Context Is the Way Back

Accountable Mediation, Agentic AI, and the New Handoff Problem

Michael Stoyanovich

Version 1.1.2 | July 2026

Abstract

Agentic AI shifts AI governance from output review to workflow accountability. As AI systems begin planning, retrieving, drafting, reviewing, remembering, and acting through tools, organizations need to understand not only what the final output says, but how it was produced. This brief argues that agentic AI creates a new handoff problem: context, evidence, uncertainty, and accountability must survive movement across agents, tools, memory, validators, and human review points. The practical answer is accountable mediation: a disciplined way to preserve a path back to evidence, traceability, consequence, and competent human ownership. This brief offers a practical test organizations can apply to agentic workflows before treating them as mature, auditable, or suitable for high-consequence reliance.

Framing Note

This is a practitioner note, not a technical evaluation of agentic AI architectures. It is written for leaders of risk, operations, IT, governance, security, privacy, compliance, and professional practice who are deciding how to incorporate agentic AI into accountable workflows.

The concern is not that all mediation is bad. Modern knowledge work has always depended on summaries, briefings, dashboards, experts, assistants, citations, procedures, institutional memory, and professional judgment. The concern is whether mediation remains visible enough to be responsibly used. This brief is meant to support conversations about when agentic workflows are ready for accountable production use.

Agentic AI makes this question more urgent. As AI systems begin planning, retrieving, drafting, reviewing, synthesizing, remembering, and acting through tools, organizations will need a reliable way back to reality: back to evidence, back to context, back to uncertainty, back to consequence, and back to the human organization that must finally own the action.

Leaders can use the Accountable Mediation Test and diagnostic questions as part of workflow design reviews, RFP assessments, risk committee discussions, or production-readiness reviews.

1. Agentic AI Changes the Governance Question

Most AI governance conversations still begin with the output. Was the answer accurate? Was the response biased? Was the source real? Did the human review it? These questions remain necessary, but they are no longer sufficient.

AI use is moving from discrete interactions with a chatbot toward workflows in which multiple agents may plan, retrieve, draft, review, validate, remember, and act through tools before a human sees the result. A person may not be reviewing "the model's answer" in any simple sense. They may be inheriting the output of a small computational workflow.

That changes the governance question.

The question is no longer only:

> Is the output right?

It is also:

> Can the organization trace the output back to evidence, context, uncertainty, tool use, and accountable human judgment?

Canhui Liu's paper, *The Organizational Behavior of Agentic AI*, gives useful language to this shift. Liu argues that agentic AI is a "partial organizational analogue." Agent collectives resemble organizations because they differentiate work, coordinate interdependence, enact routines, cross boundaries, and produce collective outcomes. But they differ because their behavior is sustained by context architecture rather than human motivation, identity, trust, socialization, or moral accountability (Liu, 2026).

This brief treats Liu's paper as a useful mechanism-oriented lens, not as settled field evidence about all agentic systems or organizational deployments.

That distinction matters. Human teams are shaped by roles, norms, experience, incentives, memory, tacit knowledge, and responsibility. Agentic systems do not participate in those forms of social life. They operate through prompts, schemas, memory, traces, validators, tools, permissions, and orchestration.

Agentic AI does not eliminate organizational problems. It relocates some of them into context architecture.[1] Context architecture means the prompts, schemas, memory designs, traces, validators, tool permissions, and orchestration rules that shape how agentic systems perform work.

Work still has to be decomposed. Evidence still has to travel. Errors still have to be detected before they cascade. Permissions still have to be bounded. Accountability still has to be reconstructed. But those problems increasingly occur inside technical workflows that may be invisible to the person receiving the final output.

This also changes how work is experienced. Some work becomes more visible because it is formalized in a trace. Some work becomes less visible because it is hidden inside prompts, memory, tooling, orchestration, and vendor-managed systems, including external orchestration and managed tooling. That shift matters later in this brief, when we turn to the Employment Game.

That is the new handoff problem.

[1] This is also the point of contact with the Wittgensteinian thread in my Four Philosophers Framework: meaning and reliability do not reside in fluent outputs alone, but in use, context, rule-governed practice, and the forms of activity in which language is embedded. In this brief, that philosophical concern is translated into a governance question: whether agentic AI workflows preserve enough evidence, context, uncertainty, traceability, and accountability for responsible organizational use.

2. The Old Problem Returns: Mediated Knowledge

AI did not invent mediated knowledge.

Professionals have always relied on things they did not directly produce or fully inspect: memos, summaries, abstracts, citations, expert opinions, dashboards, briefing books, prior workpapers, analyst notes, search results, institutional records, and inherited assumptions. Much of organizational life depends on responsible secondhand knowledge.

Pierre Bayard is useful here as a light reference point. His work on talking about books one has not read is not a governance framework and should not be stretched into one. But it does help name a familiar condition: our relationship to knowledge is often partial, mediated, social, and situated. We know some things directly, some by summary, some by reputation, some through others, and some only through traces.

Agentic AI makes Bayard's problem organizational. The question is no longer only, "Did I read the source?" It becomes:

> What kind of contact does this workflow actually have with the evidence?

The question is no longer only whether someone read the source, but how each step in the workflow touched, transformed, preserved, or left the evidence behind.

An AI-assisted workflow may retrieve a source, summarize it, compress it, pass it to another agent, synthesize it with other material, have a reviewer agent critique the result, store a version in memory, and present a polished answer to a human. At that point, the human may be several steps removed from the original evidence.

That does not make the output useless. But it does mean that responsible use requires more than confidence in fluency. It requires accountable mediation: a way to make mediation visible, bounded, reviewable, and owned.

A simple distinction helps:

> Context architecture is the designed structure.
> Context boundaries are where that structure is tested.
> Mediated work is what people experience when they rely on the resulting output.

The governance task is to make that mediation visible enough for responsible use.

3. The New Handoff Problem: Context-Boundary Risk

Agentic workflows create context boundaries. This section names the mechanics of the problem: where context crosses boundaries, how it can degrade, and why familiar handoff risk becomes harder to see in agentic workflows.

A context boundary appears whenever task information moves from one role, step, system, agent, memory store, tool, or reviewer to another. At each boundary, something happens to the work. It may be preserved, clarified, enriched, compressed, transformed, omitted, distorted, or made to look more settled than it is.

Liu calls the relevant mechanism contextual transaction cost. In agentic AI, the costly transaction is often not bargaining, contracting, or enforcement in the classic economic sense. It is the cost of making

task context usable across boundaries.

A planner's decomposition must be usable by a solver. A solver's claim must carry enough evidence for a reviewer. A memory system must preserve useful state without stabilizing error. A human manager must be able to audit the trace before the output enters an accountable workflow (Liu, 2026).

A familiar governance analogy is handoff risk. Organizations already understand that work can fail when one team hands incomplete requirements to another, when a vendor receives partial instructions, or when a reviewer sees a conclusion without the workpapers. Agentic AI introduces the same problem inside computational workflows, but faster, more quietly, and often with more polished outputs.

For example, a benefits-eligibility workflow might use one agent to interpret plan language, another to extract participant data, another to draft a determination, and another to review the answer. That may be useful. But if the first agent misreads the provision, the second summarizes incomplete data, and the reviewer sees only a polished determination, each context boundary quietly carries forward a hidden defect. The workflow looks controlled while its internal representations drift away from the actual plan and data.

In agentic workflows, context may fail in familiar but hard-to-see ways:

- A planner frames the wrong problem.
- A retriever attaches weak or incomplete evidence.
- A solver treats a compressed summary as authoritative.
- A reviewer shares the same blind spot as the original agent.
- A memory store preserves an earlier mistake.
- A final synthesis hides uncertainty behind fluent prose.
- A human receives the answer without enough trace to judge what happened.

Each item is familiar in human workflows. Agentic systems can make them quieter and faster.

The risk is not merely that context is lost. The risk is that context is lost while confidence increases.

Done Is Not a Green Check

Agentic workflows make context-boundary risk sharper because they often compress several different institutional claims into one visible status: complete, passed, resolved, ready, approved, safe, or deployable. But these are not the same claim.

Done enough to draft, done enough to merge, done enough to deploy, done enough to rely on, and done enough to announce each require different context, evidence, authority, and risk acceptance. A single completion signal can be useful as a workflow aid, but it should not be mistaken for accountable judgment.[2]

This is one reason context remains central. Context tells the organization what kind of claim is being made, who is authorized to accept it, what evidence is sufficient, and what consequence follows. Without that context, "done" becomes a user-interface signal. With context, "done" becomes an accountable institutional judgment.

[2]Dotta, in the AI Engineer YouTube talk "What Does Done Even Mean? Agents and Paperclip's Liveness Model," uses Paperclip's liveness model as a practitioner example of distinguishing different operational meanings of "done" in agentic workflows. I use the talk here as a practitioner framing example, not as formal empirical evidence.

A practical handoff test follows:

- Done for whom?
- Done for what next action?
- Done under whose authority?
- Done with what evidence?
- Done at what level of institutional risk?

This is where synthetic organizational confidence appears. Synthetic organizational confidence arises when a multi-step workflow appears thorough, deliberative, or reviewed, even though its internal representations have drifted away from the underlying evidence.

This is why final answer quality is not enough. Liu notes that a high-quality answer produced through opaque or fragile context movement may be difficult to embed in accountable work, while a slightly lower-quality answer with stronger evidence preservation may be more useful in legal, engineering, or scientific workflows (Liu, 2026).

That claim should be treated as a useful mechanism-oriented insight, not as settled field evidence about all agentic systems. The practical point remains strong: in organizational settings, the better answer may be the one that can be reviewed, explained, bounded, corrected, and responsibly used.

Governance cannot stop at the surface of the output. It has to inspect the handoffs.

4. The Risk Is Not Only Hallucination

For organizational governance purposes, the common AI risk vocabulary is still too output-centered.

Hallucination matters. Bias matters. Confidentiality matters. Security matters. Model quality matters. But agentic AI adds another risk: synthetic organizational confidence.

The next risk pattern is what happens when context-boundary failures are hidden by the appearance of structured review.

A multi-agent workflow can look like a team. It may include a planner, researcher, drafter, reviewer, critic, memory manager, tool user, and orchestrator. It may simulate deliberation, escalation, review, and consensus. But the appearance of organizational form does not guarantee organizational reliability.

This is different from simply trusting a model too much. Synthetic organizational confidence arises when the workflow structure itself manufactures the appearance of deliberation, review, and consensus.

Liu makes this point directly. Human-imitation forms can mislead. A hierarchy may become repeated context compression. A committee may produce correlated paraphrases without independent evidence. A pipeline may propagate early errors if intermediate representations are not independently checked (Liu, 2026).

This matters because organizations are likely to find human metaphors comforting. “Agent teams,” “AI managers,” “review panels,” and “committee workflows” sound familiar. But the borrowed language may hide the actual risk.

A human committee can provide different expertise, institutional memory, professional disagreement, and social accountability. An agent committee may simply remix the same prompt, retrieval base,

model assumptions, and error pattern.

That does not mean multi-agent workflows are useless. In well-designed systems, they can separate concerns, improve structured review, create better traces, and provide more reliable QA than ad hoc one-off model calls. Their value depends on whether they produce real independence, useful evidence preservation, and accountable review, rather than merely simulating those things.

The risk is not only that an AI system invents a fact.

The risk is that an agentic workflow manufactures organizational confidence around a weakly grounded result.

A workflow maintains contact with reality when it preserves evidence, shows its transformations, retains uncertainty, supports independent checking, exposes permissions and tool use, and returns the final decision to accountable human judgment.

A workflow loses contact with reality when it becomes internally coherent but externally thin: polished, plausible, multi-step, and difficult to challenge.

5. Accountable Mediation as the Governance Response

If context-boundary risk is the mechanism and synthetic organizational confidence is the risk pattern, accountable mediation is the governance response.

Accountable mediation is the discipline of designing and operating workflows so that each step of reliance on agents, tools, memory, prior work, and human review remains visible enough for someone to ask:

> What did we rely on here, and is it good enough for this consequence?

This is not the same as requiring every person to manually redo every AI-supported task. Nor is it a vague call for "human in the loop." Human review is only meaningful when the human has enough context, evidence, authority, time, and competence to review what matters.

Nor should organizations assume that a human reviewer is automatically superior to an AI-generated output. In some narrow QA settings, an AI system may be more consistent, faster, or better at detecting certain repeatable defects. The point of human review is not that humans are always better pattern-matchers. The point is that competent humans can bring contextual judgment, professional nuance, edge-case sensitivity, and accountability to consequences that agents do not bear.

That accountability matters. A human reviewer participates in the relevant professional, institutional, legal, or community practice. They can answer to a client, explain a judgment, recognize when a technically correct answer is practically wrong, and live with the consequences of the decision. An agent can simulate review, critique, or dissent, but it does not have skin in the game in the ordinary human sense: it does not belong to the practices in which responsibility, trust, harm, correction, and obligation have meaning. Human accountability is not just another workflow step. It is the point at which mediated output is returned to lived obligation and organizational consequence.

The Accountable Mediation Checklist provides a practical starting point. Its underlying concern is that professionals using AI should remain clear about what they asked, what the system produced, what sources or assumptions were involved, what was checked, what remains uncertain, and who owns the final communication or decision.

The Agentic Accountable Mediation Test extends that checklist from single AI interactions to multi-step agentic workflows. The question is no longer only how one person used AI. It is how the workflow itself mediated the work before the person relied on it.

The Test generalizes the Accountable Mediation Checklist from single interactions to multi-step workflows, shifting the focus from "what did I ask and see?" to "what did the workflow do before I saw anything?"

The Agentic Accountable Mediation Test

1. Source contact
What evidence, data, document, system, observation, or source supports the output?

2. Context continuity
What was preserved, summarized, omitted, compressed, or transformed across handoffs?

3. Uncertainty visibility
Did ambiguity, caveats, dissent, confidence limits, and unresolved assumptions survive the process?

4. Trace inspectability
Can a qualified human inspect the path from request to output?

5. Permission discipline
What tools, systems, memory stores, files, data sources, and external actions were available to the agents?

6. Independent check
Was there a genuine check using independent evidence, tools, models, methods, or human expertise?

7. Human ownership
Who is accountable for the final use, decision, recommendation, communication, or action?

These questions are not meant to turn every AI interaction into a formal audit. The level of discipline should match the consequence of the use case. A low-risk brainstorming workflow does not require the same controls as legal review, security analysis, clinical administration, benefits operations, regulatory interpretation, financial decision support, or client-facing advice.

But the basic principle holds: the higher the consequence, the more the workflow must preserve a path back.

Back to the source.
Back to the assumptions.
Back to the uncertainty.
Back to the trace.
Back to the human who owns the result.

6. Interface Organization and the Employment Game

Agentic AI is not only a technical workflow problem. It is also part of the changing employment game.

This returns us to the Employment Game: work as a socially embedded practice whose visibility, review burden, and ownership can be changed by mediation.

The usual employment question is whether AI will replace tasks, roles, or workers. That question matters, but it is too narrow for this brief. The narrower governance question is how agentic AI changes the distribution of task execution, review, judgment, visibility, and accountability.

Some work moves to agents. Some work moves to humans as review, correction, interpretation, escalation, and accountability. Some work disappears into prompts, tools, memory, orchestration, and vendor-managed systems.

That redistribution is itself a governance issue.

Liu's concept of an interface organization is useful here. An interface organization is the designed arrangement that connects human organizational behavior with agentic organizational behavior. It specifies which agent outputs can enter human workflows, what evidence must travel with them, what uncertainty must be disclosed, what permissions agents have, what traces are preserved, and where human judgment remains mandatory (Liu, 2026).

For purposes of this brief, interface organization is not a formal org chart. It is the explicit design of how human and agent workflows meet: what agents may do, what they must show, what humans must review, what evidence must travel, and who owns the result.

This is where agentic AI intersects with the Employment Game. The issue is not only whether AI performs work. The issue is how AI changes the rules by which work is recognized, reviewed, valued, and owned.

A professional may appear to be "assisted" by AI while actually becoming the accountable endpoint for a complex upstream system. The visible work may shrink: review, approve, send, sign. But the hidden work may expand: reconstruct, check, validate, test, review, correct, explain, and accept responsibility for the final use.

This is not automatically bad. In many settings, AI may reduce drudgery and improve first-draft quality. But if organizations do not notice the redistribution of work, they may misread what is happening.

They may think they have automated judgment when they have merely moved judgment to the end of the process. They may think they have reduced labor when they have converted labor into invisible review burden.

The employment question for this brief is therefore bounded: what human work becomes hidden, compressed, displaced, or converted into accountability when agentic workflows are introduced?

A human cannot meaningfully own a result if the path to that result is unavailable, unintelligible, or falsely reassuring.

7. A Practical Diagnostic for Agentic Workflows

Before relying on an agentic AI workflow, organizations should be able to answer a small set of practical questions.

Not every use case requires the same rigor. But when the output will shape advice, decisions, obligations, client communications, security actions, legal interpretations, regulated workflows, or material operational choices, these questions become important.

Before relying on an agentic AI workflow, ask:

Task & Delegation:

1. What task or decision is being supported?
   Clarify whether this is brainstorming, drafting, analysis, recommendation, decision support, or action execution.
2. What work was delegated to agents?
   Identify whether agents planned, retrieved, summarized, drafted, classified, reviewed, validated, decided, or acted.

Evidence & Uncertainty:

3. What evidence was used?
   Confirm whether the underlying sources, documents, data, systems, or observations are visible.
4. What changed across handoffs?
   Identify whether context was summarized, compressed, translated, filtered, ranked, scored, or transformed.
5. What uncertainty survived?
   Confirm whether caveats, assumptions, confidence limits, disagreements, missing data, and unresolved issues remain visible.

Checking & Trace:

6. What independent check occurred?
   Confirm whether the output was checked against independent evidence, a different method, a different model, a human expert, or authoritative source material.
7. What permissions were granted?
   Identify what files, systems, tools, APIs, memory stores, or external actions the agents could access.
8. What trace is preserved?
   Determine whether a qualified reviewer can reconstruct the path from instruction to output.

Ownership & Feedback:

9. What competent human role owns the result?
   Name who has the authority, context, and obligation to use, approve, communicate, reject, escalate, or act on the output.
10. What feedback will correct the workflow over time?
    Identify how errors, corrections, changed facts, failed assumptions, or operational consequences will be captured.

A workflow that cannot answer these questions may still be useful. But it should not be treated as mature, auditable, or suitable for high-consequence reliance.

The point is not to create paperwork for its own sake. The point is to prevent a familiar governance failure: adopting a tool faster than the organization understands the work it has changed.

8. Context Is the Way Back

Agentic AI will make some work faster, some work stranger, and some work harder to see.

The answer is not to reject mediation. Modern knowledge work has always depended on mediation. The answer is to make mediation accountable.

That means preserving the practical path back to reality.

Back to the evidence that supports a claim.
Back to the context that shaped the task.
Back to the uncertainty that should qualify the output.
Back to the trace that allows review.
Back to the consequences that test the result.
Back to the human organization that must finally own the action.

This is a modest governance claim, but an important one. It does not argue against sophisticated agentic architectures. It argues that their usefulness depends on maintaining a usable path back to evidence, context, uncertainty, and accountable human judgment.

In practice, this means designing agentic workflows with inspectable traces, structured uncertainty, explicit permission boundaries, and named human owners for consequential outputs.

In a governance stack, this note sits between technical design and policy: it asks whether agentic workflows preserve enough context, evidence, uncertainty, and trace for accountable human use before they are treated as routine infrastructure.

Context is the way back.

Related Work in This Corpus

This brief extends themes from the Four Philosophers Framework, the Accountable Mediation Checklist, and The Employment Game. The common thread is practical: AI-assisted work must remain connected to use, evidence, consequence, and accountable human judgment.

The Four Philosophers Framework provides the broader philosophical background, especially the Wittgensteinian emphasis on use, context, and rule-governed practice. The Accountable Mediation Checklist provides the human-facing discipline for visible, bounded, and owned AI assistance. The Employment Game widens the aperture to ask how AI changes task structure, review burden, competence visibility, and accountability. This brief brings those concerns together around agentic AI workflows.

References

AI Engineer. (2026, July 11). *What does done even mean? Agents and Paperclip's liveness model - Dotta, Paperclip* [Video]. YouTube. https://youtu.be/7P0elyLIxXo

Bayard, P. (2007). *How to talk about books you haven't read* (J. Mehlman, Trans.). Bloomsbury.

Liu, C. (2026). *The organizational behavior of agentic AI: Context, boundaries, and collective intelligence in human-agent workflows*. arXiv. https://arxiv.org/abs/2606.30986

Stoyanovich, M. (2026). *The accountable mediation checklist: Keeping AI-assisted work tethered to sources, evidence, and professional judgment* [Practitioner checklist].

Stoyanovich, M. (2026). *The employment game: Work as a socially embedded practice in the age of AI* (Version 1.8.1) [Working paper].

Stoyanovich, M. (2026). *Philosophy, cognitive science, and policy: Interdisciplinary perspectives on generative AI from Wittgenstein, Lewis, Dennett, and Nagel* (Version 1.24.2) [Working paper].

Ethics, Disclosure, and Acknowledgements

Ethical Considerations

This brief does not draw on private, sensitive, or personally identifiable data. Examples are hypothetical, generalized, or derived from public and professional contexts. No human-subjects research was conducted, and no institutional ethics review was required.

The ethical purpose is practical: to reduce responsibility leakage in AI-assisted and agentic workflows by keeping mediated outputs connected to evidence, uncertainty, traceability, and accountable human judgment. The brief is intended to support responsible organizational discussion, not to provide legal, regulatory, or professional advice.

Use of AI Tools

AI language models, including ChatGPT, were used during the writing process as drafting and thinking aids: for brainstorming, structure, clarity checks, citation cleanup, and editorial review. These tools can be wrong. Responsibility for all claims, framing, judgments, and conclusions remains with the human author.

Acknowledgements

Thanks to informal readers and reviewers whose questions, challenges, and editorial feedback improved the clarity, structure, and practical usefulness of this brief. All errors and omissions are the author's alone.

Disclosure Statement

This work was conducted independently, without institutional affiliation, funding, or external influence. The views expressed are the author's alone and do not represent any current or former employer. No financial or professional conflicts of interest are declared.

Trademark Notice

The Four Philosophers Framework(TM) and The 4-Philosophers Framework(TM) are unregistered trademarks of Michael Stoyanovich. The CC BY 4.0 license does not apply to these trademarks. Use of the trademarked names is permitted for scholarly citation or descriptive reference but may not be used in connection with commercial products, services, or branding without permission.

How to Cite This Brief

Stoyanovich, M. (2026). *Context Is the Way Back: Accountable Mediation, Agentic AI, and the New Handoff Problem* (Version 1.1.2). https://www.mstoyanovich.com

Version History and Document Status

This is a living document. As generative AI systems and their organizational use evolve, this brief may be periodically updated to incorporate new empirical findings, theoretical insights, governance practices, or implementation lessons. Major revisions are recorded here to preserve transparency and scholarly traceability.

Version	Date	Description
1.1.2	July 2026	Prepared publishable cleanup pass: removed formatting artifacts, compressed Section 5 opening, clarified Section 3 language, added a brief use cue in the Framing Note, block-quoted the mediated-evidence question, and strengthened the checklist bridge sentence.
1.1.1	July 2026	Tightened section signposting, clarified Liu's scope as a mechanism-oriented lens, refined the "done" discussion, moved the Dotta/Paperclip reference to a provisional footnote, and strengthened the Employment Game callback.
1.1	July 2026	Added a short clarification on "done" as a contextual governance claim in agentic workflows, including a practical handoff test and provisional reference to Dotta's YouTube discussion of Paperclip's liveness model.

Version	Date	Description
1.0	July 2026	Published practitioner brief; stable for citation.