elicitation, assumption surfacing, perspective divergence records) that reduce the

underdetermination — that narrow the space of plausible preference hypotheses — without claiming to eliminate it. Constructed Ambiguity (Pattern 4) ensures that residual underdetermination is communicated, not concealed.

4.6 3.5 — The Convergence of the Two Bridges

The two formal bridges established in this chapter are not independent. They converge on the same architectural design.

Active inference (Friston/Shannon) tells us: *The system should ask the question that maximally reduces entropy in the current model of the user's value hierarchy.* This is an information-theoretic prescription for the design of the externalization protocol.

Abductive inference (Peirce) tells us: *The user's goal-setting should proceed by identifying the most plausible present conditions needed to reach a desired future state.* This is a logical prescription for the structure of the goal-setting conversation.

Together, they describe an integrated governance architecture: - Entropy-reducing pairwise queries establish the value hierarchy (what matters most to this person) - Aspiration-Backward reasoning establishes the aspiration structure (what future state they are working toward) - The intersection of value hierarchy and aspiration structure generates the governance context for every downstream decision

Neither bridge, alone, is sufficient. A system that only reduces entropy in the preference model has a well-calibrated starting point but no goal structure. A system that only reasons backward from goals has aspirations but no value hierarchy to adjudicate between competing paths to those aspirations. The architecture requires both.

Relevance to SR7D: Triangulated Truth (Pattern 3) is the architectural expression of multi-source entropy reduction: not one questionnaire, but multiple intake modalities (narrative, quantitative, behavioral, pairwise) processed through independent analytical paths. Each path provides independent information; convergence across paths reduces uncertainty faster than any single source. Archaeological Governance (Pattern 7) is the architectural expression of temporal belief updating: the record of how value hierarchies and aspirations have changed over time provides the longitudinal data needed to measure whether the active inference loop is, in fact, reducing entropy or oscillating without convergence.

The architectural expression is no longer purely conceptual: Section 3.7 documents the Bayesian posterior layer that operationalizes the multi-source entropy-reduction claim, and Appendix E.1 provides the implementation detail. Section 4.4a documents the operational subcomponent (stat-prior confidence) through which the posterior layer feeds the measurement framework of Chapter 4.

4.7 3.6 — What the Formal Grounding Adds

One might reasonably ask: does the formal grounding matter practically? The governance gap is real whether or not pairwise elicitation can be mapped to Friston's Free Energy Principle. Decision Packets work whether or not Aspiration-Backward reasoning is formally abductive.

The formal grounding matters for three reasons.

First, it provides falsifiability. A design philosophy can be contested only at the level of intuition — you prefer a different design, I prefer this one. A formally grounded claim can be contested at the level of mathematics: show me where the mapping breaks down, show me where the entropy calculation fails, show me the counterexample to the abductive structure. Formal claims invite the kind of precise intellectual engagement that design philosophies cannot.

Second, it constrains future development. If the pairwise elicitation protocol is an active inference process, then improvements to that protocol should be evaluated by their effect on expected free energy minimization — not just by user satisfaction ratings or completion rates. The formal grounding provides evaluative criteria that are independent of the implementation.

Third, it connects Decision Governance to established research programs. Friston's framework is the subject of extensive ongoing research in neuroscience, AI, and cognitive science. Peirce's abductive logic is the subject of ongoing research in logic, philosophy of science, and AI reasoning. By establishing formal bridges to these programs, Decision Governance becomes accessible to researchers in those fields — and becomes capable of inheriting their results. This is how a framework moves from a design document to a research program.

Chapter 4 develops this research program in the specific domain of measurement: what Judgment Quality Metrics can be derived from the formal structures established here, and how they can be operationalized as empirically testable quantities.

4.8 3.7 — From Theory to Posterior Layer: Operationalization

Practitioner summary (skip the technical content below if you do not need it): The Bayesian posterior layer is the reference architecture's way of producing the confidence indicators that appear in the practitioner's interface. When the system tells you that a particular pillar weight, a particular assumption, or a particular recommendation has "high", "moderate", or "low" confidence, those categories are derived from the inference machinery described in this section. The mathematics matters because it is reproducible and auditable — but the practitioner-facing experience is the three categories, with one sentence each explaining what the category means for the recommendation in front of you. If the technical detail below is more depth than you need, skip to Section 3.8 (closing) or Chapter 4 (measurement).

Sections 3.0-3.6 established the formal grounding for the framework's information-theoretic claim: pairwise elicitation as active inference, value hierarchy as entropy reduction, aspiration-backward reasoning as abduction. The grounding was deliberately conceptual.

The reference implementation crosses the bridge from formal grounding to operational inference: the information-theoretic framework is the architecture of a Bayesian posterior layer that operates inside the analytical component. The layer takes pillar-subpoint co-occurrence counts as input, derives pointwise mutual information (PMI) using natural logarithm, constructs conditional probability tables, materializes the result as a discrete Bayesian network, and computes posteriors via belief propagation. The output is a discriminated union — each posterior carries either a max-marginal with entropy, an explicit "unobserved" marker, or an error type with diagnostic information — that eliminates the ambiguity of single-numeric confidence fields. The practical difference is the same as between a basic weather app that says "it might rain" and a professional meteorological system that shows the full radar probability model — and explicitly marks the locations where its own radar towers are offline. The former collapses uncertainty into a reassuring single guess; the latter surfaces exactly what is and is not known.

The implementation details (the five inference stages, the edge cases handled, the pgmpy version dependency, the treewidth threshold, and the rationale for natural-log over log-base-2) are documented in **Appendix E.1 — Bayesian Posterior Layer**. The connection to Sections 3.2-3.5 is direct: PMI is the empirical Shannon-mutual-information estimate from which the entropy-reduction claim of Section 3.2 is operationalized; belief propagation is the inference machinery that allows the active-inference loop of Section 3.5 to actually update beliefs as new evidence arrives.

Claim level. The connection between formal grounding and operational implementation is *proposed mapping with empirical instantiation*; full validation against decision-outcome data is pending (see Chapter 7, RQ3 and RQ5).

4.9 3.7a — Engaging the Confabulation Counter

A foundational empirical challenge to Decision Governance comes from the introspection-illusion literature (Nisbett & Wilson, 1977) and the choice-blindness paradigm (Johansson, Hall, Sikström & Olsson, 2005). Both empirical traditions document that humans routinely report reasons that did not in fact drive their choices — fluently, confidently, and with apparent sincerity. Johansson's subjects defended photographs they had not selected, articulating elaborate justifications for choices the experimenters had covertly swapped. Nisbett & Wilson's broader review found that introspective reports about one's own cognitive processes have low fidelity to the actual processes that produced the behavior.

Klein's (1998) recognition-primed decision (RPD) research sharpens the challenge in the specific direction this paper most depends on: expert decisions in domains like firefighting, intensive-care medicine, and military command are *non-articulable by construction*. Experts recognize patterns and

act on them with a fluency that resists post-hoc decomposition into "the reasons." The same Klein work the paper cites approvingly as evidence that expert judgment is real also cuts the other way: forcing articulation of non-articulable expertise may not capture the expertise — it may degrade it into a confabulated narrative that *resembles* reasoning to an auditor while diverging from the cognitive process that produced the decision.

Combined, these literatures support a structural counter to the architecture proposed in this paper:

> *Decision Packets do not preserve reasoning. They preserve post-hoc articulations that feel like reasoning to the practitioner and read like reasoning to an auditor, with cryptographic provenance attached. The hash chain guarantees the articulation was committed at time T — not that the articulation reflects the actual judgment process. The result may be worse than no record: an institutionalized confabulation layer, immune to challenge precisely because it is signed, timestamped, and chain-linked.*

This is the hardest available version of the automation-bias problem, and the paper engages it explicitly rather than concealing it.

Three architectural responses to the counter, each material and each partial:

Response 1 — Constrain what Decision Packets capture. The schema (Appendix C) does not require, and the architecture deliberately does not encourage, free-text "reasoning" fields that invite confabulation. The captured fields are structural: the pillars elicited, the alternatives presented, the assumptions surfaced, the stakeholder profiles, the time-on-question, the divergence between partners, the practitioner's confirmation timestamp, the rejection record. These are externally verifiable facts about the structure of the decision process — not internal claims about its cognitive content. Where the schema accommodates free text (`decision_reasoning`, `check_notes`, `override_notes`), the validation layer treats those fields as supplementary context, not as the canonical record of judgment. The canonical record is the structure; the prose is annotation.

Response 2 — Couple capture to outcome. The JQM feedback loop (§4.4) closes only when outcomes become observable. Until that loop closes, JQM scores are process measurements without truth-conditions — and any "reasoning" captured in Decision Packets remains hypothesis, not evidence. As outcomes accumulate, the architecture can detect *systematic confabulation patterns*: practitioners whose Decision Packets predict outcomes well are exercising genuine judgment; practitioners whose Decision Packets fail to predict outcomes are confabulating. The detection mechanism is empirical, not introspective. This is the core argument for deploying Decision Governance only in domains where outcomes are eventually observable — financial advisory, healthcare, public policy — and *not* in domains where outcomes are unobservable or arrive too late to inform calibration.

Response 3 — Preserve contestability, not certainty. Decision Packets are not certifications of correct reasoning; they are *contestable artifacts*. An auditor reviewing a Decision Packet does not ask "was this reasoning correct?" — that question is unanswerable from the record alone. The auditor asks "was the structure of the decision adequate to the stakes, given what the practitioner could have known at the time?" Confabulation degrades the value of the prose annotations but does not invalidate the structural record. Even a packet whose `decision_reasoning` field is post-

hoc rationalization still contains an honest record of *which* alternatives were considered, *which* assumptions were flagged, *whether* stakeholder divergence was captured — and these are the load-bearing structural facts that contestability requires.

What the responses do not solve. None of the three responses eliminates confabulation. They constrain its impact: by limiting what the architecture asks the human to articulate (Response 1), by deferring trust in articulations until outcomes validate them (Response 2), and by making the artifact contestable rather than authoritative (Response 3). The residual confabulation risk is real and is reported here as a known limit of the framework, not concealed. Open Research Question 8, added to Chapter 7, formalizes the empirical question: *does a Decision Packet whose `decision_reasoning` field was independently audited against the actual decision process (via concurrent verbal protocols, eye-tracking, or process-tracing) agree with the field's content above chance?* If the answer is no across a sample of practitioners and contexts, Response 1 should be strengthened — possibly to the point of removing free-text reasoning from the canonical schema entirely.

Claim level. The confabulation counter is empirically grounded and structurally robust. The three responses are architectural design choices that mitigate but do not dissolve the problem. The paper's overall claim is therefore narrower than a naive reading of Pattern 7 might suggest: Decision Packets preserve the **structure** of the decision and a **contestable record** of the practitioner's stated reasoning, with the structural fields carrying the audit weight and the prose fields carrying interpretive context that requires outcome-validation before it can be trusted as a faithful reconstruction of the actual cognitive process.

4.10 3.7b — Methodology Self-Audit: Where Producing This Paper Violates Its Own Architecture

A reflexive observation surfaced during the V1.2 review process (Doppelkritik V2 Phase 2.5 Inverted Probe Q1, May 2026): the methodology used to *produce* this paper itself violates two of the SR7D patterns that the paper *prescribes*. Disclosing the violation is the first remediation step; the structural responses are documented below.

4.10.1 Violation 1 — Pattern 1 (Digital Guardian)

The V1.2 sharpening pass orchestrated four external LLM API calls — Gemini, GPT-4o-mini, Mistral-Large, Perplexity-Sonar-Pro — each receiving the full whitepaper text (~30,000 words) plus role-specific prompts. The whitepaper content at the time of those calls was not yet published, and each API provider's terms-of-service permit some degree of input retention or training-data use. Pattern 1 (Digital Guardian) requires that user-controlled data not flow into external systems whose retention or use the user does not control. The production of this paper failed this requirement in V1.2. The mitigation step in V1.3 is to route any future external-reviewer-LLM calls through provider-tier opt-out flags (zero-retention modes) or through self-hosted models where these are technically

and economically feasible. Until that mitigation is in place, the paper carries an architectural debt against its own Pattern 1.

4.10.2 Violation 2 — Pattern 6 (Immutability First)

The four reviewer LLM outputs (`feedbacks/feedback_A.md` through `feedbacks/feedback_D.md` in the LTF-R11 folder) and the subsequent persona-agent outputs from the ce-doc-review pass were saved as mutable Markdown files on the local file system. They are not protected by a content hash, are not appended to a tamper-evident hash chain (the Pattern 6 enforcement mechanism specified in §5.3), and would not detect a post-hoc edit by the author. The same observation applies to the V1.2-Trail entries before they were extracted to the companion file. The Decision Packets the paper proposes for client-facing advisory decisions are subject to hash-chain integrity; the *meta-Decision-Packets* that produced this paper are not. The mitigation step in V1.3 is to bring the reviewer-output trail under the same Pattern-6 discipline: SHA-256 hash per file, chained with `parent_hash` pointers, validator-version stamped. Until that mitigation is in place, the paper carries an architectural debt against its own Pattern 6.

4.10.3 What This Disclosure Does and Does Not Claim

The disclosure does not claim that the V1.2 findings are invalid — the substantive findings (norm corrections in Chapter 6, math corrections in Chapter 4, confabulation-counter response in Section 3.7a, etc.) stand or fall on their own evidence, not on the integrity of their production trail. The disclosure does claim that *applying* the architecture to the *authoring* of the paper would require remediation steps the V1.2 pass did not take. The framework's Reflexive Triangulation principle (Triangulated Truth applied to the methodology itself, §3.7a) requires this self-audit to be visible to readers — concealing it would be a Pattern-7 (Archaeological Governance) violation layered on top of the Pattern-1 and Pattern-6 violations already named.

4.10.4 Remediation Tracking

The remediation is tracked as Open Research Question 9 in Chapter 7: *Can a Decision-Governance authoring pipeline be specified such that all external-reviewer-LLM calls operate under user-data-sovereignty constraints (zero retention or self-hosted), and all reviewer outputs are recorded as hash-chained immutable artifacts under the same Pattern-6 discipline the paper specifies for client-facing Decision Packets?* If the answer is yes — and a reference implementation of such a pipeline is feasible — V1.3 will adopt it and remove this self-audit's two debts. If the answer is partial (e.g., zero-retention modes available for some providers but not all), the paper will document the residual debt explicitly rather than mask it.

Claim level. The two violations are factually documented; the remediation pathway is proposed and feasible but not yet implemented. The reader should treat the V1.2 reviewer trail as a known-mutable, known-API-exposed artifact, not as a tamper-evident record. This is the same epistemic discipline the framework requires of its own Decision Packets — applied honestly to its own pro-

duction trail.

4.11 3.7c — Composite Self-Skepticism

The two preceding sections (§3.7a Confabulation Counter and §3.7b Methodology Self-Audit) and the audit-society acknowledgment that closes §6.5 each name a distinct risk that the framework licenses but does not eliminate. Read individually, each risk is bounded: the Confabulation Counter is addressed by three architectural responses with documented residual debt; the Methodology Self-Audit names two specific Pattern-1 / Pattern-6 violations and tracks remediation as RQ9; the audit-theater acknowledgment defers the detection mechanism to longitudinal calibration (RQ3). Read together, the three risks compose into a sharper challenge that no individual section licenses but that a careful reader will assemble.

4.11.1 The Composed Challenge

The composition is this. Suppose Decision Packets institutionalize confabulation rather than reasoning (§3.7a, partially mitigated but residual). Suppose further that documentation systems empirically drift toward ritual theater under sustained adoption pressure (Power 1997, §6.5 closing). Suppose finally that the architecture's own author could not apply Pattern 1 and Pattern 6 to the methodology that produced this paper (§3.7b, remediation deferred to V1.3). Now consider a deployed CFP firm: under commercial time pressure, with no audit-theater detection mechanism in production until calibration data accumulates over years, and with documented practitioner-side resistance to billable-hour-bearing documentation discipline (Appendix F Objections 1, 2, 4, and 6). The composite question follows:

> *If the framework's own author could not apply the framework faithfully to the production of the paper that specifies it, on what basis should we expect a deployed firm — operating under more commercial and time pressure than the author, with the same human cognitive constraints — to apply it any more faithfully?*

This question is not a refutation of the framework. It is the strongest available skeptical position the framework's own contents license. Naming it explicitly is a precondition for the framework to satisfy its own Constructed Ambiguity (Pattern 4) and Contestability (Ethics 3) principles applied reflexively to itself.

4.11.2 Empirical Conditions Under Which the Framework Should Be Considered to Have Failed at Its Own Standard

The composite challenge generates three empirical failure conditions. The framework should be considered to have failed at its own standard if, over a 36-month production-deployment window with at least 10 participating CFP-HNW practices:

1. **Confabulation dominance:** Independent process-tracing audits (Open RQ8) of a sample of De-

cision Packets find that the `decision_reasoning` and `check_notes` fields agree with the actual cognitive process at *or below chance levels* across practitioners and contexts. Threshold: agreement $\kappa < 0.2$ (Cohen's kappa, where 0.2 is the conventional "slight agreement" boundary). If observed: Response 1 in §3.7a (constrain capture to structural fields, drop reasoning text) must be elevated from architectural option to canonical requirement.

2. **Ritualization signature:** JQM-aggregate scores across practitioners converge to a tight high-value band (variance below 10% of theoretical range) without corresponding outcome-quality improvement on independent measures (regulatory contested-review rates, client retention, advisor-self-reported decision confidence calibrated against outcomes). If observed: §4.4 calibration loop has detected ritual theater rather than substantive process quality; the JQM weights and the GOLD constraint set both require independent re-derivation.
3. **Methodology-self-audit non-remediation:** RQ9 (the §3.7b methodology-self-audit remediation pathway) does not deliver a Pattern-1-compliant + Pattern-6-compliant authoring pipeline within the V1.3 or V1.4 release window. If observed: the framework cannot be applied to its own production at all; the Reflexive Triangulation principle (§3.7a) is not architecturally realizable; the paper's claim that the architecture is implementable should be downgraded from "instantiated (validation pending)" to "instantiable in principle, not yet demonstrated to be applicable to its own production."

If any one of these three conditions is observed, the framework's claim of being an *operational architectural discipline* should be downgraded. If two of three are observed, the framework's central architectural thesis — that Decision Governance can be enforced rather than ritualized — should be considered falsified in the deployment context studied.

4.11.3 What This Section Does and Does Not Concede

This section does not concede that the framework is currently failing on any of the three conditions; observational data is not yet available. It concedes that the three failure conditions are the empirical tests the framework has set for itself, and that the framework's adoption should be calibrated against these tests rather than against the architectural plausibility of its design. A reader who accepts the framework's design but skips this section's empirical-failure conditions has accepted only half of what the framework asks them to accept.

The composite-self-skepticism move is itself an instance of Pattern 3 (Triangulated Truth) applied reflexively: rather than relying on any single section's risk-acknowledgment, the framework's own claim to legitimacy depends on the convergence of three independent skeptical positions being honestly composed and named together. Splitting them across §3.7a, §3.7b, and §6.5 — and leaving the composition as an exercise for the skeptical reader — would let the framework optically meet its Contestability obligation while substantively understating its residual risk. This section is the explicit composition.

Claim level. The composite-risk statement above is *epistemic discipline*, not a hypothesis to be tested. The three empirical failure conditions are *risk-conditional design commitments*: the frame-

work commits in writing to specific downgrades of its own claims if specific conditions are observed in deployment. Whether the conditions will be observed is an empirical matter for the 36-month deployment window referenced above. This is the strongest form of *risk-disclosed honesty* the architecture supports without either dissolving the framework or weakening it to vacuity.

---

5 Chapter 4 — Measurement: Judgment Quality Metrics (JQM)

5.1 4.0 — The Measurement Gap

Decision science has a long history of measuring outcome quality. A portfolio's performance is measured against its benchmark. A medical diagnosis is measured against subsequent clinical findings. A legal argument is measured against the verdict. These are retrospective outcome measures: we know after the fact whether the result was good.

What the field has systematically lacked is a prospective process measure: an assessment of decision quality that does not wait for outcomes — which may be delayed by years — but evaluates the structure of the decision-making process itself, at the time the decision is made.

This gap is not an accident. It reflects a genuine philosophical difficulty: we are accustomed to measuring correctness (did the event occur as predicted?) but we lack established standards for measuring *how well the decision was made*, independent of whether it turned out to be right. A lucky guess and a well-reasoned conclusion can produce identical outcomes. A carefully structured decision process and a hasty one can both end in the same result. Outcomes alone do not tell us whether the governance was sound.

Consider the clearest illustration: a practitioner who ignores every surfaced assumption and is vindicated by a favourable outcome is indistinguishable, in result terms, from one who examined and validated each assumption before reaching the same conclusion. The difference exists entirely in the decision architecture — and is recoverable only from a record of the process itself, not from the outcome it produced.

Judgment Quality Metrics (JQM) is a proposed framework for filling this gap. It extends the measurement infrastructure of forecasting science — which has solved the related problem of measuring forecast quality prospectively — to the domain of decision governance. The extension is genuine: JQM is not Tetlock's framework renamed. It applies the core insight of proper scoring rules to a different target, and the application requires non-trivial adaptation.

5.2 4.1 — From Forecasting to Decision Quality: The Tetlock Foundation

Philip Tetlock's Superforecasting research program (2015) is the most rigorous existing framework for measuring judgment quality in prediction tasks. The core contributions are:

Calibration measures whether a forecaster's stated confidence matches their actual accuracy. A calibrator who says "I'm 80% confident" and is right 80% of the time is perfectly calibrated. A forecaster who consistently says "90% confident" but is right only 65% of the time is overconfident — a systematic miscalibration that the scoring system can detect.

Resolution measures whether a forecaster assigns meaningfully different probabilities to events that actually turn out differently. A forecaster who assigns 51% to everything has no resolution: they barely distinguish between events that will and won't occur. High-resolution forecasters make sharp predictions that subsequently discriminate.

Belief updating measures how efficiently a forecaster incorporates new information. The Good Judgment Project found that the best forecasters update more frequently and in smaller increments than average forecasters — consistent with Bayesian updating — rather than updating rarely and dramatically (anchoring) or updating constantly without signal (overreacting to noise).

The Brier score (Brier, 1950) is the primary metric that integrates these components into a single number: the mean squared error between forecast probabilities and actual outcomes. Brier scores are proper scoring rules — they incentivize forecasters to report their true probability assessments, because no strategic misreporting can improve the expected score.

The extension to Decision Quality: JQM asks whether these measurement structures can be applied not to "will event X occur?" but to "how well-structured was this decision?" The answer is yes — with important adaptations.

The key adaptation is this: forecasting events have ground truth that arrives later (the event either occurs or doesn't). Decision processes have structural properties that can be assessed immediately. A decision that was made without surfacing key assumptions is structurally worse than one that surfaced them — regardless of outcome. A decision made with no acknowledgment of stakeholder divergence is structurally worse than one that recorded it — regardless of what the client decided to do with the information. JQM measures these structural properties directly, using the same logical structure as proper scoring rules but applied to process characteristics rather than probabilistic forecasts.

5.2.1 4.1a — Triangulation as Whitepaper Methodology: A Reflexive Note

The framework's claim that triangulation across independent sources improves judgment quality creates a methodological obligation on the whitepaper itself: substantive claims should be triangulated, not asserted from a single perspective. This obligation is operationalized in the framework's authoring process through a four-role review protocol — Practitioner-Experience, Academic-Rigor,

Regulatory-Compliance, and Adversarial-Counterargument — whose independent outputs are synthesized into convergent, divergent, and orthogonal findings.

The full protocol description, the role briefs, and the source of the practitioner-objection set in Appendix F are documented in **Appendix E.2 — Four-Role Review Protocol**. The protocol's claim level: *plausibly superior* to single-reviewer review, based on convergent findings across roles, but not measured against a controlled alternative.

5.3 4.2 — Four JQM Dimensions

The four JQM dimensions are designed to be independent axes of measurement. They can each score well or poorly independently of the others. They map, as shown below, to the four principal measurement axes in Tetlock's framework.

5.3.1 4.2.1 — Assumption Coverage

Definition: The ratio of explicitly checked assumptions to total surfaced assumptions in a decision process. If the governance system surfaced twelve decision-relevant assumptions and the practitioner actively verified or challenged eight of them, Assumption Coverage = 8/12 ≈ 0.67.

What it measures: Whether the decision-maker engaged with the explicit uncertainty in their analysis, or whether they accepted the AI output's built-in assumptions uncritically. Assumption Coverage is zero for a decision made entirely on the basis of AI output accepted without examination. It is one only when every surfaced assumption was actively addressed.

Accessibility analogy: Accepting AI-generated analysis without examining its embedded assumptions is equivalent to submitting a document after dismissing every spell-checker suggestion unread. The tool surfaced the candidate issues; bypassing all of them sets Assumption Coverage to zero, regardless of how strong the underlying analysis appears. A score of one requires the practitioner to accept or reject each flagged assumption explicitly — not to trust that the AI's defaults were correct.

Tetlock mapping: Assumption Coverage maps to *resolution* in the forecasting framework. Just as a forecaster with low resolution assigns similar probabilities to very different outcomes, a decision with low Assumption Coverage treats checked and unchecked assumptions as equivalent — all are implicitly treated as valid, regardless of whether they have been examined.

SR7D connection: Constructed Ambiguity (Pattern 4) is the architectural prerequisite for Assumption Coverage: the system must surface assumptions explicitly before they can be checked. Digital Guardian (Pattern 1) provides the trust foundation: the practitioner will only engage with surfaced assumptions if they trust that the system is surfacing the genuinely relevant ones, not steering toward a preferred outcome.

5.3.2 4.2.2 — Drift Awareness

Definition: The ratio of actions taken in response to drift alerts to total drift alerts generated. If the system generated five alerts indicating that a client's stated preferences had diverged significantly from their historical behavior patterns, and the practitioner responded to four of them (by revisiting, discussing, or documenting the divergence), Drift Awareness = 4/5 = 0.8.

What it measures: Whether the practitioner is treating the client's preferences as a stable given or as a temporal trajectory that requires monitoring. Drift Awareness is zero for a practitioner who ignores all drift alerts. It is one only for a practitioner who responds to every alert — by either updating the preference model or documenting why the divergence does not require updating.

Tetlock mapping: Drift Awareness maps to *belief updating* in the forecasting framework. A practitioner with low Drift Awareness is anchoring: treating initial preference elicitation as permanent, failing to update the model when new information suggests the model is outdated.

SR7D connection: Archaeological Governance (Pattern 7) is the architectural prerequisite for Drift Awareness: without a temporal record of prior preference states, there is nothing to compare against to detect drift. The temporal dimension of Archaeological Governance — the trajectory of decisions over time — is precisely what makes Drift Awareness measurable.

5.3.3 4.2.3 — Perspective Diversity

Definition: The count of independent stakeholder profiles included in a Decision Packet for a consequential decision. For a household decision involving two partners, Perspective Diversity is maximized when both partners' value hierarchies are documented, any divergence between them is explicitly recorded, and the resolution of that divergence is part of the decision record.

What it measures: Whether the decision process incorporated multiple independent perspectives or collapsed into a single viewpoint — typically the loudest voice, the first voice, or the AI-generated synthesis that averages across perspectives without preserving their distinctness.

Tetlock mapping: Perspective Diversity maps to Tetlock's foxes-vs-hedgehogs finding: superforecasters aggregate information from multiple independent perspectives; hedgehogs rely on a single grand theory. Decisions made from a single perspective consistently underperform decisions that explicitly incorporated and reconciled multiple viewpoints.

Formal operationalization: Perspective Diversity can be formalized as the entropy of pairwise priors across user cohorts:

$$D\_perspective = H(P_1, P_2, \ldots, P_n) = -\sum p(P_i) \log_2 p(P_i)$$

where each P_i is a distinct stakeholder priority profile and $p(P_i)$ is the relative weight assigned to that profile in the decision process. High entropy means genuinely diverse perspectives were incorporated; low entropy means one perspective dominated.

The pairwise elicitation process that generates each P_i is formalized using the Bradley-Terry model (Bradley & Terry, 1952), which provides a principled maximum-likelihood method for recovering a

consistent linear ordering over alternatives from pairwise comparisons. The model was originally developed for analyzing the outcomes of sporting tournaments — specifically, the problem of ranking chess players based on game-by-game win/loss records — and was later formalized as the Elo rating system (Elo, 1978). Its application here is precise: each pairwise choice between value priorities provides an incremental update to the estimated weight vector, following the same update rules as chess player rating adjustments.

The critical property of the Bradley-Terry/Elo structure for preference elicitation is noise reduction: unlike Likert scales, which measure a point on a numerical continuum (and are subject to scale-use variance, reference-point anchoring, and mood effects), pairwise comparisons produce an ordinal tournament record from which a consistent cardinal weighting can be recovered by maximum likelihood. The method is formally equivalent to running a round-robin tournament of preferences and using the win-loss record to establish a consistent ranking.

This is the precise technical grounding for the claim in Chapter 1 that pairwise elicitation is structurally superior to Likert-scale measurement. It is not a preference — it is a consequence of the measurement model.

Caveat — transitivity assumption: Bradley-Terry imposes transitivity by model fiat: the probability that *i* beats *j* depends only on the ratio of latent strengths. Empirically, risk-preference elicitation is a domain where intransitive preferences are documented (Tversky, *Intransitive Preferences*, 1969). The Bradley-Terry fit smooths intransitivity into a consistent ordering rather than flagging the violation. In production deployments, the reference architecture monitors the empirical fit residuals; sustained high residuals indicate a stakeholder whose preferences may not be transitive in this elicitation context, and the Confidence-Band (Pattern 4) for the recovered ordering is widened accordingly. The model assumption is preserved as a working idealization, with its empirical violations surfaced rather than silently absorbed.

Caveat — multi-stakeholder aggregation and Arrow's Impossibility. The Bradley-Terry recovery above operates *within* a single stakeholder's pairwise responses. The Perspective Diversity formalization that follows (entropy across stakeholder profiles) operates *across* stakeholders. Arrow's Impossibility Theorem (Arrow, *Social Choice and Individual Values*, 1951; sharpened by Sen, *Collective Choice and Social Welfare*, 1970) establishes that no aggregation rule across ≥3 alternatives can simultaneously satisfy universal domain, Pareto efficiency, independence of irrelevant alternatives, and non-dictatorship. In a two-partner household with three or more pillars — *which is exactly the case-study context of Chapter 6* — Perspective Diversity cannot be a well-behaved aggregation under Arrow's standard axioms. The framework's response is not to claim the impossibility has been dissolved (it has not) but to *relax* the IIA axiom deliberately: the architecture preserves and *displays* the stakeholder divergence rather than aggregating it into a single household preference. The "Perspective Diversity" score measures the entropy of the divergence, not the legitimacy of a particular aggregation. This is an Arrovian relaxation, not an Arrovian dissolution; readers familiar with computational social choice should read the score in this light.

5.3.4 4.2.4 — Question Depth

Definition: The ratio of answered to generated questions in a decision process. If the governance system generated eighteen relevant questions based on the analysis and the client's stated goals, and the practitioner addressed fourteen of them in the advisory session (either by answering them, explicitly noting them as unresolvable, or documenting them as out of scope), Question Depth = 14/18 ≈ 0.78.

What it measures: Whether the advisory process genuinely explored the decision space or focused narrowly on the most salient questions while leaving important questions unaddressed. High Question Depth indicates comprehensive exploration; low Question Depth indicates selective engagement.

Tetlock mapping: Question Depth maps to the granularity finding in forecasting research: the most accurate forecasters ask more specific, finer-grained questions than average forecasters. They decompose a complex prediction into its components and address each component explicitly. Decision quality follows the same pattern: practitioners who address more of the decision-relevant questions consistently produce better-structured decisions.

SR7D connection: SPECTRADING (Pattern 2) calibrates the system's role in question generation: in low-stakes interactions, the system may generate questions automatically; in high-stakes, value-laden interactions, question generation involves the practitioner as active collaborator. Question Depth measures whether that collaboration was substantively engaged or superficial.

5.4 4.3 — Proper Scoring Rules for Decision Quality

A critical property of good measurement systems is that they incentivize honest reporting. Tetlock's Brier score is a *proper scoring rule* in the sense that a forecaster who reports their true probability estimate always maximizes their expected score, regardless of what that estimate is. Strategic misreporting always results in a worse expected score — because the score is a function of probabilistic forecasts compared against observed outcomes.

JQM is **structurally different**: its dimensions (Assumption Coverage, Drift Awareness, Perspective Diversity, Question Depth) are *process-event ratios computed from immutable Decision Packet records*, not probabilistic forecasts over outcomes. Calling JQM a "proper scoring rule in the Brier sense" would be a category error. JQM is, instead, a **tamper-resistant process measurement framework**: its honesty incentive does not come from a strategic-equilibrium property of the scoring function, but from the architectural enforcement of Immutability First (Pattern 6) and Archaeological Governance (Pattern 7). A practitioner cannot inflate AC/DA/PD/QD by misreporting; they can only improve their scores by improving their actual process, because the input data (the Decision Packet field values) is computed from append-only records they cannot retroactively edit. Practitioners should benefit from honest reporting of assumption uncertainty, genuine acknowl-

edgment of stakeholder divergence, and substantive engagement with difficult questions — not from appearing to have engaged while actually shortcutting the process.

The practical enforcement of this property depends on the measurement context. In a self-assessment context (practitioner reporting their own process quality), proper scoring rules alone are insufficient — there is no external validator. In an institutional context where JQM scores are used for performance evaluation or regulatory compliance, the scores need to be derived from independently verified records (Decision Packets), not from self-report.

This is precisely the function of Immutability First (Pattern 6) and Archaeological Governance (Pattern 7) in the context of JQM: the Decision Packet record is the ground truth from which JQM metrics are computed. A practitioner who wishes to improve their JQM scores must actually improve their decision process — because the process is recorded in an append-only, tamper-evident chain, not in a self-reported form.

The Decision Packet chain functions, in this respect, as a structurally unbribable referee: it records not only what the practitioner concluded but the sequence and timing of every interaction that produced the conclusion. A practitioner who wishes to appear to have engaged with assumptions they did not actually examine faces an architectural barrier — the append-only record distinguishes genuine engagement from retroactive appearance of engagement, because the interaction timestamps are fixed at the moment they occurred and cannot be rewritten.

Aggregate process score: A weighted process-measurement aggregate is proposed as:

$$JQM\_aggregate = \alpha_1 \cdot AC + \alpha_2 \cdot DA + \alpha_3 \cdot PD + \alpha_4 \cdot QD$$

where AC, DA, PD, QD are the four dimension scores and α_1 through α_4 are domain-specific weights that reflect the relative importance of each dimension in the decision context. **Reference-deployment values:** the reference implementation uses uniform weights $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$ pending the longitudinal-calibration study specified in RQ3. The uniform-weights choice is a deliberate placeholder, not a derived optimum: it withholds claim-of-knowledge about per-dimension importance until calibration data exists, and it preserves backwards-compatibility when RQ3 yields domain-specific weights (re-weighting a uniform baseline is mechanically simpler than re-weighting an opinionated initial vector). In financial advisory, Perspective Diversity (diverging risk tolerances between partners) and Drift Awareness (preference changes over a multi-year advisory relationship) may warrant higher weights than in a single-session context once calibration evidence is available. The weighting function is a parameter of the deployment context, not a universal constant.

Claim level: The aggregate scoring function is proposed formalization. The individual dimension definitions are operationally testable. The weight optimization is an empirical question (Research Question 3 in Chapter 7).

5.5 4.4 — JQM as a Living Feedback System

The most important property of JQM is not its ability to score any individual decision, but its ability to improve governance quality over time through structured feedback.

The feedback loop operates as follows:

1. **Decision Packet created:** practitioner documents AI analysis, human decision, key assumptions, stakeholder perspectives, open questions
2. **JQM computed:** four dimension scores derived from the Decision Packet record
3. **Outcome recorded:** the actual result of the decision (portfolio performance, client satisfaction, regulatory compliance status) is appended to the packet chain when available
4. **Calibration:** the relationship between JQM scores and decision outcomes is measured over accumulating records, providing empirical evidence for or against the validity of each dimension
5. **Adjustment:** dimension weights and scoring parameters are updated based on the calibration evidence

This is the decision quality equivalent of the forecasting calibration loop: assess the quality of the prediction → observe the outcome → update the scoring model → produce better predictions. The difference is that the calibration target is the decision process, not a point forecast, and the outcome variable is multi-dimensional.

The feedback loop has a practical implication that extends beyond individual practitioners: at the institutional level, an organization with accumulating Decision Packet records across multiple advisors has a unique empirical asset — the data needed to determine which aspects of decision process quality actually predict good decision outcomes in their specific domain. No external research study can provide this, because the relationship between process quality and outcome depends on the specifics of the domain, the client population, and the institutional context. The longitudinal JQM record is the institution's own calibration data.

Non-Normativity (Ethics 1) is essential in this context: JQM is a mirror, not a judge. The system does not evaluate whether the practitioner made the right decision — there is no external standard for that. It evaluates whether the process was well-structured. The difference matters: a practitioner who made a well-structured decision that turned out badly has produced evidence against the weighting model, not evidence of personal failure. Agency Preservation (Ethics 2) ensures that the practitioner retains full authority over what to do with the JQM output — whether to adjust their process, contest the measurement, or document why their specific context warrants deviation from the standard scoring model.

5.6 4.4a — Operational Subcomponent: Stat-Prior Confidence

Practitioner summary (skip the technical content below if you do not need it): Stat-prior confidence is the reference architecture's way of telling you whether a statistical assumption baked into an AI-generated analysis is well-grounded or weakly-grounded. It does so by checking three things: how much data the prior was calibrated on (more data → higher confidence), whether the unit system matches your context (monthly vs annual, euros vs dollars — mismatch produces an immediate hard-zero that blocks the calculation), and whether the calibration window covers your decision's reference period (out-of-window also produces a hard-zero). What you see as a practitioner: a confidence flag on each prior used in the analysis. What the system refuses to do: silently combine priors with mismatched units. The mathematics below documents the formula and the failure-mode handling.

The four JQM dimensions developed in Sections 4.2-4.4 are the framework's primary measurement layer. Below them, an operational subcomponent has been instantiated in the reference architecture: a per-prior confidence score that allows the Assumption Coverage dimension (Section 4.2.1) to be computed from actual numerical priors rather than from binary checked/unchecked classification alone.

5.6.1 4.4a.1 — Functional Form

For each statistical prior used in a Decision Packet's analytical component, a confidence score `c` in `[0, 1]` is computed from three quantities:

- *n* = the sample size on which the prior is calibrated
- *u* in {0, 1} = a unit-consistency flag (binary; the prior either matches the analytical context's unit system or it does not)
- *t* in {0, 1} = a temporal-validity flag (binary; the prior's calibration window either covers the decision's reference period or it does not)

The reference implementation computes:

$$c = u \cdot t \cdot \min(1.0, \max(0.0, (\log(n) - \log(n_{min})) / (\log(n_{max}) - \log(n_{min}))))$$

where `n_min` and `n_max` are deployment-configurable bounds (the reference deployment uses `n_min = 30`, `n_max = 10000`). The upper bound `n_max` addresses a non-obvious failure mode: very large datasets can create *false precision* — a prior calibrated on a million data points appears maximally reliable, but if a substantial fraction of those points pre-date a material shift in the underlying distribution (market regime change, regulatory amendment, client-population demographic shift), the dataset volume amplifies the confidence of a stale model rather than validating a current one. The logarithmic scaling in the formula above deliberately caps confidence gains above `n_max`, preventing practitioners from substituting data volume for distributional currency. **Value-level justification:** `n_min = 30` reflects the conventional CLT-heuristic threshold for sample-mean approximation under typical financial-advisory prior distributions (returns, risk-tolerance, household-budget items — all

approximately normal at n ≥ 30 in calibration data). `n_max = 10000` is the empirical knee in the reference deployment's confidence-vs-sample-size curve: above 10,000 samples, marginal confidence gains drop below 0.5%/log-decade and a flat ceiling avoids spurious confidence-precision claims at sample sizes where calibration drift dominates true measurement gain. Both values are conservative starting points; deployments with domain-specific priors (e.g., insurance actuarial datasets with much larger calibration samples) should re-derive the values rather than inherit them. The hard zeros on `u` and `t` reflect the design judgment that unit mismatches and temporal-window mismatches are non-recoverable failures, not soft-weightable factors.

5.6.2 4.4a.2 — Why Hard-Zero on Unit Mismatch

The hard-zero behavior on the unit-consistency flag is the consequential design choice. It originated from a specific failure mode observed during calibration: a prior calibrated on monthly income figures was incorrectly applied in an annual-income context, producing an analytical output that was technically valid but semantically meaningless. Soft-weighting the unit mismatch would have masked the failure; hard-zeroing surfaces it as an immediate `c = 0`, which prevents downstream calculations from completing and forces the practitioner to resolve the mismatch before proceeding.

This failure mode is itself an Archaeological Governance artifact: the original Decision Packet, the corrected packet referencing it via parent_hash, and the documented confidence-score modification together form an audit trail demonstrating that the architecture detected its own measurement failure rather than concealing it. This is the pattern the JQM framework is designed to support: not infallibility, but traceable correction.

5.6.3 4.4a.3 — JQM-Connection and Claim Level

Assumption Coverage (Section 4.2.1) was originally defined as the binary ratio of checked-to-surfaced assumptions. The stat-prior confidence score allows a graded refinement: an assumption verified against a high-confidence prior contributes more to Assumption Coverage than one verified against a low-confidence prior.

Claim level. Stat-prior confidence as defined here is an operationally tested measurement component (the reference implementation includes edge-case tests for `n = 0`, `u = 0`, `t = 0`, and the boundary `n = n_min`). Its use as a graded extension of Assumption Coverage is *proposed formalization* pending validation against decision-outcome data — see Chapter 7, RQ3.

5.7 4.5 — What JQM Adds to the Governance Architecture

JQM closes the measurement gap identified in Section 4.0: it provides a prospective process measure that does not require waiting for outcomes. In doing so, it transforms Decision Governance from a documentation system into a learning system.

Without JQM, the governance architecture can ensure that decision processes are documented (Archaeological Governance), that records are immutable (Immutability First), and that uncertainty is communicated (Constructed Ambiguity). These are valuable properties. But they are static: the documentation improves the quality of the record, not the quality of the decision.

With JQM, the governance architecture acquires a feedback mechanism. Each decision produces not only a record but a score. The score makes the quality of the decision process visible — to the practitioner, to the institution, and, in appropriate contexts, to the client. Visible quality can be improved. Invisible quality can only be hoped for.

Chapter 5 addresses the enforcement side of this architecture: how deterministic mechanisms ensure that the JQM measurement cannot be gamed, and how the governance guarantees of Immutability First and Principled Inefficiency create the conditions under which JQM scores are trustworthy rather than performative.

---

6 Chapter 5 — Deterministic Cores in AI-Abundant Environments

---

6.1 5.0 — The Verification Problem

The governance architecture developed in preceding chapters requires validation. The Governance Gap (Chapter 1) demands that the transition from AI output to human decision be structured. The SR7D framework (Chapter 2) specifies ten constraints that govern this transition. The information-theoretic foundation (Chapter 3) grounds the externalization protocols in formal epistemics. JQM (Chapter 4) provides measurement.

None of this means anything if the governance constraints can be violated without detection. A Non-Normativity requirement that depends on developer discipline is not a constraint — it is a guideline. An Immutability First guarantee that can be overridden by an administrator is not immutability — it is a soft recommendation. A Contestability commitment that has no mechanism for verifying provenance is not contestability — it is theater.

This chapter addresses the enforcement layer of Decision Governance. The central claim is architectural: the governance constraints specified in the SR7D framework must be enforced by deterministic mechanisms, not by probabilistic ones. This claim has a specific technical meaning and a specific practical implication, both of which deserve careful development.

6.2 5.1 — Why Probabilistic Evaluation Is Insufficient

The dominant paradigm in AI evaluation as of 2026 is probabilistic: use AI systems to evaluate AI systems. LLM-as-judge frameworks ask a capable language model to assess whether another model's output meets specified quality criteria. The EQUATOR framework, DeepEval, and related tools formalize this approach, providing rubrics that guide LLM evaluators toward consistent scoring.

These frameworks represent genuine progress. But they have a structural limitation that makes them insufficient as the sole enforcement mechanism for decision governance.

The compounding uncertainty problem: A probabilistic evaluator assessing probabilistic output compounds the uncertainty of both. If the generating model has some probability p_1 of producing normative language in a given context, and the evaluating model has some probability p_2 of detecting normative language given that it is present, the probability of an undetected normative output is approximately $p_1 \times (1 - p_2)$. Neither p_1 nor p_2 is zero, and neither is one. The product is nonzero. In a system that processes thousands of decision interactions, a nonzero false-negative rate translates to a guaranteed rate of governance failures — violations that the evaluation layer missed.

This is not a criticism of probabilistic evaluation tools. They are appropriate for many tasks. It is a statement about the requirements of governance enforcement: in safety-critical systems, a nonzero false-negative rate for safety violations is unacceptable. Consider why we trust a suspension bridge: not because the engineers hoped everyone followed the specifications, but because the mathematical tension of steel cables makes failure physically impossible — structural certainty is baked into the geometry of the system, not into the goodwill of its builders. Aviation does not rely on probabilistic evaluation of whether the landing gear deployed — it deploys sensors that deterministically confirm or deny gear extension. The evaluation is binary, not probabilistic.

The LLM inconsistency problem: Language models exhibit prompt sensitivity, context sensitivity, and version-to-version variation. An LLM evaluator that correctly identifies normative language in 98% of cases in one deployment context may perform differently with a different system prompt, a different model version, or a different context window. Governance enforcement cannot have this property: the behavior of the enforcement layer must be stable and predictable across all contexts.

The appropriate conclusion is not that probabilistic evaluation is worthless — it is that governance enforcement requires a *deterministic floor* beneath the probabilistic systems. Probabilistic tools evaluate quality, generate insights, and flag potential issues; deterministic mechanisms enforce the non-negotiable constraints.

6.3 5.2 — The Safety Engineering Precedent

Safety engineering has spent decades developing architectures for exactly this problem: how do you build reliable systems from components that individually fail? The approach is well-established and the principles transfer directly to decision governance.

Safety Cases are formal documents that argue, systematically, for the safety of a given system. Originally developed in aviation (EASA, FAA certification frameworks) and nuclear power, a Safety Case makes a structured claim: "This system is safe to operate in this context, and here is the evidence for each component of that claim." The structure requires explicit statements of assumptions, explicit identification of hazards, explicit documentation of mitigations, and explicit acknowledgment of residual risks.

The structural isomorphism with Decision Packets is exact:

Safety Case Element	Decision Packet Equivalent
System purpose and operational context	Advisory context and client situation
Hazard identification	Risk factors and assumption vulnerabilities
Mitigation evidence	Surfaced assumptions + stakeholder perspectives
Residual risk acknowledgment	Confidence bands and unresolved open questions
Argument structure	Reasoning chain from AI analysis to human decision
Sign-off and authorization	Practitioner commitment + JQM record

The aviation industry does not rely on pilots remembering to document safety considerations — it designs systems in which the documentation is structurally required, architecturally enforced, and impossible to bypass without triggering detectable anomalies. Decision Governance imports this design philosophy: the Decision Packet is not a form to be filled in when convenient. It is the artifact that the governance architecture produces as an invariant of every consequential decision process.

Hollnagel's Safety-II perspective adds an important nuance (Hollnagel, 2014). Safety-I (traditional safety engineering) focuses on preventing failures: identify hazards, specify requirements, certify compliance. Safety-II recognizes that complex sociotechnical systems are too dynamic for complete hazard specification — safety in practice comes from systems that adapt and recover, not only from systems that prevent. The implication for Decision Governance is that deterministic enforcement should focus on the invariants that are genuinely non-negotiable (Non-Normativity, Immutability), while JQM feedback mechanisms (Safety-II style) handle the adaptive, contextual quality improvements that cannot be fully specified in advance.

6.4 5.3 — The Deterministic Shell Architecture

The architecture that integrates these principles can be described as a Deterministic Shell around a Probabilistic Core. The pattern has four layers:

6.4.1 Layer 1: Probabilistic Generation (LLM Core)

The language model generates scenario analyses, pillar extractions, Monte Carlo projections, narrative summaries, and question sets. This generation is irreducibly probabilistic: the model's outputs are samples from a learned probability distribution over text, conditioned on the input context. This is appropriate — the value of generative AI in decision support is precisely its ability to explore the solution space, surface unexpected considerations, and produce fluent natural language explanations.

No deterministic enforcement operates at this layer. Generation is free to be probabilistic.

6.4.2 Layer 2: Deterministic Validation (Constraint Enforcement)

Before any AI-generated output reaches the practitioner interface, it passes through a deterministic validation layer that enforces the non-negotiable governance constraints.

The primary enforcement mechanism is a Pydantic contract — a schema validation specification that the reference architecture calls `sr7d_output_contract`. This contract enforces, at minimum:

1. **Non-Normativity check:** Scan output for normative language markers from the forbidden lexicon: `["should", "must", "recommend", "you need to", "the best option is", "you ought to", "I suggest", "advise"]`. Any output containing a match is rejected before display — not flagged for review, rejected. The practitioner sees a system notification that the output was blocked for Non-Normativity violation, and the blocked output is logged for audit purposes. The system then requests a regeneration without the normative framing.
2. **Schema conformance check:** Verify that the output contains all required Decision Packet fields (see Appendix B for the full field specification). Outputs missing required provenance, confidence indicators, or assumption lists are rejected with a specific missing-field error.
3. **Completeness checks:** Verify that quantitative claims are accompanied by confidence intervals, that risk assessments identify both upside and downside scenarios, and that multi-stakeholder contexts include separate profiles where more than one stakeholder is present.

This validation layer is deterministic in the technical sense: for any given input, the validation produces a binary pass/fail result and a deterministic error message. There is no probabilistic threshold, no LLM-judged "is this normative enough?", no contextual override. The check runs, the result is binary, the enforcement is unconditional.

Accessibility analogy: The validation layer functions as a structurally unfeeling bouncer at a high-security facility. The bouncer does not evaluate the nuance of an argument for why the visitor forgot their ID badge — if the badge is absent, entry is refused. There is no probabilistic threshold,

no 'close enough', no contextual override. An AI output missing a required Decision Packet field is refused at the gate; the practitioner sees a specific missing-field notification, not a softly degraded output.

The full Validator Spec is documented in Appendix B.

6.4.3 Layer 3: Cryptographic Security (Tamper-Evident Hash Chain)

Once a Decision Packet passes validation and is committed by the practitioner, it is written to an append-only tamper-evident hash chain. Each packet contains:

- A content hash of the packet's data
- A `parent_hash` referencing the preceding packet in the chain
- A timestamp verified against an external time source
- The validator version that processed the packet

The chain integrity property is: any modification to a historical packet breaks the hash chain at the modification point, making tampering structurally detectable. To be concrete: if a practitioner modifies even a single character in a historical Decision Packet — a comma, a digit, a space — the content hash of that packet changes. Because that hash was embedded in the fingerprint of the next packet in the chain, and that fingerprint in the packet after that, the modification propagates as a structural break through every subsequent record. The tampering is not detected by an auditor reviewing documents — it is detected by basic arithmetic at the point of chain verification. This is not a blockchain — it requires no distributed consensus — but it provides the same integrity guarantee for a single-institution deployment: Immutability First (Pattern 6) is enforced not by policy but by mathematics.

6.4.4 Layer 4: Deterministic Measurement (JQM Computation)

JQM metrics are computed from the Decision Packet record using deterministic functions. Given the same Decision Packet, the computation always produces the same score. There is no LLM evaluation at this layer — the metrics are derived from the structured fields in the packet (count of surfaced assumptions vs. checked assumptions, count of drift alerts vs. responses, number of distinct stakeholder profiles, count of generated questions vs. addressed questions).

The deterministic nature of JQM computation is what makes the scores trustworthy. A practitioner who wants to improve their Assumption Coverage score must actually increase the ratio of checked to surfaced assumptions in their Decision Packets. There is no other path to improving the score.

6.4.5 Production Hardening: Frozen Constraint Sets

Layer 2's deterministic validation (the `sr7d_output_contract` Pydantic specification) describes the enforcement mechanism. Production deployment surfaces a second-order requirement: the con-

straint set itself must be governed.

A reference implementation operates with a frozen-constraint pattern: a designated constraint set (informally referred to as the "GOLD" set in the reference deployment) is checked into version control, signed by a designated reviewer, and treated as immutable for the duration of a deployment cycle. Modifications to the GOLD set are themselves Decision Packets: a proposed change is documented with rationale, the architectural implication of the change is analyzed, and the change is accepted only through an explicit governance gate (not through a routine pull-request review).

This pattern addresses a failure mode that becomes visible only in production: the constraint set is itself a governance artifact. If a normative-lexicon entry can be silently removed by a developer who finds the constraint inconvenient, the deterministic enforcement layer is no longer deterministic in the governance sense — it has become a configuration option. Treating the constraint set as a versioned, signed, change-controlled artifact closes this loop. The implication is governance all the way down: not only must the AI outputs be governed, but the governance rules themselves must be governed, and the governance of the governance rules must be governed. Each layer is a Decision Packet; each change to a lower layer is itself subject to the constraints of the layer above.

The pattern has been audit-tested in retrospective compliance reviews (the reference deployment uses BaFin-style retro-audit drills, in which a sample of historical outputs is re-validated against the current and prior GOLD sets to detect any constraint drift that would have allowed a previously rejected output to now pass). A drift detection without corresponding governance-gate documentation is treated as an audit finding.

This is a deployment pattern, not an architectural extension to the Deterministic Shell described in Section 5.3. It is documented here because the gap between "the architecture supports deterministic enforcement" and "the enforcement actually remains deterministic in production over time" is the kind of gap that governance frameworks frequently fail to close.

6.5 5.4 — Machine-Enforceable Non-Normativity: A Proof of Concept

The enforcement of Non-Normativity is the most distinctive governance capability in this architecture and deserves specific treatment.

Non-Normativity (Ethics 1) is the ethical guardrail that prevents the governance system from substituting its judgment for the user's. It is the hard line between a decision support system — which structures the decision — and a recommendation engine — which makes the decision. As Chapter 2 established, the moment the system recommends, it has substituted AI judgment for human judgment, and the governance architecture fails at its most fundamental purpose.

The standard approach to enforcing Non-Normativity is developer training and code review: developers are instructed not to include prescriptive language, and outputs are reviewed for compliance. This approach depends on human attention and has a nonzero failure rate.

The architecture enforces Non-Normativity at the schema level: every AI-generated output that includes text intended for the practitioner interface must pass through `sr7d_output_contract` before display. The contract's normative lexicon check runs deterministically, at every output, without exception. No configuration switch disables it. No "expert mode" bypasses it. The check is not a feature — it is an invariant.

This is unique in the field. Responsible AI guidelines routinely specify "do not make prescriptive recommendations to users." No existing framework, to the authors' knowledge, enforces this specification at the architectural level as a machine-checkable structural constraint. The reference architecture does.

Why machine enforcement matters beyond technical compliance: The Non-Normativity invariant has a second function beyond the immediately obvious one. When practitioners know that the system cannot produce recommendations — not "is designed not to" but "physically cannot, as a structural constraint" — their relationship with the system's outputs changes. They know, with certainty, that every AI output is a structured presentation of information, not a steered choice. This knowledge is itself a governance property: it preserves the practitioner's sense of authorship over their decisions. Agency Preservation (Ethics 2) is not only a behavioral goal — it is architecturally enabled by the certainty that the system cannot be making choices on the practitioner's behalf.

6.6 5.5 — What Cannot Be Made Deterministic

Intellectual honesty requires acknowledging the limits of deterministic enforcement.

Context judgment cannot be made deterministic. Whether a specific AI output is, in context, practically normative even if it passes the lexical Non-Normativity check — whether a framing that avoids the forbidden words nonetheless steers the user toward a predetermined conclusion — requires human judgment. The lexical check catches the clear cases. Sophisticated edge cases require review. A sufficiently capable model can frame the downside of one option catastrophically and the upside of another brilliantly — leading the practitioner to a predetermined conclusion without ever triggering the forbidden-lexicon check. This is the practical limit of lexical enforcement: it catches the lazy violation, not the sophisticated one. Human vigilance, not architectural constraints, is the defence at this layer.

Quality improvement cannot be made deterministic. JQM can measure the structure of decision processes; it cannot determine what the optimal structure is for a given decision in a given context. High JQM scores are necessary but not sufficient for good decisions. They are measures of process discipline, not measures of wisdom.

Stakeholder judgment cannot be delegated to the enforcement layer. Whether the client's stated preferences in a Decision Packet actually reflect their genuine goals — whether Drift Awareness alerts indicate genuine preference change or measurement artifacts — requires practitioner judgment that the architecture supports but cannot replace.

The Deterministic Shell architecture is not a claim that governance can be automated. It is a claim that the non-negotiable invariants — Non-Normativity, Immutability, provenance tracking — can and should be enforced mechanically, freeing the practitioner to exercise judgment in the domains where judgment is irreplaceable.

6.7 5.6 — The Key Differentiator

The observation bears restating as a summary:

Every major technology provider and professional services firm currently publishes "responsible AI" principles. These principles are policy documents. They specify what the system should and should not do. They depend on compliance.

The reference architecture described in this paper enforces its governance constraints architecturally. Non-Normativity is not a principle — it is a Pydantic contract that runs at every output. Immutability is not a commitment — it is a tamper-evident hash chain that makes tampering cryptographically detectable. JQM is not a self-assessment — it is computed from append-only records.

This is the distinction Chapter 2 drew between policy and architecture. Policy says "decisions should be transparent." Architecture makes opacity structurally difficult. In safety-critical domains — and financial advisory is safety-critical, with real consequences for clients who receive poorly governed advice — the architectural approach is not a premium option. It is the appropriate standard.

Chapter 6 demonstrates this architecture in the domain where the governance gap is most measurable and where the regulatory requirements most directly align with the constraints the SR7D framework already enforces.

---

7 Chapter 6 — Case Study: Financial Advisory under the EU AI Act

---

7.1 6.0 — Why Financial Advisory First

The SR7D framework and its associated architecture are claimed to be general — applicable wherever consequential decisions are made with AI involvement and where accountability, traceability, and human oversight are structurally required. Chapter 7 will address the case for generalization across domains.

Before making that case, the framework requires validation in a specific domain. Financial advisory is the appropriate first domain for several reinforcing reasons.

Regulation is concrete and current. The EU AI Act (in force 2024), BaFin MaComp (Circular 5/2018), and MiFID II provide specific, enforceable requirements for transparency, documentation, and human oversight in AI-assisted financial advisory. These requirements are not future projections — they are current obligations. A framework that claims to meet these requirements can be tested against them today.

The stakes are individual and measurable. Unlike organizational or policy decisions, financial advisory decisions have direct, measurable consequences for identifiable individuals. A portfolio recommendation that fails to account for genuine risk tolerance impairs real retirement security. An AI that processes a mortgage application in ninety seconds and produces a mathematically precise result — while the human professional who overrides or accepts that output leaves no structured record of why — illustrates both the stakes and the gap in a single event. This is not a thought experiment.

The Governance Gap is uniquely visible. In financial advisory, the gap between AI-generated analysis and human-committed advice is documented — not because practitioners keep good records, but because regulation requires some record-keeping, which makes the inadequacy of current records visible. The gap between what is required and what most advisory practices actually produce is the clearest existing evidence that the governance infrastructure problem is not theoretical.

The fiduciary duty creates structural alignment. Financial advisors bound by fiduciary duty are legally and ethically required to act in their clients' best interests. Digital Guardian (Pattern 1) is, architecturally, fiduciary duty translated into a system invariant. The domain's existing ethical infrastructure creates a foundation on which the governance architecture can build.

Segment scope of the first validation. Within financial advisory, this paper's reference implementation targets the high-net-worth (HNW) segment served by Certified Financial Planner (CFP) practitioners as its initial validation context. *Definitional anchors:* HNW is operationalized for this study as households with investable assets above the local segment threshold used in the calibration data (in the German market, this corresponds approximately to households above 500,000 EUR in liquid financial assets, though the precise threshold varies by jurisdiction); CFP refers to the Certified Financial Planner credential as issued by national CFP boards (CFP Board in the US, FPSB-member organizations elsewhere) and recognized within the EU MiFID II advisory framework.

The choice of this segment is empirical, not architectural: HNW-CFP engagements are characterized by (a) extended advisory relationships that produce sufficient decision history for Drift Awareness (JQM Dimension 2, Chapter 4) to be measurable, (b) multi-stakeholder household decisions that exercise Perspective Diversity (JQM Dimension 3), and (c) regulatory documentation requirements that already apply pressure on the existing record-keeping infrastructure, making the governance-gap pathology observable in current practice rather than hypothetical. Statistical priors used in the reference implementation are calibrated on this segment; their direct transfer to mass-market

or institutional advisory contexts would require re-calibration and is not claimed in this paper. *Reader navigation:* the rest of Chapter 6 can be read in two complementary orders. (a) Linear: §6.1 regulatory framework → §6.1a competitive landscape → §6.2 CFP scenario → §6.2a practitioner-objections summary → §6.5 legal-forensics → §6.6 cross-domain → §6.6a companion-documents → §6.7 discovery-call decision page. (b) Practitioner-direct: §6.2 (scenario for self-recognition) → §6.2a (objections to compare against your own) → Appendix F (full objection catalogue) → §6.7 (decision page). The §2.0a Pain-to-Pattern Map provides a third entry-point if you prefer pattern-driven navigation. *Regulatory-supervision caveat:* the regulatory mapping in §6.1 and §6.5 assumes BaFin/MaComp applicability (WpHG-licensed institutions). DACH practitioners operating under § 34h GewO (Honorar-Finanzanlagenberater) or § 34f GewO (Finanzanlagenvermittler) are supervised by Industrie- und Handelskammern (IHKs) under FinVermV; for these practitioners, the Decision-Packet architecture remains regime-neutral but the specific norm-references in this paper must be substituted with the corresponding FinVermV obligations (§ 12-22 FinVermV). The case for cross-segment generalization is taken up in Chapter 7.

7.2 6.1 — The Regulatory Framework

The regulatory landscape for AI in financial advisory has three layers with distinct but complementary requirements.

7.2.1 EU AI Act (Regulation (EU) 2024/1689 — Articles 12, 13, 14)

The EU AI Act (Regulation (EU) 2024/1689) does **not** classify all AI used in financial services as high-risk by default. Annex III enumerates the high-risk application categories; relevant for financial services are specifically (i) credit scoring of natural persons (Annex III §5(b)) and (ii) life- and health-insurance risk assessment and pricing (Annex III §5(c)). AI-assisted investment advisory is **not** explicitly listed in Annex III. Where AI-assisted investment advisory materially affects access to an essential private service it may, under Article 6(3) and Annex III §5 read together, become subject to high-risk obligations; routine portfolio advisory does not automatically trigger them. The Decision-Governance architecture anticipates regulatory expansion in this direction while remaining applicable across the risk spectrum.

For systems that *are* classified as high-risk, three Articles together establish the documentary requirements the architecture targets:

Article 12 — Record-keeping (automatic logging). Article 12(1) requires high-risk AI systems to be designed to allow for the automatic recording of events (logs) over their lifetime. Article 12(2) specifies the purposes of these logs: identification of risk situations (Article 79), post-market monitoring (Article 72), and supervision of operations (Article 26(5)). For Annex III §1(a) systems, Article 12(3) lists minimum log contents (use period, reference databases, input data with matches, identity of verifying persons). Article 12 — not Article 14 — is the locus of the logging obligation.

Article 13 — Transparency. Article 13(1) requires high-risk AI systems to provide information sufficient for users to understand the system's capabilities, limitations, and intended operational context. Article 13(3)(b)(ii) requires that information accompanying the system include the level of accuracy, robustness and cybersecurity against which the system has been tested — a direct mandate for Constructed Ambiguity (Pattern 4) and confidence-band communication.

Article 14 — Human Oversight. Article 14(1)–(4) requires high-risk AI systems to be designed to enable effective human oversight. Article 14(4) lists the capabilities oversight personnel must possess: understanding the system's capabilities and limitations (lit. a), remaining aware of automation bias (lit. b), correctly interpreting outputs (lit. c), being able to decide not to use the output or to override it (lit. d), and being able to intervene or stop the operation (lit. e). The Decision Packet (Appendix C) is the architectural artifact that operationalizes (a)–(e): the packet captures what the practitioner understood about the AI output, what alternatives were considered, and on what basis the output was accepted, modified, or rejected.

These obligations together — automatic logging (Article 12), transparency about accuracy (Article 13), and human-oversight artifacts (Article 14) — are not satisfied by end-to-end encrypted model logs. They require human-interpretable records that connect AI-generated analysis to human-committed decisions. That is precisely the function of the Decision Packet architecture.

7.2.2 BaFin MaComp (Circular 5/2018)

Germany's Federal Financial Supervisory Authority's minimum standards for compliance function in investment services (MaComp) establish organizational requirements for advisory documentation that go beyond the EU AI Act's transparency focus. MaComp requires that investment advice be documented in a way that allows reconstruction of the advisory process by a third party — including the information base used, the client's stated situation and goals, and the recommendation made with its rationale.

This reconstruction requirement is Archaeological Governance (Pattern 7) as a regulatory mandate. The regulatory requirement predates the AI Act by six years and is in some respects more specific: not just "log the AI output" but "document the full reasoning chain from client situation to recommendation."

7.2.3 MiFID II Suitability Requirements

MiFID II Article 25 requires that investment firms assess the suitability of investment advice for the individual client, including the client's risk tolerance, investment objectives, financial situation, and investment knowledge. Critically, MiFID II requires that the suitability assessment be documented and provided to the client.

The suitability documentation requirement is a direct mandate for the pillar extraction and value hierarchy documentation embedded in the reference architecture's onboarding flow. A suitability assessment that documents "client stated moderate risk tolerance" is MiFID II compliant; a suitability assessment that documents the full pairwise elicitation process, the entropy-weighted value

hierarchy, and any divergence between the two partners' profiles is substantively more informative and more defensible. The current operational reference for Art. 25 implementation in EU practice is **ESMA Guidelines on Suitability ESMA35-43-3172** (2022 update; superseding the 2018 version) — these guidelines specify the per-partner suitability requirement for shared advisory relationships referenced in the Legal Forensics table (§6.5).

7.2.4 Fiduciary Duty

Beyond statutory regulation, financial advisors bound by fiduciary duty face an ethical and legal obligation to act in the client's interests — not the firm's, not the AI system's, not any third party's. The fiduciary standard is higher than the suitability standard: it requires affirmative action in the client's interest, not merely avoiding recommendations that are not in the client's interest.

Digital Guardian (Pattern 1) is, architecturally, fiduciary duty as a system constraint. A system that cannot optimize for engagement metrics, cannot steer users toward preferred products, and cannot operate with hidden incentives is a system whose architecture enforces the fiduciary obligation at the structural level rather than relying on practitioner discipline alone.

7.3 6.1a — Related Work and Competitive Landscape: Why AI Explainability Is Not Decision Governance

Few tools today focus explicitly on *documenting and archiving the human decision over AI-supported outputs* for financial advisors, without generating recommendations. The market is dominated by solutions for recording, audit trails, and compliance monitoring that use AI for analysis and logging. This subsection maps representative tools and shows that even those positioned for "AI decision audit" address a different layer of the chain — AI explainability and record-keeping — leaving the governance gap (human decision over AI output) unfilled.

MiFID-II–focused tools. ASC Technologies (Recording Insights, AI Policy Templates) automates call recording and analysis, multi-year archiving, and violation detection; it checks content for suitability, risk disclosures, and fee communication and produces exportable audit documents. Clarity AI supports documentation of ESG preferences in suitability processes with automated data handling and API integration. Both are MiFID-II–oriented and EU-focused.

SEC and cross-regulatory tools. IntelliHuman AI provides full audit trails for AI-assisted decisions, including reasoning, citations, and knowledge archiving to meet recordkeeping and related regulatory expectations; it emphasizes transparency in financial decisions without normative recommendations. Neurons Lab's ARKEN offers compliant audit trails for client communication and portfolio messaging, aligned with SEC and MiFID II, with a focus on documentation rather than generation.

Three distinctions that define the gap. (1) *What is being documented.* IntelliHuman-style audit trails answer: "Why did the *AI* produce this output? Which data, reasoning, and sources?" That is

AI explainability. The Decision Packet answers: "Why did the *advisor* accept this recommendation for this client? Which pillar weights, which advisor assumptions, which trade-offs were explicitly considered?" That is decision governance at the human–AI boundary. IntelliHuman makes the AI transparent; a governance architecture makes the *human decision over the AI output* transparent — two different links in the chain: AI produces → (explainability: why did AI produce?) → human decides → (governance: why did the human accept?). Current AI explainability tooling gives practitioners, precisely, a flight recorder for the autopilot's code — capturing what the AI calculated, which sources it weighted, and why it produced a given output. What is structurally absent is the flight recorder for the pilot's brain: the record of why the human practitioner decided to follow the recommendation, override it, or — critically — silently defer without examining it. (2) *Compliance tool vs. governance architecture.* Tools like IntelliHuman integrate with existing compliance platforms and support record-keeping and auditability; they do not ask "Are the assumptions behind the recommendation transparent?" but "Is the recommendation documented and traceable?" Assumption-surfacing and approval gates in the workflow are governance-layer concerns that record-keeping alone does not address. (3) *Regulatory focus.* IntelliHuman is positioned for SEC, FINRA, and US banking regulation; the EU AI Act gap — and the question of whether AI-generated financial advice falls under high-risk classification — is the domain this framework and the EU-focused case study address.

Competitive landscape (summary). The following table locates representative tools and the unfilled governance gap. The named tools are illustrative; some are documented from public sources, some are characterizations of broader tool categories rather than precise product descriptions. Readers performing competitive analysis should validate current product positioning directly with vendors.

Tool	What it audits	What is missing (governance gap)
ASC Technologies	Call recording, policy compliance	No decision provenance; no assumption transparency
IntelliHuman AI	AI reasoning, citations, record-keeping	No client-side assumptions, no non-normative governance layer
Clarity AI	ESG preference matching, documentation	No Decision Packet, no governance gate
Neurons Lab ARKEN	Client/messaging audit trails	No assumption transparency

Tool	What it audits	What is missing (governance gap)
Governance architecture (this framework)	**Human decision over AI output**	**The missing middle link**

These tools validate demand for auditability and transparency in AI-assisted finance; they do not fill the structural gap between AI output and human-committed decision. A framework that governs the *human* decision over the AI output — with structured artifacts for assumptions, value hierarchies, and reasoning — occupies a distinct position: IntelliHuman (and similar) make the AI legible; analysis tools make the analysis efficient; a decision-governance architecture makes the human judgment in between auditable. Together they close the chain; the framework contributes the link that today is missing.

7.4 6.2 — The CFP Scenario: Where the Governance Gap Appears

The following scenario is not hypothetical — it represents a class of situation that occurs in financial advisory practices daily. The specific details are illustrative.

A certified financial planner meets with a couple in their mid-forties: one partner in corporate finance, one working part-time in education. They are preparing for an advisory session to review their retirement strategy. Before the meeting, the advisor's AI system has processed their uploaded financial documents and generated a comprehensive analysis in under 90 seconds: a Monte Carlo simulation across 10,000 scenarios for their target retirement age and income, a tax-loss harvesting opportunity analysis identifying €4,800 in realizable losses, an insurance gap assessment flagging underinsurance in long-term disability coverage, and a projected comparison of three portfolio rebalancing approaches across their defined contribution accounts.

All four analyses are technically correct. The AI's performance is unambiguous: analytical capacity that would have required two hours of manual work has been generated in 90 seconds, with greater consistency and more scenarios than any human analyst would have computed.

Now the advisory session begins.

The partner in corporate finance expresses strong confidence in equities. His rationale is that his parents invested through the 1970s and 1980s and did well. He has absorbed an inherited disposition toward market optimism — not an analysis, but a formed view that feels like knowledge.

The partner in education hesitates at the equity allocation in scenario 2. She does not explain

her hesitation clearly. When pushed, she says: "I'm not sure — it just feels like a lot." This is not irrational: her income is variable, her employment security is lower than her partner's, and her sense of what "a lot" means is calibrated to different loss aversion parameters than the AI's risk tolerance questionnaire has captured.

The AI knows neither of these things. It has processed income statements, account balances, and a five-point risk tolerance questionnaire answer. It does not know that the husband's confidence is inherited rather than analyzed. It does not know that the wife's hesitation reflects genuine differential loss aversion that the questionnaire failed to elicit. It has produced four technically correct analyses, none of which are actionable without the context that only the practitioner, in conversation with the clients, can supply.

Which portfolio allocation should the advisor recommend?

There is no answer to this question that does not require human judgment. The AI can compute the optimal allocation given stated risk tolerance. It cannot determine whether the stated risk tolerance is genuine for this specific couple, on this specific day, given the context the advisor is observing. The advisor could know — but only with the right infrastructure to structure the conversation, capture the divergence between the partners' value hierarchies, and document the reasoning that connected the AI analysis to the final recommendation.

Without Decision Governance infrastructure, this conversation produces a PDF with the recommended allocation, a CRM note ("discussed options, clients chose moderate growth allocation"), and no recoverable record of why scenario 2 was not chosen, what the wife's hesitation meant for the final decision, or how the advisor resolved the divergence between the partners' expressed and revealed risk tolerances.

This is the governance gap in its most concrete form: the most important reasoning in the advisory session — the human judgment that made the technical analysis meaningful for this specific couple — is invisible to any record system.

7.5 6.2a — Field-Validated Practitioner Objections

The architectural arguments above have been tested against practitioner objections collected during the framework's iterative development. Seven objections recurred with sufficient frequency to merit explicit architectural response: skepticism about whether current documentation is already sufficient; concerns about client friction; redundancy with AI-explainability tooling; the absence of billable hours attached to Decision Packet work; client confusion about confidence indicators; technology-cost barriers in smaller practices; and liability concerns when documented agreements turn on incorrect AI output.

The full objection catalogue, the architectural counter-evidence for each, and the methodological status of the collection process are documented in **Appendix F — Practitioner Objection Index**.

The objections are reproduced there as architectural challenges that recurred with sufficient frequency to merit treatment, not as resolved positions. Their representativeness across the broader practitioner population is identified as a research extension within RQ4 (Chapter 7).

7.6 6.3 — What the Framework Delivers

The reference architecture's implementation in a financial advisory context produces the following outputs from the scenario described above:

Narrative intake → Atoms: The advisory conversation is structured to produce Atoms — structured records that combine hard data (income, account balances, insurance coverage) with narrative context (the husband's inherited market view, the wife's variable income risk) and emotional context (the wife's hesitation about equities, the source of the husband's confidence). Atoms are not CRM notes — they are structured artifacts with explicit fields for data type, narrative context, confidence level, and source.

Pillar extraction → Value hierarchy per stakeholder: The pairwise elicitation protocol processes each partner's responses separately, producing separate value hierarchies. The husband's hierarchy: income stability → retirement security → housing security → liquidity → education legacy. The wife's hierarchy: housing security → income stability → education legacy → retirement security → liquidity. The divergence is explicit and quantified. This divergence is a primary input to the advisory recommendation, not an afterthought.

Decision Packet: The advisory outcome is recorded as a Decision Packet that includes: the AI-generated analyses (model version, input data, outputs, timestamps), the human decision (chosen allocation, rejected alternatives), the key assumptions (husband's risk tolerance grounded in inherited view, wife's loss aversion higher than questionnaire elicited), the reasoning chain (why the selected allocation was preferred over alternatives given the stakeholder divergence), and the open questions (whether wife's loss aversion should be revisited in the next session given her employment trajectory).

Temporal trajectory → Drift detection: When this couple returns in 12 months, the governance system surfaces the change in their portfolio and employment situation against the recorded baseline. If the wife's employment security has changed, the system generates a drift alert. The practitioner sees a specific comparison: original value hierarchy vs. current responses, with divergences highlighted. This is not a generic "markets have changed" notification — it is a specific alert about preference drift for this specific couple relative to their documented baseline.

JQM → Measurable advisory quality: The session produces JQM scores: Assumption Coverage (7 of 9 surfaced assumptions addressed), Drift Awareness (n/a for first session — will be measurable at follow-up), Perspective Diversity (2 distinct profiles, divergence explicitly documented), Question Depth (11 of 14 generated questions addressed in session). These scores are not judgments of the advisor's quality as a person — they are structural measurements of the session's governance

discipline.

7.7 6.4 — Tail Risk Disclosure as a Structural Requirement

The CFP scenario illustrates the governance gap at the individual decision level. There is a second dimension of risk that standard advisory practice — even with good documentation — systematically fails to capture: the structural correlation between value priorities under stress.

Conditional Value at Risk: CVaR (also called Expected Shortfall) is the standard risk metric for tail exposure in quantitative finance (Rockafellar & Uryasev, 2000). While Value at Risk (VaR) answers "What is the worst-case loss at the 95th percentile of scenarios?", CVaR answers "In the worst 5% of scenarios, what is the *average* loss?" CVaR is a more conservative and more informative metric because it characterizes the tail, not merely its threshold.

In financial advisory, CVaR is typically applied to portfolio returns: the expected portfolio loss in the worst 5% of market scenarios. This is valuable but incomplete. For household financial planning, the relevant tail is not a market portfolio — it is the household's total financial exposure, which includes multiple correlated risks.

Cross-Pillar Tail Risk: The wife and husband's value hierarchies — housing security, income stability, retirement adequacy, education legacy — are treated as independent dimensions in standard risk models. But they are not independent. A sudden job loss for either partner simultaneously threatens income stability (direct), housing security (through mortgage serviceability), retirement adequacy (through contribution pause and potentially early distribution), and education legacy (through reallocation of savings). These pillars are structurally correlated: a single adverse event propagates through all of them simultaneously.

Standard Monte Carlo simulations that model each pillar independently miss this correlation. They understate tail risk by assuming diversification across the value priorities themselves — when the reality is that a single catastrophic event can activate all priorities simultaneously.

Epidemiology isomorphism: This cross-pillar propagation behaves structurally like disease contagion. Compartmental models — specifically the SIR (Susceptible-Infected-Recovered) framework from epidemiology — model the spread of infection through a connected population. When one individual in a well-connected network is infected, the expected spread depends on the network structure, not just the individual infection probability.

Applied to cross-pillar tail risk: when one priority "pillar" is shocked (income stability collapses), the probability that adjacent pillars are subsequently shocked is higher than their independent base rates, because the pillars are connected by the household's budget constraint. The SIR formalism provides a mathematical structure for modeling this propagation that standard portfolio simulation does not capture.

Note on claim level: CVaR as applied to portfolio returns is established quantitative finance. The

cross-pillar application — treating a household's value priorities as a correlated network and applying SIR-style propagation models to estimate joint tail risk — is a proposed extension. The mathematical substrate exists; the specific application to household pillar networks is a research question, not a validated method. Chapter 7 includes this as an open research direction. What can be claimed today is the structural insight: cross-pillar correlation makes household tail risk higher than independent models suggest, and governance infrastructure should make this correlation visible to the advisory conversation.

7.8 6.5 — Legal Forensics: The Decision Packet as Regulatory Artifact

The most direct demonstration that Decision Governance infrastructure meets existing regulatory requirements comes from a field-by-field mapping of Decision Packet contents to regulatory article requirements.

The following table maps the Decision Packet structure to current regulatory requirements as of 2026. The mapping demonstrates that a fully populated Decision Packet **supports compliance** with major documentary requirements across the EU AI Act, BaFin MaComp, and MiFID II — not by designing to each requirement separately, but because the requirements converge on the same underlying need: a structured, human-interpretable record that connects AI-generated analysis to human-committed decisions. The Decision Packet does not formally substitute for a Konformitätsbewertung (Conformity Assessment under EU AI Act Article 43) or for the regulatorily required documentation forms (e.g. § 64 Abs. 4 WpHG Geeignetheitserklärung); it is a complementary architectural artifact that captures the reasoning-chain regulators need to reconstruct, alongside (not in place of) the statutory documents.

Decision Packet Field	EU AI Act Art.13	EU AI Act Art.14	BaFin MaComp	MiFID II
System purpose + capabilities/limitations	✓ (Art. 13.1 transparency information)	—	—	—
Human oversight log (practitioner ID, confirmation timestamp)	—	✓ (Art. 14.4 oversight capabilities a-e)	✓ (MaComp BT 7 Geeignetheitsprüfung)	✓ (Art. 25 Abs. 2 i.V.m. Art. 54 DV (EU) 2017/565)

Decision Packet Field	EU AI Act Art.13	EU AI Act Art.14	BaFin MaComp	MiFID II
Assumption provenance (which assumptions, source, confidence)	—	✓ (Art. 14.4 lit. a understanding)	✓ (MaComp BT 7)	✓ (Art. 25 suitability assessment)
Confidence bands on AI-generated estimates	✓ (Art. 13.3(b)(ii) accu-racy/robustness metrics)	—	—	✓
Stakeholder divergence record (separate profiles, divergence documented)	—	✓ (Art. 14.4 lit. c interpretation)	✓ (MaComp BT 7)	✓ (Art. 25 Abs. 2 i.V.m. ESMA Guidelines on Suitability ESMA35-43-3172)
Tamper-evident hash-chain audit trail	—	✓ (Art. 12 Record-keeping, separate from Art. 14)	✓ (audit trail)	✓ (Art. 16 Abs. 6/7 record-keeping)
Commission / conflict-of-interest disclosure	—	—	✓ (MaComp BT 5 i.V.m. § 63 Abs. 2 WpHG)	✓ (MiFID II Art. 23 / Art. 24 Abs. 9 Inducements)
Rejected alternatives with documented reasoning	—	✓ (Art. 14.4 lit. d override capability)	✓	✓
AI model version and input data snapshot	✓ (Art. 13.1: system characteristics)	✓ (Art. 12 Record-keeping)	✓	—

Note: Table reflects regulatory requirements as of 2026-03-01. Regulatory requirements evolve; the mapping should be reviewed against current guidance on any deployment.

The significance of this table is not that the reference architecture was designed to comply with each regulation individually. It is that the regulations and the governance architecture converge on

the same requirement from different starting points. Regulators, approaching from accountability and client protection, specify what records must exist. The SR7D framework, approaching from epistemic governance and noise reduction, specifies what records should exist. The overlap is near-complete.

This convergence is evidence that the governance architecture is not solving a compliance problem — it is solving the underlying epistemic problem that regulatory compliance is also trying to address.

Two adjacent literatures the framework engages but does not resolve. Power (*The Audit Society: Rituals of Verification*, 1997) and Strathern ("The Tyranny of Transparency", 2000) document a recurring pattern in governance-documentation systems: they evolve over time from substantive verification into ritual theater, with measured proxies systematically diverging from the underlying quality they were originally designed to preserve (a Goodhart's-Law-style failure). Decision Governance is structurally vulnerable to this pattern; the framework does not claim immunity. The architectural mitigation — calibration loops (§4.4) that test whether process scores actually predict decision-outcome quality over time — is a partial response. Whether the response is sufficient is an empirical question that only longitudinal deployment data can answer (Chapter 7, RQ3). **Detection heuristic available before calibration data exists:** ritualization-as-it-happens shows a measurable signature — variance compression in JQM-aggregate scores across practitioners *without* corresponding outcome-quality improvement on independent measures. If, in a multi-practitioner deployment, all practitioners converge to a tight high-value JQM band (variance below 10% of theoretical range) within the first 12 months, that pattern is empirically distinguishable from genuine quality improvement (which preserves practitioner-to-practitioner variance reflecting case-mix and skill differences). *Practitioner-facing illustration:* a deployment across 50 advisors in which every practitioner converges to near-perfect documentation scores within 12 months is the statistical signature of compliance theater — because genuinely variable human judgments on genuinely variable client situations preserve score spread. Uniform high scores indicate box-ticking, not active governance. The architectural commitment that follows: deployments instrument variance compression as a leading indicator and treat sustained low-variance-high-score patterns as triggers for the GOLD-constraint re-derivation governance gate (§5.3a).

Adjacent in the AI-safety literature, the corrigibility research program (Hadfield-Menell, Russell & Christiano, *Cooperative Inverse Reinforcement Learning*, 2017; subsequent work on off-switch and intervention games) formalizes the trust-calibration problem between an AI system and a human supervisor. Pattern 1 (Digital Guardian) and Ethic 2 (Agency Preservation) cover similar territory architecturally, arriving at structurally similar conclusions from a different starting point (regulatory and ethical, rather than RL-theoretic). The convergence is encouraging; full integration between the two formal traditions remains an open research direction. **Architectural commitment that follows from this acknowledgment:** the validator (Appendix B) preserves an explicit intervention-capability invariant — every deterministic-shell rejection is reversible by a practitioner override (logged as `override_notes`); no architectural state forces a practitioner to accept a system-blocked output. This matches the corrigibility-literature requirement that the human-supervisor channel remain operative under all system states. **Architectural commitment from the Power audit-society**

acknowledgment: the GOLD-constraint pattern (§5.3a) makes the constraint set itself a Decision-Packet-class governance artifact subject to retro-audit drills. This is the structural anti-ritualization commitment: the constraint set cannot drift toward theater unnoticed because its modifications are themselves logged, signed, and subject to governance review.

7.9 6.6 — Domain Extension

Financial advisory is the primary case study because the regulatory mandate is current, the stakes are individual and measurable, and the governance gap is largest. But the underlying structure — AI generates analysis, human commits to consequential decision, governance gap exists between them — is present in several adjacent domains.

Healthcare: The informed consent process in medicine requires that patients understand the risks, benefits, and alternatives to a proposed treatment before agreeing to it (Faden & Beauchamp, *A History and Theory of Informed Consent*, 1986 — the canonical philosophical-medical-ethics treatment). In AI-assisted diagnostic and treatment contexts, the equivalent of the Advisory Decision Packet is the informed consent record: AI assessment of relevant evidence, physician interpretation, explicit treatment alternatives presented, patient choice with documented understanding. The governance gap in healthcare is visible in the documented problem of patients agreeing to procedures without genuine understanding of risk — a process quality failure with direct clinical consequences.

HR and Employment: Algorithmic hiring and performance evaluation systems operate in a domain with high noise (Kahneman's research on interview-based hiring is among the most robust findings in the noise literature) and high regulatory exposure (GDPR, EU AI Act Annex III §4 explicit high-risk classification for employment decisions). The governance gap here — between AI-generated candidate scores and human hiring decisions — is particularly consequential because the audit requirements are already present in employment law, but the current infrastructure for meeting them is inadequate (Raghavan, Barocas, Kleinberg & Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, 2020, FAccT — the canonical accountability-gap mapping in algorithmic hiring).

Public Policy: Urban planning, social program design, and policy evaluation decisions involve AI models that project outcomes across populations. The governance gap in policy contexts is the absence of structured records connecting model outputs to adopted policies — making it difficult or impossible, years later, to understand why a policy was adopted, what assumptions it was based on, and whether those assumptions were subsequently validated. Tetlock-style forecasting infrastructure (belief updating, calibration, explicit assumption tracking) applied to policy decisions would constitute a form of civic Decision Governance (Eubanks, *Automating Inequality*, 2018 — documents the governance gap in algorithmically mediated public-services delivery; reading this work motivates the cross-domain extension claim above more concretely than the abstract framing alone).

The common structure across all three domains: AI capability has expanded faster than governance infrastructure. The governance gap exists because no one designed the infrastructure to fill it. The architecture described in this paper is the first systematic attempt to specify that infrastructure. Until it exists, the pattern is the same across all three domains: AI is not really serving the people it touches. It is just happening to them.

7.10 6.6a — Companion-Document Map

This paper is one artifact in a family of documents. The other documents are not required reading for the primary persona, but the map below names them so readers know where related material lives and what is in scope for each.

Document	Audience	Scope	Status
This whitepaper (WP-01)	CFP-HNW-DACH primary persona	Architectural and epistemic case for Decision Governance; reference implementation overview	V1.2 — published
Research Program / Noise RCT Protocol	Academic methodology readers	Full OSF-style pre-registration of the proposed RCT (Appendix A in this paper, future external companion)	Draft within Appendix A; planned migration to standalone companion
Engineering Specification	Implementer / technical adopter	Detailed validator-spec, Decision-Packet JSON schema, Bayesian posterior layer implementation, locale extensions, integration interfaces	Currently distributed across Appendices B, C, and E; planned consolidation into standalone engineering companion

Document	Audience	Scope	Status
Methods Handbook / Editorial Process	Whitepaper-authors and reviewers of subsequent versions	Editorial Checklist (10-point gate), reference compliance tooling, four-role review protocol, revision-trail conventions	Currently Appendix D + D-Annex; planned migration to internal authors' handbook
Practitioner Playbook	CFP-HNW practitioners adopting the architecture	ROI worked examples, integration sketches for common CFP-firm IT stacks (DATEV, wealthnet, Salesforce Financial Services, German Robo-Advisors), German-locale forbidden-lexicon details, change-management considerations	Not yet drafted; planned post-V1.2
Investor Materials	Investors performing due diligence on Steerable as a company	Commercial framing, TAM estimation, GTM hypothesis, unit economics	Explicitly out of scope for this paper family; lives in separate channels

Readers who find themselves wanting depth on a topic that this paper treats only at architectural level should consult the corresponding companion (or note its absence and request it). The deliberate division of material across companions is itself an instance of Pattern 2 (SPECTRADING): different readers have different agency-and-information needs, and a single document cannot serve all of them equally well.

7.11 6.7 — Discovery-Call Decision Page (Practitioner-Facing)

This section is the practical exit-ramp from this paper. If you have reached this section and want to evaluate whether Decision Governance applies to your advisory practice, the structure below names the components, the integration question, and the discovery-call agenda.

7.11.1 What you would integrate

Two operationally instantiated components are mature enough for a discovery conversation:

SIC-026 — Stat-Prior Confidence Subcomponent. A per-assumption confidence indicator that flags whether the statistical priors underlying an AI-generated analysis are well-grounded (large sample, matching units, current data) or weakly-grounded (small sample, unit mismatch, stale calibration). In your practice, this addresses the "the AI gave me a number but I don't know whether to trust it" pain. Plain-language outputs (high / moderate / low confidence), with hard-zero on unit-mismatch to prevent silent compounding of bad priors. Detail: Section 4.4a + Appendix E.1.

SIC-028 — Bayesian Posterior Layer. The inference machinery that produces the confidence indicators above and propagates them through the pillar-subpoint reasoning graph. More technically demanding to integrate; useful where your existing AI tooling produces structured pillar weights that should be aggregated into a coherent posterior rather than averaged or top-1-selected. In your practice, this addresses the "different tools give different scores and I don't know how to combine them" pain. Detail: Section 3.7 + Appendix E.1.

7.11.2 Which one to ask about — quick guide

- *If your AI stack is one consumer-grade tool (ChatGPT, Copilot) plus your existing CRM and a single Robo-Advisor*: SIC-026 is the right first conversation. It plugs into the existing tool outputs and adds a confidence-flag layer without replacing anything.
- *If your AI stack includes multiple analytical tools producing potentially conflicting pillar weights (Robo-Advisor + ESG-rating tool + retirement-projection tool)*: SIC-028 is the right first conversation. It is the layer that turns multiple structured AI outputs into a single coherent posterior the practitioner can act on.
- *If you are unsure*: SIC-026, then revisit SIC-028 once stat-prior confidence is operational.

7.11.3 Agenda template for the discovery call (30 minutes)

A productive discovery call covers the following four topics in approximately the times indicated; if any topic does not resolve in its allotted time, the conversation is unlikely to converge in a single call.

Time	Topic	Expected outcome
0:00-0:10	Live demo of a Decision Packet built around an anonymized version of one of your real client cases (you bring the case)	You can see what the packet looks like with your own data structure

Time	Topic	Expected outcome
0:10-0:20	Integration sketch into your current stack (CRM, portfolio analytics, your AI tools, your compliance archive). The discussion names concrete data flows and identifies the integration friction	You leave with an estimate of integration effort in weeks, not in vague "we'll figure it out"
0:20-0:25	ROI conversation grounded in your own practice numbers (number of clients, average advisory engagement duration, frequency of contested-review-grade incidents in your firm or your peer firms). Hedged, not deterministic	You leave with a defensible cost/benefit framing for your own internal decision, not a pitched number
0:25-0:30	Pricing and implementation timeline. If neither converges to something workable, the call ends. If both do, a follow-on technical conversation is scheduled	You leave with a yes/no/needs-more-info decision, not a soft "let us send you materials"

7.11.4 What this section is not

This section is not a sales pitch. It is the explicit, practical exit-ramp the rest of this paper deliberately avoids constructing in its main argument — because the rest of the paper is concerned with the architectural and epistemic case, not the commercial offering. The two operationally instantiated components named here exist; the integration questions named here are real; the call structure named here is what a productive first conversation looks like. Whether to schedule the call is the reader's decision. The paper does not recommend that you do. If the rest of the paper has earned a discovery conversation, the section above tells you how to get the most out of thirty minutes; if not, this section is the last paragraph you needed to read.

8 Chapter 7 — Conclusion and Research Agenda

8.1 7.0 — What This Paper Has Argued

This paper opened with an observation about abundance. AI has solved the analytical bottleneck that constrained professional judgment for most of human history. Analysis — generating scenarios, processing data, synthesizing research, producing projections — is now fast, cheap, and increasingly accurate. The capability constraint that once justified the premium attached to professional analytical work is gone.

The observation led to a problem. Abundance without governance is noise. The elimination of analytical noise — the benefit AI delivers — creates new noise at the human-AI decision boundary: model noise, prompt noise, context noise, and the compounding effect of automation bias. The governance gap — the structural absence of infrastructure at the transition from AI-generated output to human-committed decision — is the space where accountability disappears, variance compounds, and audit trails end.

Six chapters have developed a response to this problem.

Chapter 1 established the governance gap as the central problem: not an AI alignment issue, not a regulatory compliance issue, but an infrastructure issue — the absence of structured artifacts at the point where professional judgment is exercised.

Chapter 2 proposed the SR7D framework as the architectural specification for filling that gap: ten constraints — seven core patterns and three ethical guardrails — that are not aspirational principles but implementable, testable, and in several cases machine-enforceable design requirements.

Chapter 3 grounded two of the framework's central mechanisms in formal epistemics: pairwise value elicitation as an active inference process (Friston/Shannon), and Aspiration-Backward reasoning as abductive inference (Peirce). These are not metaphors — they are structural claims about the formal properties of the mechanisms, offered with explicit acknowledgment of their current confidence level.

Chapter 4 proposed Judgment Quality Metrics as a measurement framework that extends Tetlock's forecasting science to the domain of decision process quality. The four JQM dimensions — Assumption Coverage, Drift Awareness, Perspective Diversity, Question Depth — are derived from the structure of the decision process and provide prospective, process-level quality assessment that does not require outcome data.

Chapter 5 described the enforcement layer: the Deterministic Shell architecture that places deterministic validation, cryptographic integrity, and deterministic measurement around the probabilistic AI core. The central claim — that Non-Normativity and Immutability can be enforced at the structural level, not merely specified as guidelines — distinguishes this architecture from responsible AI frameworks that rely on policy compliance.

Chapter 6 validated the framework in the domain most directly suited to it: financial advisory under the EU AI Act. The CFP scenario demonstrated where the governance gap appears in practice. The Legal Forensics Schema demonstrated that a fully populated Decision Packet satisfies current regulatory requirements across EU AI Act, BaFin MaComp, and MiFID II simultaneously. The tail risk discussion extended the framework's contribution to a gap in current advisory practice that regulation has not yet caught up to.

8.2 7.1 — The Pacioli Parallel

In 1494, Luca Pacioli published a section on double-entry bookkeeping in his mathematical compendium *Summa de Arithmetica, Geometria, Proportioni et Proportionalità*. Pacioli did not invent double-entry bookkeeping — the method had been in use in northern Italian merchant houses for at least a century before he formalized it. His contribution was to systematize, document, and disseminate a practice that had been working but was not yet universally understood.

What double-entry bookkeeping did was structurally transformative: it made financial state legible to strangers. Before double-entry, a merchant's financial position was reconstructable primarily by the merchant and those who had worked with them directly. After double-entry, it was reconstructable by any literate person with access to the ledger — a partner, an investor, a regulator, a future generation of managers. The ledger created the possibility of trust between strangers at scale, and that trust was the infrastructure on which capital markets developed.

The structural parallel to Decision Governance is exact.

Before Decision Governance, a practitioner's reasoning is reconstructable primarily by the practitioner and those who were present in the advisory conversation. After Decision Governance, it is reconstructable by any observer with access to the Decision Packet — a colleague reviewing the case, a regulator conducting an audit, a client who wants to understand why a recommendation was made three years ago, or the practitioner themselves, revisiting a decision to learn from it.

Decision Governance creates the possibility of trust between strangers in AI-assisted judgment. A client who has never met an advisor can, in principle, evaluate the governance quality of the advisor's decision process by examining the Decision Packet record. A regulator who was not present in the advisory session can reconstruct the reasoning chain from the append-only audit trail. An institution that wants to improve its advisory quality over time has, in the accumulated Decision Packet record, the empirical data to do so.

Pacioli's contribution was not a new way of doing business. It was a new way of making business visible. Decision Governance is not a new way of making decisions. It is a new way of making decisions visible.

Consider a more immediate frame: the middle-school teacher who marks a math test with *show your work* in red ink. The student used a calculator, wrote the correct answer, and received zero

credit — because the answer alone is not the argument. The AI's output is the answer on the page. The client, the regulator, or the practitioner revisiting a case three years later does not want only the answer; they need the steps, the dependencies, the alternatives discarded. The Decision Packet is the structural mechanism for showing your work at professional scale.

The missing ledger is not financial. It is epistemic.

8.3 7.2 — Open Research Questions

The framework proposed in this paper makes several claims that require empirical validation, formal development, or cross-domain testing. Intellectual honesty requires distinguishing between what has been argued and what remains open. The following research questions are offered not as admissions of weakness but as invitations — the specific open questions that the framework's structure generates, and that would most strengthen or challenge its claims.

RQ1 — Topological Resilience: The assumption graphs embedded in Decision Packets have a network structure: assumptions are connected to each other through logical dependencies. Can algebraic connectivity — specifically, the Fiedler value (the second-smallest eigenvalue of the graph Laplacian) — serve as a structural health metric for assumption coverage? A well-connected assumption graph (high Fiedler value) would imply that no single assumption can be removed without disconnecting the reasoning chain; a sparse graph (low Fiedler value) would identify fragile points where a single false assumption collapses the entire case. *Practitioner-facing analogy:* a low Fiedler value is a Jenga tower of logic — pull one block (one false assumption) and the case collapses because the entire argument rested on a single dependency chain. A high Fiedler value is a brick wall — remove any single assumption and the reasoning still stands, because the logic is cross-verified and supported from multiple directions.

This is a research question, not a solved problem. The mathematical structure exists; the empirical validation — whether Fiedler value actually predicts decision quality outcomes in practice — requires longitudinal data.

RQ2 — Cross-Domain Validation: This paper validates the SR7D framework in financial advisory. The claim is that the 7+3 constraints hold across any domain involving consequential AI-assisted decisions with accountability requirements. Healthcare, HR, and public policy are named as candidate domains. Whether the constraint set transfers without modification, with modification, or with domain-specific additions is an empirical question that requires domain experts and validation data in each context.

A specific edge case worth investigating: does Non-Normativity (the prohibition on system prescriptions) hold in emergency medicine, where decision latency is itself a patient safety variable? In a time-critical triage context, a system that surfaces multiple options without guidance may be structurally unsuited to the stakes — the governance constraints that protect a pension fund look different from those in a trauma bay; the difference, at its starkest, is between picking a mutual fund

and stopping a bleed. This is not a claim that Non-Normativity fails in emergency medicine — it is a flag that the application requires careful examination.

RQ3 — JQM Calibration: What is the empirical relationship between JQM scores and decision outcome quality over time? The four JQM dimensions are defined with theoretical justification — they map to validated components of forecasting quality (Tetlock) and behavioral decision research (Kahneman). But the empirical relationship between process scores and outcomes in financial advisory specifically has not been measured.

A longitudinal study would require: a cohort of practitioners using the reference architecture, Decision Packet records accumulated over 3–5 years, and outcome measures linked to each recorded decision. The challenge is defining outcome quality for financial advisory in a way that separates governance quality from market performance — clients with identical governance quality can have different financial outcomes due to factors entirely outside the advisory relationship. Developing outcome measures that are responsive to governance quality while not confounded by market factors is itself a research contribution.

RQ4 — Abductive Formalization: Chapter 3 proposed that Aspiration-Backward reasoning is formally abductive inference in Peirce's sense. The structural parallel is sound; the full formal development — specifying the inference procedure, the goal representation, the hypothesis generation mechanism, and the plausibility evaluation function in terms that admit mathematical analysis — remains to be done. A complete formalization would either strengthen the claim or identify where the parallel breaks down.

RQ4-Extension — Practitioner-Objection Validation: A subsidiary question within Cross-Domain Validation: do the seven practitioner objections catalogued in Appendix F generalize across advisory-firm cohorts, regulatory jurisdictions, and practice sizes? A structured survey instrument with documented sampling methodology — covering practitioner role (CFP, RIA, EU vs. US, firm size, AI-tool exposure), exposure to AI-assisted advisory workflows, and elicitation of objections in both supported-recall and free-response formats — would test whether the objection catalogue is representative of broader resistance or an artifact of the convenience-sample methodology by which it was originally collected. The study design is non-trivial: objection elicitation is sensitive to framing, and practitioners may articulate different objections in confidential survey conditions than in iterative product-development conversations. Pre-registration of the survey instrument and the analysis plan is essential to avoid post-hoc rationalization of any findings.

RQ5 — Friston-Shannon Bridge Tightening: Chapter 3 proposed that pairwise value elicitation is formally an entropy-reduction process under Shannon's information measure, and that the active inference framework (Friston) provides a theoretical account of why this process is effective. The full mathematical development — specifying the generative model over pillar weights, the update rule for pairwise comparisons, and the convergence proof — is open. The intuition is robust; the mathematical treatment should be developed in parallel with empirical validation.

RQ6 — Automation Deference Experiment: Does Top-k scenario presentation reduce automation deference compared to single-recommendation AI output?

The experimental design is straightforward and pre-registerable. Participants are randomly assigned to one of three conditions: (a) single-recommendation output from an AI system, (b) Top-3 scenario presentation with equivalent information, (c) Top-3 presentation with explicit assumption surfacing and stakeholder divergence records. Primary measures: decision confidence ratings, stated vs. revealed risk preferences (comparing questionnaire responses to actual portfolio choices), decision latency. Secondary measures: regret rates at three-month follow-up, belief updating between initial decision and follow-up session.

The hypothesis — that Top-k presentation reduces automation deference (Parasuraman & Manzey, 2010) by forcing genuine choice rather than ratification — is well-grounded in the choice architecture literature. The experiment would provide direct evidence for or against the core behavioral claim underlying SPECTRADING (Pattern 2) and Principled Inefficiency (Pattern 5).

RQ7 — Federated JQM Aggregation: At the institutional level, JQM scores accumulated across practitioners provide calibration data for improving decision process quality within that institution. The next question is whether JQM signals can be aggregated *across* institutions — across advisory firms, across hospital systems, across jurisdictions — without requiring the centralization of sensitive decision data.

Differential privacy techniques (Dwork, McSherry, Nissim, & Smith, 2006) provide formal guarantees for this class of problem: statistics can be computed over aggregated private data with mathematically bounded privacy loss for any individual record. Applied to JQM: could a privacy-preserving protocol aggregate anonymized JQM signals across institutions to calibrate dimension weights — determining, empirically, whether Assumption Coverage or Drift Awareness is more predictive of good outcomes in a given domain — without any individual institution's records being visible to the others?

The technical prerequisites exist. The practical implementation would require agreeing on a minimum schema, a privacy budget, and k-anonymity thresholds per cohort. This is a network effects problem: the value of federated calibration increases with the number of participating institutions, but adoption requires sufficient trust in the privacy guarantees. It is also a prerequisite for cross-institution learning at scale — the mechanism that would allow the governance architecture to improve faster than any single institution can improve on its own.

RQ8 — Decision-Packet Reasoning-Field Audit (Confabulation Counter Operationalization): Does a Decision Packet whose `decision_reasoning` field was independently audited against the actual decision process — via concurrent verbal protocols, eye-tracking, or process-tracing methods drawn from cognitive psychology (Klein 1998; Ericsson & Simon 1980; Nisbett & Wilson 1977) — agree with the field's content above chance? The empirical threshold is articulated in §3.7c: agreement $\kappa \geq 0.2$ (Cohen's kappa) sustained across practitioners and contexts. If the answer is no, the architectural option named in §3.7a Response 1 (constrain Decision Packets to externally verifiable structural facts; remove free-text reasoning fields from the canonical schema) must be elevated from option to requirement. This is the load-bearing falsifiability test for the framework's central claim that Decision Packets preserve reasoning rather than confabulating it.

RQ9 — Methodology-Self-Audit Remediation (Pattern-1 + Pattern-6 Compliance for the Authoring Pipeline): Can a Decision-Governance authoring pipeline be specified such that all external-reviewer-LLM calls operate under user-data-sovereignty constraints (zero-retention provider modes or self-hosted models) and all reviewer outputs are recorded as hash-chained immutable artifacts under the same Pattern-6 discipline the paper specifies for client-facing Decision Packets? The §3.7b self-audit documents that V1.2's production violated Pattern 1 and Pattern 6. The empirical test for RQ9 is whether V1.3 or V1.4 ships an authoring pipeline that eliminates both violations. If neither version ships such a pipeline, the Reflexive Triangulation principle (§3.7a) is not architecturally realizable, and the paper's claim that the architecture is implementable should be downgraded per §3.7c.

8.4 7.3 — Closing

AI creates abundance. Abundance without governance is noise.

This is not a pessimistic claim about AI — it is a structural observation about what happens when a capability grows faster than the infrastructure surrounding it. The capability — analysis, simulation, synthesis, pattern recognition — is genuinely transformative. The infrastructure problem — how consequential decisions made with AI involvement are governed, documented, and improved — is genuinely unsolved.

The companies and institutions that build the governance infrastructure build the infrastructure for the next economic era. Not because governance is a competitive advantage in the usual sense, but because trust is, and governance is the architecture of trust.

Pacioli's ledger made it possible for a Venetian merchant to do business with a Florentine banker who had never met him, because both could read the same record and both could trust that it accurately represented financial state. Decision Governance makes it possible for a client to trust an advisory practice they have only recently engaged, for a regulator to verify that AI-assisted advice was sound, for a practitioner to learn from their own decision history rather than repeating the same errors — because all of them can read the same Decision Packet and trust that it accurately represents the reasoning that led to the decision.

The missing ledger is not financial. It is epistemic. And building that ledger is the defining infrastructure opportunity of the AI-abundant era.

---

9 Appendices

Note: As of V1.2 (2026-05-24), Appendices C-G have been relabeled to B-F to close the previously vacant Appendix B slot (an artifact of an earlier draft). Section/subsection numbers within each appendix have been updated accordingly. Earlier references in V1.2 Revision Trail entries (which document findings against the pre-relabel state) preserve the prior letter labels — those are historical-audit records and intentionally not updated.

10 Appendix A — Noise RCT Protocol (Pre-Registration Draft)

10.1 Decision Governance Overlay and Inter-Rater Variance Reduction in Financial Advisory

Protocol version: Draft 1.0 — 2026-03-01 Status: Pre-registration draft — for submission to OSF (Open Science Framework) prior to data collection

---

10.2 A.1 — Study Title and Registration Intent

Full title: Does a Decision Governance Overlay Reduce Inter-Rater Variance in Financial Advisory Recommendations? A Three-Arm Randomized Controlled Trial

Pre-registration platform: Open Science Framework (OSF) — https://osf.io **Pre-registration intent:** This protocol is designed for pre-registration before data collection begins. Pre-registration locks the primary hypothesis, endpoints, and analysis plan, preventing outcome-contingent reporting.

10.3 A.2 — Background and Motivation

Chapter 1 of the accompanying whitepaper documents the governance gap in AI-assisted financial advisory: the structural absence of infrastructure between AI-generated analysis and practitioner-committed recommendations. One consequence of this gap is elevated inter-rater variance: two equally qualified advisors, given the same client case, may produce recommendations that differ in meaningful ways — not because one is more competent, but because the unstructured nature of the decision process allows idiosyncratic factors (framing effects, anchoring, automation bias) to introduce noise.

Kahneman, Sibony, and Sunstein (2021) documented inter-rater variance in professional judgment broadly. This study applies the noise audit methodology specifically to AI-assisted financial advisory, testing whether a structured Decision Governance overlay reduces variance.

10.4 A.3 — Primary Hypothesis

H1 (Primary): Advisors operating with a full Decision Governance overlay (Arm C) will produce recommendations with significantly lower inter-rater variance than advisors operating without an overlay (Arm A) or with a partial overlay (Arm B).

Operationalization: Inter-rater variance is measured as the standard deviation of recommended equity allocations (as a percentage of total portfolio) across advisors presented with the same client vignette. Lower standard deviation = lower noise.

Null hypothesis: The Decision Governance overlay has no effect on inter-rater variance (H1_null: $\sigma_C = \sigma_A$).

Expected direction: $\sigma_C < \sigma_B < \sigma_A$

10.5 A.4 — Study Design

10.5.1 Arms

Arm A — Baseline: Participants receive a standardized client vignette (demographic information, financial situation, stated goals, AI-generated portfolio analysis) with no additional structure. They produce a written recommendation as they would in their current practice.

Arm B — Ghost Overlay: Participants receive the same client vignette plus an AI-generated assumption list and a structured question set generated by the Decision Governance system. They are asked to review the assumptions and answer the questions before producing their recommendation. No Decision Packet documentation requirement.

Arm C — Full Framework + Decision Packet: Participants receive the same client vignette, the assumption list, the structured question set, AND are required to complete a Decision Packet before submitting their recommendation. The Decision Packet completion requires: selecting the client's value hierarchy from the pairwise elicitation output, marking which assumptions they verified, noting any open questions, and providing a brief reasoning statement connecting the AI analysis to their recommendation.

10.5.2 Assignment

- Randomization unit: individual practitioner
- Randomization method: block randomization, stratified by years of advisory experience (< 5 years, 5–15 years, > 15 years)
- Allocation ratio: 1:1:1

10.6 A.5 — Participants

Inclusion criteria: - Licensed financial advisors or certified financial planners - Active client-facing practice at time of enrollment - No prior exposure to the reference architecture

Exclusion criteria: - Advisors in training or under direct supervision for cause - Advisors involved in the development of the reference architecture

Recruitment: Professional associations (CFP Board, FPSB member organizations), financial advisory firm partnerships. Participants recruited via professional networks, not through advertising to clients.

Sample Size Rationale:

The study is powered to detect a 25% reduction in inter-rater standard deviation in Arm C relative to Arm A.

Assumptions: - Baseline standard deviation of equity allocation: $\sigma_A = 12$ percentage points (based on Kahneman et al. 2021 noise audit benchmarks for professional judgment) - Target detectable effect: $\sigma_C = 9$ percentage points (25% reduction) - Power: 0.80 - Significance level (two-tailed): $\alpha = 0.05$ - Test: F-test for equality of variances (Levene's test)

Estimated n per arm: 85 advisors (255 total)

Note: The sample size calculation is a preliminary estimate. The final calculation should be reviewed by a biostatistician before pre-registration submission. The variance estimate ($\sigma = 12$ pp) is drawn from the closest available analog (general professional judgment noise audits); domain-specific pilot data would improve precision.

10.7 A.6 — Primary Endpoint

Measure: Standard deviation of recommended equity allocation (as percentage of total portfolio) across advisors within each arm, for the primary client vignette.

Timing: Single cross-sectional measurement at the end of the experimental session (approximately 45 minutes after vignette presentation).

Analysis: Levene's test for equality of variances across three arms. If the overall test is significant ($p < 0.05$), pairwise comparisons (A vs. C, B vs. C) are conducted with Bonferroni correction ($\alpha = 0.025$ per comparison).

10.8 A.7 — Secondary Endpoints

Endpoint	Measure	Timing
Decision Latency	Time from vignette delivery to recommendation submission (minutes)	During session
Confidence Calibration	Self-reported confidence in recommendation (0–100) vs. actual inter-advisor agreement	During session
Stated vs. Revealed Risk Preference	Difference between advisor's stated belief about client risk tolerance and the allocation they recommend	During session
Regret Rate	Proportion of advisors who, at 3-month follow-up, would change their recommendation given the same vignette	3-month follow-up survey
JQM Scores (Arm C only)	Assumption Coverage, Question Depth from Decision Packet	During session

---

10.9 A.8 — Vignette Specification

All arms receive identical client vignettes. The vignettes are developed using the following specifications:

- **Primary vignette:** Household in mid-40s, two partners with diverging income trajectories, three financial goals with implicit tradeoffs (housing security, retirement adequacy, education funding), AI-generated portfolio analysis with three allocation options
- **Complexity level:** Sufficient to require genuine judgment (not resolvable by algorithm) but not so complex that advisor time commitment exceeds 45 minutes
- **Ambiguity:** Vignette includes deliberate ambiguity about true risk tolerance (stated moderate, behavioral indicators of lower tolerance)
- **Vignette validation:** Pilot test with 5 advisors per arm (n=15) to confirm vignette complexity and timing calibration before main study

---

10.10 A.9 — Pre-Registration Checklist

The following items must be completed before any data collection begins:

☐ Protocol deposited on OSF with timestamp lock
☐ Analysis plan (statistical code) pre-committed with protocol
☐ Sample size calculation reviewed by biostatistician
☐ Vignette validated in pilot (n=15)
☐ IRB approval obtained (if required by institutional affiliation of study coordinators)
☐ Participant consent form reviewed by legal counsel
☐ Randomization sequence generated by independent statistician
☐ Primary vignette locked (modifications require protocol amendment)
☐ Primary endpoint definition locked (no changes after pre-registration)

10.11 A.10 — Limitations

Ecological validity: Laboratory conditions may not fully replicate the time pressures and client relationship dynamics of real advisory practice. Results should be interpreted as evidence of the framework's potential under favorable conditions, not as proof of effectiveness in all deployment contexts.

Vignette coverage: A single primary vignette cannot cover the full range of advisory situations. The study design includes two secondary vignettes to test replication, but domain coverage remains limited.

Self-selection: Advisors who volunteer for the study may be systematically different from the general population of advisors — potentially more reflective, more open to structured processes, or more technically inclined.

Follow-up attrition: The 3-month follow-up for regret rate measurement will experience attrition. Analysis will be conducted on available follow-up data with multiple imputation for missing values.

Protocol status: Draft for pre-registration review. Not for data collection until OSF pre-registration is confirmed and IRB approval obtained.

---

11 Appendix B — Validator Specification v0.1

11.1 Decision Governance Architecture: Deterministic Constraint Enforcement

Version: 0.1 — 2026-03-01 This specification evolves with production deployment. v0.1 represents the minimum viable enforcement layer.

11.2 B.1 — Purpose

This appendix specifies the deterministic validation layer described in Chapter 5 (Section 5.3). The validator is the enforcement mechanism for the non-negotiable architectural constraints of the SR7D framework. It runs on every AI-generated output before that output reaches the practitioner interface.

The validator enforces two properties: 1. **Non-Normativity compliance:** Output does not contain prescriptive language 2. **Decision Packet schema conformance:** Output contains all required fields for a compliant Decision Packet

---

11.3 B.2 — Required Decision Packet Fields

A compliant Decision Packet must contain the following fields. Missing fields cause a schema validation failure and the packet is not written to the audit chain.

```
class DecisionPacket(BaseModel):
    # Identity
    packet_id: str                # UUID v4
    parent_hash: Optional[str]    # SHA-256 of parent packet; None for first in chain
    created_at: str               # ISO 8601 timestamp (UTC)
    validator_version: str        # e.g. ”v0.1” — version of this spec used

    # Context
    decision_context: str         # Plain text: what decision is being governed
    domain: str                   # e.g. ”financial_advisory”, ”healthcare”, ”hr”
    practitioner_id: str          # Anonymized practitioner identifier

    # AI Component
    ai_model_id: str              # Model name + version that generated the analysis
    ai_input_hash: str            # SHA-256 of the input data sent to the model
    ai_output_text: str           # The AI-generated analysis (pre-validation)
    ai_output_hash: str           # SHA-256 of ai_output_text
```

```
    # Governance Fields
    assumptions: List[Assumption]           # Minimum 1; see C.3
    confidence_band: ConfidenceBand         # Required; see C.4
    stakeholder_profiles: List[StakeholderProfile]  # Minimum 1; see C.5
    open_questions: List[str]               # Questions not resolved in this session
    rejected_alternatives: List[str]        # Named alternatives explicitly not chosen

    # Human Decision Record
    human_decision: str          # Plain text: what the practitioner decided
    decision_reasoning: str      # Plain text: why this decision, given the analysis
    human_confirmed_at: str      # ISO 8601 timestamp of practitioner confirmation

    # Chain Integrity
    packet_hash: str             # SHA-256 of all above fields (computed at write time)
```

11.3.1 C.2.1 — Assumption Record

```
class Assumption(BaseModel):
    assumption_id: str           # UUID v4
    assumption_text: str         # Plain text description
    source: str                  # "ai_generated" | "practitioner_surfaced" | "client_stated"
    confidence: float            # 0.0 – 1.0; 0.5 if genuinely unknown
    checked: bool                # True if practitioner actively verified or addressed
    check_notes: Optional[str]   # Required if checked=True
```

11.3.2 C.2.2 — Confidence Band

```
class ConfidenceBand(BaseModel):
    estimate: float              # Point estimate (the headline number)
    lower_bound: float           # Lower bound of confidence interval
    upper_bound: float           # Upper bound of confidence interval
    confidence_level: float      # 0.0 – 1.0; typically 0.90 or 0.95
    basis: str                   # Plain text: what drives the uncertainty range
```

11.3.3 C.2.3 — Stakeholder Profile

```
class StakeholderProfile(BaseModel):
    stakeholder_id: str          # Anonymized identifier
    value_hierarchy: List[str]   # Ordered list of pillars from most to least important
    elicitation_method: str      # "pairwise" | "narrative" | "questionnaire" | "combined"
    divergence_from_partners: Optional[str]  # Required if multiple stakeholders present
```

11.4 B.3 — Forbidden Normative Lexicon

The following terms, when found in `ai_output_text`, cause a Non-Normativity validation failure. The output is rejected and logged; the practitioner sees a system notification that the output was blocked.

```
FORBIDDEN_NORMATIVE_LEXICON: List[str] = [
    ”should”,
    ”must”,
    ”recommend”,
    ”I recommend”,
    ”we recommend”,
    ”you need to”,
    ”you ought to”,
    ”the best option is”,
    ”the right choice is”,
    ”I suggest”,
    ”I advise”,
    ”advise you to”,
    ”it is advisable”,
    ”you should consider”,
    ”you must”,
    ”strongly suggest”,
    ”highly recommend”,
]
```

Matching rules: - Case-insensitive matching - Whole-word and substring matching (to catch “recommended”, “recommends”, etc.) - Context is not evaluated — no “charitable reading” of normative language - The check is unconditional: no configuration switch disables it

Locale extension — German forbidden lexicon. Deployments serving German-speaking advisory practice require the equivalent forbidden lexicon in German. The reference implementation includes the following entries (extending FORBIDDEN_NORMATIVE_LEXICON when the system locale is `de-DE` or `de-AT`):

```
FORBIDDEN_NORMATIVE_LEXICON_DE: List[str] = [
    ”empfehle”,  # ”ich empfehle”, ”wir empfehlen”
    ”empfehlen”,
    ”empfohlen”,
    ”Empfehlung”,  # noun form
    ”sollten”,  # ”Sie sollten”, ”Sie sollten erwägen”
    ”müssen Sie”,
```

```
    "müssen wir",
    "ist optimal",
    "die beste Option ist",
    "die richtige Wahl ist",
    "raten",  # "wir raten Ihnen"
    "raten wir",
    "ratsam ist es",
    "ratsam wäre",
    "die optimale Allokation",
    "empfehlenswert",
]
```

Additional locale extensions (fr-FR, it-IT, es-ES) follow the same pattern; the reference implementation supports per-locale extension without architectural change. Locale-specific false-positive cases (e.g. "die regulatorisch *empfohlene* Risikoklassifizierung" where "empfohlen" is a quotation of a regulatory term, not a system recommendation) follow the same handling as English: rephrase the prompt to avoid normative constructions; do not weaken the rule.

On false positives: The forbidden lexicon will occasionally block outputs that use normative language in a non-prescriptive way (e.g., "the regulatory framework *requires* this disclosure"). These cases should be addressed by rephrasing the AI prompt to avoid normative constructions, not by weakening the validation rule. The enforcement cost of occasional rephrasing is lower than the governance cost of normative outputs reaching the practitioner interface unchecked.

11.5 B.4 — Tamper-Evident Hash Chain Integrity

Decision Packets are written to an append-only tamper-evident hash chain. The chain structure is:

```
{
  "packet_id": "uuid-v4-here",
  "parent_hash": "sha256-of-previous-packet-or-null",
  "packet_hash": "sha256-of-this-packet-all-fields",
  "created_at": "2026-03-01T14:32:00Z",
  "validator_version": "v0.1",
  "... all other DecisionPacket fields ..."
}
```

Chain integrity rules: 1. `packet_hash` is computed over all fields including `parent_hash` before writing 2. `parent_hash` of each packet must match `packet_hash` of the preceding packet 3. Any modification to a historical packet breaks the hash chain at that point 4. Chain integrity verification is a deterministic operation: compare `packet_hash` against recomputed hash of all fields

Note: This is not blockchain. No distributed consensus is required. The integrity guarantee is local: given a chain of packets, chain integrity can be verified by any party with access to the chain data, without requiring trust in a third party. This is the same integrity principle as Git's object model, applied to decision records.

11.6 B.5 — Validation Failure Handling

Failure Type	System Response	Audit Action
Non-Normativity violation	Reject output; display notification to practitioner; request regeneration	Log rejected output with violation markers and timestamp
Missing required field	Reject packet; display field-specific error message	Log incomplete packet attempt with missing fields identified
Confidence band absent	Reject; display specific error: "Confidence band required for quantitative claims"	Log
No assumptions listed	Reject; display: "Minimum 1 assumption required per Decision Packet"	Log
Chain hash mismatch	Alert (not rejection): "Chain integrity anomaly detected at packet [ID]"	Log with high-severity flag; flag for review

---

11.7 B.6 — Version Flag and Evolution Path

This specification is version 0.1. Known limitations to be addressed in subsequent versions:

- **Semantic Non-Normativity:** The current lexical check catches explicit normative language. Semantically normative outputs that avoid the forbidden words are not caught. v0.2 should include context-sensitive validation for common normative framings.
- **Domain-specific required fields:** The current schema is domain-generic. v0.2 should support domain profiles (e.g., `financial_advisory` profile with additional required fields for suitability documentation; `healthcare` profile with informed consent fields).
- **Multi-language support:** The forbidden lexicon currently covers English only. Deployments in German, French, Spanish require lexicon extensions.

- **Confidence band validation:** The current spec requires a confidence band but does not validate whether the band is appropriately calibrated. v0.2 should include checks for degenerate bands (lower = upper = estimate).

Flag: This spec is v0.1 — machine-enforceable constraints evolve with production deployment and empirical evidence from validation failures.

---

12 Appendix C — Decision Packet JSON Schema

12.1 Minimum Viable Schema v0.1

Version: 0.1 — 2026-03-01 This schema is compatible with the Validator Spec (Appendix B) and the Legal Forensics Schema (Chapter 6, Section 6.5).

12.2 C.1 — Schema Overview

A Decision Packet is the primary artifact produced by the Decision Governance architecture. It is a self-contained, structured record that documents a consequential decision: the AI-generated analysis that informed it, the human judgment that committed to it, the assumptions that underpinned it, and the context that shaped it.

The schema below is expressed as a JSON Schema draft-07 specification. All fields marked `required` must be present for a packet to pass validation (see Appendix B). Fields marked `optional` are conditionally required based on context (e.g., `stakeholder_divergence` is required when `stakeholder_count > 1`).

12.3 C.2 — JSON Schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "$id": "decision-packet-v0.1",
  "title": "DecisionPacket",
  "description": "SR7D-compliant decision record. Append-only. Part of tamper-evident hash chain.
  "type": "object",
  "required": [
    "packet_id",
```

```
    "created_at",
    "validator_version",
    "decision_context",
    "domain",
    "practitioner_id",
    "ai_component",
    "governance",
    "human_record",
    "chain"
  ],
  "properties": {

    "packet_id": {
      "type": "string",
      "format": "uuid",
      "description": "UUID v4. Globally unique identifier for this packet."
    },

    "created_at": {
      "type": "string",
      "format": "date-time",
      "description": "ISO 8601 UTC timestamp of packet creation."
    },

    "validator_version": {
      "type": "string",
      "description": "Version of the Validator Spec used to validate this packet. E.g. 'v0.1'."
    },

    "decision_context": {
      "type": "string",
      "minLength": 10,
      "description": "Plain text description of what decision is being governed."
    },

    "domain": {
      "type": "string",
      "enum": ["financial_advisory", "healthcare", "hr", "public_policy", "other"],
      "description": "Domain classification for reporting and domain-specific validation profiles
    },
```

```
    "practitioner_id": {
      "type": "string",
      "description": "Anonymized identifier for the practitioner who committed this packet."
    },

    "ai_component": {
      "type": "object",
      "required": ["model_id", "input_hash", "output_text", "output_hash"],
      "properties": {
        "model_id": {
          "type": "string",
          "description": "Model name and version. E.g. 'claude-sonnet-4-6' or 'gpt-4o-2024-08-06'
        },
        "input_hash": {
          "type": "string",
          "description": "SHA-256 of the full input context sent to the model."
        },
        "output_text": {
          "type": "string",
          "description": "The AI-generated analysis. Must have passed Non-Normativity validation.
        },
        "output_hash": {
          "type": "string",
          "description": "SHA-256 of output_text."
        },
        "generation_timestamp": {
          "type": "string",
          "format": "date-time",
          "description": "Timestamp of model output generation."
        }
      }
    },

    "governance": {
      "type": "object",
      "required": ["assumptions", "confidence_band", "stakeholder_profiles", "open_questions"],
      "properties": {

        "assumptions": {
          "type": "array",
          "minItems": 1,
```

```
      "items": {
        "type": "object",
        "required": ["assumption_id", "assumption_text", "source", "confidence", "checked"],
        "properties": {
          "assumption_id": { "type": "string", "format": "uuid" },
          "assumption_text": { "type": "string", "minLength": 5 },
          "source": {
            "type": "string",
            "enum": ["ai_generated", "practitioner_surfaced", "client_stated"]
          },
          "confidence": {
            "type": "number",
            "minimum": 0.0,
            "maximum": 1.0,
            "description": "Practitioner's confidence that this assumption is valid. 0.5 if g
          },
          "checked": {
            "type": "boolean",
            "description": "True if practitioner actively verified, challenged, or addressed
          },
          "check_notes": {
            "type": "string",
            "description": "Required if checked=true. Summary of how the assumption was addre
            "field_role": "annotation",
            "audit_caveat": "Per §3.7a Response 1, annotation only. The 'checked' boolean abo
          }
        }
      }
    },

    "confidence_band": {
      "type": "object",
      "required": ["estimate", "lower_bound", "upper_bound", "confidence_level", "basis"],
      "properties": {
        "estimate": { "type": "number" },
        "lower_bound": { "type": "number" },
        "upper_bound": { "type": "number" },
        "confidence_level": {
          "type": "number",
          "minimum": 0.0,
          "maximum": 1.0,
```

```
        "description": "Confidence level for the interval. Typically 0.90 or 0.95."
      },
      "basis": {
        "type": "string",
        "description": "Plain text: what drives the uncertainty range."
      }
    }
  },

  "stakeholder_profiles": {
    "type": "array",
    "minItems": 1,
    "items": {
      "type": "object",
      "required": ["stakeholder_id", "value_hierarchy", "elicitation_method"],
      "properties": {
        "stakeholder_id": { "type": "string" },
        "value_hierarchy": {
          "type": "array",
          "items": { "type": "string" },
          "description": "Ordered list of pillars from most to least important for this sta
        },
        "elicitation_method": {
          "type": "string",
          "enum": ["pairwise", "narrative", "questionnaire", "combined"]
        },
        "divergence_from_partners": {
          "type": "string",
          "description": "Required if multiple stakeholders. Documents disagreements in val
        }
      }
    }
  },

  "open_questions": {
    "type": "array",
    "items": { "type": "string" },
    "description": "Questions generated during the session that were not resolved. May be e
  },

  "rejected_alternatives": {
```

```
        "type": "array",
        "items": { "type": "string" },
        "description": "Named alternatives explicitly considered and not chosen, with brief rea
      }
    }
  },

  "human_record": {
    "type": "object",
    "required": ["human_decision", "decision_reasoning", "human_confirmed_at"],
    "properties": {
      "human_decision": {
        "type": "string",
        "minLength": 10,
        "description": "Plain text: what the practitioner decided."
      },
      "decision_reasoning": {
        "type": "string",
        "minLength": 10,
        "description": "Plain text: why this decision, given the analysis and stakeholder conte
        "field_role": "annotation",
        "audit_caveat": "Per §3.7a Response 1, this field is treated as supplementary annotatio
      },
      "human_confirmed_at": {
        "type": "string",
        "format": "date-time",
        "description": "Timestamp of practitioner confirmation. Must be after ai_component.gene
      },
      "override_notes": {
        "type": "string",
        "description": "Optional. If practitioner overrode a system suggestion, reasoning is do
        "field_role": "annotation",
        "audit_caveat": "Per §3.7a Response 1, this field is annotation, not canonical record.
      }
    }
  },

  "chain": {
    "type": "object",
    "required": ["parent_hash", "packet_hash"],
    "properties": {
```

```
        "parent_hash": {
          "type": ["string", "null"],
          "description": "SHA-256 of parent packet. Null for the first packet in a chain."
        },
        "packet_hash": {
          "type": "string",
          "description": "SHA-256 of all fields in this packet (computed at write time, before se
        }
      }
    }

  }
}
```

12.4 C.3 — Regulatory Mapping

The schema fields map to regulatory requirements as documented in Chapter 6, Section 6.5 (Legal Forensics Schema). Key correspondences:

Schema Path	Regulatory Requirement
`ai_component.model_id` + `ai_component.input_hash`	EU AI Act Art.13.1 (system characteristics) + Art. 12 (record-keeping) (logging)
`human_record.human_confirmed_at`	EU AI Act Art.14.1 (human oversight log)
`governance.assumptions`	EU AI Act Art.14.4 (assumption provenance) + BaFin MaComp
`governance.confidence_band`	EU AI Act Art.13.3b (reliability/accuracy communication) + MiFID II
`governance.stakeholder_profiles` + `divergence_from_partners`	MiFID II Art.25 (per-partner suitability) + EU AI Act Art.14.4
`chain.parent_hash` + `chain.packet_hash`	EU AI Act Art. 12 (record-keeping) + BaFin MaComp (audit trail)
`governance.rejected_alternatives`	EU AI Act Art.14.4 (intervention capability documentation)

12.5 C.4 — Usage Notes

Immutability: Decision Packets must never be modified after `packet_hash` is computed. Corrections are appended as new packets referencing the original via `parent_hash`. The correction packet includes `decision_context: "Correction of packet [original_packet_id]"` and documents what changed and why.

Practitioner ID privacy: `practitioner_id` must be anonymized before any cross-institutional aggregation. Recommended: HMAC-SHA256 of the practitioner's unique identifier with an institution-specific secret key. This allows within-institution practitioner tracking while preventing cross-institution identification.

Domain profiles: This is the base schema. Domain-specific profiles (v0.2) will extend the schema with additional required fields for each domain (e.g., `suitability_assessment` for `financial_advisory`, `informed_consent_record` for `healthcare`).

12.6 C.5 — Field-Role Discipline (V1.2.1 Schema-Implementation of §3.7a Response 1)

Per the Confabulation Counter response in §3.7a Response 1, the schema differentiates two field-role classes for human-record content:

`field_role: "structural"` — externally verifiable facts about the decision process. The canonical record. Examples in this schema: `human_decision`, `human_confirmed_at`, `packet_id`, `parent_hash`, `packet_hash`, `validator_version`, `created_at`, `domain`, `practitioner_id`, `ai_component.model_id`, `ai_component.input_hash`, `governance.assumptions[].checked` (the boolean), `governance.stakeholder_profiles[].value_hierarchy`, `governance.confidence_band.estimate/lower_bound/upper_bound/confidence_level`, `governance.rejected_alternatives[]` (when present as enumerated alternatives rather than free-text reasoning).

`field_role: "annotation"` — practitioner-articulated free-text fields that carry confabulation risk per Nisbett-Wilson 1977 / Johansson 2005 / Klein 1998 RPD. *Not* the canonical record; supplementary interpretive context. Examples: `decision_reasoning`, `check_notes`, `override_notes`, `governance.confidence_band.basis` (free-text basis description), `governance.assumptions[].assumption_text`, `ai_component.output_text` (the AI-generated prose, distinct from `ai_component.output_hash` which is structural).

12.6.1 Validator Discipline

The reference validator implementation enforces:

1. **Structural-field absence is a HARD validation failure.** If any field with `field_role: "structural"` is missing from a Decision Packet, the validator rejects the packet (same severity as schema-conformance failure).
2. **Annotation-field absence is a soft warning, not a rejection.** Practitioners who choose not to populate `decision_reasoning` or `check_notes` produce a Decision Packet with reduced interpretive context but a complete structural record. The validator notes the absence; it does not reject the packet.
3. **Audit-trail discipline differs by field role.** Hash-chain integrity (Pattern 6) protects structural and annotation fields equally — both are part of the cryptographically signed content. But downstream consumers of the packet (auditors, JQM-score-computers, re-loadable DecisionPacket-renderers) treat annotation fields as evidence-of-articulation, not evidence-of-cognitive-process. JQM scores are derived primarily from structural-field content; annotation fields contribute interpretive context but do not raise JQM scores by themselves.
4. **RQ8 trigger condition.** If RQ8 (§7.2) empirical evidence demonstrates that annotation fields agree with actual cognitive process below the $\kappa \geq 0.2$ threshold specified in §3.7c, the canonical schema must be modified to remove annotation-class fields entirely (per §3.7a Response 1 elevation from option to requirement). The `field_role` enum is the architectural prerequisite that makes this modification mechanically straightforward — removal of all `field_role: "annotation"` fields preserves the schema's structural integrity.

12.6.2 Migration from V1.2 Schema (Version 0.1 → 0.1.1)

V1.2.1 is a backwards-compatible schema update. Packets validated against schema v0.1 remain valid against v0.1.1 — all V1.2-shaped packets are accepted. The `field_role` keys are *added as schema annotations*, not enforced as required keys. Validators upgraded to v0.1.1 begin distinguishing structural-vs-annotation failure-severity but do not reject pre-V1.2.1 packets.

---

13 Appendix D — Editorial Checklist for Chapter Authors

13.1 Decision Governance Whitepaper: 10-Point Quality Gate

Version: 1.0 — 2026-03-01 Apply to every chapter before submission. All 10 points must pass.

13.2 How to Use This Checklist

Complete this checklist for each chapter before it is considered ready for the consistency pass (Phase 5). Mark each item as PASS, FAIL, or N/A with a brief note. FAIL items must be resolved; N/A requires

justification.

A chapter is ready for the final assembly only when all applicable items are PASS.

13.3 The 10-Point Checklist

13.3.1 1 — SR7D Pattern Names: Canonical and Complete

Check: Every reference to an SR7D pattern uses the canonical name, including the pattern number.

Canonical reference: - Core Patterns: Digital Guardian (Pattern 1), SPECTRADING (Pattern 2), Triangulated Truth (Pattern 3), Constructed Ambiguity (Pattern 4), Principled Inefficiency (Pattern 5), Immutability First (Pattern 6), Archaeological Governance (Pattern 7) - Ethical Guardrails: Non-Normativity (Ethics 1), Agency Preservation (Ethics 2), Contestability (Ethics 3)

Fail condition: "Temporal Depth" appears as a standalone principle. "Pattern X" without the name. Any invented or shortened name. Fewer than 2 SR7D references in a chapter.

How to check: Search for each of the 10 canonical names. Search for "Temporal Depth." Verify each reference includes both the name and the number.

13.3.2 2 — No Product Name in Prose

Check: The string "Steerable" does not appear in the chapter body as a product name.

Acceptable references: "the reference architecture," "the proposed framework," "the governance architecture," "the SR7D framework."

Fail condition: "Steerable" appears as a product name. "the Steerable system." "Steerable's approach."

How to check: Search for "Steerable" (case-sensitive and case-insensitive).

13.3.3 3 — Scientific References: Specific, Not Vague

Check: Every empirical or scientific claim is attributed to a specific reference: Author(s), Year, and the specific concept or finding cited.

Fail condition: "Research shows…", "Studies indicate…", "Experts believe…", "It has been argued…" without citation. Reference to a field without a specific author/year.

How to check: Scan every paragraph for empirical claims. Verify each has an author-year citation. Cross-reference against the Citation Shortlist (20 references in Master Outline).

13.3.4 4 — Overclaim Control: Research Questions Flagged as Such

Check: Every claim that has not been empirically validated or formally proven is explicitly flagged at the appropriate confidence level.

Acceptable flagging language: - "proposed formalization, not proven theorem" - "proposed mapping" - "research question — not solved" - "claimed as generalizable, validated only in [domain]" - "preliminary estimate — requires pilot data"

Fail condition: A formally unproven mathematical claim presented as proven. A validated result from one domain stated as universal. An empirical finding from a different context claimed to apply without qualification.

How to check: Identify every claim that involves mathematical formalization, cross-domain generalization, or future empirical validation. Confirm each has an explicit confidence-level flag.

13.3.5 5 — Word Count: 3,000–4,000 Words (Ch.1–Ch.6) / 1,500–2,500 Words (Ch.7)

Check: Chapter body (excluding quality-gate notes and header lines) falls within the specified range.

Fail condition: Chapter is below 2,800 words (insufficient development) or above 4,500 words (excessive length).

How to check: Count words in the chapter body, excluding the header, section headings, and quality-gate note block at the end.

13.3.6 6 — Structure: Definition → Argument → Scientific Grounding → Architectural Implication → Relevance

Check: Each major section follows the standard structure established in Chapter 2 as the style reference.

Structure elements: - **Definition:** What is the concept being introduced? - **Argument:** Why does it matter? What is the core claim? - **Scientific grounding:** What established research supports this claim? - **Architectural implication:** How does this translate into the SR7D framework or the reference architecture? - **Relevance to Decision Governance:** Why does this belong in this paper?

Fail condition: A section makes a claim without scientific grounding. A section provides scientific grounding but does not connect it to the governance architecture. A section has an architectural implication but no relevance statement.

How to check: For each section, verify that all five elements are present, even if not labeled explicitly. Not every section needs a separate subsection for each element — but all five must be addressed.

13.3.7 7 — Governance Gap Definition Referenced

Check: The formal Governance Gap definition (introduced in Chapter 1) is referenced or recalled in any chapter that invokes the governance gap concept.

Definition: *"The Governance Gap is the structural absence of infrastructure at the transition from AI-generated output to human-committed decision. It is the space where accountability disappears, variance compounds, and audit trails end."*

Fail condition: A chapter discusses the governance gap without anchoring to this definition, leaving the reader without a clear referent.

How to check: Search for "governance gap" (case-insensitive). Verify that the first substantive use in each chapter is grounded in the formal definition, either by quoting or citing "as defined in Chapter 1."

13.3.8 8 — No Standalone "Temporal Depth"

Check: The temporal dimension of the framework is attributed to Archaeological Governance (Pattern 7), not to a standalone principle called "Temporal Depth."

Acceptable: "the temporal dimension of Archaeological Governance," "Archaeological Governance's trajectory analysis," "temporal reconstructability [Pattern 7]."

Fail condition: "Temporal Depth" appears as a named principle. "The Temporal Depth principle." "Following our Temporal Depth framework."

How to check: Search for "Temporal Depth" and "temporal depth" (case-insensitive).

13.3.9 9 — Citation Cross-Reference

Check: Every scientific reference cited in the chapter appears in the Citation Shortlist (Master Outline v1.1, 20 references). No references are cited in the chapter that do not appear in the shortlist.

Exception: Domain-specific regulatory references (EU AI Act, BaFin MaComp, MiFID II) are always acceptable. References added to the shortlist as part of the N1–N10 merge (entries 16–20) must be used in the correct chapters.

Fail condition: A citation appears in the chapter body that is not in the shortlist. A shortlist reference that the chapter outline specifies as belonging to this chapter is absent from the chapter.

How to check: List all author-year citations in the chapter. Compare against the 20-entry shortlist. Verify that chapter-specific references (e.g., Bradley & Terry and Elo in Ch.4; Rockafellar & Uryasev in Ch.6) are present.

13.3.10 10 — Non-Normativity Self-Check

Check: The chapter itself does not tell the reader what to do, what to decide, or what is "right." It describes, argues, grounds, and implies — it does not prescribe.

Fail condition: "Practitioners should…", "Organizations must…", "The right approach is…", "You need to…", "We recommend that…", "The correct answer is…"

Note: This self-check applies the Non-Normativity constraint (Ethics 1) to the paper itself. The whitepaper is a governance architecture for Decision Governance — it must model the principles it describes. A paper that tells practitioners how to make decisions is internally inconsistent with its own Non-Normativity requirement.

How to check: Scan for imperative constructions and normative claims. Verify that prescriptive-sounding statements are framed as architectural specifications (what the system enforces) or empirical claims (what the evidence shows), not as direct instructions to the reader.

13.4 Checklist Summary Table

#	Checkpoint	Status	Notes
1	SR7D canonical names + numbers		
2	No product name "Steerable"		
3	Scientific references specific		
4	Overclaims flagged		
5	Word count in range		
6	Structure: Def→Arg→Sci→Arch→Rel		
7	Governance Gap definition anchored		
8	No standalone "Temporal Depth"		
9	Citations cross-referenced to shortlist		

#	Checkpoint	Status	Notes
10	Non-Normativity self-check		

Chapter ready for assembly: All applicable items PASS.

13.5 Appendix DANNEX — Reference Compliance Tooling

Version: 1.0 — proposed extension to the 10-point manual checklist Status: reference implementation available; manual checklist (E.1) remains canonical for editorial sign-off

The manual checklist (E.1) establishes the editorial discipline. In production deployments, items 3 (Scientific References), 4 (Overclaim Control), and 9 (Citation Cross-Reference) are amenable to deterministic verification. A reference implementation is described here as proof-of-concept; the manual checklist remains canonical for editorial sign-off.

13.5.1 D-Annex.1 — Scope of Automation

Three checklist items have machine-checkable structure:

Item	What is automated	Residual manual judgment
3 — Scientific References	Detection of empirical-claim language without (Author, Year) suffix	Whether the citation actually supports the claim
4 — Overclaim Control	Detection of unflagged formal claims (e.g., “proven”, “always”, “guaranteed”) in chapters known to contain proposed-not-proven content	Whether the flag chosen is appropriate to the claim’s actual confidence level
9 — Citation Cross-Reference	Mechanical diff between chapter citations and the Citation Shortlist	Whether absent shortlist references are legitimately absent or missing

Items 1, 2, 5, 6, 7, 8, 10 require semantic judgment that exceeds reliable automation as of this writing and remain manual.

13.5.2 D-Annex.2 — Reference Compliance Tooling

A reference implementation of the automated subset (items 3, 4, 9) is maintained as a separate utility within the framework's operations layer. The tool reads chapter source files, applies pattern matchers for each automatable check, and produces a per-chapter audit report identical in structure to the E.1 Summary Table — with PASS/FAIL/UNCERTAIN status per item.

The tool's design follows three constraints that mirror the framework's own ethical guardrails:

- **Non-Normative output (Ethics 1):** The tool reports findings; it does not auto-correct chapters or recommend rewrites. Editorial decisions remain with the author.
- **Constructed Ambiguity (Pattern 4):** UNCERTAIN is a first-class outcome alongside PASS/FAIL — used when the matcher cannot determine compliance with high confidence (e.g., ambiguous citation format).
- **Archaeological Governance (Pattern 7):** Every audit run produces a timestamped record linked to the chapter's content hash. Re-running the check against an unchanged chapter produces an identical report; any difference indicates either rule evolution or chapter modification, both of which are traceable.

13.5.3 D-Annex.3 — Documented Failure Mode

The reference implementation has one documented failure mode worth surfacing: pattern matchers for item 4 (Overclaim Control) require maintenance as the formal-grounding language in the paper evolves. A claim originally flagged as "proposed mapping" that is later substantiated by empirical work should not continue to be flagged as an overclaim — but the automation does not know that the empirical work has occurred unless the chapter language is updated. Manual review remains the source of truth for confidence-level changes.

This failure mode is itself an instance of the broader pattern: automation handles deterministic structure; humans handle semantic shifts. The 10-point checklist remains canonical precisely because some judgments cannot be safely delegated.

13.5.4 D-Annex.4 — Cadence

The reference tooling is intended for pre-merge use: run the check before integrating a chapter revision, fix UNCERTAIN flags by clarifying the chapter language (not by suppressing the matcher), and treat FAIL findings as blockers requiring revision before merge. The cadence is therefore tied to chapter-merge events, not to a fixed time schedule.

---

14 Appendix E — Reference Implementation Notes

Version: 1.0 — 2026-05-24 Companion to Chapters 3 and 4: implementation details for the Bayesian posterior layer and the four-role review protocol.

14.1 E.1 — Bayesian Posterior Layer: From Theory to Inference

This appendix documents the operational instantiation of the information-theoretic bridge developed in Chapter 3 (Sections 3.2 through 3.5). The reference implementation runs in production within the framework's reference architecture; the description here is at the level required for reproducibility and architectural audit, not at the level of code listings.

14.1.1 E.1.1 — The Five-Stage Inference Path

Stage 1 — Co-occurrence to PMI. Pillar-subpoint co-occurrence counts from the reference architecture's elicitation history are transformed into pointwise mutual information (PMI) scores. The base of the logarithm is a documentation choice, not a substantive one: natural log and log base 2 differ by a constant factor (1/ln 2) that any downstream normalization absorbs. The reference implementation uses natural log because the subsequent CPT-construction stage performs an inverse exponential (i.e., `exp()`) — using natural log avoids an explicit base-conversion factor in the inversion code. This is an engineering preference for code readability, not a claim that natural log is mathematically superior. Deployments may legitimately use log base 2 with no analytical consequence as long as the same base is used throughout the pipeline.

Stage 2 — PMI to CPT. Conditional probability tables (CPTs) are constructed from PMI scores by inverse exponential transformation, with normalization to ensure CPT rows sum to 1.0. Three edge cases are handled explicitly, each corresponding to a documented failure mode discovered during implementation:

- *Bidirectional negative PMI:* When PMI(A, B) is negative in both directional senses, the analytical implication (A and B are anti-correlated) requires non-trivial CPT construction. The unit test catalogue includes a test for this case because a directional asymmetry was observed in calibration data.
- *Multi-parent CPT overwrite:* An earlier implementation silently lost CPT entries when multiple parents shared a child node. The current implementation includes a guard that fails closed (raises an exception) rather than silently overwriting.
- *Near-zero PMI clipping:* PMI values near zero are clipped to a configurable epsilon to prevent ill-conditioned exponentials.

Stage 3 — Topology check. The graph structure assembled from PMI-derived edges is checked for treewidth using a min-fill-in heuristic. Belief propagation has computational hardness scaling exponentially with treewidth; the heuristic catches structures that would have rendered the posterior

computation intractable. The exact treewidth threshold is deployment-configurable; the reference implementation defaults to a treewidth of 8, which keeps single-node inference under one second on commodity hardware.

Stage 4 — Model build (pgmpy). The validated graph is materialized as a `pgmpy.models.DiscreteBayesianNetwork`. The pgmpy choice (version 0.1.x at time of writing) was constrained by three requirements: open-source license, well-documented API, and amenability to test-driven development against analytically tractable known-posterior cases. The framework's CI suite includes a known-posterior test that validates the full pipeline against hand-computed posteriors for a three-node chain, ensuring that any pgmpy-version upgrade does not silently change inference behavior.

Stage 5 — Belief propagation and inference. Posteriors are computed via belief propagation. The output is structured as a discriminated union: each posterior carries either (a) a max-marginal value with entropy, (b) an explicit "unobserved" marker indicating that the node lacks sufficient evidence, or (c) an error type with diagnostic information. The discriminated-union design — refined through architecture review — eliminates the ambiguity of a single numeric field that conflates "low confidence", "no data", and "computation failed". Each consumer of the posterior is forced to handle each case explicitly.

14.1.2 E.1.2 — Integration Points with the SR7D Framework

The posterior layer is exposed through three integration points:

- **As Constructed Ambiguity (Pattern 4):** Max-marginal entropy is the raw computational input from which the user-facing confidence indicators (high / moderate / low) are derived. See Chapter 6, Section 6.2a and Appendix F, Objection 5.
- **As input to Assumption Coverage (JQM Dimension 1):** Posterior probability of each surfaced assumption being load-bearing on the final recommendation can be computed; this allows Assumption Coverage to be weighted by load rather than counted as binary checked/unchecked. See Section 4.4a for the operational subcomponent.
- **As Triangulated Truth verification (Pattern 3):** When multiple independent intake paths (narrative, quantitative, behavioral) feed into the posterior layer, their convergence or divergence is recovered from the marginals rather than asserted by the application layer.

14.1.3 E.1.3 — Claim Level and Open Questions

The posterior layer is operational in the reference implementation. Its connection to the formal bridges established in Sections 3.2-3.5 is direct: PMI is the empirical Shannon-mutual-information estimate from which the entropy-reduction claim of Section 3.2 is operationalized; belief propagation is the inference machinery that allows the active-inference loop of Section 3.5 to actually update beliefs as new evidence arrives.

The following remain open empirical questions: (a) whether the posterior layer's confidence calibra-

tion matches outcome calibration in deployment over time, (b) whether the discriminated-union representation reduces user misinterpretation of confidence as compared with a single numeric field (a UX-research question), and (c) whether the natural-log PMI choice generalizes beyond the elicitation-domain calibration on which it was tested. These are tracked in Chapter 7.

Claim level. The instantiation described above is implemented and tested in the reference architecture. The connection to the formal bridges of Sections 3.2-3.5 is *proposed mapping with empirical instantiation*; full validation against decision-outcome data is pending.

14.2 E.2 — Four-Role Review Protocol (Whitepaper Authoring Methodology)

The framework's claim that triangulation across independent sources improves judgment quality (Pattern 3, Triangulated Truth; Chapter 3, Section 3.5) creates a methodological obligation on the whitepaper itself: substantive claims should be triangulated, not asserted from a single perspective.

This obligation is operationalized in the framework's authoring process through a four-role review protocol. Each substantive section is read by four reviewer instances assigned distinct epistemic positions:

- **Role A — Practitioner-Experience:** reads with attention to whether the architectural claim survives contact with the realities of advisory-firm operations
- **Role B — Academic-Rigor:** reads with attention to logical structure, citation adequacy, and resistance to overreach
- **Role C — Regulatory-Compliance:** reads with attention to whether the claim is consistent with EU AI Act, MiFID II, BaFin MaComp, and adjacent regulatory frameworks
- **Role D — Adversarial-Counterargument:** reads with the goal of producing the strongest available counter to the claim

The four reviews are produced independently — each role receives only the source artifact and its role-specific brief, not the outputs of the other roles. The four outputs are synthesized in a structured triage that identifies convergent findings (high confidence in revision priority), divergent findings (requiring further investigation), and orthogonal findings (independent additions to the original claim space).

The protocol's purpose is not consensus — it is the surfacing of disagreement that a single reviewer would not produce. The four-role design is calibrated against Tetlock's finding (Section 4.1) that aggregating independent expert judgments outperforms individual expert judgment, with the qualification that the independence must be genuine and the roles sufficiently distinct.

This methodology is the source of the practitioner-objection set in Appendix F: the seven objections are convergent findings from Role A across multiple review runs. The roles' raw outputs are preserved as Archaeological Governance artifacts (Pattern 7) of the whitepaper's own production process and remain available for inspection in the framework's documentation archive.

Claim level: The four-role triangulation protocol is the whitepaper's working methodology; its effectiveness as a quality-improvement mechanism, relative to single-reviewer protocols, is *claimed as plausibly superior* based on convergent findings across multiple roles, but has not been measured against a controlled alternative and remains an open empirical question.

---

15 Appendix F — Practitioner Objection Index

Version: 1.0 — 2026-05-24 Companion to Chapter 6: field-collected practitioner objections and architectural counter-evidence.

15.1 F.1 — Scope and Methodology

The seven objections catalogued below were collected during the framework's iterative development through unstructured practitioner conversations with financial advisory professionals (CFP-credentialed, German and EU practice context) over the period 2026-Q1 to 2026-Q2. The conversations were not a controlled study: the sample was a convenience sample drawn from the framework authors' existing professional network and from feedback received via the framework's lead-magnet distribution pipeline. Sample size is small (n in the low double-digits, exact count not recorded). The objections are reproduced here as architectural challenges that recurred with sufficient frequency to merit treatment; their representativeness across the broader practitioner population is an empirical question identified in Chapter 7 (see RQ4-Extension).

The intent of this appendix is twofold: (a) to demonstrate that the architectural arguments in Chapter 6 have been tested against field-level resistance rather than only against internal critique, and (b) to surface the objections as questions for the broader research community, not as resolved positions.

15.2 F.2 — The Seven Objections

15.2.1 Objection 1: "My documentation is already compliant — this is overengineering."

The objection conflates two questions: *Does current documentation satisfy minimum regulatory inspection?* and *Does current documentation enable reconstruction of the advisory reasoning by a third party years later?* MiFID II Article 25 and BaFin MaComp formally require the second; in practice, most firms produce documentation that satisfies the first. The governance gap is not regulatory failure — it is the inadequacy of current artifacts to support the regulator's actual reconstruction requirement, which only becomes visible during contested reviews.

15.2.2 Objection 2: "Clients won't tolerate the additional friction."

Principled Inefficiency (Pattern 5) addresses this directly: friction is introduced only at consequential decision moments, not throughout the advisory workflow. Field observation suggests that clients distinguish between friction that signals seriousness (mandatory reflection at the moment of portfolio commitment) and friction that signals process inefficiency (repeated re-entry of data the system should have retained). The former increases trust; the latter erodes it.

15.2.3 Objection 3: "We already use AI explainability tools — this is redundant."

Section 6.1a addresses this in detail. AI explainability documents *why the model produced output X*. Decision Governance documents *why the practitioner accepted output X for this client given this context*. The two are complementary, not competitive. Most current tools fill the first need; the governance gap is the second.

15.2.4 Objection 4: "The Decision Packet adds work without billable hours attached."

The objection assumes the Packet is overhead. The Packet is the auditable record of work the advisor is already required to perform — the surfaced assumptions, the noted client divergences, the reasoned departure from the AI's recommendation. Without the Packet, this work is performed but not preserved. The cost is in preserving it, not in performing it.

The billable-hours question is itself revealing: in well-governed advisory practice, the time a practitioner spends surfacing assumptions, recording stakeholder divergence, and documenting reasoning is not separate from the consulting work; it is the consulting work. Asking "where are the billable hours for the Decision Packet?" is structurally similar to asking "where are the billable hours for the Geeignetheitserklärung?" — the answer in both cases is that the work is the documentation; the billable hour is in the conversation that produces both. Hedged ROI framing (not a sales claim): in contested-review contexts, a Decision Packet record may materially affect the firm's liability allocation in ways that an unstructured CRM note cannot. Whether that translates to a defensible ROI estimate depends entirely on (a) the firm's contested-review base-rate, (b) the average exposure per incident, and (c) the marginal time-cost per packet. The framework neither asserts nor recommends a specific ROI multiplier; that calculation belongs to each adopting firm's own risk-and-time analysis.

15.2.5 Objection 5: "Bayesian confidence indicators will confuse clients."

Constructed Ambiguity (Pattern 4) addresses this. Confidence indicators are not exposed to clients as Bayesian posteriors; they are exposed as one of three calibrated categories (high / moderate / low) with a one-sentence explanation of what the category means for the recommendation in front of them. The internal mathematics is invisible. What clients see is whether the system itself considers its recommendation strongly or weakly grounded in the available evidence — information clients already attempt to infer from advisor tone, often inaccurately.

15.2.6 Objection 6: "Our smaller practice cannot afford the technology investment."

The architecture is intentionally implementable at multiple scales. The reference implementation runs on commodity infrastructure with open-source components (see Appendix E). The cost is in the editorial discipline of producing Decision Packets, not in the technology. Practices that consider documentation a cost center will resist this; practices that consider documentation an asset will adopt it earliest.

15.2.7 Objection 7: "What happens when the AI is wrong and we documented our agreement?"

The objection inverts the actual risk. Without the Packet, an AI error becomes the advisor's error by default — there is no record of the reasoning that connected AI output to advisor decision. With the Packet, an AI error remains traceable to its source: the AI's model version, input data, and output are preserved; the advisor's reasoned acceptance or modification is also preserved. The Packet does not transfer liability; it makes liability allocation reconstructable. In current practice, the absence of records does not protect the advisor — it merely makes contested allocations arbitrary.

15.3 F.3 — Status and Open Empirical Question

The objections above were collected through unstructured conversation and informal feedback. They are not validated through a controlled survey with documented sampling methodology, statistical power analysis, or representativeness assessment. Wider validation across an advisory-firm sample is identified as a research extension (Chapter 7, RQ4-Extension).

Researchers seeking to extend or refute this catalogue are encouraged to publish their methodologies and findings; the framework authors will treat replication studies as primary input to subsequent versions of this appendix.