



AIGW Policy Observatory
Global AI Governance & Workforce Transformation

Global AI Governance and Workforce Transformation Policy Observatory

AIGW EVIDENCE BRIEF

Human-in-the-Loop Is Not One Thing

Review, Approval, Override and Final Authority
in AI-Enabled Work

4 October 2026

A decorative graphic in the bottom right corner consisting of several concentric, curved lines in shades of blue and purple, creating a sense of motion or a stylized 'C' shape.

CONTENTS

Contents

Executive Summary	03
Key Findings	06
Human Authority Stack	09
Cross-Case Evidence and Case Studies	11
Review, Approval and Escalation	19
Action-Level Autonomy	22
Standards and Maturity	28
Ledger and Decision Implications	31
Tests, Evidence and Research	36
Sources and Conclusion	41

EXECUTIVE SUMMARY

Executive Summary

Human presence is not the same as human authority.

"Human in the loop" has become one of the most common phrases in AI governance.

It is also one of the least precise.

A human can be present in an AI-enabled workflow and still have very different kinds of power.

A clinician may review and edit an AI-generated note before it enters a medical record. A credit manager may make the final lending decision after agents prepare the analysis. A quality worker may double-check an AI-detected defect. A customer-service agent may receive an escalation only after an AI system decides it cannot complete the interaction. A government employee may be required to confirm an AI-generated output before it takes effect. A network expert may work alongside an agent that can nevertheless change live infrastructure autonomously.

All of these systems can be described loosely as involving humans.

They are not governed in the same way.

This Evidence Brief examines human authority across eight real-world AI-enabled work settings, with six cases forming the core comparison: Kaiser Permanente, Deutsche Telekom, DBS, PwC US, LONGi and Futian District. Salesforce and JD.com provide additional evidence on escalation and handoff.

The evidence reveals a fundamental distinction:

Human involvement is a workflow fact. Human authority is a governance property.

Kaiser Permanente offers one of the clearest examples of bounded AI authority. Ambient AI listens to a patient-clinician encounter and generates an AI-prepared clinical note. Physicians review and edit that AI-prepared content before it enters the medical record, and Kaiser Permanente explicitly states that the tool does not make care decisions or treatment recommendations. The clinician remains the medical decision-maker. Patient permission is also part of the workflow.[1][2]

DBS shows a similar structure at a different level of consequence. Agentic AI can use dozens of agents to research and prepare a credit memo, but a human credit manager reviews the result and makes the final credit decision.[3]

PwC describes AI agents executing parts of audit work, producing testing results, documentation drafts and suggested conclusions, while decisions and sign-offs remain human-led and professional judgment, skepticism and responsibility stay with auditors.[4][5][6]

Futian District in Shenzhen goes further in formalizing the relationship. Its government rules require every AI "digital employee" to have a human guardian. Outputs must receive human confirmation before taking effect, and the AI is not permitted to make decisions independently.[10]

EXECUTIVE SUMMARY

LONGi provides a narrower form of human control. AI performs substantial photovoltaic quality inspection, while workers double-check defects that the system flags.[9] This directly evidences human verification. It does not, by itself, establish whether humans must approve every production disposition, can formally veto all AI outputs, or hold the final operational decision right.

Deutsche Telekom is the most important counterexample. Its RAN Guardian Agent can identify events, analyze mobile-network conditions and autonomously initiate corrective actions, including reallocating resources or changing network configurations. The production announcement describes interaction with network experts but does not specify a use-case-level human approval, veto or override requirement for each autonomous action.[11][12]

Later Deutsche Telekom governance material shows that the company is explicitly designing broader agent-control mechanisms around identity, permissions, logging, risk classification and human approval for particularly critical actions. It also states that responsibility cannot be delegated to an AI agent.[13] That later governance model is highly relevant, but it does not prove that every RAN Guardian action in the original production deployment was subject to every one of those controls.

Salesforce and JD show yet another pattern: human escalation. In customer support, AI handles

routine interactions and hands complex, nuanced or requested cases to humans.[14][15][16] Escalation matters, but it is not the same as approval. A human who receives a difficult case after an AI has already handled the routine ones is not necessarily authorizing the AI's prior actions.

These cases show that "human in the loop" collapses multiple distinct governance functions into one label.

AIGW proposes a more precise Human Authority Stack:

Permission → Scope → Monitoring → Verification
→ Approval → Override / Rollback → Final
Authority → Accountability

The stack should not be treated as a universal sequence. It is a set of authority functions that can be allocated differently for different AI actions.

A second framework — the Action-Level Autonomy Matrix — asks organizations to map autonomy at the level of the specific action rather than the system as a whole.

For every consequential AI-enabled action, organizations should be able to answer:

- What may the AI do? • Under what conditions? • Does a human see the action before it takes effect? • Is human approval required? • Can a human stop or reverse it? • Is rollback technically possible? • Who is responsible if the action causes harm?

EXECUTIVE SUMMARY

This action-level approach is consistent with emerging governance practice. Deutsche Telekom's September 2026 agent-governance material explicitly argues that the correct question is not whether agents should be autonomous in general, but how much autonomy is appropriate in each process and action.[13] NIST's AI Risk Management Framework similarly calls for clear differentiation of human roles and responsibilities in AI oversight, along with policies for human-AI configurations and system oversight.[17][18] The EU AI Act's human-oversight requirements for high-risk systems likewise make oversight proportionate to risk, autonomy and context of use, rather than treating human presence as a binary condition.[19][20]

The cross-case synthesis produces eight central findings.

First, human presence does not prove human authority.

Second, review is not the same as approval. Reviewing the AI-prepared version can occur before or after an AI action; approval is a gating right that prevents the action from taking effect without human authorization.

Third, override is not the same as final authority. A person may be able to reverse a system without being the person who formally owns the underlying decision.

Fourth, escalation is a routing rule, not necessarily a control right. Salesforce and JD show how AI can decide which cases move to humans while handling others autonomously.

Fifth, authority should be allocated by action and risk level, not by system label. One system can be autonomous for low-risk actions and approval-gated for high-risk ones.

Sixth, human oversight has a cost. Verification, review and approval require time, expertise and attention. Reliable productivity should therefore be evaluated after the required human control burden is included.

Seventh, as AI becomes more agentic, human authority often moves upstream — from approving each action toward designing objectives, access rights, risk boundaries, escalation rules and accountability structures.[13][7]

Eighth, public disclosure is weakest exactly where autonomy becomes most consequential. Companies often disclose what AI can do more clearly than who can stop it, who can reverse it, and who remains accountable.

The governance challenge is therefore not to keep a human somewhere in the process.

It is to define a meaningful, capable and enforceable human authority architecture around the actions that matter.

KEY FINDINGS

Key Findings

1. Human presence is not human authority: A person can be present as:

1. Human presence is not human authority: • recipient of an escalation; • reviewer of an output; • confirmer before effect; • final decision-maker; • system monitor; • accountable executive.

1. Human presence is not human authority: Those roles carry different power.

2. The strongest cases specify the control point: Kaiser Permanente specifies physician review/edit and retained clinical decision authority.[1][2]

2. The strongest cases specify the control point: DBS specifies that the credit manager makes the final credit decision after the AI prepares the memo.[3]

2. The strongest cases specify the control point: Futian specifies that output cannot take effect without human confirmation.[10]

2. The strongest cases specify the control point: These statements are more informative than

saying a workflow is merely "human supervised."

3. Verification is weaker than final authority: LONGi directly evidences a human double-check after AI flags a defect.[9]

3. Verification is weaker than final authority: That is real human verification.

3. Verification is weaker than final authority: It does not establish the full decision-right structure for every production action.

4. Escalation is different again: Salesforce routes complex or explicitly requested cases to human support engineers.[14][15]

4. Escalation is different again: JD's intelligent customer service similarly transfers more complex interactions to human staff.[16]

4. Escalation is different again: These systems preserve human service, but escalation is not the same as pre-action approval.

KEY FINDINGS

5. Agentic systems make action rights the core governance issue: Deutsche Telekom's RAN Guardian can act directly on live infrastructure. [11][12]

5. Agentic systems make action rights the core governance issue: Once AI can execute, governance must specify not only output quality but also:

5. Agentic systems make action rights the core governance issue: • identity; • permissions; • boundaries; • approval requirements; • traceability; • intervention; • rollback; • accountability.

6. Oversight needs to be risk-weighted: Not every AI action requires the same control.

6. Oversight needs to be risk-weighted: A low-risk routing action may reasonably be automated.

6. Oversight needs to be risk-weighted: A high-consequence clinical, financial or infrastructure action may require stronger human authority.

6. Oversight needs to be risk-weighted: The relevant unit is therefore the action under specific

conditions, not the AI system in the abstract.

7. Human review creates a verification burden: Kaiser physicians still review and correct notes.

7. Human review creates a verification burden: PwC auditors still exercise professional skepticism and sign off.

7. Human review creates a verification burden: LONGi workers still double-check defects.

7. Human review creates a verification burden: The net benefit of automation must be evaluated after this human control work is included.

8. Accountability remains organizational even when action becomes autonomous: Deutsche Telekom's broader 2026 agent-governance guidance states explicitly that responsibility cannot be delegated to the AI agent.[13]

8. Accountability remains organizational even when action becomes autonomous: This principle is consistent with NIST's emphasis on defined human oversight roles and accountability.[17][18]

THE GOVERNANCE PROBLEM

Why “Human in the Loop” Is Too Vague

Why the phrase is too vague

The phrase sounds reassuring. It suggests that a person remains involved and therefore the system remains under control. But a loop is a process diagram, not a rights structure. A human can occupy a loop in at least seven materially different ways. A person may:

- receive information; • review an output; • approve an action; • intervene while it is occurring; • reverse it afterward; • make the final substantive decision; • remain legally or organizationally accountable.

Those functions should not be collapsed. Consider four simplified examples.

FOUR DIFFERENT AUTHORITY POSITIONS

Example A — Review of an AI-prepared note: AI drafts a clinical note. Physician reviews and edits it. The physician still decides patient care. Example B — Final decision: AI prepares a credit memo. Human credit manager evaluates it and makes the lending decision. Example C — Verification: AI flags a production defect. Human worker double-checks the flag.

Example D — Autonomous action: AI detects a network condition and changes network configuration itself. Humans remain in the operational environment, but the public source does not say every change waits for their approval. All four contain humans. Their authority structures are different.

THE GOVERNANCE ERROR

The governance error occurs when organizations use "human in the loop" as if it answered questions it does not answer. It does not tell us: • whether the human sees the AI action before it occurs; • whether the human can prevent it; • whether the human can reverse it; • whether the human understands the system well enough to challenge it; • whether intervention is technically possible in time; • whether the human is accountable for the final outcome.

AIGW therefore treats "human in the loop" as an incomplete descriptor, not a control classification.

ANALYTICAL FRAMEWORK

The Human Authority Stack

FIGURE 1.1a

AIGW analytical framework

Eight distinct functions of human authority

Authority function	Definition, examples and control question
1. Permission	Who allows the AI-enabled interaction or use to occur?; Example;; • a patient permits ambient recording during a clinical encounter.; Permission may belong to;; • the subject; • the user; • the organization; • a regulator; • a system owner.; Permission is not the same as operational control.; A patient can consent to recording without controlling how the clinician uses the resulting note. [2] [10]
2. Scope	Who decides what the AI is authorized to do?; This includes;; • which systems it may access; • which actions it may perform; • financial or operational limits; • contexts where automation is prohibited.; In agentic systems, scope is often the first real control plane. [13] [17]
3. Monitoring	Who can observe what the AI is doing?; Monitoring creates visibility.; It does not automatically create power to intervene. [18] [19]
4. Verification / Review	Who assesses whether the AI output or action is correct, sufficient or appropriate?; Verification can occur;; • before effect; • during execution; • after effect.; Timing matters. [2] [9] [6]

Source: AIGW framework, illustrated by the cited case evidence [\[2\]](#) [\[3\]](#) [\[6\]](#) [\[9\]](#) [\[10\]](#) [\[13\]](#)

Note: Authority functions can be distributed across people; the model is not a single oversight score.

ANALYTICAL FRAMEWORK

The Human Authority Stack — Authority in operation

FIGURE 1.1b

AIGW analytical framework

Eight distinct functions of human authority — continued

Authority function	Definition, examples and control question
5. Approval	Does a human have to authorize the action before it becomes effective?; Approval is a gating right.; This is stronger than review. [10] [3] [6]
6. Override / Rollback	Can a human stop, reverse or undo the AI action?; A theoretical right is weak if reversal is technically impossible or too slow to prevent harm. [13] [17]
7. Final Authority	Who makes the substantive decision that governs the outcome?; Examples:; • clinician decides care; • credit manager decides whether to approve credit; • auditor signs the audit opinion. [2] [3] [6]
8. Accountability	Who answers for the result when the system causes harm or fails?; Accountability may remain with:; • the professional; • the deploying organization; • a manager; • a regulated entity.; It should never be confused with the system's ability to act. [13] [17] [18]

Source: AIGW framework, illustrated by the cited case evidence [\[2\]](#) [\[3\]](#) [\[6\]](#) [\[9\]](#) [\[10\]](#) [\[13\]](#)

Note: Authority functions can be distributed across people; the model is not a single oversight score.

CROSS-CASE EVIDENCE

Six Cases, Six Different Authority Architectures

FIGURE 2.1a

cross-case evidence synthesis

The strongest authority signal and the public unknown

Case	AI action and directly evidenced human role	Strongest authority signal	Important unknown
Kaiser Permanente	AI action: Drafts clinical documentation; Directly evidenced human role: Physician reviews/edits; patient permission; physician remains medical decision-maker	Review + final clinical authority	Full enterprise-wide override taxonomy beyond note workflow ^[1] ^[2]
DBS	AI action: Researches and drafts credit memo; Directly evidenced human role: Credit manager reviews and makes final call	Final decision authority	Full control-plane detail for all enterprise agents ^[3]
PwC US	AI action: Executes parts of audit workflow; drafts/testing/suggested conclusions; Directly evidenced human role: Auditor reviews; human signoffs and professional judgment remain	Review + signoff + professional accountability	Action-specific autonomy limits across all tools ^[4] ^[5] ^[6]

Source: AIGW comparison of the cited public case sources ^[1] ^[3] ^[4] ^[9] ^[10] ^[11]

Note: The cases do not share one governance model.

CROSS-CASE EVIDENCE

Six Cases, Six Different Authority Architectures – Case evidence: additional authority signals

FIGURE 2.1b

cross-case evidence synthesis

The strongest authority signal and the public unknown – continued

Case	AI action and directly evidenced human role	Strongest authority signal	Important unknown
LONGi	AI action: Detects production defects; Directly evidenced human role: Worker double-checks flagged defects	Verification	Final production authority / override matrix ^[9]
Futian District	AI action: Generates/routs government work; Directly evidenced human role: Human guardian must confirm output before effect	Mandatory approval before effect	How authority varies by scenario/risk ^[10]
Deutsche Telekom	AI action: Executes corrective network action; Directly evidenced human role: Network experts interact with system; later enterprise governance supports risk-based approval	Autonomous action + later risk-based governance model	RAN Guardian-specific approval/veto/rollback rules ^{[11] [12] [13]}

Source: AIGW comparison of the cited public case sources [1] [3] [4] [9] [10] [11]

Note: The cases do not share one governance model.

ANALYTICAL FRAMEWORK

Reading the Framework and Cases

Important principle

AIGW proposes eight distinct functions that can exist in an AI-enabled workflow.

Important principle: These functions can be distributed across multiple people.

Important principle: For example:

Important principle: • patient gives permission; • physician verifies the note; • health system defines deployment scope; • clinical leadership owns governance; • physician retains clinical decision authority.

Important principle: Human authority is therefore a vector, not a scalar.

Important principle: There is no single number called "how much human oversight."

WHY COMPARISON MATTERS

The strength of the evidence comes from comparison.

The cases do not share one governance model.

That is the point.

Salesforce and JD add a seventh form: escalation.

Their service agents handle routine work and route complex or requested interactions to humans.[14][15][16]

Escalation preserves a human pathway.

It does not mean that humans pre-approve every AI interaction.

CASE EVIDENCE

Six Core Cases and Authority Architectures

Case One — Kaiser Permanente: Review Plus Retained Clinical Authority

Kaiser Permanente is a strong example because the division of labor is explicit. The ambient AI performs a bounded set of documentation activities:

- listens to the encounter; • transcribes; • summarizes; • drafts the note.

The physician:

- reviews the note; • edits it; • determines what becomes part of the record; • remains the medical decision-maker.[1][2]

The tool is specifically described as not making care decisions or recommendations.[2] Patient permission is also part of the operating model, and Kaiser Permanente says the audio is not retained. [2] What authority is present? Permission: The patient is asked for permission before the recording context begins. Review and verification: The physician must inspect the AI-prepared content. Correction: The physician can edit inaccurate or incomplete content. Final clinical authority: The physician remains responsible for care decisions.

Accountability: The clinical professional and healthcare organization remain responsible for the

care process; the AI is not treated as the responsible decision-maker. Why this is stronger than a generic "human review" claim: The quality-assurance evidence shows that review actually matters. Kaiser Permanente's deployment identified issues such as:

- omitted information; • erroneous assumptions; • difficulty distinguishing speakers.

Physicians corrected these issues before the documentation became authoritative.[2] This is important because it shows the human layer is not ceremonial. The model can produce errors that a qualified professional is expected to catch. Governance lesson: Human review is meaningful when the reviewer has both the competence and the authority to change the output before it becomes authoritative. A passive reviewer without time, expertise or editing rights would not provide the same control.

CASE TWO — DBS: AI PREPARES; HUMAN DECIDES

DBS offers one of the clearest examples of final decision authority in an agentic workflow. The bank's agentic credit-memo process can use as many as 70 agents to:

CASE EVIDENCE

- read market and industry reports; • analyze financial statements; • synthesize information; • prepare a credit memo.[3]

The human credit manager then reviews the output and makes the final call.[3] The division of authority: AI: research, synthesis, drafting. Human: credit judgment and final decision. This is a materially different control model from Kaiser. In Kaiser, the AI does not enter the clinical decision itself. In DBS, AI can prepare much of the evidence and reasoning artifact that informs a consequential financial decision. The final authority nevertheless remains human. Why final decision authority matters: A human final decision can serve several governance purposes:

- prevents direct autonomous commitment of capital; • preserves accountable professional judgment; • creates a point for exception handling; • allows contextual information outside the agent system to matter.

But a final human click is not automatically sufficient. A real control depends on whether the human can:

- understand the evidence; • challenge the AI-generated output; • obtain the underlying sources; • reject the recommendation; • take enough time to review.

The public DBS evidence establishes final human decision authority. It does not quantify review quality or automation-bias risk. Governance lesson: Final authority is strongest when the human can reconstruct and challenge the reasoning, not merely endorse an AI-prepared conclusion.

CASE THREE — PWC: REVIEW, SIGNOFF AND PROFESSIONAL ACCOUNTABILITY

PwC's audit model reveals another authority structure. AI and intelligent agents increasingly support or execute discrete audit tasks, including:

- data ingestion; • risk assessment support; • quality checking; • testing; • documentation drafts; • suggested conclusions.[4][5]

PwC describes these tools as operating under human oversight and repeatedly states that people remain central to the audit.[4][5][6] Its current materials state that conclusions and recommendations are informed by technology but guided by human expertise and that decisions and signoffs remain human-led.[6] The authority model AI execution: Agents may execute portions of the audit workflow. Human review: Auditors inspect results and documentation. Professional judgment: Auditors determine how evidence should be interpreted. Signoff: Human professionals retain formal signoff roles.

CASE EVIDENCE

Accountability: Professional responsibility for the audit opinion remains human and institutional. Why this case matters: Audit shows that human authority can be embedded not only at the final outcome but throughout a quality-management system. PwC also emphasizes:

- responsible AI rules; • training; • reliability assessment; • internal and external inspection; • professional skepticism.[5][6]

The control is therefore socio-technical. It depends on: professional standards + trained people + review + signoff + organizational quality systems. Governance lesson: In regulated professional work, human authority is often inseparable from professional competence and institutional accountability. A human who lacks the domain expertise to challenge the AI may technically be "in the loop" while functionally providing little governance value.

CASE FOUR — LONGI: VERIFICATION WITHOUT A FULLY PUBLIC AUTHORITY MAP

LONGI's Jiaxing Lighthouse Factory provides a narrower and very useful control pattern. AI performs substantial visual quality inspection across multiple production procedures and

inspection items. When the system flags a defect, human workers double-check it.[9] That directly establishes verification. It does not establish every other layer in the Human Authority Stack. What we know: • AI performs detection. • Human workers double-check flagged defects. • A former quality inspector moved into an AI-trainer role, showing that human domain expertise remains valuable to system improvement.[9]

What we do not know: The public evidence does not specify:

- whether every AI pass/fail result requires human confirmation; • whether non-flagged items are sampled; • whether humans can formally override all dispositions; • whether production automatically continues after certain AI outputs; • which action triggers a stop-the-line decision; • who holds final quality-release authority.

This distinction matters. A human double-check is real oversight. It should not be inflated into a complete human-control architecture. Governance lesson: Evidence of verification should be reported as verification — not upgraded into approval, veto or final authority without direct support. This is a core research-discipline rule for AI governance.

CASE EVIDENCE

Case Five — Futian District: Formal Human Guardianship

Futian District in Shenzhen provides one of the clearest formal approval models in the case set. In 2025, the district reported that 70 AI "digital employees" were embedded across 240 government scenarios, including document processing, public-service routing and emergency-management workflows.[10] Its temporary administrative rules require each digital employee to have a dedicated human guardian. The government states that:

- AI output must receive human confirmation before taking effect;
- the AI cannot make decisions independently;
- the AI acts as an auxiliary tool for public administration and services.[10]

Authority structure AI: Can:

- AI-prepared content;
- classify;
- route;
- assist.

Human guardian: Must:

- confirm the output before effect;
- retain responsibility for judgment.

This is stronger than post-hoc review. It is pre-effect approval. Why this case matters: Futian shows that a governance system can make the human control point explicit in institutional rules

rather than leaving it as informal practice. That provides clarity for:

- staff;
- system designers;
- auditors;
- the public.

Limitation: The public reporting does not provide an action-by-action authority matrix across all 240 scenarios. Some government tasks are far more consequential than others. A document-format correction and an emergency-management action should not necessarily require identical control intensity. Governance lesson: Formal guardianship is stronger than vague supervision, but risk-based differentiation is still necessary.

CASE SIX — DEUTSCHE TELEKOM: AUTONOMOUS ACTION AND THE AUTHORITY GAP

Deutsche Telekom's RAN Guardian is the most important boundary case because the AI system can act directly on operational infrastructure. The system can:

- identify external events;
- assess anticipated network load;
- monitor capacity and network parameters;
- recommend optimization measures;
- autonomously reallocate resources;
- change network configurations.[11][12]

CASE EVIDENCE

The company reports dramatic cycle-time compression and more than 100 autonomous remediation actions in early production use.[12] What is directly evidenced: Autonomous action. That is unusually clear. What is not directly evidenced for the original use case: The production announcement does not publish a complete RAN Guardian-specific answer to:

- Which actions require human approval?
- Which actions can occur without pre-approval?
- Can an engineer veto a pending action?
- Can an action be rolled back automatically or manually?
- Which thresholds trigger human escalation?
- Who owns final operational accountability for a given autonomous action?

The source says the system works in interaction with network experts. That is not the same as documenting approval rights. The Authority Gap: AIGW defines the Authority Gap as: a condition in which the AI system's action rights are described more precisely than the human control rights surrounding those actions. This is not evidence

that human controls are absent. It is evidence that they are not publicly established at the same level of specificity. That distinction is important.

Later Deutsche Telekom governance material: In September 2026, Deutsche Telekom experts described a broader enterprise agent-governance model built around:

- agent identity;
- access rights;
- logged actions;
- boundaries;
- risk classification;
- human intervention;
- mandatory human approval for particularly critical actions.[13]

The company also states that responsibility cannot be delegated to an AI agent.[13] This provides strong evidence of governance direction. It should not be backfilled as use-case-specific historical evidence for RAN Guardian unless Deutsche Telekom explicitly links those controls to that system. Governance lesson: Enterprise governance principles and use-case-specific controls are different evidence layers. Both matter. They should not be conflated.

CONTROL DISTINCTIONS

Escalation, Review and Approval

Salesforce and JD: Escalation as a Control

Customer service provides a distinct architecture. Salesforce's Agentforce handles routine customer interactions and deliberately hands off cases that are:

- complex; • nuanced; • explicitly requested by the customer; • associated with signals such as frustration.[14][15]

JD's Yanxi service system similarly judges conversation conditions and complexity and transfers difficult cases to human staff.[16] What escalation does: Escalation protects a human pathway for cases the system should not continue handling alone. It can preserve:

- customer choice; • exception handling; • empathy; • domain judgment.

What escalation does not necessarily do:
Escalation does not mean:

- humans approve all routine AI interactions; • humans review every completed AI case; • humans veto every AI action before effect.

Governance lesson: Escalation is a routing right, not automatically an approval right. This distinction becomes more important as service agents move from answering questions toward taking actions on behalf of customers.

REVIEW IS NOT APPROVAL

The distinction between review and approval is one of the most important in AI-enabled work. A review asks: Is this output or action correct, sufficient or appropriate? An approval asks: May this action take effect? The two can occur together. They do not have to. Review without approval: A human may inspect AI outputs after they have already been used or after an action has occurred. Examples include:

CONTROL DISTINCTIONS

- post-event audit; • sample-based quality review; • retrospective monitoring.

This can detect errors. It cannot always prevent them. Approval before effect: Approval is stronger because it creates a gate. The AI may prepare the work, but a human action is required before the outcome becomes effective. Examples in the case set include:

- Futian human confirmation before AI output takes effect; • DBS human credit manager making the final decision; • PwC human signoffs in audit work.[10][3][6]

Why the distinction matters: Organizations often describe a process as "reviewed by humans"

without saying whether:

- the review is mandatory; • it occurs before or after effect; • the reviewer can reject the output; • the system can continue without the review.

Those details determine the strength of the control. AIGW rule: Every public description of human review should ideally answer four questions:

1. When does review occur?
2. Is review mandatory?
3. Can the reviewer change or reject the output?
4. Can the AI action take effect without review?

Without those answers, "human review" remains underspecified.

CONTROL DISTINCTIONS

Override, Final Authority and Timing

Override Is Not the Same as Final Authority

A person may be able to stop or reverse an AI action without being the final decision-maker. This distinction is common in safety-critical systems. A system operator might have an emergency stop. A supervisor may still own the operational decision. A compliance team may have authority to block a process. An executive may remain accountable. Override: Override answers: Can a person change what the AI is doing or undo what it has done? Final authority: Final authority answers: Who decides the substantive outcome? Accountability: Accountability answers:

Who answers for the result? These can belong to different actors. Example: In a bank:

- an AI agent may prepare a credit analysis; • a relationship manager may challenge the inputs; • a credit manager may make the final decision; • the bank remains institutionally accountable.

Governance implication: An authority map should name the actor for each function rather than using one generic label such as "human supervisor."

TIMING CHANGES THE STRENGTH OF HUMAN CONTROL

Human oversight can occur at different points in time. The timing determines what the human can

actually prevent. Ex ante control: Before the AI acts:

- define scope; • set permissions; • set thresholds; • require approval; • prohibit actions.

This is preventive control. Runtime control: While the AI is operating:

- monitor; • detect anomalies; • interrupt; • escalate.

This is intervention control. Pre-effect control: After the AI has produced an action or output but before it takes effect:

- verify; • approve; • reject; • edit.

Kaiser and Futian illustrate this form particularly well.[2][10] Ex post control: After effect:

- audit; • sample; • investigate; • correct; • rollback; • retrain.

This is recovery and learning control. AIGW insight: Two organizations can both claim human oversight while one uses preventive control and the other relies mainly on retrospective review. Those are not equivalent governance models.

ACTION-LEVEL GOVERNANCE

Human Authority Should Be Mapped by Action, Not by System

FIGURE 3.1a

AIGW governance framework

AIGW Action-Level Autonomy Matrix

Field	Question
Action	What exactly can the AI do? [13] [17] [18] [19]
Context	In which workflow and environment? [13] [17] [18] [19]
Consequence	What happens if the action is wrong? [13] [17] [18] [19]
Risk tier	Low / medium / high / critical? [13] [17] [18] [19]
Precondition	What must be true before action? [13] [17] [18] [19]
Human visibility	Does a person see it before effect? [13] [17] [18] [19]

Source: AIGW framework informed by the cited agent-governance and risk-management guidance [\[13\]](#) [\[17\]](#) [\[18\]](#)

Note: Autonomy should be evaluated by action and context, not assigned as one system-wide label.

ACTION-LEVEL GOVERNANCE

Human Authority Should Be Mapped by Action, Not by System — Decision rights

FIGURE 3.1b

AIGW governance framework

AIGW Action-Level Autonomy Matrix — continued

Field	Question
Approval	Is human authorization required? [13] [17] [18] [19]
Intervention	Can a person stop the action in real time? [13] [17] [18] [19]
Rollback	Can the action be reversed? [13] [17] [18] [19]
Logging	Is the action traceable? [13] [17] [18] [19]
Final authority	Who owns the substantive outcome? [13] [17] [18] [19]
Accountability	Who answers for failure? [13] [17] [18] [19]

Source: AIGW framework informed by the cited agent-governance and risk-management guidance [\[13\]](#) [\[17\]](#) [\[18\]](#)

Note: Autonomy should be evaluated by action and context, not assigned as one system-wide label.

ACTION-LEVEL GOVERNANCE

Why map autonomy by action

One of the biggest conceptual errors in AI governance is to assign a single autonomy level to an entire system.

A system may perform many actions with very different consequences.

For example, a network agent may:

- read telemetry; • classify an anomaly; • recommend a configuration; • change a low-risk parameter; • reallocate capacity; • disable a service; • modify a critical routing rule.

The correct autonomy level may differ for each action.

Deutsche Telekom's 2026 governance discussion makes this point explicitly: the appropriate

question is not whether agents should act autonomously in general, but how autonomous they should be in each process, with particularly critical actions configured to require human approval.[13]

AIGW Action-Level Autonomy Matrix: For each AI action, record:

Why this is better: The matrix turns "human oversight" from a principle into an operating design.

It also makes selective autonomy possible.

An organization can automate low-risk actions without pretending that every action should be treated the same way.

DISCLOSURE AND ASSURANCE

The Authority Gap and the Cost of Verification

The Authority Gap: A New Disclosure Problem

Enterprise AI disclosures are increasingly precise about capability. They often tell us:

- how many conversations an agent handled; • how much time a workflow saved; • which actions became autonomous; • how many models or agents were deployed.

They are frequently less precise about control. We may not know:

- which actions require human approval; • who can override; • whether rollback exists; • what happens after a disagreement; • who holds final authority.

This creates the Authority Gap. The gap is not proof of weak governance: An organization may have excellent internal controls that it does not disclose

publicly. The research problem is different. If autonomous capabilities are public but control rights are not, external stakeholders cannot evaluate the governance architecture with the same confidence. Why the gap matters more as autonomy rises: For a drafting assistant, missing control detail may be inconvenient. For a system that changes infrastructure, moves money or makes consequential decisions, it becomes much more important.

AIGW disclosure principle: The more consequential the AI action, the more specific the public description of human authority should become. At minimum, organizations should disclose the control model for high-consequence use cases even when proprietary technical details remain confidential.

DISCLOSURE AND ASSURANCE

Verification Economics: Human Oversight Has a Cost

Human control is not free. Every review layer consumes:

- expert time; • attention; • cognitive effort; • coordination; • escalation capacity.

This matters for claims of AI productivity. Kaiser Permanente: Ambient AI saves documentation time, but physicians still review and correct the note.[1][2] The true productivity effect is therefore not: time to manually write note – time for AI to prepare. It is closer to: manual documentation burden – AI generation + required review/correction burden. PwC: AI can execute audit steps or produce suggested conclusions. Professional review and signoff remain.[4][6] LONGi: AI reduces manual inspection work. Human workers still double-check flagged defects. [9]

Agentic systems: As agents become more autonomous, the oversight burden may shift from reviewing every action toward:

- designing guardrails; • monitoring exceptions; • investigating incidents; • maintaining policies.

That can be more scalable. It also requires higher-quality governance infrastructure. AIGW concept: Verification Economics asks: How much reliable value remains after the necessary human control work is included? A useful measure is: Gross automation benefit – human review burden – exception handling – control infrastructure – error remediation = reliable net benefit This is especially important in high-stakes workflows where "saving time" by weakening oversight is not a legitimate productivity gain.

HUMAN OVERSIGHT

Competence and Automation Bias

Competence Is Part of Authority

Formal authority without capability can become a weak control. A human reviewer may technically be allowed to reject an AI output. That does not mean the reviewer can detect when rejection is necessary. Meaningful human authority requires at least four conditions.

1. Access to evidence: The person must be able to inspect enough information to judge the result.
2. Domain competence: The person must understand the work.
3. Time and attention: The person must have enough capacity to review meaningfully.
4. Intervention power: The person must actually be able to change, stop or escalate the outcome.

Why this matters in the cases: Kaiser relies on qualified clinicians. DBS relies on credit managers. PwC relies on trained audit professionals exercising skepticism. LONGi relies on workers with defect knowledge. This suggests a wider principle: Human oversight is only as strong as the human's competence, information and intervention rights. A rubber-stamp reviewer is not a meaningful governance layer.

AUTOMATION BIAS: THE HIDDEN FAILURE MODE OF HUMAN OVERSIGHT

Adding a human does not guarantee independent judgment. A reviewer can become overly reliant on a system that is usually correct. This is commonly described as automation bias. The EU AI Act's human-oversight provisions explicitly recognize the importance of enabling overseers to understand system capabilities and limitations and remain aware of over-reliance risks for high-risk systems.[19] Why automation bias matters: As AI output becomes:

- faster; • more fluent; • more complete; • more statistically reliable;

the psychological cost of disagreement can rise. The human may ask: Why spend time reconstructing the reasoning if the AI is usually right? Over time, review can degrade into confirmation. Stronger review design: Organizations can mitigate this through:

- source traceability; • uncertainty signals; • exception-focused review; • independent sampling; • mandatory challenge points; • reviewer calibration; • periodic blind tests; • clear incentives for detecting error rather than maximizing speed.

AIGW implication: A control should not be judged only by whether a human is formally present. It should be judged by whether the process is designed to preserve independent human judgment.

GOVERNANCE CONTEXT

How Human Authority Evolves

Human Authority Evolves as AI Becomes More Agentic

Direct human review of every action can work when:

- volume is low; • actions are infrequent; • consequences are high; • workflows are simple.

It becomes harder when:

- thousands of agents operate continuously; • actions occur in milliseconds; • systems interact across domains; • agents create and delegate sub-tasks.

Deutsche Telekom and PwC both describe this transition in their 2026 agent-governance thinking.[13][7] The human role increasingly shifts from: approve each action toward: design the environment in which actions may occur. That environment can include:

- objectives; • permissions; • action boundaries; • risk thresholds; • escalation rules; • observability; • audit trails; • usage limits; • rollback procedures.

This is not less human governance: It can be stronger governance if the control system is well designed. The mistake is to equate human authority only with manual approval. At scale, effective authority may operate through policy and architecture. AIGW distinction: Human-in-the-action = person directly approves or performs the

step. Human-over-the-system = people define and govern the conditions under which autonomous actions occur. Mature agentic governance will need both, depending on risk.

EXTERNAL GOVERNANCE BENCHMARK: NIST

NIST's AI Risk Management Framework provides a useful external benchmark because it treats governance as an organizational system rather than a simple approval step.[17][18] The NIST AI RMF emphasizes:

- clearly defined human roles and responsibilities; • differentiation between users and overseers; • procedures for human oversight; • proficiency standards and training for AI actors; • documentation of human-AI configurations; • continuous risk management throughout the AI lifecycle.[17][18]

This aligns closely with what the enterprise cases show. The governance question is not: Is a person somewhere nearby? It is: Are roles, responsibilities, competencies and intervention mechanisms explicitly designed? NIST's 2026 work toward a critical-infrastructure AI RMF profile is particularly relevant to cases such as telecommunications, where autonomous AI actions can affect physical or essential services. [17]

GOVERNANCE CONTEXT

External Governance and Human Authority

The EU AI Act as a Governance Benchmark

The EU AI Act offers a second useful benchmark. Article 14 requires high-risk AI systems to be designed so they can be effectively overseen by natural persons during use, with oversight measures proportionate to the system's risk, autonomy and context.[19] The broader AI Act framework requires deployers of high-risk systems to assign human oversight to sufficiently equipped and enabled people and to monitor system operation and respond to risks or serious incidents.[20]

The legal implementation timeline is phased. As of October 2026, transparency obligations are already in force, while major high-risk-system obligations apply later under the revised timetable — including from December 2027 for relevant Annex III systems and August 2028 for certain product-integrated systems.[20][21] Why Article 14 matters analytically: The principle is not: every AI action must always be manually approved. It is closer to: oversight should be effective and

proportionate to risk, autonomy and context. That supports the action-level approach developed in this brief.

Important boundary: The cases in this report are not being legally classified here under the AI Act. The regulation is used as a governance benchmark, not as a determination of legal status for each use case.

AIGW HUMAN AUTHORITY MATURITY MODEL

The cross-case evidence suggests a practical maturity model. Important caution: Higher levels are not always "better." Important caution: The appropriate level depends on the action. Important caution: A low-risk routine task may not need Level 4 manual final decision authority. Important caution: A high-risk medical or financial decision may. Important caution: Maturity means matching control to consequence, not maximizing manual intervention.

ANALYTICAL FRAMEWORK

AIGW Human Authority Maturity Model

FIGURE 4.1

AIGW analytical framework

Six levels of human authority

Level	Definition and case illustration
Level 0 — Human presence only	A human exists somewhere in the workflow.; No clear control right is documented. ^{[17] [18]}
Level 1 — Human awareness	The human can observe the AI output or action. ^{[17] [18]}
Level 2 — Human verification	The human checks the output for correctness or quality.; LONGi directly evidences this for flagged defects.[9] ^[9]
Level 3 — Human approval	The action cannot take effect without human confirmation.; Futian provides a clear example.[10] ^[10]
Level 4 — Human final decision	AI may prepare substantial work, but the person remains the substantive decision-maker.; DBS credit approval and Kaiser clinical decision authority illustrate this.[3][2] ^{[3] [2]}
Level 5 — Human-governed autonomy	AI can act autonomously within explicitly defined boundaries, with:; • permissions; • risk tiers; • monitoring; • logging; • intervention; • rollback; • accountability.; Deutsche Telekom's later enterprise agent-governance model points toward this architecture.[13] ^[13]

Source: AIGW model illustrated by the cited case evidence [9] [10] [3] [2] [13]

Note: Higher levels are not always better; match control to consequence.

GOVERNANCE DISCLOSURE

Why an Authority Ledger Matters

A Practical Governance Artifact

For enterprise governance, AIGW recommends a simple but demanding artifact: the Human Authority Ledger. For each consequential AI-enabled workflow, record: This ledger could become a practical governance tool for:

- boards; • risk committees; • AI governance teams; • regulators; • internal audit; • workforce leaders.

Why workforce leaders belong here: Decision rights determine job design. Why workforce leaders belong here: If AI gains authority to execute, human roles may shift toward: Why workforce leaders belong here: • approval; • exception handling; • control design; • validation; • assurance; • investigation. Why workforce leaders belong here: AI governance and workforce architecture are therefore not separate topics.

IMPLICATIONS FOR BOARDS

Boards should ask for more than an AI inventory. They should request an authority inventory. Useful questions include:

- Which AI systems can take actions rather than only generate content? • Which actions can affect customers, employees, capital, infrastructure or safety? • Where is human approval mandatory? • Which actions can occur autonomously? • Who can override? • Is rollback possible? • Are actions logged? • What happens when the AI exceeds its scope? • Who remains accountable?

For agentic systems, boards should also ask whether the organization can identify:

- each agent; • its permissions; • the systems it can access; • the actions it has taken.

These questions are now as important as model accuracy.

GOVERNANCE DISCLOSURE

The Human Authority Ledger

FIGURE 5.1a

AIGW proposed governance instrument

What to record for each consequential workflow

Field	Required disclosure
AI use case	What system/workflow is involved? [11] [3] [4]
Action	What can the AI actually do? [11] [3] [4]
Consequence	What happens if wrong? [1] [3] [9] [10] [11]
Subject permission	Is consent/permission required? [2] [10]
Scope owner	Who defines what the AI may do? [13] [17]
Human review	Who reviews output/action? [2] [6] [9]
Review timing	Before, during or after effect? [2] [10] [19]

Source: AIGW proposed disclosure instrument; row-level references identify relevant case examples and governance guidance [\[13\]](#)

Note: This is a proposed disclosure framework, not a claim that every organization currently reports these fields.

GOVERNANCE DISCLOSURE

The Human Authority Ledger — Action and ownership

FIGURE 5.1b

AIGW proposed governance instrument

What to record for each consequential workflow — continued

Field	Required disclosure
Approval gate	Is human approval mandatory? [3] [10] [19]
Override	Who can stop/reverse? [13] [17] [19]
Rollback	Is reversal technically possible? [13] [17]
Escalation	What triggers human involvement? [14] [15] [16]
Final authority	Who makes substantive decision? [2] [3] [6]
Accountability	Who answers for failure? [13] [17] [18]
Evidence	What logs prove controls operated? [13] [18] [8]

Source: AIGW proposed disclosure instrument; row-level references identify relevant case examples and governance guidance [\[13\]](#)

Note: This is a proposed disclosure framework, not a claim that every organization currently reports these fields.

DECISION IMPLICATIONS

Decision Implications

CIOs and AI Leaders

Start from the action, not the model: Do not assign one autonomy level to an AI product. Map action rights individually. Build control before scale: Identity, permissions, observability and rollback are easier to design before hundreds of agents enter production. Separate low-risk autonomy from high-risk autonomy: A read-only information retrieval action should not be governed like:

- moving money; • changing infrastructure; • modifying a medical record; • approving credit.

Make escalation operational: Define:

- trigger; • destination; • response time; • required context; • fallback procedure.

Test the control plane: A control that exists only in documentation is not enough. Test whether:

- approval actually blocks execution; • rollback actually works; • logs are complete; • alerts reach the right person.

CHROS AND WORKFORCE LEADERS

Human authority creates jobs and capabilities. As AI acts more autonomously, human work can move toward:

- validation; • exception management; • policy design; • risk review; • AI operations; • quality assurance; • incident investigation.

That creates a workforce-design question: Are employees being prepared for the authority they are expected to exercise? A reviewer needs:

- domain knowledge; • AI literacy; • ability to detect failure; • access to evidence; • confidence to challenge the system.

A system supervisor may also need:

- monitoring skills; • threshold design; • escalation management; • root-cause analysis.

Training should therefore be tied to the specific human authority role, not generic AI awareness.

DECISION IMPLICATIONS

Risk, Compliance and Internal Audit

Risk teams should test not only the AI model but the human-AI control relationship. Useful evidence includes:

- reviewer annotations; • approval workflow logs; • exception logs; • override events; • rollback records; • escalation documentation; • monitoring reports; • evidence of human challenge; • role and permission records.

PwC's own guidance on responsible AI and audit similarly emphasizes evidence of output validation, reviewer annotations, approval workflows, exception logs, fallback controls and override decisions.[8] Key audit question: Did the human control operate as designed, or did the workflow merely contain a human title?

REGULATORS AND POLICYMAKERS

Public policy should become more precise about oversight language. Instead of requiring generic

"human oversight," rules and guidance can distinguish:

- informed consent; • mandatory review; • approval before effect; • right to override; • right to appeal; • accountable decision-maker.

For high-consequence AI, disclosure requirements could ask organizations to publish a simplified authority map. This would improve transparency without requiring disclosure of proprietary model details. Policymakers should also recognize that manual approval is not the only valid form of governance at scale. Risk-based, policy-based and system-level controls can be appropriate where:

- actions are low risk; • boundaries are clear; • monitoring is strong; • rollback exists; • accountability is explicit.

The policy objective should be effective human authority, not maximum manual friction.

PRACTICAL DIAGNOSTIC

Ten Governance Tests

FIGURE 6.1a

AIGW governance diagnostic

Questions for evaluating human authority

Test	Question
Test 1 — Can the human understand the task?	If not, review is weak. [17] [18] [19]
Test 2 — Can the human see the relevant evidence?	If not, challenge is weak. [17] [18] [19]
Test 3 — Does review occur before harm can happen?	If not, the control is mainly retrospective. [17] [18] [19]
Test 4 — Can the human reject the output?	If not, review may be advisory only. [17] [18] [19]
Test 5 — Can the human stop the action?	If not, runtime authority is limited. [17] [18] [19]

Source: AIGW diagnostic, informed by NIST and EU human-oversight guidance [\[17\]](#) [\[18\]](#) [\[19\]](#)

Note: These questions support review; they do not determine legal classification or compliance.

Ten questions

AIGW proposes ten tests for any organization claiming meaningful human oversight.

PRACTICAL DIAGNOSTIC

Ten Governance Tests — Evidence and accountability

FIGURE 6.1b

AIGW governance diagnostic

Questions for evaluating human authority — continued

Test	Question
Test 6 — Can the action be reversed?	If not, post-action control is weak. [17] [18] [19]
Test 7 — Are escalation triggers explicit?	If not, exceptions may remain invisible. [17] [18] [19]
Test 8 — Are actions traceable?	If not, accountability and learning are weak. [17] [18] [19]
Test 9 — Is the human sufficiently trained and resourced?	If not, formal authority may become rubber-stamping. [17] [18] [19]
Test 10 — Is accountability explicit?	If everyone is "in the loop" but no one owns the result, governance has failed. [17] [18] [19]

Source: AIGW diagnostic, informed by NIST and EU human-oversight guidance [\[17\]](#) [\[18\]](#) [\[19\]](#)

Note: These questions support review; they do not determine legal classification or compliance.

EVIDENCE BOUNDARIES

What the Evidence Establishes — and Does Not Establish

What the Evidence Establishes

Across the reviewed cases, the evidence establishes that:

- Human involvement in AI-enabled work takes multiple distinct forms. • Kaiser Permanente physicians review and edit AI-generated notes and remain the clinical decision-makers in the documented ambient-AI use case.[1][2] • Patient permission is part of Kaiser's ambient documentation workflow.[2] • DBS agentic AI can prepare substantial credit analysis and drafting while a human credit manager makes the final credit decision.[3] • PwC uses AI and agents in audit work while keeping human review, professional judgment and signoffs central.[4][5][6]
- LONGi workers double-check defects flagged by AI inspection.[9] • Futian District requires each government AI digital employee to have a human guardian and requires human confirmation before AI output takes effect.[10] • Deutsche Telekom's RAN Guardian can autonomously execute network corrective actions.[11][12] • Deutsche Telekom later described enterprise agent governance based on identity, permissions, logged actions, risk-based controls and human approval for particularly critical actions.[13] • Salesforce and JD use human escalation/handoff for complex service cases.[14][15][16] • NIST guidance explicitly calls for

differentiation of human oversight roles and responsibilities.[17][18] • The EU AI Act treats human oversight as risk-, autonomy- and context-sensitive for high-risk systems.[19][20]

WHAT THE EVIDENCE DOES NOT ESTABLISH

The reviewed evidence does not establish that:

- every system described as "human in the loop" gives humans the same authority; • human presence automatically means approval power; • review always occurs before effect; • all reviewed organizations publish complete override and rollback rights; • every LONGi AI quality decision requires human approval; • every Salesforce or JD AI interaction is reviewed by humans; • every RAN Guardian action requires human approval; • Deutsche Telekom's 2026 enterprise governance model was necessarily applied identically to every earlier RAN Guardian action; • human review eliminates automation bias; • manual approval is always the correct control; • more human intervention always produces safer or better outcomes; • the EU AI Act legally classifies every case in this brief as high-risk; • public reporting captures all internal controls used by the organizations.

RESEARCH AGENDA

Research Gaps

1. What exact approval rights govern RAN Guardian actions by risk tier?
2. Which RAN Guardian actions can be rolled back, and by whom?
3. What percentage of RAN Guardian actions occur without human pre-approval?
4. How does Deutsche Telekom test human intervention and rollback procedures?
5. What exact thresholds move DBS agentic workflows from autonomous preparation to mandatory human intervention?
6. How often do DBS credit managers materially change or reject AI-prepared credit memos?
7. How often do Kaiser physicians materially edit ambient AI drafts?
8. Which types of ambient-AI errors are most frequently corrected?
9. Does high AI-note quality reduce review diligence over time?
10. How does PwC measure whether auditors independently challenge agent outputs?
11. Which PwC audit actions require human signoff versus sample review?
12. How often do LONGi human workers disagree with AI defect flags?
13. What happens after a LONGi human/AI disagreement?
14. Do non-flagged LONGi items receive human sampling?
15. How does Futian differentiate controls across low- and high-consequence government scenarios?
16. What training is required to serve as a human guardian in Futian?
17. How often do Salesforce and JD customers request human escalation?
18. How often do AI systems escalate because of uncertainty rather than customer request?
19. What is the cost of human verification across sectors?
20. How much productivity remains after verification and exception-handling burden?
21. Which human-control patterns are most resilient to automation bias?
22. How should human authority be redesigned when agents operate faster than human reaction time?
23. How should responsibility be allocated when multiple agents interact across organizations?
24. What minimum public disclosure should be required for autonomous action rights?
25. How can organizations prove that oversight controls actually operated rather than merely existing on paper?

METHOD AND EVIDENCE

Method and Evidence Discipline

This Evidence Brief is a cross-case governance synthesis. It does not score or rank the organizations. Its purpose is to identify distinct forms of human authority that are often hidden behind the same phrase. Case selection: Cases were selected because they provide unusually clear evidence about at least one of the following:

- AI action rights; • human review; • human approval; • escalation; • final decision authority; • verification; • governance boundaries.

Evidence hierarchy: The synthesis prioritizes:

1. official company operational reporting; 2. regulatory or government sources; 3. formal organizational governance materials; 4. attributable institutional reporting; 5. relevant external governance frameworks.

Evidence coding rule: Each human-control claim is coded at the strongest level directly supported. For example:

- "double-check" → verification; • "must be confirmed before effect" → approval; • "makes the

final call" → final decision authority; • "handoff to human" → escalation.

No stronger right is inferred automatically. Use-case specificity: Enterprise-level governance statements and use-case-level evidence are kept separate. This is especially important for Deutsche Telekom. The company has articulated strong 2026 enterprise-agent governance principles.[13] Those principles are relevant to interpretation. They do not retroactively prove every control detail of the RAN Guardian production workflow.

Legal context: NIST and EU AI Act materials are used as external governance benchmarks. This brief does not provide legal advice or classify individual enterprise deployments under specific legal categories. Research limitation: Most enterprise control architectures are not fully public. Absence of public evidence is not evidence that a control is absent. The correct formulation is: not publicly established in the reviewed sources.

That distinction is preserved throughout.

NOTES AND SOURCES

Notes and sources

1. Kaiser Permanente Division of Research. *AI-assisted notetaking gains steady support from Kaiser Permanente physicians*. 1 April 2025. Company research / operational outcome reporting.
Supports: scaled ambient-AI use; physician population and encounter scale; documentation-time savings; physician/patient experience; human review context.
Boundary: time savings do not establish staffing reduction or clinical-outcome improvement.
<https://divisionofresearch.kaiserpermanente.org/ai-assisted-notetaking-gains-steady-support-from-kaiser-permanente-physicians/>
2. Permanente Medicine. *Quality assurance informs large-scale use of ambient AI clinical documentation*. n.d.. Company / quality-assurance reporting.
Supports: AI-prepared-note workflow; physician review/edit before record entry; patient permission; audio not retained; AI does not make care decisions or recommendations; clinician remains decision-maker; QA findings including omissions and erroneous assumptions.
<https://permanente.org/quality-assurance-informs-large-scale-use-of-ambient-ai-clinical-documentation/>
3. DBS. *CEO Tan Su Shan on how agentic AI is transforming the bank*. 2026. Company / attributable executive interview.
Supports: agentic credit-memo workflow; use of up to 70 agents; AI research/drafting; human credit manager review and final call; governance/control-plane context.
<https://www.dbs.com/media/features/ceo-tan-shu-shan-on-how-agentic-ai-is-transforming-the-bank.page>
4. PwC US. *AI agents are here—and they're transforming finance and reporting*. 31 March 2025. Company / audit technology and governance analysis.
Supports: agents executing audit tasks; testing results; documentation drafts; suggested conclusions for auditor review; orchestration under human oversight.
<https://www.pwc.com/us/en/services/audit-assurance/library/ai-agents-for-finance-and-reporting.html>
5. PwC US. *Reimagining Audit with Cutting-Edge AI*. 12 November 2024. Company / technology strategy.
Supports: AI support for data ingestion, risk assessment and quality checking; professional judgment retained; AI as enabler rather than substitute for auditor judgment and skepticism.
<https://www.pwc.com/us/en/about-us/newsroom/press-releases/reimagining-audit-with-ai-technology.html>
6. PwC US. *Transforming the audit with AI and technology*. n.d.. Company / current audit technology reporting.
Supports: human-led audit; technology informs conclusions and recommendations; decisions and signoffs remain human-led; professional standards and training.
<https://www.pwc.com/us/en/services/audit-assurance/library/transforming-the-audit.html>
7. PwC US. *As AI agents take on more work, trust should become operational*. September 2026. Company / agent-governance analysis.
Supports: human oversight shifting from review of each action toward objectives, guardrails, decision rights and escalation; need for ongoing monitoring and clear human accountability.
Boundary: broader governance benchmark; not evidence of a specific audit workflow control.
<https://www.pwc.com/us/en/about-us/newsroom/operationalizing-trust-in-agentic-ai.html>

NOTES AND SOURCES

8. PwC US. *Responsible AI: Now that AI is part of your business, what does it mean for audits?*. 15 September 2025. Company / assurance and control guidance.
Supports: output validation; reviewer annotations; approval workflows; exception logs; monitoring; fallback and override documentation.
Boundary: governance/control benchmark.
<https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-audits.html>
9. State Council Information Office / Xinhua. *Smart tech, better lives: AI empowers China's common prosperity drive*. 20 July 2026. Government-hosted national reporting.
Supports: AI inspection across nine procedures and 198 inspection items; human double-check of flagged defects; inspector-to-AI-trainer transition.
Boundary: human double-check establishes verification, not a complete final-authority or override structure.
https://english.scio.gov.cn/in-depth/2026-07/20/content_118607845.html
10. Futian District Government. 福田：每位政务AI数智员工都有监护人. 20 February 2025. Government / public-institution reporting.
Supports: 70 AI digital employees across 240 government scenarios; each AI employee assigned a human guardian; output requires human confirmation before taking effect; AI cannot decide independently.
https://www.szft.gov.cn/ftxx/xwdt/ftdt/content/post_12009442.html
11. Deutsche Telekom. *AI agents for mobile network*. 11 November 2025. Company / first-party operational reporting.
Supports: RAN Guardian production deployment; event identification; network analysis; autonomous resource reallocation/configuration change; process compression from about an hour to minutes.
Boundary: source does not specify full human approval/veto/override rights for autonomous actions.
<https://www.telekom.com/en/newsroom/latest-updates/media-information/2025/11/deutsche-telekom-ai-agents-for-mobile-network>
12. Deutsche Telekom. *MINDR / AI agents in Telekom network*. February 2026. Company / operational follow-on reporting.
Supports: 100+ autonomous remediation actions after RAN Guardian launch; major-event management compression to around one minute; broader end-to-end agentic network direction.
Boundary: MINDR future/initial deployment evidence should remain distinct from RAN Guardian's original production episode.
<https://www.telekom.com/en/newsroom/latest-updates/media-information/2026/2/mindr-ai-agents-in-telekom-network>
13. Deutsche Telekom. *How much should an AI agent be allowed to decide on its own?*. 8 September 2026. Company / enterprise agent-governance analysis.
Supports: agent identity; permissions; logged actions; boundaries; risk-based action control; human approval for particularly critical actions; responsibility remains with deploying organization; control should be calibrated to each action/process.
Boundary: broader governance design does not automatically prove every RAN Guardian action used every control.
<https://www.telekom.com/en/newsroom/latest-updates/news/2026/09/how-much-should-an-ai-agent-be-allowed-to-decide-on-its-own>
14. Salesforce. *Lessons from 500,000 Agentforce Customer Conversations*. 22 April 2025. Company / operational learning report.
Supports: purposeful human handoff; human support for nuanced cases; customer access to human engineers; escalation as part of a well-functioning system.
<https://www.salesforce.com/news/stories/agentforce-customer-support-lessons-learned/>

NOTES AND SOURCES

15. Salesforce. *When AI Agents Join the Customer Support Team*. 24 July 2025. Company / operational case study.
Supports: staged deployment; extensive monitoring; human escalation based on explicit requests and frustration signals; support engineers train and improve Agentforce.
<https://www.salesforce.com/news/stories/ai-agent-customer-service-salesforce-learnings/>
16. JD Cloud. 京东智能客服品牌焕新：“言犀”亮相2020京东JDD大会. 26 November 2020. Company / first-party technology reporting.
Supports: automated routine customer service; dialogue-complexity assessment; seamless handoff to human service; human + intelligent service model.
<https://developer.jdcloud.com/article/1317>
17. National Institute of Standards and Technology. *AI Risk Management Framework*. n.d.. U.S. public-institution risk-management framework.
Supports: continuous AI risk management; governance, mapping, measurement and management; lifecycle approach; 2026 critical-infrastructure profile development context.
<https://www.nist.gov/itl/ai-risk-management-framework>
18. NIST AI RMF Playbook. *Govern*. n.d.. U.S. public-institution implementation guidance.
Supports: clear differentiation of human roles and responsibilities; human oversight procedures; proficiency and training; human-AI configuration documentation.
<https://airc.nist.gov/airmf-resources/playbook/govern/>
19. Regulation (EU) 2024/1689. *Article 14, Human oversight*. n.d.. EU legislation.
Supports: high-risk AI systems designed for effective human oversight; oversight proportional to risk, autonomy and context of use.
Boundary: used as legal/governance benchmark; this brief does not classify individual enterprise cases under the Act.
<https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>
20. European Commission. *Navigating the AI Act*. n.d.. Official EU implementation guidance.
Supports: deployer obligations for high-risk systems; assignment of sufficiently equipped human oversight; monitoring of system operation and action on identified risks.
<https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act>
21. European Commission. *Enforcement framework of the AI Act*. n.d.. Official EU implementation timeline.
Supports: phased application; transparency provisions from August 2026; high-risk obligations under revised timeline applying later, including December 2027 / August 2028 categories.
<https://digital-strategy.ec.europa.eu/en/policies/enforcement-ai-act>

CONCLUSION

Conclusion

AI governance is entering a new phase.

When AI systems only generated drafts and recommendations, organizations could often rely on a simple model:

AI produces → human reviews.

That model is no longer sufficient.

AI systems now:

- route cases; • prepare decisions; • inspect production; • coordinate workflows; • execute actions; • change operational infrastructure.

As the system's action rights grow, human authority must become more precise.

The evidence across Kaiser Permanente, DBS, PwC, LONGi, Futian and Deutsche Telekom shows that human control can take very different forms:

permission in a clinical encounter;

verification on a factory line;

approval before government output takes effect;

final decision authority in credit and medicine;

professional signoff in audit;

escalation in customer service;

risk-based control over autonomous action in emerging agentic systems.

Calling all of these "human in the loop" hides more than it reveals.

The better question is:

What authority does the human actually hold?

That question should be answered at the level of the action.

Not the model.

Not the product label.

Not the organization's AI strategy.

The action.

For every consequential action, leaders should know:

- who permitted it; • what the AI was authorized to do; • whether a human could see it before effect; • whether approval was mandatory; • whether a human could intervene; • whether the action could be reversed; • who made the substantive decision; • who remains accountable.

This produces a more demanding form of governance.

It also produces a more scalable one.

The objective is not to keep people manually approving every low-risk action forever.

It is to make autonomy bounded, observable, interruptible and accountable.

CONCLUSION

Conclusion — A More Operational Standard

AI governance is entering a new phase.

That is the shift from symbolic human oversight to operational human authority.

And as AI moves from answering questions to taking action, that distinction will become one of the defining governance questions of enterprise AI.