TACIT FRAGMENTS

OPERATIONALISING TACIT KNOWLEDGE AS A GOVERNED MEMORY LAYER FOR AGENTIC AI

Arsalan Shahid[1,*], Gordon Suttie[1], Philip Black[1], Antonio Garzón-Vico[2]

[1]Brightbeam AI, The Minaun, Faithlegg, Waterford, Ireland

[2]School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, Ireland

[*] Correspondence: `arsalan.shahid@brightbeam.com`

ABSTRACT

For more than half a century, tacit knowledge has explained why skilled organisational work cannot be reduced to written rules or formal procedures. Competent action relies on situated judgement, perceptual cues, and experience-shaped discrimination. Historically, AI systems could not access this know-how because they required explicit rules, labelled examples, or articulated document corpora. The shift to agentic AI changes this practical reality: agents can now use tools, maintain memory, observe workflow traces, and interact with human workers directly. This capability raises a question: which aspects of tacit practice leave enough behavioural traces and contextual conditions to become useful computational artefacts? This paper presents *Tacit* as a governed memory layer for agentic AI that complements procedural, semantic, and episodic memory. To operationalise it, we introduce Metis, a reference architecture specifying how systems can responsibly capture, represent, validate, govern, and retrieve "tacit fragments". We delineate an *exogenous* pathway that captures fragments from situated human practice, the paper's primary focus, alongside an *endogenous* pathway where agents surface latent competence from their own operational traces. We set out how agentic systems can turn situated human practice into governed memory an agent can use, provided each fragment remains strictly bound to its conditions, provenance, confidence, and human validation.

Keywords Tacit knowledge · Agentic AI · Memory architecture · Knowledge governance · Collaborative human-agent protocol (CHAP) · Human-AI collaboration

1 Introduction

Organisational work depends heavily on knowledge that people use but cannot fully articulate. A technician may hear that a machine sounds wrong before any warning light appears; an analyst may sense that a sample looks off before any test confirms it. Manuals and standard operating procedures do not capture this know-how because much of it appears only in practice through judgement, perception, and timing. This is tacit knowledge.

While expert systems elicited rules, machine-learning models inferred patterns from logged data, and retrieval-augmented systems connected models to document corpora [1, 2], they faltered where knowledge was embodied or perceptual. They assumed designers could state the relevant content explicitly. Agentic systems make bridging this gap urgent. As agents take on memory, planning, and action inside real workflows [3–12], they must operate where the decisive knowledge has never been written down. The design question is not whether a system can understand tacit knowledge exactly as a person does. Instead, we must ask whether selected aspects of tacit practice can be made computationally useful under real accountability.

Two pathways could build this tacit competence into an agent. The exogenous pathway captures fragments from observed human practice. The endogenous pathway infers tacit competence from the agent's own operational traces. Because operational settings demand accountability, this paper advances the exogenous pathway as its primary contribution.

Code and reference implementation: `https://github.com/BrightbeamAI/metis`

We propose the fourth stratum: a governed tacit-memory layer for agentic AI. It does not stand apart from procedural, semantic, and episodic memory as a fourth silo; what distinguishes it is its governance contract rather than its data, and it draws on the other three rather than replacing them. Rather than treating tacit knowledge as impenetrable mystique or a database awaiting extraction, the stratum stores partial, situated, and inspectable fragments of practice. Each fragment remains inextricably tied to its provenance, the conditions under which it applies, its uncertainty, and its human validation state.

Agentic AI cannot support skilled work reliably by merely adding documents to a language model; it needs a disciplined, governed way to handle the tacit residue of practice. We refer to this layer as *Tacit*, the fourth stratum; the reference architecture that populates and governs it is Metis, named after *mētis* (μῆτις), the classical Greek term for situated, practical intelligence [13, 14]. Our argument is not tied to any single sector. The same problem, and the same architecture, apply wherever skilled work leaves behavioural traces and presents natural decision points: a manufacturing line, a hospital ward, a financial underwriting desk, or a software operations team. Most worked examples in this paper come from regulated industrial and laboratory settings, where the cues are concrete and the governance stakes are explicit; the architecture itself is domain-general.

2 Tacit Knowledge

Organisations do not run only on documents, rules, procedures, and recorded data. They also depend on workers knowing what to notice, when to pause, how to adjust, and when a written rule no longer fits the situation. Current AI systems are powerful, but they remain poorly equipped to work with this kind of know-how. Much of it is difficult to state in advance and becomes visible only in practice. This is the intuitive problem of tacit knowledge: the practical understanding that helps people act competently even when they cannot fully explain the basis of their judgement.

Tacit knowledge has interested many disciplines because it names a simple but difficult problem: people often know more than they can say. Polanyi made this point directly, arguing that skilled action depends on things we rely on without being able to bring them fully into words [15, 16]. Ryle expressed a similar idea through the distinction between "knowing-how" and "knowing-that" [17], and Dreyfus later used it to challenge the view that expertise can be reduced to following explicit rules [18]. The argument then moves from the individual to the social setting in which expertise is learned and used. Collins showed that expertise is often transmitted through situated practice rather than formal instruction alone [19], while Tsoukas warned that articulating tacit knowledge does not simply preserve it, but also changes it [20]. Organisational research extends this insight by showing that tacit knowledge is embedded not only in individuals, but also in routines, communities, tools, and everyday coordination [21–25].

This is why tacit knowledge is central to the knowledge-based view of the firm. Organisations compete not only through the information they possess, but through their ability to use knowledge in skilled and coordinated ways. A senior worker, for example, may not be able to give a complete rule for when a process is drifting, but may still recognise the pattern earlier than a novice. Tacit knowledge can therefore support durable organisational capability precisely because it is difficult to observe, copy, and transfer [26–29].

2.1 A Working Definition

To explain how AI agents might incorporate aspects of tacit knowledge, we first need a working definition. We define tacit knowledge as experience-based know-how that shapes competent action but cannot be fully expressed as rules, facts, or instructions at the moment it is used. It is acquired mainly through situated engagement with tasks, tools, people, and environments, and it loses some of its meaning when translated into words, rules, or formal representations. This leads to our first two working principles that will guide our technical approach to capturing tacit knowledge.

Operational principle 1: *Tacit knowledge is never fully captured; any translation into words, rules, or data is necessarily partial and may lose practical meaning.*

Operational principle 2: *Tacit knowledge should therefore be approached through triangulation, combining expert explanation, situated observation, and evidence from outcomes.*

Our definition highlights three characteristics of tacit knowledge. First, it is experience-based and develops through practice rather than formal instruction alone. Second, it guides competent action even when individuals cannot fully explain how they know what to do. Third, it resists complete codification, meaning that some of its meaning is lost when translated into explicit rules, procedures, or representations.

This definition also helps clarify two common misconceptions about how, and whether, tacit knowledge can be handled. The first is to treat tacit knowledge as an impenetrable mystery: so deeply embedded in human practice that AI systems

can never interact with it. The second is to treat tacit knowledge as merely undocumented explicit knowledge: a hidden set of rules waiting to be extracted with the right method. Both views are misleading. Some aspects of skilled practice may permanently resist articulation, but others can be partially surfaced through careful elicitation, observation, comparison, and reflection. At the same time, tacit knowledge should not be understood as a latent body of explicit rules waiting to be uncovered. Rather, it is expressed through situated judgement, experience, and practice, and can only be approximated through representations that capture selected aspects of that practice. The important point is that these surfaced elements should not be mistaken for tacit knowledge itself. They become useful for AI agents only when represented with their conditions, limits, source, and validation status.

Operational principle 3: *Any representation of tacit knowledge should remain tied to its source, conditions of use, limits, confidence, and validation status.*

In our working definition, tacit knowledge appears in different forms. Different research traditions highlight different aspects of the phenomenon. Phenomenological and practice-based accounts emphasise embodiment, tools, and situated action [18, 30–32]. Cognitive research distinguishes declarative knowledge from procedural and non-declarative forms of memory [33–35]. Naturalistic decision-making research shows how expertise often appears through pattern recognition, sensory cues, anomaly detection, and mental simulation under pressure [36–38]. Organisational research adds that tacit knowledge is also embedded in routines, absorptive capacity, communities of practice, and the interaction between individual and collective intelligence [21, 24, 39–41].

Following Collins [19], these differences can be understood by asking why the knowledge remains tacit. Some tacit knowledge is relational: it remains tacit because the right explanation, shared context, or social relation is missing, and may therefore become more explicit through interaction. Other tacit knowledge is somatic: it is tied to the body, perception, and skilled action, and may never be fully articulated even though it leaves observable traces. Organisational work also involves collective forms of tacit knowledge, where competence is distributed across routines, roles, tools, and coordination patterns rather than held by one individual alone.

Together, these perspectives lead to an important operational conclusion that should inform any attempt to identify, codify, and insert tacit knowledge: tacit knowledge is not one uniform reality located only inside individuals. It is a family of situated competences that differ in how visible, social, embodied, and codifiable they are. For example, a troubleshooting heuristic, the felt resistance of a tool, the timing of a material process, and the rhythm of a team handover are all forms of tacit competence, but they cannot be identified, codified, or inserted into AI agents in the same way. Existing categories in the literature provide a useful starting point for this work, but they should not be treated as final. As AI systems become more deeply embedded in organisational practice, new distinctions may be needed, informed both by theory and by the practical difficulties encountered when trying to capture tacit knowledge in real work settings.

Operational principle 4: *Different forms of tacit knowledge require different methods of identification, codification, and insertion into AI agents.*

Operational principle 5: *Codified representations of tacit knowledge should use a governed, human-readable semantic vocabulary, grounded where applicable in the organisation's regulatory and quality language and adapted to its domain.*

Given this understanding of tacit knowledge, there are two broad ways to approach the problem of building AI agents that can work with it. The first is a developmental or naturalistic approach. Instead of extracting tacit knowledge from humans, this approach asks how agents might acquire functionally similar competence through structured engagement with tasks, tools, environments, and feedback over time. In human learning, tacit knowledge develops through the interaction between internal dispositions, such as attention, motivation, personality, and prior knowledge, and external conditions, such as repeated practice, social learning, feedback, and exposure to variation. A naturalistic approach would therefore seek to recreate some of these developmental conditions for AI agents. This is an important direction, but it lies beyond the scope of this paper.

The second approach is an extraction approach. It starts from existing human expertise and asks how aspects of tacit knowledge might be identified, codified, and made available to AI agents. This involves three steps: identifying the parts of tacit practice that matter for performance, translating them into representations that can be stored and inspected, and inserting those representations into agents so that they improve action in similar situations. This paper develops the extraction approach. The naturalistic approach remains a complementary direction for future work.

2.2 The Extraction Approach

The extraction approach raises three challenges. First, we must identify which aspects of human practice actually matter for competent performance. Second, we must codify those aspects into a form that can be stored, inspected, and reused. Third, we must make that representation available to the agent in a way that improves action in similar situations. Each step introduces loss. At the identification stage, important cues may be missed or the wrong elements may be treated as causally important. At the codification stage, meaning may be distorted when know-how is translated into words, rules, examples, or data. At the insertion stage, the agent may apply the representation in contexts where it no longer fits. The danger, therefore, is that the system inherits only a thin or distorted version of human expertise: one that appears competent under familiar conditions but performs poorly when the context changes.

2.2.1 Identification

Identification means deciding which parts of human practice actually matter for performance. This is difficult because the causes of skilled action are often unclear. Even when an expert acts successfully, it may be hard to know which cue, habit, judgement, or environmental condition made the difference. Asking the expert does not fully solve the problem. People may struggle to explain why they acted as they did, or they may offer a plausible explanation that was not the real basis of their action. There is therefore a gap between doing something well and knowing why one did it well. This echoes Wittgenstein's view that skilled action often involves knowing how to go on within a practice, rather than following an explicit rule [42]. For AI agents, the implication is clear: identifying relevant tacit knowledge cannot rely on unstructured self-report alone. Structured elicitation interviews, particularly cognitive task analysis methods such as the Knowledge Audit and the Critical Decision Method, are among the effective ways to surface relevant tacit knowledge. The limitation is not interviews as such, but interview quality, method, domain context, and validation. Such interviews should therefore not stand alone; they are most reliable when treated as part of a triangulated, continuous process that compares what people say, what they do, and what happens as a result (Operational principle 2).

2.2.2 Codification

Codification is the second challenge. Even if we identify aspects of practice that seem relevant to skilled action, we still need to translate them into a form that an AI agent can store and use. This translation is not neutral. A cue that matters in practice may lose meaning when expressed as a sentence, a rule, a label, a prompt, or a training example. Quine's idea of translational indeterminacy is useful here as an analogy: meaning does not move perfectly from one system of representation to another [43]. The same problem applies to tacit knowledge. When a worker says that a sample "looks wrong," or that a handover "does not feel complete," the phrase only makes sense against a background of experience, context, and shared practice. Codification must therefore do more than record the statement. It must preserve, as far as possible, the conditions under which the judgement was made, the cues that supported it, the limits of its use, and the evidence that connects it to a better outcome. Otherwise, the agent receives an isolated representation without the practical context that made it meaningful.

Codification must also do more than preserve that context: it must place the representation inside a shared semantic structure. Organisations need a standard semantic vocabulary for codification, a controlled, human-readable language that defines the categories, conditions, cues, evidence types, confidence levels, authority states, escalation rules, and validation statuses used across the system. In regulated industries, this vocabulary should be grounded in part in the language of the applicable regulatory and quality system, giving a conforming default for risk, control, deviation, evidence, validation, review, authority, escalation, and change control. It cannot be only regulatory, however; it must be extended to the organisation's specific products, markets, equipment, workflows, roles, and decision points. The aim is that human workers, domain experts, quality reviewers, and AI agents operate from the same controlled vocabulary wherever possible, minimising ambiguity, ideally to near zero in safety-critical or regulated workflows, while allowing the vocabulary to evolve as new practices and contexts emerge. Without such a vocabulary, codified records risk becoming isolated free-text statements, local idioms, or agent-generated paraphrases that different readers interpret differently, weakening retrieval, validation, auditability, and operational trust (Operational principle 2).

2.2.3 Insertion

Insertion is the third challenge. Even if relevant tacit knowledge has been identified and partially codified, the agent still needs to use it correctly. This is not simply a matter of adding the representation to a prompt, memory store, or training set. The agent must know when it applies, when it does not, how much authority it should have, and when to defer to a human. A statement such as "wait longer when the material looks unstable" may be useful in one setting and dangerous in another. If the agent retrieves it too broadly, treats it as a rule, or applies it without checking the context, the original expertise can become a source of error. The problem is therefore not only how to put tacit knowledge into an agent, but

how to control its use. Safe insertion requires strict applicability conditions, confidence boundaries, and escalation rules, so that the agent treats codified tacit knowledge as situated guidance rather than universal instruction.

These three challenges require clearer terminology. The extraction approach does not transfer tacit knowledge itself into an AI agent. Rather, it works through two intermediate concepts. ***Tacit practice*** refers to the observable way that tacit competence appears in situated work. ***Tacit fragments*** refer to partial representations of that practice that may be made available to AI agents. A fragment is not a free-floating note or general lesson. It is a bounded representation linked to its source, the conditions under which it applies, its confidence level, and its review status.

For example, a worker in a manufacturing setting may have an internalised sense that a machine is beginning to behave unusually. That sense is tacit knowledge: the worker may not be able to articulate it fully, and any explanation may be incomplete. Reducing the load before an alarm is triggered is tacit practice. A record stating that a particular vibration pattern under high load may warrant reducing throughput is a tacit fragment, but only if the record also specifies where it came from, when it applies, how reliable it is, and who has reviewed it.

We give a tacit fragment a precise formal structure in Section 3.2, alongside the fourth stratum that stores it (Section 3.3).

2.3 Two Pathways to Source Tacit Fragments

The extraction approach can source candidate tacit fragments from two pathways. Sourcing here refers to the identification stage of the extraction approach; each pathway surfaces candidate fragments, which then pass through codification and insertion. Both feed the same fourth stratum and pass through the same qualification and governance: the checks that decide whether a candidate fragment is admitted to the stratum and the controls on how it may subsequently be used, set out in Sections 4.4 and 4.5.

2.3.1 Exogenous Pathway

The exogenous pathway sources fragments from situated human practice. Candidate fragments are surfaced through observation and in-flow probing at natural workflow pauses, and worker confirmation. Worker confirmation does not by itself guarantee fidelity, since experts cannot always fully articulate their practice and may offer post hoc rationalisations; a stronger validation layer, such as triangulating across multiple experts so that fidelity does not depend on a single practitioner's rationalisation, re-eliciting an explanation at different times and in different formats, and drawing on research into cognitive biases, remains an open challenge we return to in Section 5.1. This is the primary pathway developed in this paper and its architectural detail follows in Sections 3 and 4.

2.3.2 Endogenous Pathway

The endogenous pathway sources candidate fragments by trace-analytic inference over the agent's own operational logs. Rather than creating new competence, it detects competence already latent in the agent's history by reading the three established memory strata, procedural, semantic, and episodic memory, against one another. Procedural memory records what the agent was supposed to do; episodic memory records what it actually did, under what conditions, and with what result.

The system identifies where tacit competence has accumulated by comparing the intended procedure, the specified rule recorded in procedural memory, against the enacted event. We formalise this detection as a difference signal: $\Delta = \text{enacted} \ominus \text{specified}$. Here $\ominus$ denotes a structured difference, not arithmetic subtraction: the Δ operator performs a structured comparison of two event sequences over episodes matched on task type and operating conditions, capturing the steps, timings, conditional pauses, and choices in the enacted episode that differ from the specified rule. Semantic memory supplies the conceptual frame for interpreting these gaps; a gap may reflect a genuine adaptation, but it may equally reflect an error, drift, or a contaminated input, and only recurrence across cases together with later validation can tell these apart. A candidate fragment is proposed only when a specific deviation recurs and stabilises across multiple cases. In this sense, the endogenous pathway turns the same work-as-imagined versus work-as-done diagnostic [44, 45] that the exogenous pathway applies to human practice inward onto the agent's own operational history. The two pathways are best read not as competing approaches but as two sources of data feeding a single work-as-imagined versus work-as-done diagnostic: the exogenous drawn from situated human practice, the endogenous from the agent's operational history. It meets analogues of the three challenges in Section 2.2: the difference signal does the work of identification, the structured fragment record does that of codification, and the authority layer governs insertion.

The difference signal is the starting point, not the end state. Once surfaced, a stable deviation becomes a candidate tacit fragment exactly as an observed human practice does, and it enters the fourth stratum only by passing through the same qualification and governance as any exogenous fragment: it is recorded with its provenance and conditions of

application, tested against evidence, and assigned an authority level under human validation. The endogenous pathway changes how a fragment is found; it does not change what a fragment must satisfy to be trusted.

The validation problem bites harder on the endogenous pathway than on the exogenous one. A stable deviation may encode genuine competence, or it may encode a spurious regularity, a contaminated input, or an artefact of training, and that distinction cannot be drawn inside the same loop that produced the deviation. Candidate fragments surfaced this way must therefore enter the stratum strictly as hypotheses, flagged with endogenous provenance and held to a higher evidence bar. The architecture does not permit the agents to promote them to controlled authority on its own assessment.

3 From Tacit Practice to Tacit Fragment

3.1 Why Ordinary Agent Memory Is Not Enough

Watch anyone do skilled work and the same loop repeats: the worker looks at how things stand, does something, and sees what follows, then looks again. The three parts have plain names. The worker observes a state, takes an action, and gets an outcome. An AI agent works the same way: it reads what it can of a situation, acts through the tools it has, and takes in the result before acting again. This is the picture reinforcement learning has always used, and the recent work on agents builds on it [3–12]. It fits a person and a program equally well. A line worker, a clinician at a bedside, and a software agent reading a case file are all choosing actions in a situation they can see only in part.

Within this loop, a novice and an expert differ in one specific way. The novice does what the procedure says: in a given state, the novice takes the action the manual specifies. The expert does more. The expert notices cues the manual never mentions, such as a tool that is starting to bind or a sample that simply looks wrong, and in some states takes a different action that no written rule calls for. This is tacit knowledge at work. It shows up in what the expert does, in the gap between the action the procedure specifies and the action the expert takes, and not in anything the procedure records.

An agent's memory already comes in three established dimensions, and we can ask which of them records this gap. Procedural memory records what the agent was supposed to do, the specified rule. Episodic memory records what it actually did, under what conditions, and with what result. Semantic memory supplies the conceptual frame for interpreting them. Each dimension is real and useful, yet none records the gap itself: the action the expert takes when the rule does not fit appears in practice, but in no rule, history, or concept that these three dimensions hold. That residue is what a fourth dimension of memory would have to carry.

A foundation-model agent dropped into a real workflow starts on the novice's side of this gap. It is general, fluent, and well documented, yet it does not know this particular environment: not its edge cases, not the feel of its equipment, not the timing its experts have learned. The familiar ways of briefing an agent, context engineering, retrieval, and standard operating procedures, all give it the action the procedure specifies. None of them give it the action the expert takes when the procedure does not fit.

Foundation models change the tools system designers possess, but they do not erase the underlying epistemic limits. While these models can interpret complex language, execute tool calls, plan subtasks, and interact fluidly within human environments [3–5], this very fluency introduces danger precisely because it masks the gap: a system on the novice's side can still sound like an expert. Language models can hallucinate competence, infer too much from weak behavioural traces, and dangerously blur the line between providing advisory support and assuming operational authority [46–50].

A language model alone is not a tacit-knowledge system, nor is a standard document retrieval system. The distinctive opportunity instead lies in the agentic combination of interaction, memory, and situated context. An agent embedded within a workflow can notice discrepancies between formal procedures and actual practice, ask a bounded clarification question at an appropriate moment, store the human's response with clear provenance, and retrieve that memory under similar future conditions. This controlled process does not achieve full articulation. Rather, it captures selected traces of practice, the divergences between what was expected and what actually occurred, as candidates for organisational memory. A divergence becomes reviewable organisational memory only once elicitation or another identification method establishes why it occurred.

3.2 The Tacit Fragment as the Unit of Capture

That reviewable memory has a unit, the tacit fragment, and the fragment needs a precise structure before an agent can store, retrieve, and inspect it. Each fragment is governed in the sense that it carries an explicit wrapper of provenance, conditions, and review status (Operational principle 3). For example, in a manufacturing setting, a worker's internalised feel for when a pump's vibration signals a failing bearing is tacit knowledge; easing back the load before the alarm

trips is tacit practice; and a governed record that "high-load operation plus low-frequency vibration warrants reducing throughput," tagged with its provenance, conditions, and confidence, is a tacit fragment.

We treat a tacit fragment as a record with fixed schema rather than free text: instead of a prose note such as "watch the pump under heavy load," each fragment is a defined set of typed fields. Formally, each fragment f is the tuple

$$f = \langle \text{content}, \ \text{provenance}, \ \text{conditions}, \ \text{confidence}, \ \text{authority}, \ \text{validation-state} \rangle \tag{1}$$

In this structure, *content* defines the specific cue or behavioural adjustment, and need not be textual: it may be a worker's phrasing, an annotated exemplar, or a learned representation of a signal such as an image or audio segment. *Provenance* logs the exact origin of the observation. *Conditions* restrict the context in which the fragment applies. *Confidence* quantifies the strength of the supporting evidence. *Authority* dictates the operational weight the agent may assign to the fragment, and *validation-state* records its formal human review status.

Each field mirrors a property of human tacit knowledge established above. Content captures the cue itself, here the low-frequency vibration under high load, because such knowledge is experience-based and acquired through situated engagement. Provenance records the engagement the cue came from, such as the shift on which the worker first linked the vibration to a failing bearing. Conditions fix the context in which the knowledge holds, here high-load operation rather than all running states, because it loses meaning when generalised. Confidence grades how strongly the evidence supports it, low after a single shift and higher once the same vibration-and-load pattern recurs across many shifts, because tacit knowledge is enacted with varying degrees of sureness. Authority sets how much weight an agent may place on it, here enough to let the agent suggest reducing throughput but not to cut it on its own, because expertise is distributed unevenly. Validation-state records whether a human has reviewed it before reliance, here unreviewed when first captured and confirmed only after a reviewer checks it, because externalised tacit knowledge is partial and fallible. This structure formalises the governance wrapper around a fragment, not the tacit content itself, which remains partial by definition. The taxonomy below, the governance layers in Section 4, and the two pathways in Section 2.3 all operate on fragments of the form given in Equation (1).

3.3 The Fourth Stratum

With the fragment's structure fixed, we name the layer that stores it. We propose the Tacit as the fourth stratum (Figure 1). This tacit-memory layer complements procedural, semantic, and episodic memory by representing selected fragments of situated practice as governable knowledge objects. We build this framework on three core design commitments. First, tacit knowledge remains partially representable but explicitly not fully extractable (Operational principle 1). Second, systems must capture representations context-first, because a fragment's meaning is conditional: a cue that is reliable under one set of circumstances can mislead under another, so a representation stripped of its originating context invites misapplication. The database must therefore store each fragment in full, every field of Equation (1), keeping its content together with the provenance, conditions, confidence, authority, and validation-state, that is, the exact circumstances under which workers observed, confirmed, and approved it for use. Third, organisations must strictly govern how systems use these fragments. A fragment should never influence action merely because a statistical model retrieves it.

Panel b of Figure 1 traces the two integration pathways: exogenous fragments from human practice and endogenous inferences from agent memory both populate the governed tacit store. We treat the fourth stratum as a distinct dimension, not a tag laid over what already sits in the other three memory dimensions. What makes it distinct is its governing contract, not its data source: every entry carries conditions of applicability, an authority level, a validation state, and a review obligation, and no entry may influence action except through that contract. The first commitment navigates between the two conceptual traps discussed above. Following Collins [19] (the relational, somatic, and collective bands introduced in Section 2.1), we recognise that relational tacit knowledge remains tacit for contingent reasons and can become explicit, whereas somatic tacit knowledge is inscribed in the body yet leaves observable correlates. Only collective tacit knowledge, held in the social fabric, genuinely resists representation [19]. The fourth stratum explicitly targets the first two bands. A fragment encodes the articulable cue, the observable adjustment, and the conditions under which they held. The agent consumes this partial representation as situated guidance rather than as a definitive rule. The fragment does not reproduce competence in full, nor does it claim to capture bodily residue without loss. The governance wrapper prevents the agent from mistaking the fragment for comprehensive expertise or applying it where its original conditions no longer hold.

3.4 Why Taxonomy Matters

Because tacit practice takes so many forms, developers must reject any universal capture method. Engineers must systematically ask what kind of tacit content is at play, how it physically appears in practice, what evidence validates it, and when the AI must unconditionally defer to human judgement. To prevent dangerous conflations across these

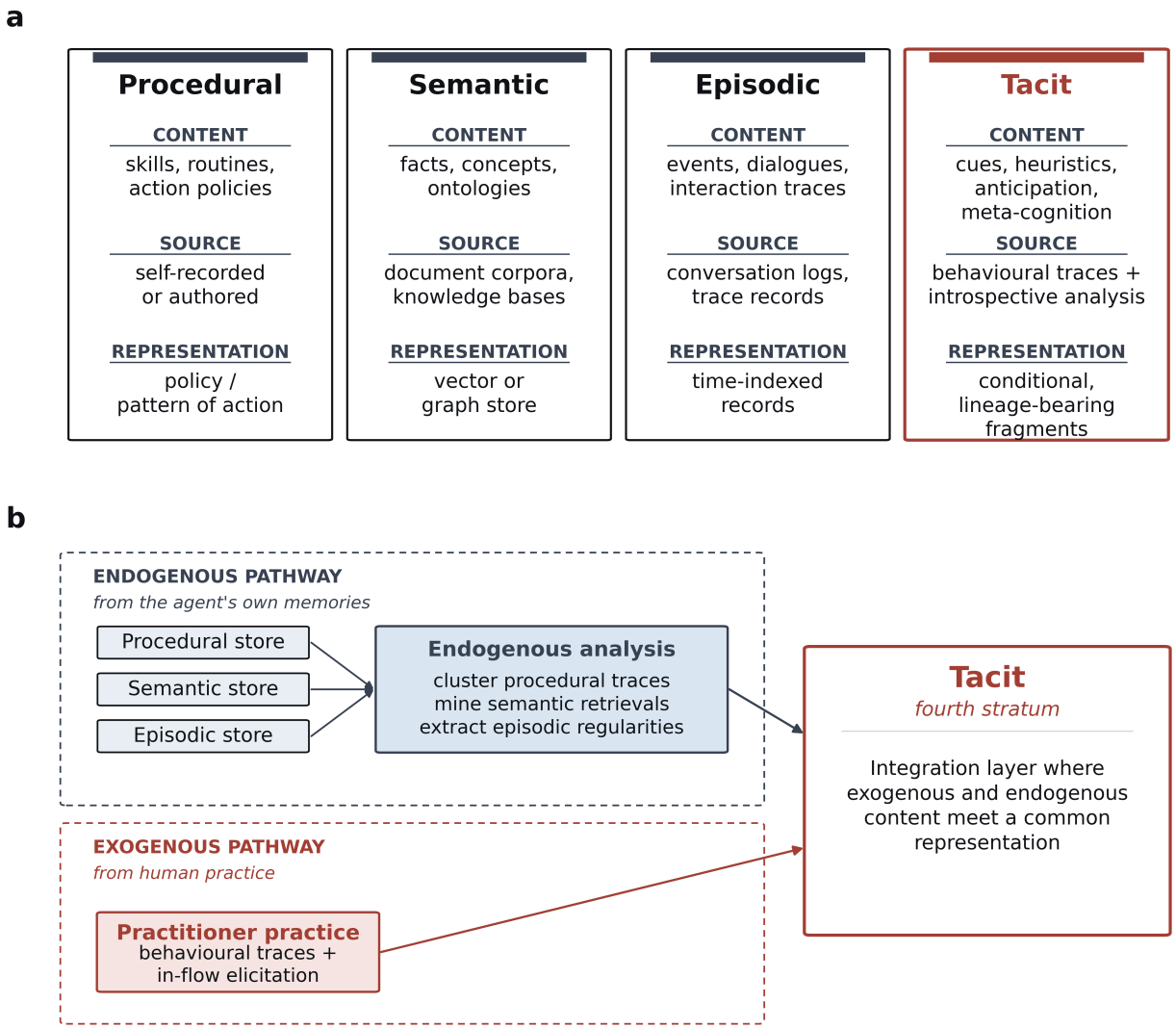

Figure 1. The fourth stratum. a, The proposed tacit-memory layer complements procedural, semantic, and episodic memory. **b,** Integration pathways showing how exogenous fragments from human practice and endogenous inferences from agent memory populate the governed tacit store.

different forms of practice, we require a clear working taxonomy (Operational principle 4). Figure 2 places these working categories on the conscious/automatic and individual/social dimensions drawn from Spender's organisational knowledge typology [24]. Figure 3 groups the same categories by the form of cognition or practice they primarily involve; the aim is design discrimination rather than classificatory completeness, since each category implies different capture and validation choices. Table 1 sets out the resulting working design vocabulary: for each domain, from procedural–embodied through social–normative, it pairs the tacit content at stake with its computational implication, so that forms of practice needing different capture and validation choices are kept apart.

The taxonomy outlined in Figures 2 and 3, and detailed in Table 1, provides this necessary design vocabulary. The complete multi-domain computational taxonomy and its associated capture-pathway mappings are retained in Appendix A (Table A1) and Appendix B (Table A2), respectively, to ensure the framework's architectural detail remains inspectable without disrupting the primary narrative.

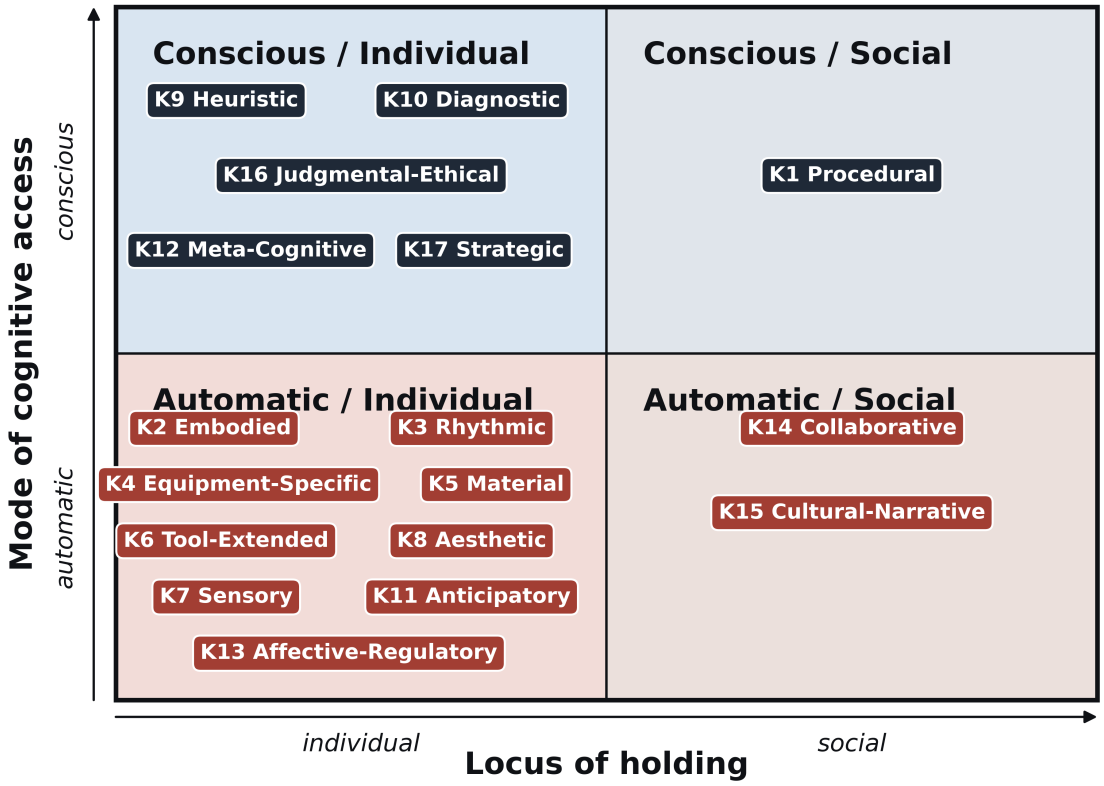

Figure 2. Tacit categories in relation to Spender's typology. The figure places the working categories against conscious/automatic and individual/social dimensions derived from Spender's organisational knowledge typology [24].

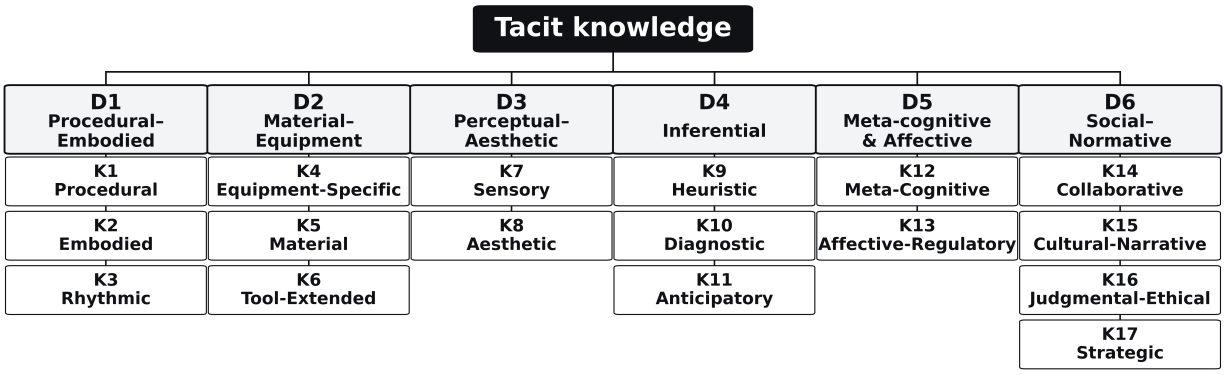

Figure 3. Working taxonomy for computational treatment. Categories are grouped by the form of cognition or practice they primarily involve. The purpose is not classificatory completeness but design discrimination: each category implies different capture and validation choices.

Table 1. Working design vocabulary for computational tacit fragments. The taxonomy is intended to separate forms of practice that require different capture and validation choices, not to provide a final ontology.

Domain	Tacit content	Computational implication
Procedural–embodied	Procedural routines, embodied skill, and timing of action	Compare formal procedure with observed action; capture timing and sequence under stated conditions.
Material–equipment	Equipment-specific adaptations, material feel, and tool-extended skill	Link fragments to equipment, material lots, tool substitutions, and local operating conditions.
Perceptual–aesthetic	Sensory cues and quality distinctions learned through exemplars	Use cue narration, exemplar annotation, and multimodal evidence; avoid converting perception into unsupported rules.
Inferential	Heuristics, diagnostic reasoning, and anticipatory judgement	Capture exception reasoning, incident reconstruction, and pre-event signals with human explanation.
Meta-cognitive and affective	Escalation judgement, confidence boundaries, and regulation under pressure	Treat pause, help-seeking, and composure as operational signals requiring careful validation.
Social–normative	Handoffs, cultural narratives, ethical judgement, and strategic sense-making	Use interviews, handoff analysis, and governance review; avoid direct automation of normative judgement.

4 The Metis Process: From Situated Practice to Governed Memory

The fragment's schema sets out what a finished record must contain. It does not settle who produces that record, who checks that it is faithful and safe, or who sets the limits on what it may do. These are separate jobs, and a system that lets one part do all three cannot be held accountable for any of them. Metis keeps them apart. It is a reference architecture that carries a fragment from a worker's situated practice into governed memory, assigning capture, validation, and enforcement to different units so that none can both create a fragment and grant it authority. Metis is built on the Collaborative Human-Agent Protocol (CHAP) [51]: the worker-agent exchanges it depends on, each confirmation, correction, validation, and handoff, are recorded as structured, append-only evidence rather than as untracked interactions in application code or chat threads. A reference implementation is available at `https://github.com/BrightbeamAI/metis`.

Metis does not assume an agent observing a person in a room. It assumes the work runs through systems the agent can read, for example a work-order tool, a clinical record, a case-management platform, or a code repository, and that the work has natural moments where a worker can answer a one-line question. On that basis the architecture is domain-general; the examples below are drawn from regulated settings only because the cues and stakes there are concrete.

Metis operates as a deliberately conservative framework. Rather than granting agents autonomous operational control, the architecture positions them strictly as bounded observers, pattern detectors, prompt designers, memory organisers, and safety monitors. The system does not make the agent the arbiter of expertise. Instead, human workers, supervisors, quality owners, and domain experts retain absolute responsibility for confirmation, interpretation, validation, and authority.

4.1 Overview of the Process

A candidate fragment travels from situated practice to governed memory, and the subsections that follow trace that path: capture detects where practice diverges from procedure and elicits the worker's explanation (Section 4.2); storage records the fragment as a structured object (Section 4.3); validation establishes its fidelity, relevance, and normative alignment (Section 4.4); governance assigns the authority layer that controls its use (Section 4.5); and retrieval applies it under the conditions it records, or escalates (Section 4.6). These stages meet the three challenges of Section 2.2: detection and elicitation do the work of identification, storage does that of codification, and governance and retrieval govern insertion.

To execute this safely, the architecture relies on three organisational primitives that physically separate capture from validation, and validation from policy enforcement. This separation remains critical because tacit knowledge becomes

computationally risky when the exact same system observes, interprets, authorises, and acts. We define these primitives concretely:

- **The Capture Cell:** this unit serves as the local hub of capture. It encompasses a worker or small team, the specific equipment or workflow in scope, and narrowly bounded agents that observe, infer, probe, confirm, and store.
- **The Mission Group:** this unit operates as the accountable review board. It examines candidate fragments across multiple cells, evaluates evidence, resolves conflicts, decides on promotion or rejection, and requests further elicitation.
- **The Runtime Orchestrator:** this unit functions as the policy enforcer. It strictly dictates permissions, audit lineages, safety rules, prompt budgets, retrieval constraints, and escalation requirements.

These three primitives map onto the validation and authority machinery introduced with Figure 4 below: a candidate fragment is checked in two validation tiers (Section 4.4) and may hold one of three authority layers, Evidence, Advisory, or Controlled (Section 4.5).

Figure 4 brings the three primitives together with the work they do, and the path a fragment travels through them. Validation runs in two tiers: Tier 1 is a local fidelity check in the Capture Cell, and Tier 2 is the Mission Group's review of evidence and normative alignment (both detailed in Section 4.4). Panel a traces one candidate fragment from situated practice to governed memory. The Capture Cell observes the work and runs a short local loop that turns a raw signal into a candidate fragment, and Tier 1 confirms on the spot that the description is faithful to what the worker meant. The Mission Group then applies Tier 2, weighs the evidence, and decides what role the fragment may play. The Runtime Orchestrator enforces the limits that role carries. Panel b sets out the three layers of authority a fragment can hold, and the gate it must clear to move up: Tier 2 to leave the Evidence Layer, and change control to reach the Controlled Layer. The two panels line up layer by layer, with the Capture Cell at the Evidence Layer, the Mission Group at the Advisory Layer, and the Runtime Orchestrator at the Controlled Layer. Because capture, review, and enforcement sit in separate primitives, no single unit can both produce a fragment and grant it authority, and a fragment advances only by passing the check that each stage requires. The Evidence, Advisory, and Controlled layers are defined in Section 4.5.

4.2 Detection and Elicitation

Capturing what a worker knows forces a trade-off between two methods, each weak where the other is strong. A deep interview produces real and largely durable knowledge, since much of what makes an expert valuable is built on processes that change slowly, but it is costly: it pulls the worker off routine work and cannot practically be run on every worker, decision, and shift, so a single pass can also miss changes that emerge later. Continuous observation is cheap and always current, but it yields signals rather than knowledge: it records that something happened, not what the worker understood. Metis escapes this trade-off by splitting capture into two activities that are usually run as one. Detection is the cheap, continuous half: it watches the work and notices when something is worth a closer look, and nothing more is asked of it. Elicitation is the deep half: it asks the worker to explain, but it runs only where detection points, so the costly step is spent on the few cases that warrant it rather than on every worker and every shift. The split turns observation into the trigger for an interview instead of a poor substitute for one, and it lets capture run as a continuous process within the organisation: deep interviews stay rare enough to be affordable, while re-elicitation is targeted to where change actually occurs rather than assumed everywhere.

A short example shows the shape of this. In a quality-control laboratory, an analyst reviews instrument output against release specifications. The procedure says to accept any result inside the limits, yet for one product the analyst routinely orders a re-run when a particular peak looks irregular, even when the numbers pass. The agent reads the review system, not the analyst, and notices the recurring re-run. At the next pause, after the analyst signs the batch record, it surfaces one line in the same system: "You re-ran this product despite an in-spec result. Was the peak shape the reason?" The analyst confirms and adds a sentence. That confirmed sentence, bound to the product, the instrument, and the analyst who gave it, is a candidate tacit fragment. No interview took place, the analyst lost a few seconds, and nothing guides any decision until the Mission Group reviews it. The same loop runs, with different signals, when a nurse adjusts the timing of a step on a clinical pathway or a handler reroutes a class of cases in a claims system.

Cognitive task analysis (CTA) is the established way to do elicitation [52]. Two CTA techniques sit at opposite ends of a coverage-and-depth trade-off: the Knowledge Audit (KA) covers a wide range of cognitive tasks but stays shallow on each, while the Critical Decision Method (CDM) probes a single decision incident at length [52, 53]. Both are themselves structured interviews [54]; they differ in coverage and depth. For lighter, recurring re-elicitation we use a mini-CDM: as narrow as CDM, focused on a single incident, but deliberately less deep, stopping short of CDM's full chain-of-decisions reconstruction (Figure 5). Each of these methods asks a practitioner to articulate, in a controlled

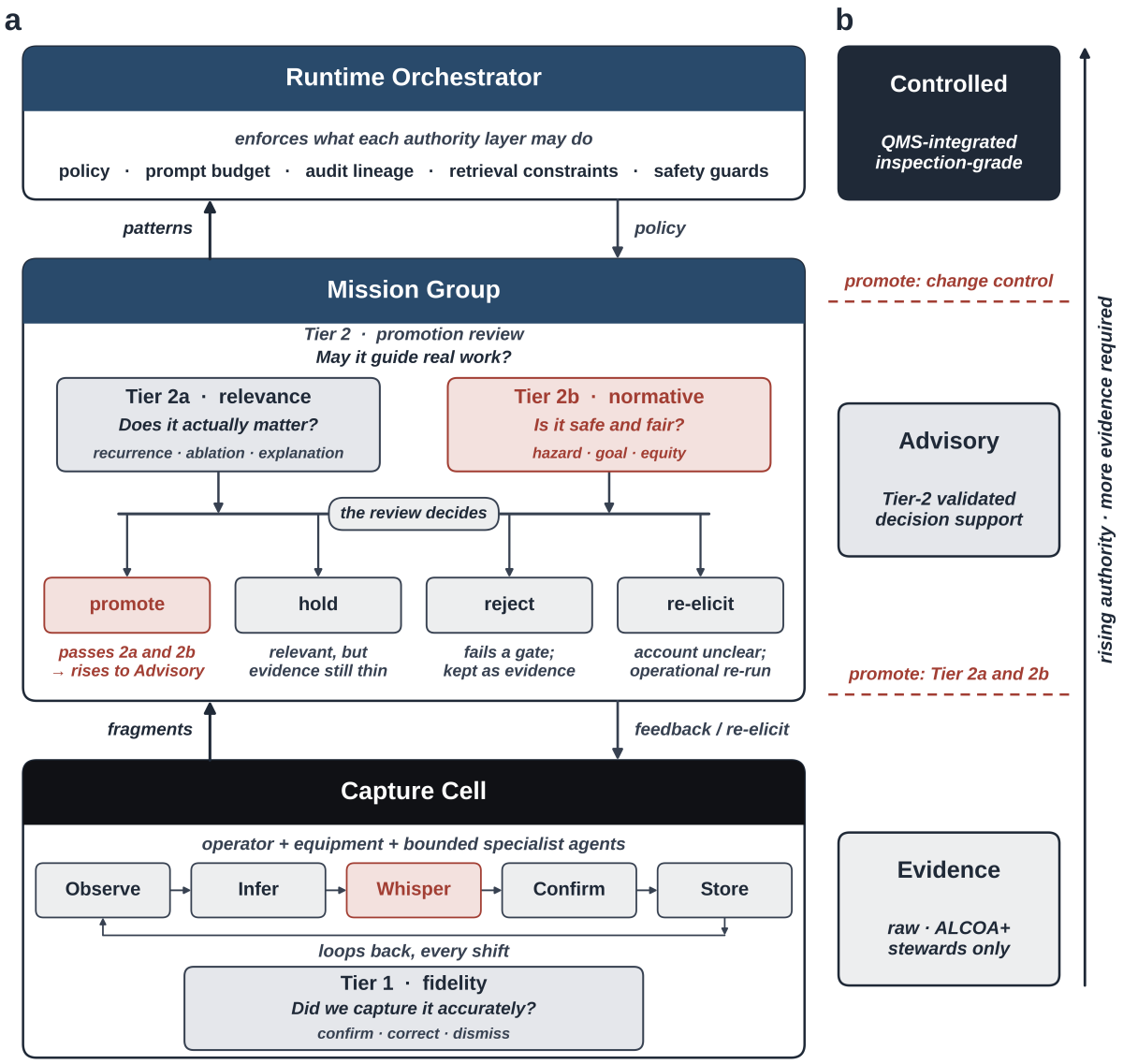

Figure 4. Metis, from situated practice to governed memory. a, The three organisational primitives and the path a candidate fragment travels. The Capture Cell pairs a worker or small team, the equipment or workflow in scope, and bounded agents that observe, infer, probe, confirm, and store; Tier 1 establishes description fidelity at the point of capture. The Mission Group reviews candidate fragments from across many Capture Cells: Tier 2 tests operational relevance, through recurrence, ablation, and expert explanation, and normative alignment, through hazard, goal, and equity review, then issues one of four outcomes, promote, hold, reject, or re-elicit. The Runtime Orchestrator enforces what each authority layer permits. Candidate fragments rise for review; policy and feedback flow back down. The term whisper denotes a minimal contextual probe, not a conversational intervention. **b,** The three-layer governance lifecycle. A fragment moves from the Evidence Layer to the Advisory Layer and the Controlled Layer, and the gate at each step names what it must pass: Tier 2 to leave the Evidence Layer, and change control to reach the Controlled Layer. Each primitive sits level with the authority layer it works at.

setting, the reasoning behind expert behaviour, and produces propositions that a knowledge engineer can validate and codify.

Metis's in-flow whispers are detection, not elicitation. The system watches the work, notices when a worker's actions diverge from what the procedure expects, and asks one short, contextual question at a natural pause. The question reaches the worker through the interface they already use for the task, the same screen on which the work is recorded, as a lightweight inline prompt rather than a separate conversation or a new tool to open. The worker confirms, corrects, or dismisses it in seconds. When a fragment needs more depth than a confirm-or-correct response can carry, the system flags it for follow-up elicitation rather than extracting more through whispers. Detection finds where elicitation should happen; elicitation produces the proposition (Figure 5a).

The two layers run on three cadences (Figure 5b). *Discovery* runs KA and CDM at multi-year intervals or whenever a major reconfiguration changes the work, seeding the model with both breadth and depth. *Refresh* runs in-flow whispers continuously, surfacing the small daily divergences that signal drift. *Re-elicitation* runs a focused mini-CDM when Refresh signals show drift, conflict, or novelty the current model does not cover. The cycle is not a pipeline: what workers know changes with the equipment, the products, and the people, so the three cadences run concurrently rather than in sequence.

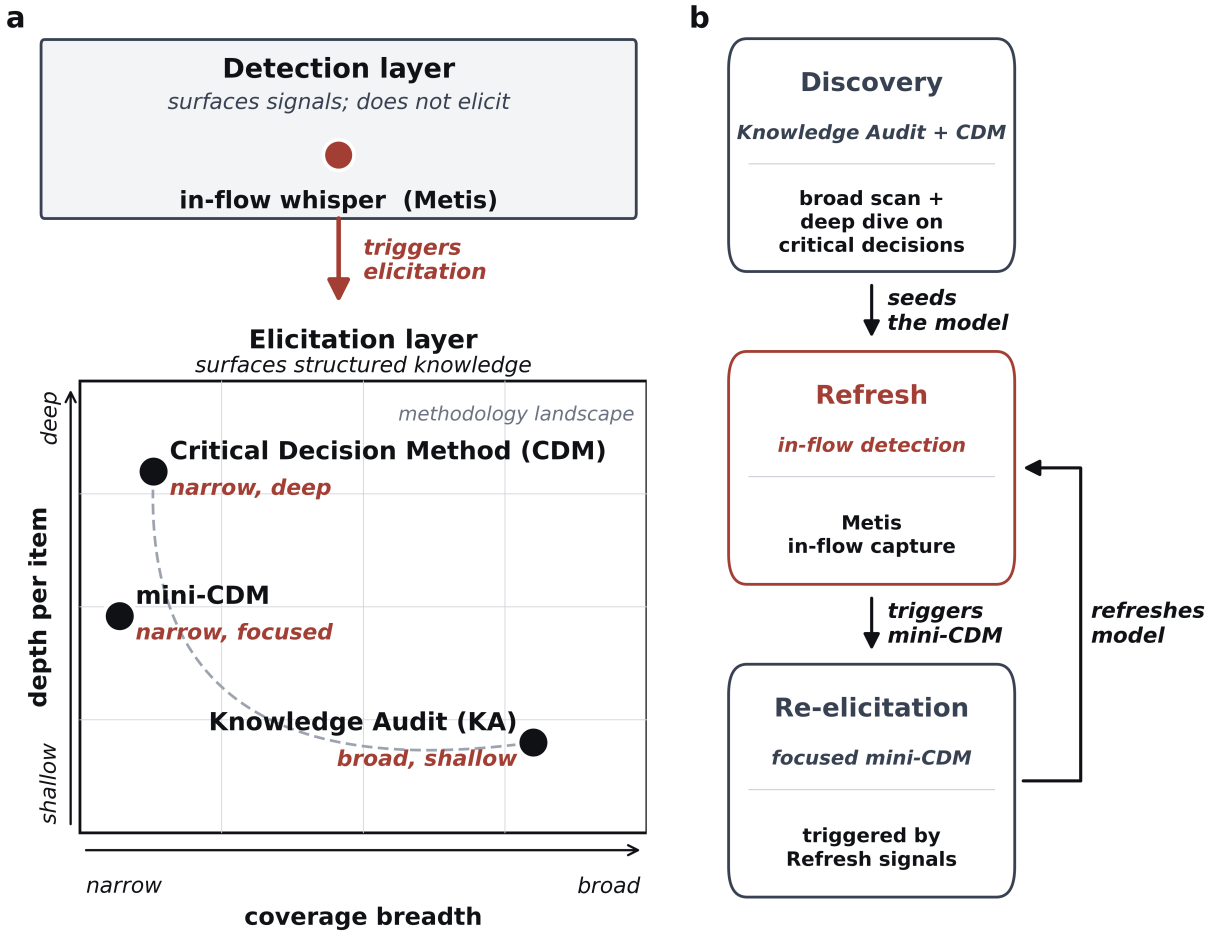

Figure 5. Capture methodology. a, The capture stack has two layers. The detection layer surfaces signals but does not produce structured knowledge; in Metis, detection comes from in-flow whispers, short prompts the system surfaces when a worker's actions diverge from what the procedure expects. Detection triggers elicitation. The elicitation layer contains the canonical CTA methods, plotted on a coverage-and-depth Pareto trade-off: Knowledge Audit (broad, shallow), Structured Interview (medium), and Critical Decision Method (narrow, deep), with a mini-CDM positioned directly below CDM, as narrow as CDM but focused, so the trade-off frontier curves toward the origin rather than running straight. **b,** Three cadences run concurrently. Discovery uses KA and CDM to seed the model. Refresh runs in-flow whispers continuously, surfacing drift signals as work proceeds. Re-elicitation runs a focused mini-CDM when Refresh signals show drift, conflict, or novelty the current model does not cover.

Detection has a second source that requires no interview. The endogenous pathway of Section 2.3.2 runs a trace-analytic monitor over the agent's own procedural, semantic, and episodic memory, computing the difference signal $\Delta = \text{enacted} \ominus \text{specified}$ and surfacing deviations that recur and stabilise. This monitor sits in the Capture Cell alongside the in-flow whisper, but it reads the agent's own history rather than the worker's work. A stable deviation it surfaces is a candidate tacit fragment, not a confirmed one: it carries endogenous provenance, enters the Evidence Layer only, and passes through the same Mission Group validation as any exogenous fragment before it can gain authority. Detection thus draws on two streams, human practice and the agent's own traces, that feed the same fourth stratum through the same qualification and governance.

4.3 Structured Storage

Within a Capture Cell, the continuous operating loop relies on five actions, shown in panel a of Figure 4: Observe, Infer, Probe (Whisper), Confirm, and Store (Remember). Observation detects signals from logs, sensors, documents, interaction traces, or task outcomes. Inference evaluates whether a signal reflects a meaningful difference in practice. The probe, denoted functionally as a "whisper" to emphasise its role as a minimal contextual intervention rather than a conversational disruption, asks the smallest useful question at a natural workflow pause. Confirmation records whether the worker recognises the description and verifies its faithfulness. Finally, storage records the candidate fragment alongside its category, conditions, provenance, validation state, authority layer, review date, and use constraints.

Crucially, free-text notes fail entirely for this purpose. A captured fragment becomes computationally useful to an agent only when the system can retrieve it under the right conditions, apply it with appropriate confidence, and rigorously inspect it after the fact. Therefore, a fourth-stratum record must minimally include the structured fields specified in Table 2, containing the operational record that expands the Equation (1) fields with lineage, consent, and revocation.

Table 2. Minimum structured record for a governed tacit fragment.

Field	Purpose
Content	Worker-confirmed articulation, exemplar, cue, or operational description.
Category	K1 to K17 category, with domain and cross-walk to the conceptual taxonomy (full K1–K17 list in Appendix A, Table A1).
Conditions of applicability	Structured context such as role, equipment family, product family, shift pattern, operating mode, risk class, and exclusion conditions.
Lineage	Originating observation, prompt, response, validation steps, dates, reviewers, and state changes.
Attribution and consent	Worker or group attribution, consent setting, visibility rule, and withdrawal constraints.
Confidence and evidence	Summary of recurrence, comparison, explanation, uncertainty, and validation outcome.
Authority layer	Evidence, advisory, or controlled state, including permitted uses and escalation requirements.
Validity and review	Review date, expiry triggers, supersession links, and retirement conditions.
Revocation status	Active, superseded, withdrawn, rejected, or under re-elicitation, with reasons preserved.

Representing these fragments as graph nodes rather than isolated text entries allows edges to connect fragments to equipment, products, roles, shifts, incidents, procedures, hazards, other fragments, and validation records [55]. This architecture allows a human reviewer to traverse provenance effortlessly. Crucially, it allows the system to evaluate conditions computationally before retrieval.

4.4 Validation and Responsible Use

When evaluating a captured fragment, reviewers must not ask whether the text sounds plausible. Instead, they must determine what organisational role the fragment is allowed to play. Metis divides this evaluation into three distinct judgments. First, description fidelity asks whether the system accurately represented the worker's account or observed behaviour. Second, operational relevance asks whether the fragment connects to a meaningful outcome or decision. Third, normative alignment asks whether the practice is safe, compliant, fair, and consistent with legitimate organisational goals. These judgements can easily diverge. An operational shortcut may be highly effective but normatively unacceptable, whereas a team ritual may be harmless but operationally irrelevant. Panel a of Figure 4 places each judgment where it happens.

- **Tier 1:** this evaluation occurs locally in the Capture Cell. It strictly concerns fidelity, verifying whether the candidate fragment accurately represents what the worker meant.
- **Tier 2:** this evaluation occurs in the Mission Group and concerns promotion. Reviewers verify whether there is sufficient evidence for the fragment to be used in a defined way. Evidence at this stage includes recurrence

across workers, historical comparison, ablation, expert explanation, and relation to process outcomes. In parallel, normative review checks whether the fragment introduces safety, quality, compliance, equity, or surveillance risks.

In the figure, the two parts of Tier 2 meet at a single review, which issues one of four outcomes: promote, hold, reject, or re-elicit. Rejection never means deletion. Rejected fragments remain highly useful as governed evidence of what the organisation considered and explicitly denied from guiding practice.

4.5 Governance and Regulatory Alignment

To map these validation tiers to actual operational authority, Metis utilises a three-layer governance lifecycle, shown in panel b of Figure 4.

- **The Evidence Layer:** this baseline layer holds raw observations, worker confirmations, rejected fragments, and early hypotheses. It supports system learning, but designers explicitly prevent it from directly influencing operational decisions.
- **The Advisory Layer:** this intermediate layer holds Tier-2 validated fragments. The system restricts their use strictly to conditional decision support.
- **The Controlled Layer:** this authoritative layer holds fragments formally incorporated into procedures, quality systems, or operating rules. The organisation manages these exclusively through rigorous change control.

A fragment does not drift between these layers. It rises only by clearing the gate that Figure 4b marks, and the gate names the validation it requires: Tier 2 to enter the Advisory Layer, and formal change control to enter the Controlled Layer. System architects must ensure that movement across layers remains recorded, justified, and reversible. The Runtime Orchestrator additionally enforces epistemic-vigilance triggers: when a candidate retrieval falls outside a fragment's stated conditions of applicability, it escalates the decision to human oversight rather than returning a confident-looking precedent, and any consequent change to a fragment's conditions or authority is agreed collectively by the Mission Group review panel, not by a single expert, before it takes effect (see Section 5.4). While this structure is computationally heavier than a single memory store, it provides the necessary foundation for accountable organisational use in regulated environments. Pharmaceutical quality systems already require strict knowledge management protocols, as outlined by ICH Q10 [56]. Meanwhile, frameworks like GAMP 5 and the ISPE AI Guide mandate risk-based validation for computerised GxP systems [57, 58]. Concurrently, broader AI governance frameworks, including the FDA and EMA guiding principles [59, 60], NIST AI RMF [50], ISO/IEC 42001 [61], and the EU AI Act [62], consistently demand technical transparency, robust risk management, data governance, and guaranteed human oversight.

Ultimately, these requirements also bear on worker dignity, not only on regulatory compliance. A system that observes work seamlessly becomes an extractive surveillance tool if it records practice without consent, converts workers into training data without attribution, or uses tacit capture for performance monitoring. Responsible use therefore requires role-sensitive consent, visibility of captured material, contestability, revocation where feasible, and a strict separation between decision support and compliance monitoring. These requirements are not merely ethical additions. They technically determine whether workers will participate honestly and whether the organisation can trust the resulting knowledge database.

4.6 Retrieval and Use by the Agent

The earlier steps produce and govern a fragment; the closing step is the agent using it when it acts again, under the conditions the fragment records (Section 3.1). Retrieval is conditional before it is similarity-based: the system evaluates a fragment's conditions of applicability computationally and filters on them before any semantic ranking, so a fragment reaches a decision only where its recorded conditions hold. Where they hold, the agent applies the fragment with the authority layer it carries, as situated guidance rather than a rule, and within the limits the Runtime Orchestrator enforces (Section 4.5). The Runtime Orchestrator additionally enforces epistemic-vigilance triggers: when a candidate retrieval falls outside a fragment's stated conditions of applicability, it routes the decision to human deliberation rather than returning a confident-looking precedent (see Section 5.4). The fragment is thus used, withheld, or escalated by the same governance wrapper it carried into memory.

5 Discussion, Implications and Research Questions

This framework demands that system designers treat tacit knowledge in agentic AI as a rigorously bounded design problem, rather than a metaphor or a mystique. By defining the fourth stratum and the Metis architecture, we can concretely evaluate the engineering challenges of capturing situated practice.

5.1 Challenges and Options for the Exogenous Pathway

The exogenous pathway captures fragments directly from human practice. The primary challenge here involves translation loss and worker burden. When a system interrupts a worker to codify a sensory cue or heuristic, it risks distorting the situated, fluid knowledge into a rigid text rule. Furthermore, excessive prompting fatigues the worker, which degrades the quality of the capture.

To mitigate these risks, designers must connect the capture process directly to the working principles of the continuous operating loop. By deploying the minimal contextual probe, the in-flow whisper, the system gathers confirmation at natural workflow pauses without breaking the worker's concentration. Additionally, the architecture must strictly enforce the first core design commitment: tacit knowledge remains only partially representable. The Runtime Orchestrator must never treat the confirmed fragment as a universal rule. Instead, it must enforce the recorded conditions of applicability, ensuring the agent uses the fragment strictly as situated guidance.

5.2 Challenges and Options for the Endogenous Pathway

The endogenous pathway detects latent competence by auditing the agent's own operational logs. This route faces a severe engineering challenge: the high risk of spurious pattern recognition. Without grounding in human physical reality, an agent might identify a recurring deviation and falsely encode a software glitch, a contaminated input, or a statistical artefact as a valid tacit fragment.

To solve this, developers must apply our strict governance principles. The system cannot draw the distinction between genuine competence and spurious correlation inside the same computational loop that produced the deviation. Designers must therefore force endogenous derivations to enter the fourth stratum strictly as operational hypotheses. The architecture must flag these machine-generated fragments with an endogenous provenance tag and hold them in the Evidence Layer. The system must never auto-promote an endogenous fragment to the Advisory or Controlled layers; human reviewers in the Mission Group must explicitly validate its operational relevance and normative alignment. To keep this review tractable, the Mission Group does not assess every machine-generated fragment in isolation. The system groups related endogenous fragments and ranks them by recurrence, so that the most frequently observed deviations are surfaced first in a prioritised review queue, while isolated or rare derivations wait until they recur or are batched with similar cases.

5.3 Challenges and Options for Combining Pathways

An advanced agentic system will ideally combine both approaches, observing human workers while simultaneously auditing its own task execution. The architectural challenge here involves data contamination. If a system casually mixes human-validated perceptual cues with unverified machine inferences, it destroys the integrity and accountability of the organisational memory.

The option for resolving this lies in the structured record of the fourth stratum. The system must rigorously track the provenance and lineage of every graph node. When combining pathways, the Mission Group must use exogenous human fragments to ground and verify the endogenous machine inferences. For example, the system can cross-reference an agent's trace-analytic deviation against a human worker's confirmed heuristic. The architecture merges them into a unified operational strategy only when human reviewers explicitly approve the synthesis under Tier 2 validation.

5.4 Implications for Theory and Engineering

The knowledge-based view of the firm clearly explains why tacit knowledge drives sustainable competitive advantage. Agentic AI fundamentally shifts the practical question. Researchers must now ask how fragments of situated practice become computationally available without detaching from the social and normative conditions that give them meaning.

Methodologically, researchers must empirically compare capture routes rather than assuming a single optimal method. Deep cognitive task analysis, in-flow observation, incident reconstruction, expert annotation, handoff analysis, and agent-memory analysis will likely surface vastly different kinds of fragments. We predict that capture quality will

heavily depend on category fit, the worker burden, prompt timing, validation clarity, and the trust workers place in the governance system.

Furthermore, expertise carries a well-documented dark side. The exact tacit competence that makes experts highly performant under familiar conditions can become catastrophically miscalibrated under unfamiliar ones. Confidence calibrated to a stable distribution of cases often persists when the distribution shifts. This overconfidence has contributed to well-documented organisational failures, including the Challenger and Columbia losses and Air France 447 [63–65]. This dark side dictates two strict engineering consequences. First, fragments captured from expert practice inherently encode the expert's operating envelope, including their blind spots. Second, developers must build explicit epistemic-vigilance triggers into the governance layer. When a trigger activates because a situation falls outside a fragment's stated envelope, the system must route the decision to human deliberation rather than retrieving a confident-looking precedent.

5.5 Limitations

This paper contains clear limitations. The endogenous pathway, while architecturally specified, remains empirically untested and inherits the spurious-pattern-recognition risk discussed in Section 5.2. This paper scopes itself to the extraction approach; a naturalistic alternative remains outside its scope. Furthermore, the framework currently focuses on skilled, repetitive, and regulated operational work. Designers will likely need to alter it to support highly fluid, strategic, or creative knowledge work. Finally, this paper does not resolve the critical issues of knowledge ownership, worker compensation, labour relations, or the cross-site transfer of worker expertise. We view these as foundational issues for the broader research agenda rather than peripheral concerns.

5.6 Propositions and Research Agenda

From this framework, several key propositions emerge as formal hypotheses for future study. We outline the core propositions here and preserve the complete set of thirteen hypotheses, alongside the expanded research agenda, in Appendix C.

First, tacit categories will differ materially in the specific capture methods and validation evidence they require. Second, agentic systems equipped with governed tacit memory will handle operational exceptions more reliably than systems limited to explicit documents and ordinary episodic memory, provided organisations strictly enforce authority boundaries. Third, exogenous capture from human practice and endogenous analysis of agent traces produce fundamentally different candidate fragments; engineers must never conflate them. Fourth, validated tacit fragments only prove useful when systems explicitly represent their conditions of use and computationally check those conditions at retrieval time. Fifth, worker trust, consent, and contestability represent definitive determinants of system quality, rather than mere post-hoc adoption factors.

6 Conclusion

Tacit knowledge became central to organisational theory because written documents never fully encapsulate skilled work. It remained a persistent challenge for artificial intelligence because earlier systems depended overwhelmingly on explicit representation, formalised rules, labelled data, or retrievable documents. Foundation-model agents do not dissolve this epistemological difficulty, but they fundamentally change the practical conditions under which system designers can address it. Because agents can now operate with memory, tools, workflow context, and human interaction, researchers face a renewed imperative to revisit tacit knowledge as a computational problem, provided they rigorously bound their claims.

We propose the fourth stratum as a framework for naming and navigating this bounded problem. The stratum does not contain tacit knowledge in full; rather, it contains governed fragments. These fragments serve as partial representations of situated practice, inextricably tied to conditions, provenance, validation, authority, and review. Metis operationalises this concept, describing how such fragments might safely move from situated practice to governed memory through strict capture, confirmation, validation, and controlled use.

The ultimate value of this framework lies less in any single technical component than in the architectural discipline it imposes. System designers must not treat tacit knowledge as mere extraction data. They must not let agents infer operational authority from linguistic plausibility. They must not extract worker expertise without robust governance. Finally, they must not integrate digital intelligence into organisational workflows while ignoring the deeply situated knowledge through which human beings actually accomplish work.

Whether the proposed stratum demonstrably improves operational learning, quality, resilience, and human-agent collaboration remains an open empirical question. By providing a precise vocabulary and architecture, this framework ensures the scientific community can study the problem rigorously, without ever overstating what computation can safely capture.

Code Availability: A reference implementation of Metis is openly available at https://github.com/BrightbeamAI/metis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, pages 9459–9474, 2020.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. https://arxiv.org/abs/2312.10997, 2023.

[3] Rishi Bommasani et al. On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258, 2021.

[4] Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901, 2020.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837, 2022.

[6] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. https://arxiv.org/abs/2210.03629, 2022.

[7] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36*, 2023.

[8] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[9] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. https://arxiv.org/abs/2308.08155, 2023.

[10] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. https://arxiv.org/abs/2310.08560, 2024.

[11] Zhiheng Xi et al. The rise and potential of large language model based agents: A survey. https://arxiv.org/abs/2309.07864, 2023.

[12] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025. doi: 10.1145/3748302. Article 155.

[13] Marcel Detienne and Jean-Pierre Vernant. *Cunning Intelligence in Greek Culture and Society*. Harvester Press, Hassocks, Sussex, 1978. Translated by Janet Lloyd.

[14] James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT, 1998.

[15] Michael Polanyi. *Personal Knowledge: Towards a Post-Critical Philosophy*. Routledge and Kegan Paul, London, 1958.

[16] Michael Polanyi. *The Tacit Dimension*. Routledge and Kegan Paul, London, 1966.

[17] Gilbert Ryle. *The Concept of Mind*. Hutchinson, London, 1949.

[18] Hubert L. Dreyfus and Stuart E. Dreyfus. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press, New York, 1986.

[19] Harry M. Collins. *Tacit and Explicit Knowledge*. University of Chicago Press, Chicago, 2010.

[20] Haridimos Tsoukas. Do we really understand tacit knowledge? In Mark Easterby-Smith and Marjorie A. Lyles, editors, *The Blackwell Handbook of Organizational Learning and Knowledge Management*, pages 410–427. Blackwell, Oxford, 2003.

[21] John Seely Brown and Paul Duguid. Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1):40–57, 1991.

[22] Jean Lave and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, 1991.

[23] Etienne Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, 1998.

[24] J.-C. Spender. Making knowledge the basis of a dynamic theory of the firm. *Strategic Management Journal*, 17 (S2):45–62, 1996.

[25] Scott D. N. Cook and John Seely Brown. Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science*, 10(4):381–400, 1999.

[26] Jay Barney. Firm resources and sustained competitive advantage. *Journal of Management*, 17(1):99–120, 1991.

[27] Robert M. Grant. Toward a knowledge-based theory of the firm. *Strategic Management Journal*, 17(S2):109–122, 1996.

[28] David J. Teece, Gary Pisano, and Amy Shuen. Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7):509–533, 1997.

[29] Linda Argote. *Organizational Learning: Creating, Retaining and Transferring Knowledge*. Springer, New York, 2nd edition, 2013.

[30] Martin Heidegger. *Being and Time*. Harper & Row, New York, 1962. Translated by J. Macquarrie and E. Robinson.

[31] Lucy A. Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge, 1987.

[32] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York, 1983.

[33] Endel Tulving. How many memory systems are there? *American Psychologist*, 40(4):385–398, 1985.

[34] Larry R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2):195–231, 1992.

[35] Larry R. Squire. Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82(3):171–177, 2004.

[36] Gary Klein. *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA, 1998.

[37] Gary Klein. Naturalistic decision making. *Human Factors*, 50(3):456–460, 2008.

[38] Raanan Lipshitz, Gary Klein, Judith Orasanu, and Eduardo Salas. Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, 14(5):331–352, 2001.

[39] Wesley M. Cohen and Daniel A. Levinthal. Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1):128–152, 1990.

[40] Thomas O. Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, 26:125–173, 1990.

[41] Ikujiro Nonaka. A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37, 1994.

[42] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953. Translated by G. E. M. Anscombe.

[43] Willard Van Orman Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

[44] Erik Hollnagel. *The ETTO Principle: Efficiency–Thoroughness Trade-Off*. Ashgate, Farnham, 2009.

[45] Sidney Dekker. *The Field Guide to Understanding Human Error*. CRC Press, Boca Raton, FL, 3rd edition, 2014.

[46] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-AI interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. Paper 3.

[47] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 7, pages 2–11, 2019.

[48] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1369–1385, 2023.

[49] OpenAI. GPT-4 technical report. https://arxiv.org/abs/2303.08774, 2023.

[50] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023.

[51] Arsalan Shahid, Gordon Suttie, and Philip Black. Collaborative human-agent protocol (CHAP): An open protocol for auditable, structured multi-human and multi-agent collaboration. Technical Report arXiv:2606.09751, Brightbeam AI, 2026. URL https://arxiv.org/abs/2606.09751.

[52] Beth Crandall, Gary Klein, and Robert R. Hoffman. *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. MIT Press, Cambridge, MA, 2006.

[53] Robert R. Hoffman, Beth Crandall, and Nigel Shadbolt. Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors*, 40(2):254–276, 1998.

[54] Jan Maarten Schraagen, Susan F. Chipman, and Valerie L. Shalin, editors. *Cognitive Task Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[55] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

[56] International Council for Harmonisation. ICH Q10 pharmaceutical quality system. https://database.ich.org/sites/default/files/Q10%20Guideline.pdf, 2008.

[57] ISPE. *GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems*. International Society for Pharmaceutical Engineering, Tampa, FL, 2nd edition, 2022.

[58] International Society for Pharmaceutical Engineering. ISPE GAMP guide: Artificial intelligence. https://ispe.org/publications/guidance-documents/gamp-guide-artificial-intelligence, 2025.

[59] U.S. Food and Drug Administration and European Medicines Agency. Guiding principles of good AI practice in drug development. https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development, 2026.

[60] European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle-scientific-guideline, 2024.

[61] International Organization for Standardization. ISO/IEC 42001:2023 information technology — artificial intelligence — management system. https://www.iso.org/standard/42001, 2023.

[62] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024.

[63] Diane Vaughan. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago, 1996.

[64] Columbia Accident Investigation Board. Report volume 1. Technical report, NASA, Washington, DC, 2003.

[65] Bureau d'Enquêtes et d'Analyses. Final report on the accident on 1 june 2009 to the airbus A330-203 registered F-GZCP. Technical report, BEA, Paris, 2012.

A Full Computational Taxonomy

The following taxonomy serves as a conceptual bridge between tacit knowledge theory and agentic memory design. It applies three design principles: operational distinguishability (categories must differ in capture modality or validation), cognitive coherence (categories must recruit a related form of cognition), and computational consequence (categories must imply an engineering choice).

Three categories are especially important because they are frequently under-specified in operational treatments of expertise: rhythmic knowledge (action timing and process cadence), meta-cognitive knowledge (knowing the limits of competence and when to escalate), and affective-regulatory knowledge (emotional composure and reading affective states during situated judgement).

Table A1. Full computational taxonomy of tacit knowledge across domains, categories, Spender mapping, and primary capture modality.

Domain	Category	Primary Spender quadrant	Primary capture modality
D1. Procedural–Embodied	K1 Procedural	Conscious / Social	Document ingestion; procedure analysis
	K2 Embodied	Automatic / Individual	Multimodal observation; expert confirmation
	K3 Rhythmic	Automatic / Individual	Action-timing analysis; pause/tempo comparison
D2. Material–Equipment	K4 Equipment-specific	Automatic / Individual–Social	Maintenance logs; cross-equipment comparison
	K5 Material	Automatic / Individual	Outcome correlation with material lots; worker narration
	K6 Tool-extended	Automatic / Individual	Tool-use traces; substitution events
D3. Perceptual–Aesthetic	K7 Sensory	Automatic / Individual	Multimodal cue capture; on-cue narration
	K8 Aesthetic	Automatic / Individual	Expert annotation of exemplars
D4. Inferential	K9 Heuristic	Conscious–Automatic / Individual	In-flow prompt; exception logging
	K10 Diagnostic	Conscious / Individual	Critical decision method; incident reconstruction
	K11 Anticipatory	Automatic / Individual	Pre-event prompting on detected divergence
D5. Meta-cognitive and affective	K12 Meta-cognitive	Conscious / Individual	Reflective probing; help-seeking pattern analysis
	K13 Affective-regulatory	Automatic / Individual	Consented affective cues; post-event reflection
D6. Social–Normative	K14 Collaborative	Automatic / Social	Analyse handoffs, dependencies, and cross-actor coordination
	K15 Cultural-narrative	Automatic / Social	Long-form interview; story collection
	K16 Judgemental–ethical	Conscious / Individual	Post-event walk-through; structured reflection
	K17 Strategic	Conscious / Individual–Social	Strategic interview; cross-level alignment review

B Capture-Pathway Mapping

A captured fragment only becomes useful to an agent when it can be retrieved under precise conditions, applied with appropriate confidence, and fully inspected after the fact. The fourth stratum imposes five design requirements for this memory: situated capture, conditional representation, human confirmation, independent validation, and governed retrieval.

Table A2. Capture pathway under Metis by tacit category.

Category	Spender quadrant	Primary pathway	Loop role
K1 Procedural	Conscious / Social	Already documented	Reference baseline; not primary tacit target
K2 Embodied	Automatic / Individual	In-flow	Observe action; prompt for confirmation at natural pause
K3 Rhythmic	Automatic / Individual	In-flow	Detect cadence, waiting, tempo, and timing differences
K4 Equipment-specific	Automatic / Individual–Social	In-flow	Compare equipment-specific deviations and adaptations
K5 Material	Automatic / Individual	In-flow	Prompt on lot, batch, or material-state changes
K6 Tool-extended	Automatic / Individual	In-flow	Detect tool substitutions and workarounds
K7 Sensory	Automatic / Individual	In-flow	Capture cue-triggered noticing; request on-cue narration
K8 Aesthetic	Automatic / Individual	Re-elicit	Use exemplar annotation and expert comparison
K9 Heuristic	Conscious–Automatic / Individual	In-flow	Prompt at exception, threshold, or deviation moment
K10 Diagnostic	Conscious / Individual	Re-elicit	Trigger mini-CDM or incident reconstruction
K11 Anticipatory	Automatic / Individual	In-flow	Prompt before predicted event or divergence
K12 Meta-cognitive	Conscious / Individual	Re-elicit	Examine hesitation, help-seeking, and escalation patterns through reflection
K13 Affective-regulatory	Automatic / Individual	In-flow / Re-elicit	Use consented cues and post-event reflection; avoid covert inference
K14 Collaborative	Automatic / Social	In-flow	Analyse handoffs, dependencies, and cross-actor coordination
K15 Cultural-narrative	Automatic / Social	Discovery	Capture through stories, interviews, and community interpretation
K16 Judgemental–ethical	Conscious / Individual	Re-elicit	Use structured reflection and normative review
K17 Strategic	Conscious / Individual–Social	Discovery	Capture through leadership engagement and cross-level alignment review

C Full Proposition Set and Research Agenda

The following 13 propositions are stated as hypotheses for future empirical and engineering work:

P1. Tacit knowledge is multidimensional: a single-dimensional or single-quadrant decomposition is insufficient for engineering capture systems in skilled organisational work.

P2. Meta-cognitive and affective-regulatory knowledge are distinguishable categories whose omission under-represents safety-critical dimensions of expertise.

P3. The fourth memory stratum is populated by two distinct pathways: exogenous capture from human practice and endogenous extraction from analysis of the agent's own procedural, semantic, and episodic traces.

P4. Agent-memory architectures relying strictly on procedural, semantic, and episodic stores will structurally under-represent consequential operational knowledge in environments dependent on skilled practice.

P5. Tacit retrieval must be conditional before it is similarity-based; applicability conditions must actively filter fragments before semantic ranking occurs.

P6. The gap between formal expectation and observed practice acts as a first-line detector for candidate tacit content, though not all detected differences represent valid signals.

P7. Deep Cognitive Task Analysis (CTA) and in-flow capture are complementary rather than substitutable methodologies.

P8. Worker dignity, consent, visibility, contestability, and attribution are core determinants of capture quality, not merely post-hoc adoption considerations.

P9. Operational relevance and normative alignment must be validated entirely separately.

P10. The architectural separation of capture, validation, and operationalisation is strictly necessary for accountable deployment in regulated settings.

P11. Determinism on safety-critical paths must be enforced architecturally through controlled retrieval and escalation, rather than assumed as an inherent property of a language model.

P12. In-flow capture is highly plausible for automatic-individual and inferential categories, but fundamentally weaker for cultural-narrative, strategic, ethical, and substantial meta-cognitive categories.

P13. Captured fragments must be codified into structured, conditional, and lineage-bearing records to be simultaneously consumable by AI agents and fully auditable by human organisations.

The research agenda logically follows four lines:

- **Theory:** test the stability of the taxonomy across sectors and clarify the precise relation between tacit fragments, routines, communities of practice, and overall organisational memory.
- **Empirical method:** for the exogenous pathway, compare full CTA, mini-CTA, incident reconstruction, expert annotation, and in-flow capture on yield, fidelity, burden, and fragment decay, and run controlled studies that capture practice from a small cohort of professionals, map it to an agentic system, and compare the human expert against the tacitly-augmented agent side by side. For the endogenous pathway, measure how reliably trace-analytic inference separates genuine latent competence from spurious regularity, contamination, and training artefact, test whether the higher evidence bar and evidence-layer quarantine prevent false promotion, and characterise how endogenous candidate fragments differ in kind from exogenous ones.
- **Systems engineering:** extend the initial reference implementations of Capture Cells, the trace-analytic monitor, Mission Groups, Runtime Orchestrators, graph-based fragment stores, and conditions-first retrieval models toward production-grade systems.
- **Governance and ethics:** rigorously study attribution, consent, withdrawal, compensation, labour relations, surveillance risk, and the net effect of tacit-capture systems on organisational accountability.