

When Permission Outlives Its Justification

A Temporal Ontology for Meta-Level Authority Control in Agentic AI

Qinlei Wang

Social Sciences graduate, King's College London, London, United Kingdom

Abstract Autonomous agents increasingly act under delegated authority, yet conventional authorization mechanisms represent credential validity more readily than the continuing grounds that make an action appropriate after the agent's world model changes. This paper frames authority maintenance as a cognitive task of meta-level control. It proposes a temporal ontology in which each grant depends on explicit justificatory propositions whose epistemic status, freshness, and revalidation conditions are represented symbolically. World-model updates can therefore change actionability while the stored permission remains unchanged. Neural components may interpret documents, dialogue, or perceptual inputs and propose belief updates, while the symbolic layer validates those updates and retains execution control. Five epistemic-control states distinguish supported, stale, contradicted, suspended, and revoked authority. A comparative evaluation design tests whether this representation improves safe execution, avoids unnecessary interruption, and produces explanations grounded in the conditions that actually governed action. The resulting framework connects world modelling, metacognitive monitoring, and high-level execution within a concrete challenge problem for symbolically orchestrated agents.

1. Authority Maintenance as a Cognitive Task

An agent acting for another party must repeatedly determine whether the reasons that made an action permissible remain supported by its current beliefs. Authenticated delegation makes principals, scopes, and accountability chains inspectable (South et al., 2025), while recent token and receipt proposals formalise attenuation, time windows, and verifiable delegation records (Niyikiza, 2026; Nelson, 2026). A distinct problem appears when the credential is intact and a contextual proposition supporting its use has changed. A purchasing agent can retain a valid grant after a supplier loses approval, while a robotic agent can retain an action capability after a safety condition expires. The cognitive task maps a proposed action, a represented authority grant, and the current world model onto an explicit actionability state that guides execution and explanation. This places authority inside high-level control, where world-model revision can alter the status of an otherwise unchanged permission. Content-centric cognitive systems already emphasise structured knowledge, actionable interpretation, and inspectable reasoning (Nirenburg, McShane, & English, 2020; McShane et al., 2025), while recent cognitive-robotics work similarly assigns structured architectures a primary role in reasoning and control around foundation models (Sridharan, 2025).

2. A Temporal Ontology of Justified Authority

We represent an authority grant as $A = \langle P, Ag, Act, Obj, S, J, T, R \rangle$, where P denotes the principal, Ag the authorised agent, Act the action class, Obj the target, S the explicit scope, J the justificatory propositions, T the temporal interval, and R the rules for revalidation and revocation. The representational novelty lies in J : each proposition is linked to epistemic status, evidential provenance, freshness requirements, and an authorised revalidation source. A procurement grant can therefore encode a spending ceiling in S while J records that the supplier remains approved, the project remains active, and the applicable risk status remains acceptable. Five states organise meta-level control over this relation. Valid and justified indicates that every required proposition remains supported; valid but stale marks beliefs whose permitted epistemic age has elapsed; valid but contradicted marks evidence incompatible with a required proposition; suspended pending revalidation blocks execution while directing information seeking to a specified authority source;

revoked records explicit termination. These distinctions let the system respond to changing knowledge through targeted revalidation, diagnosis, escalation, or termination without collapsing different epistemic conditions into one generic denial.

3. Symbolic Meta-Level Control with Bounded Neural Tools

The architecture assigns authority assessment to a symbolic orchestration layer containing an authority model, a justification model, a temporal model, and an execution gate. The authority model represents principals, agents, actions, resources, and delegation relations; the justification model links grants to the world-model propositions on which their actionability depends; the temporal model tracks credential intervals, belief freshness, and event-triggered checks; and the execution gate evaluates a proposed action against these representations before downstream tool use or robotic execution. Neural components remain useful in bounded interpretive roles. An LLM can extract a candidate supplier-status change from a document or interpret an instruction, while a perceptual model can propose that an environmental condition has changed. Each output enters as a candidate belief update with source metadata, after which symbolic validation determines whether the authoritative world model changes. The resulting control decision is recoverable from the same knowledge structures that generated it, giving explanation a direct connection to the agent's actual reasoning process and preserving symbolic epistemic control over execution.

4. Comparative Evaluation and Contribution

A compact benchmark can vary formal scope and contextual justification independently across multi-step workflows, allowing three architectures to be compared under identical task sequences: a scope-only authorization baseline, an LLM contextual gate, and the proposed symbolic temporal model. Primary measures are unsafe executions after relevant justificatory change, unnecessary suspensions after irrelevant change, correct routing of uncertain cases to revalidation, and explanation fidelity against the proposition that actually changed. Further manipulations can introduce delayed evidence, conflicting sources, nested delegation, and incomplete knowledge. The strongest test concerns cognitive behaviour across time: the system should preserve an authorised plan while its supporting beliefs remain current, detect when a relevant belief changes the plan's authority basis, identify the proposition responsible, and select an appropriate meta-level response. The contribution is therefore an architectural account of delegated authority as a temporally maintained relation between permission and an explicit world model. It gives a symbolically orchestrated agent the capacity to represent why it is presently authorised to act, monitor whether those grounds remain epistemically adequate, and revise actionability when its beliefs change. This challenge problem joins symbolic orchestration, knowledge validation, genuine explanation, trust and safety, and comparative evaluation within one integrated cognitive task.

References

- McShane, M., Nirenburg, S., English, J., & Oruganti, S. (2025). Pursuing actionable perception interpretation in cognitive robotic systems. *Proceedings of the Twelfth Annual Conference on Advances in Cognitive Systems*, 367-386.
- Nelson, R. (2026). Delegation Receipt Protocol for AI Agent Authorization. IETF Internet-Draft draft-nelson-agent-delegation-receipts-10.
- Nirenburg, S., McShane, M., & English, J. (2020). Content-centric computational cognitive modeling. *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*.
- Niyikiza, N. A. (2026). Attenuating Authorization Tokens for Agentic Delegation Chains. IETF Internet-Draft draft-niyikiza-oauth-attenuating-agent-tokens-01.
- South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Chan, A., & Pentland, A. (2025). Position: AI agents need authenticated delegation. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 82211-82231.
- Sridharan, M. (2025). Toward an architecture for robots in the era of foundation models. *Proceedings of the Twelfth Annual Conference on Advances in Cognitive Systems*, 21-32.