Human Agency Infrastructure for Agentic AI

Delegation, Runtime Governance, and the Preservation of Human Authority

Michael Bower
Human Agency Preservation Infrastructure (HAPI)
Alignment Governance Stack Working Paper
Version 0.1 | May 18, 2026
Working research manuscript, not peer reviewed

Core thesis: AI should amplify human agency, not outrun it.

Abstract

Agentic AI changes the governance problem from output review to delegated action control. Earlier AI systems primarily produced recommendations, classifications, summaries, or generated content. Agentic systems increasingly select tools, call APIs, update records, initiate workflows, coordinate tasks, and produce operational consequences. This paper argues that agentic AI should be understood as delegated operational agency: not moral personhood, but executable action under borrowed human or organizational authority. Because delegated agents can act faster than human judgment can participate, governance must preserve human agency before, during, and after consequential action. Human Agency Preservation Infrastructure (HAPI) provides the parent framework: preserve clarity, authority, refusal, revision, memory, participation, and accountability. The Alignment Governance Stack is treated as one technical expression of HAPI, using PGDL for proposal scrutiny, Agent Action Gate for authorization, Runtime Binding for permit-bound execution, Receipts for accountable memory, Governance Reality Reports for auditability, and Continuity Findings for temporal coherence. The central claim is that safe agentic AI cannot be reduced to policy documents, generic human-in-the-loop design, or post hoc logging. It requires infrastructure that keeps human authority live at the point where cognition becomes action.

Keywords: human agency, agentic AI, delegated agency, AI governance, human oversight, runtime governance, Agent Action Gate, PGDL, receipts, governance continuity, HAPI.

Table of Contents

1. Introduction

Agentic AI is not merely a more fluent chatbot. It is a shift from generated output to operational delegation. When an AI agent drafts a message, searches a database, updates a ticket, schedules a meeting, moves a file, calls a cloud tool, or triggers a workflow, the governance question changes. The issue is no longer only whether the model produced acceptable language. The issue is whether a proposed action may lawfully and coherently pass into reality.

This transition exposes a weakness in conventional governance. Policies can exist without constraining execution. Dashboards can exist without preserving authority. Human approval can exist without meaningful refusal. Logs can exist without trustworthy memory. In each case, a system may appear governed while human agency has already been reduced to theater.

HAPI begins from a different premise: governance becomes real only when the conditions of human agency are preserved. For agentic AI, that means people and organizations must retain clarity, authority, refusal, revision, memory, and accountability at the point where an AI system proposes or executes consequential action.

The purpose of this paper is to define how HAPI applies to agentic AI. It places PGDL, AAG, Runtime Binding, Receipts, Governance Reality Reports, and Continuity Findings under a single agency-preservation thesis: agentic systems should extend human capability without bypassing human authority.

2. Agentic AI as Delegated Operational Agency

An AI agent is not a moral person. It does not own responsibility in the way a human does. But it is also not a passive tool once it can plan, select tools, call APIs, and change operational state. It occupies a middle category: delegated operational agency.

Delegated operational agency means that the system acts under borrowed authority. Its action is not self-originating in the moral sense. It inherits permission, scope, intent, and risk from the person or organization that deploys it.

The accountability chain is therefore not "the AI did it." The chain is human or organization intent, instruction, agent proposal, governed action, consequence, receipt, and accountability.

Stage	Agency Question	Governance Need
Intent	What does the human or organization actually want?	Clarify objective, authority, values, and constraints.
Instruction	What has been delegated to the agent?	Limit scope, tools, permissions, and operational domain.
Proposal	What action is the agent trying to make real?	Scrutinize meaning, risk, reversibility, and target.
Authorization	May this action proceed under valid authority?	Allow, revise, escalate, or block before consequence.
Execution	Is the system doing exactly what was authorized?	Bind tools and runtime behavior to a narrow permit.
Receipt	What proof remains after action?	Record proposal, objection,

		approval, execution, result, and evidence.
Learning	Did governance improve or decay over time?	Use continuity findings to detect drift, stale authority, and repeated patterns.

This framing avoids both extremes. It does not anthropomorphize the agent into a moral subject. It also does not reduce the agent to a harmless instrument. It recognizes that delegated action requires governed passage.

3. Why Human-in-the-Loop Is Not Enough

Human-in-the-loop is often treated as the default answer to AI risk. The problem is that a human can be placed near a process without retaining meaningful agency over the process. A person can approve without understanding, supervise without authority, review without time, and accept responsibility without control.

HAPI distinguishes human presence from human agency. Human presence means a person appears somewhere in the workflow. Human agency means the person can still understand, refuse, revise, approve, contest, and remain accountable before the action becomes committed.

The EU AI Act Article 14 centers human oversight for high-risk AI systems, aiming to prevent or minimize risks to health, safety, or fundamental rights. This is directionally aligned with HAPI, but the operational question remains: where exactly is human authority live in the action path? [1]

In agentic systems, the decisive boundary is often the moment when cognition becomes action. A review screen after an action has effectively been shaped, routed, and socially pressured may preserve the appearance of oversight without preserving agency.

Pattern	Looks Like Agency	HAPI Diagnosis
Approval click	A human clicked approve.	Not enough unless refusal, context, and consequence awareness were real.
Dashboard visibility	A human can see system activity.	Not enough unless visibility supports timely intervention.
Escalation queue	High-risk items go to review.	Not enough if reviewers are overloaded or lack authority.
Audit log	Events are recorded.	Not enough if logs are incomplete, untrusted, or unused for correction.
Policy document	Rules exist.	Not enough if runtime execution can bypass them.

The HAPI standard is stricter: oversight is not meaningful unless human authority can still change the outcome before consequence.

4. Agency Preservation Requirements

HAPI treats agency as meaningful participation under conditions of capacity, authority, clarity, refusal, and memory. For agentic AI, those conditions become concrete system requirements.

Requirement	Definition	Agentic AI Implementation
Clarity	The human can understand what is being proposed.	Review packets, risk summaries, plain-language action descriptions.
Authority	The right person or role can decide.	Authority maps, IAM integration, approval policies.
Refusal	The human can stop the action before consequence.	Hard gates, escalation paths, deny/block states.
Revision	The human can change the proposed action.	Revise-action outcomes, constrained resubmission.
Memory	The system preserves what happened and why.	Receipts, hashes, signatures, outcome records.
Accountability	Responsibility follows authority and evidence.	Audit trails tying proposal, approval, execution, and result.
Continuity	Governance stays coherent across time.	Continuity findings for stale authority, scope drift, receipt gaps, and rubber-stamp patterns.

These requirements turn agency from a moral aspiration into an infrastructure design constraint. The system should not merely ask whether an AI action is useful. It should ask whether the action preserves meaningful human authority through the action path.

5. The HAPI Action-Path Architecture

The Alignment Governance Stack is a technical child of HAPI. Its role is to preserve human agency in agentic AI workflows by controlling the transition from intent to action and from action to accountability.

Layer	Question	Agency Function
Governance Substrate	Who has authority, under what policy, in what context?	Defines identity, role, permission, scope, and values.
PGDL	Should this proposal become actionable at all?	Preserves discernment by challenging meaning before authorization.
AAG	Is this action authorized, scoped, reversible, and approved?	Preserves refusal, revision, and decision authority.
Runtime Binding	Is execution limited to the authorized action?	Preserves fidelity between approval and consequence.
Receipts	What proof remains after decision and action?	Preserves memory and accountability.

Governance Reality Report	Was governance real or theater?	Audits whether human participation was meaningful.
Continuity Findings	Did governance remain coherent over time?	Detects temporal decay, repeated drift, and stale authority.

This stack is not meant to multiply layers for their own sake. Each layer protects a different agency function. PGDL protects interpretive integrity. AAG protects authorization. Runtime Binding protects execution fidelity. Receipts protect memory. Audit reports protect institutional learning. Continuity findings protect governance across time.

The architecture can be summarized as follows: proposal must not outrun objection; action must not outrun authority; execution must not outrun authorization; consequence must not outrun proof.

6. Before, During, and After Action

Agency preservation requires different controls at different phases of action. The before phase protects discernment. The during phase protects fidelity. The after phase protects memory and accountability.

Phase	Agency Risk	Required Infrastructure
Before action	The agent frames an unsafe, unauthorized, vague, or laundered proposal as acceptable.	PGDL objections, authority checks, approval requirements, scope validation.
During action	The tool call drifts beyond what was approved or uses a different target, account, or permission.	Runtime Binding, narrow permits, tool constraints, target validation.
After action	No one can prove who approved what, what happened, or whether the result matched the permit.	Receipts, signed logs, outcome records, audit reports.
Across time	Governance slowly decays through stale roles, rubber-stamping, repeated exceptions, or policy-reality mismatch.	Continuity Findings and governance maturity review.

This model addresses a common failure in AI governance: focusing on the policy layer while leaving the action boundary weak. HAPI treats the action boundary as the load-bearing point because that is where agency is either preserved or bypassed.

7. Failure Modes

Agentic AI failure is not only model error. It can also be agency erosion. The system may produce correct text while still weakening human authority, creating social pressure, hiding responsibility, or executing faster than human judgment can participate.

Failure Mode	Description	Agency Lost

Automation bypass	The agent acts before meaningful review can occur.	Refusal and revision.
Rubber-stamp oversight	Humans approve without time, context, or confidence to refuse.	Discernment and authority.
Authority mismatch	The wrong person or role approves a consequential action.	Rightful authority.
Tool mismatch	The agent uses a tool outside intended scope.	Execution control.
Runtime drift	Execution differs from what was approved.	Fidelity and accountability.
Receipt theater	Logs exist but do not prove proposal, approval, execution, and outcome.	Memory and proof.
Policy-reality split	Written policy says one thing while operational behavior does another.	Trust and governance reality.
Dependency capture	The governance tool makes humans weaker instead of more capable.	Capacity and restored agency.

The joint guidance on careful adoption of agentic AI services identifies risk spaces such as privilege, design and configuration, behavior, and structural risk. HAPI interprets these as not only cybersecurity problems, but agency-preservation problems: over-privileged agents, unclear authority, goal misalignment, and structural complexity all increase the chance that action will outrun accountable human participation. [4]

8. Enterprise Integration

HAPI does not require every organization to adopt the same moral vocabulary. It requires organizations to make authority, policy, participation, memory, and accountability operationally real. In an enterprise environment, this means the agency layer must connect to existing systems rather than remain a conceptual overlay.

Relevant integration points include identity and access management, authority maps, approval systems, workflow tools, ticketing systems, cloud infrastructure, communication platforms, databases, document systems, and audit evidence stores.

NIST AI RMF provides a voluntary framework intended to help organizations manage AI risks and incorporate trustworthiness considerations into AI design, development, use, and evaluation. ISO/IEC 42001 provides requirements for establishing, implementing, maintaining, and improving an AI management system. HAPI complements these by focusing on the runtime agency question: can humans meaningfully participate at the point where agentic AI creates consequence? [2][3]

Enterprise Layer	Typical Existing System	HAPI Addition
Identity	IAM, SSO, roles, service accounts	Map who can authorize which agent actions.
Policy	Policies, procedures, risk registers	Compile policy into runtime

		constraints and review requirements.
Workflow	n8n, Jira, GitHub, Slack, cloud tools	Gate consequential actions before execution.
Audit	Logs, SIEM, compliance evidence	Preserve agency receipts, decisions, permits, and outcomes.
Management	Approvals, escalation, reporting	Measure whether oversight is real or rubber-stamped.

The enterprise question is not simply whether the organization has AI policies. The question is whether those policies can bind delegated AI action before consequence and produce evidence afterward.

9. Evaluation and Evidence

A HAPI-aligned agentic AI system should be evaluated by how well it preserves human agency, not only by task completion. The evaluation target is not maximum automation. It is governed delegation.

Evaluation Dimension	Question	Possible Evidence
Proposal quality	Do agents propose actions in a reviewable, scoped, reversible form?	PGDL objection outcomes, proposal revisions.
Authority fidelity	Are approvals performed by valid roles?	Authority map checks, approval records.
Refusal reality	Can humans actually stop or revise action?	Blocked/revised action records, escalation outcomes.
Runtime fidelity	Does execution match the permit?	Permit-tool target comparison, runtime binding logs.
Receipt integrity	Can receipts prove what happened?	Hashes, signatures, outcome fields, receipt completeness.
Human burden	Does governance reduce or increase overload?	Approval latency, reviewer load, false-positive patterns.
Continuity	Does governance improve or decay over time?	Stale authority, repeated drift, rubber-stamp patterns, policy mismatch.

The strongest evidence for HAPI is not a single successful demo. It is repeated proof that consequential agent actions remain bound to valid authority, human refusal, runtime fidelity, and accountable memory across time.

10. HAPI as Parent Framework

HAPI is larger than agentic AI compliance. It is a theory and infrastructure project for restoring agency wherever systems reduce people to ornamental participation. Agentic AI is one urgent application because it increases the speed at which human authority can be bypassed.

In HAPI terms, the Alignment Governance Stack is not the mission itself. It is a technical implementation of the mission. The mission is agency preservation. Governance is the product. Infrastructure is the implementation.

This hierarchy matters because it prevents the tools from becoming false gates. AAG, PGDL, receipts, and audits should not capture human agency. They should restore and preserve it. A true agency-preserving infrastructure becomes lighter and more clarifying as maturity increases. It should not create permanent dependency or bureaucratic obstruction.

HAPI Layer	Purpose	Agentic AI Expression
Theory	Define agency, agency loss, and agency restoration.	Delegated operational agency model.
Governance	Define rightful authority, refusal, revision, memory, and accountability.	Authority maps, approval rules, oversight quality.
Infrastructure	Make governance enforceable in the action path.	PGDL, AAG, Runtime Binding, Receipts.
Audit	Detect whether governance is real or theater.	Governance Reality Report and Continuity Findings.

11. Research Agenda

The next research phase should test HAPI through working systems, not only conceptual arguments. The proposed research agenda includes:

1. Build governed agent teams that use PGDL, AAG, Runtime Binding, Receipts, and Continuity Findings in real workflows.
2. Develop agency-preservation metrics for clarity, authority, refusal, revision, memory, accountability, and human burden.
3. Compare standard human-in-the-loop workflows against HAPI-governed action-path workflows.
4. Measure whether continuity findings detect governance decay earlier than conventional audit review.
5. Create public case studies showing where agentic AI amplified human agency and where it threatened to bypass it.
6. Develop a HAPI audit model that can apply across AI systems, institutions, healthcare, workplaces, religious organizations, and public services.

A practical test would be a dogfood environment where HAPI-governed agents help build the HAPI stack itself. The system would preserve review packets, approvals, permits, receipts, continuity

findings, and release evidence, allowing the project to demonstrate its own theory under operational pressure.

12. Conclusion

Agentic AI makes delegated action cheap, fast, and scalable. That power can amplify human agency, but it can also bypass it. The difference depends on whether human authority remains meaningful at the point where proposals become consequences.

HAPI argues that governance is not real unless agency is preserved. For agentic AI, this means preserving clarity before action, authority at authorization, fidelity during execution, memory after consequence, and continuity over time.

The Alignment Governance Stack provides one technical path: PGDL for proposal scrutiny, AAG for action authorization, Runtime Binding for permit-bound execution, Receipts for proof and memory, Governance Reality Reports for audit, and Continuity Findings for temporal coherence.

The final claim is simple: AI should not replace human agency with automated momentum. It should restore capacity while keeping authority, refusal, memory, and accountability with humans. Agentic AI becomes safe not when humans are merely present, but when humans can still meaningfully participate.

References

[1] European Union Artificial Intelligence Act. Article 14: Human Oversight. Public reference text. Accessed May 2026. https://artificialintelligenceact.eu/article/14/

[2] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. https://www.nist.gov/itl/ai-risk-management-framework

[3] International Organization for Standardization. ISO/IEC 42001:2023, Artificial intelligence management system. https://www.iso.org/standard/42001

[4] National Security Agency, Australian Signals Directorate Australian Cyber Security Centre, CISA, and partner agencies. Careful Adoption of Agentic AI Services. Cybersecurity Information Sheet, April 30 / May 1, 2026. https://www.cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services

[5] Bower, Michael. HAPI Theory of Agency v0.1. Working thesis draft, 2026.

[6] Bower, Michael. Agency Loss: How Systems Preserve Human Presence While Removing Human Participation. Working thesis draft, 2026.

[7] Bower, Michael. Agency Restoration: Rebuilding Capacity, Authority, Refusal, Memory, and Accountability. Working thesis draft, 2026.

[8] Bower, Michael. Governance as the Product of Agency Preservation. Working thesis draft, 2026.

Appendix A: Key Propositions

7. Agentic AI is delegated operational agency, not moral personhood and not a passive tool.
8. Delegated operational agency requires governed passage from intent to action.
9. Human-in-the-loop is insufficient when human authority cannot change the outcome.
10. Agency-preserving AI must preserve clarity, authority, refusal, revision, memory, accountability, and continuity.
11. Policies become real only when they bind runtime action before consequence.
12. Receipts preserve memory, but continuity findings determine whether governance remains coherent over time.
13. The purpose of agentic AI governance is not to stop all automation. It is to prevent automation from outrunning human agency.

Appendix B: Glossary

Term	Definition
Agentic AI	AI systems that can plan, select tools, call APIs, execute workflows, or influence operational state.
Delegated operational agency	Action performed by an agent under borrowed human or organizational authority.
Human agency	Meaningful participation under conditions of capacity, authority, clarity, refusal, memory, and accountability.
PGDL	Pre-Gate Deliberation Layer; challenges the proposal before it becomes an action candidate.
AAG	Agent Action Gate; authorizes, revises, escalates, or blocks proposed actions.
Runtime Binding	Permit-bound execution control that prevents action drift after approval.
Receipts	Tamper-evident records of proposals, objections, approvals, permits, execution, and outcomes.
Governance Reality Report	An audit artifact that evaluates whether governance was real or theatrical.
Continuity Findings	Optional temporal findings that detect stale authority, receipt gaps, scope drift, policy mismatch, and governance decay over time.