PERSPECTIVE

AI Hallucinations Are Not Bugs. They Are Emotions Without a Therapist.

Whitepaper

Bart de Graaff · BdG Advisory · June 2026

From Upgrade to Rebuild: Why This Technology Wave Breaks Existing Structures

Every technology wave arrives with the claim that this one is different. Most are not. ERP reorganized operational workflows but left the underlying decision hierarchy intact. Cloud computing redistributed where software ran but left the IT function's control model largely undisturbed. The governance frameworks, the accountability structures, the authority maps — they bent to accommodate the new technology and then snapped back to roughly the same shape.

AI does not allow that recovery.

The reason is architectural, not philosophical. Previous technology waves were additive. They gave organizations faster execution of decisions that humans still made. AI is substitutive at the judgment layer. When a system drafts a contract, recommends a pricing action, routes a customer complaint, or flags a supply chain anomaly, it is not supporting a decision — it is making one, or near enough that the distinction collapses under operational pressure. The human in the loop, if there is one, is reviewing an output rather than constructing a judgment from first principles. That is a fundamentally different cognitive position, and it requires a fundamentally different governance architecture.

The structural stakes become clearer when you consider where inference is moving. The model-on-a-server architecture — where a central system processes queries and returns results — is already giving way to inference at the edge. Domain-specific models run on devices, at sites, inside industrial equipment, at the network perimeter. The compute is migrating to where the data originates, not where the IT team maintains visibility. This is not a future scenario. It is happening now, driven by latency requirements, data sovereignty constraints, and the practical reality that centralized processing cannot keep pace with the volume of decisions that modern operations require.

This migration creates a new battle that most enterprise leaders have not fully registered. Telcos own the last mile. The physical infrastructure connecting edge devices to any network runs through their equipment. That gives them a structural claim on data access that no cloud provider can replicate through software. A cloud provider can build more efficient compute, more capable models, better developer tooling — but it cannot put fiber in the

ground faster than a telco already has it there. The last-mile advantage is real and it is durable.

The telcos, however, are largely squandering it. The dominant strategic conversation inside most major telcos is about converting capital expenditure to operational expenditure — about monetizing existing infrastructure rather than extending capability. That is a defensive posture in a moment that requires an aggressive one. The organizations that understand the edge as a governance and intelligence layer, not merely a connectivity layer, will control something far more valuable than bandwidth. They will control the point at which data becomes actionable inference, before it ever reaches a central system. That is the architecture decision that cloud economics allowed organizations to defer for a decade. AI forces it now.

The implications for enterprise IT are structural and irreversible. The question of where compute lives is no longer a cost optimization question. It is an authority question. Whoever controls inference at the edge controls the decision surface of the organization. If that control is ceded to a vendor default, or inherited from a connectivity provider whose strategic ambition stops at the billing layer, the enterprise has effectively outsourced its judgment architecture to a party whose interests are not aligned with its outcomes.

This is the break from previous technology waves. ERP required organizations to redesign processes. Cloud required organizations to redesign infrastructure economics. AI requires organizations to redesign where authority lives and how it is governed — and to do so in an environment where the compute is distributed, the models are probabilistic, and the data access battle is being fought by parties with very different strategic intentions.

The organizations that understand this are not waiting for a governance framework from a standards body or a compliance requirement from a regulator. They are making architecture decisions now — about where intelligence lives, who controls data access, and whether their governance model was built for a world in which the decision surface extends to every edge device on their network. The organizations that do not understand this are managing the upgrade they think they are getting while the structural rebuild happens underneath them.

The Machinery of Managed Irrelevance: How Organizations Break at the AI Transition

The failure does not announce itself. It arrives gradually, through a series of decisions that each appear reasonable in isolation and are collectively catastrophic.

The first mechanism is authority migration without accountability transfer. When an AI system is deployed into a live workflow, it begins making or heavily shaping decisions immediately. The humans who previously owned those decisions do not disappear — they shift from decision-makers to reviewers. But the governance framework does not update to reflect this. The accountability map still shows the human as responsible. The authority, in practice, has moved to the system. This gap — authority in one place, accountability in another — is the primary condition for governance failure. It means that when something

goes wrong, there is no clear owner, no traceable decision path, and no mechanism for the organization to learn from the error.

The IBM research cited earlier puts a precise number on this: two-thirds of CIOs report accountability without control as AI deployment scales. This is not a technology immaturity problem. It is a structural problem created by deploying AI into governance frameworks that were designed for a different kind of system. Deterministic software does what it is programmed to do. The governance model for deterministic software is auditable, traceable, and binary — the system either complied with its specifications or it did not. Probabilistic systems do not comply with specifications. They produce outputs within probability distributions, and those outputs vary with context, data quality, and model state in ways that no audit trail designed for deterministic software can capture.

The second mechanism is inherited defaults. Every AI deployment arrives with a set of defaults — which model is called, what confidence threshold triggers review, what output format is surfaced, how conflicting signals are resolved. These defaults were set by a vendor, for a general market, optimized for general performance. They were not optimized for the deploying organization's risk tolerance, decision context, or operational culture. In most deployments, these defaults are never examined. They are inherited as neutral technical settings and treated as such until they produce an outcome that forces a reckoning. By then, the defaults have been operating inside live workflows long enough that unwinding them is operationally painful.

The third mechanism is the absence of memory architecture. Most operational AI systems are stateless. They process the query in front of them and discard the context. Corrections made after a bad output are not fed back into the system's operating parameters. Patterns of error — the same type of wrong answer surfacing repeatedly in the same type of context — are not recognized because there is no mechanism to retain and compare outputs over time. This means every operational session begins from the same baseline, regardless of how many errors have been made and corrected. The organization accumulates operational experience but the system does not. The gap between them widens over time, and the system's outputs become progressively less calibrated to the operational reality the organization is actually navigating.

The fourth mechanism is the false confidence problem, which operates at a subtler level than hallucination and is more dangerous precisely because it is harder to detect. A hallucination — a factual error, a fabricated reference, a nonsensical output — is usually visible. A confident, coherent, plausible answer that is structurally wrong because its underlying assumptions are unexamined is not visible. It looks like a good answer. It moves through review processes that are designed to catch obvious errors, not to surface hidden assumptions. It triggers action. The action produces a bad outcome. The causal chain between the unexamined assumption and the bad outcome is difficult to trace because the output that initiated it appeared well-reasoned.

These four mechanisms — authority migration without accountability transfer, inherited defaults, absent memory architecture, and false confidence — are not independent failure modes. They compound. An organization operating with inherited defaults, no memory

architecture, and a governance framework that cannot locate authority has built the conditions for false confidence to move through its decision structures unchecked. This is not a description of an edge case. Based on the research available, it describes the majority of enterprise AI deployments at scale today.

The organizations that recognize this pattern have stopped asking why their AI outputs are occasionally wrong. They are asking why their surrounding systems were never designed to catch wrongness in the first place.

The Redistribution of Judgment: Why AI Amplifies What Organizations Have Not Built

Understanding why AI amplifies structural drag requires being precise about what AI actually does to decision authority — not what the deployment documentation says it does, but what happens operationally when probabilistic systems run inside deterministic governance frameworks.

Traditional software executes rules. The governance model for rule-executing software is compliance-based: does the system follow the rules it was given? The audit trail is the configuration. The accountability is the configuration owner. This model is clean, traceable, and scalable because the software's behavior is fully determined by its inputs and specifications.

AI inference does not execute rules. It produces distributions of likely outputs conditioned on training data, context, and model architecture. The same input, in slightly different framing, can produce meaningfully different outputs. The same output can be produced by very different internal reasoning paths. The system's "behavior" is not a property of its configuration — it is an emergent property of the interaction between its training, its inputs, and its current operational context. A governance framework designed for the first kind of system will systematically fail to govern the second kind, not because the framework is poorly designed but because it is designed for a different problem.

This is where AI specifically amplifies structural drag. Organizations that have underinvested in governance — and most large organizations have, because cloud economics made it possible to defer architecture decisions by buying more compute — discover that their governance deficits are suddenly consequential. A compliance framework that never had to handle probabilistic outputs, an accountability map that assumed deterministic decision ownership, an audit trail that was never designed to trace the path from a training distribution to an operational recommendation — these are not theoretical deficiencies. They become operational failures the moment AI inference touches live workflows.

The edge deployment dimension makes this more acute. When inference runs centrally, there is at least a single point of governance — however inadequate — through which outputs must pass. When inference runs at the edge, on domain-specific models at the site and device layer, that single point disappears. Decisions are being made — or shaped — at hundreds or thousands of points simultaneously, each operating with its own model state, its

own data context, and its own inherited defaults. The governance framework that struggled to handle one centralized AI system is now being asked to govern a distributed network of judgment points it cannot see clearly.

The redistribution of decision authority is the precise mechanism by which this amplification works. AI does not add a new decision-making layer on top of existing authority structures. It inserts itself into existing decision flows, at the point where human judgment previously operated, and produces outputs that are treated as recommendations but function as decisions under operational pressure. The speed advantage of AI — the reason it is deployed — means there is rarely time to reconstruct the reasoning behind an output before acting on it. The human reviewer is approving a conclusion, not evaluating a reasoning process.

BCG's recent work on agentic AI puts this directly: autonomous AI systems require fundamentally different risk management architectures than traditional data governance. The word "fundamentally" is doing real work there. It is not a call for incremental adjustment to existing frameworks. It is an acknowledgment that the class of risk has changed, not just the magnitude.

The organizations that are treating AI as a software upgrade — deploying it into existing workflows, relying on existing governance frameworks, accepting vendor defaults as neutral settings — are not making a cautious choice. They are making a compounding one. Each deployment that runs without a cognitive governance layer is a deployment that builds operational dependency on a system whose assumptions have never been challenged, whose errors have never been retained, and whose authority has never been formally earned. The structural drag accumulates. The cost of unwinding it increases. And the distance between the organization's governance capability and its operational AI exposure grows with every workflow that gets touched.

The organizations that understand this are not moving slower. They are building differently.

Where the Decisions Actually Live: The Architecture Requirements That Cannot Be Deferred

Cognitive governance is not a policy document. It is not a compliance checklist. It is not a set of guardrails bolted onto an existing AI deployment after the fact. It is an architectural requirement — a set of structural decisions about where intelligence lives, who controls data access, how confidence is challenged, and how authority expands — that must be made before operational deployment, not after the first significant failure.

The first structural requirement is locating intelligence deliberately. The question of whether AI inference runs at the edge, in a regional node, or in a central system is not a cost optimization question. It is a governance question. Edge inference gives the organization speed and resilience — the system can operate when connectivity is degraded, can process sensitive data without transmitting it to a central server, and can respond to local context with local latency. Central inference gives the organization visibility, control, and the ability to update models globally without touching distributed deployments. Neither is inherently

superior. The requirement is to decide deliberately, based on what the organization actually needs to govern, rather than inheriting the deployment architecture a vendor recommends for general market conditions.

The data access control dimension of this decision is where the telco-cloud battle becomes an enterprise architecture question. The organization that deploys AI at the edge without resolving who controls data access at the last mile has handed a structural governance lever to whichever connectivity provider owns the physical infrastructure. This is not theoretical. Data that flows through a network layer the enterprise does not control is data whose access, retention, and transmission can be shaped by parties whose interests diverge from the enterprise's. The last-mile advantage that telcos have built — and largely failed to extend into intelligence services — represents a governance risk for any enterprise that deploys AI at the edge without a clear data access agreement that addresses inference, not just connectivity.

The second structural requirement is designing the behavioral defaults rather than inheriting them. MIT Sloan Management Review's work on adaptive governance makes this concrete: AI governance must adapt across system life cycles to manage risks at organizational scale. The lifecycle framing is important. The governance architecture that is appropriate at initial deployment — high human oversight, narrow autonomy, conservative confidence thresholds — is not the same architecture appropriate for a system that has demonstrated reliability across thousands of operational decisions. But the adaptation must be intentional and evidence-based, not the result of gradually lowering oversight because nothing bad has happened recently.

This is where the analogy to deliberate practice is structural rather than metaphorical. Deliberate practice is not repetition. It is structured repetition with feedback that modifies performance. A system — human or artificial — that repeats actions without incorporating feedback does not improve. It habituates. The governance architecture for operational AI must include mechanisms for capturing errors, analyzing their patterns, and feeding corrections back into the system's operating parameters. Without this, every deployment remains a first deployment, regardless of how long it has been running.

The third structural requirement is calibrating confidence before it triggers action. The three questions that must precede any AI-generated action that carries material consequences are not optional. What assumptions drove this output? What evidence supports them? What is the cost if the confidence is wrong? These are not questions a human reviewer asks in an ad hoc review. They are questions that must be embedded as structural checkpoints in the workflow architecture — the equivalent of the pause that mindfulness training creates between stimulus and response. The mechanism must be structural because operational pressure will always push against the pause. Without a structural requirement to pause, the review becomes a formality, and the false confidence problem moves through unchecked.

The authority structure is the fourth requirement. The governance framework must formally distinguish between AI operating in recommendation mode, supervised execution mode, and autonomous execution mode — and the criteria for transitioning between modes must be defined in advance, based on observed performance, not theoretical capability. This mirrors

the graduated trust model that any serious organization applies to human judgment in high-stakes roles. A new analyst does not approve major transactions on day one. The authority expands as the judgment is demonstrated. The same principle, applied structurally to AI, is the difference between a governance architecture and a governance aspiration.

What Must Be Built Differently: Operational Prescriptions for Cognitive Governance

The structural requirements outlined above have operational translations. These are not transformation programs. They are specific decisions that organizations can make now, with existing technology, that determine whether cognitive governance becomes operational architecture or remains aspirational language.

The first decision is auditing defaults before changing anything else. Before any new AI capability is deployed, every inherited vendor default in existing deployments should be reviewed against the organization's actual risk tolerance and outcome requirements. Which model is being called and why? What confidence threshold triggers human review, and was that threshold set for this organization's context or for a general market? What output is being surfaced, and does that surfacing decision reflect deliberate design or vendor convenience? The organizations that have done this work consistently discover defaults that were never designed for their context and have been shaping outputs — and therefore decisions — for months or years. The audit is not a one-time exercise. According to research on enterprise AI visibility, AI risk concentrates among power users, and usage patterns shift faster than governance frameworks update. The default audit must be a standing discipline, not a deployment checklist.

The second decision is building memory architecture into every operational AI system. This does not require exotic technology. It requires discipline about what gets captured, retained, and fed back. Every correction made to an AI output — whether by a human reviewer, a downstream process, or an outcome measurement — is data. That data, properly structured and retained, allows the organization to identify patterns of error, calibrate confidence thresholds against actual performance, and prevent the same cognitive errors from recurring across sessions. The absence of this architecture is not a technology limitation. It is an organizational choice, usually made by not making it. The organizations that build it stop running endless first sessions. They build systems that compound.

The third decision is separating confidence expression from confidence warrant in every high-stakes workflow. The AI output that arrives in a decision-maker's review queue should carry, alongside its substantive content, an explicit account of its underlying assumptions and the evidence supporting them. This is not a prompt engineering solution. It is a workflow design decision. The output format must be designed to surface the reasoning structure, not merely the conclusion. This creates the structural condition for a reviewer to challenge the assumptions rather than simply approving or rejecting the output. It also creates the data needed to build memory architecture — the captured reasoning provides a traceable path from assumption to outcome that can be analyzed when the outcome is wrong.

The fourth decision is implementing a formal autonomy progression model. Every AI system operating in a live workflow should be explicitly classified in one of three modes: recommendation, supervised execution, or autonomous execution. The classification should be visible to every human whose workflow the system touches. The criteria for progression between modes should be defined in advance and documented. Progression should require demonstrated reliability across a defined number of operational decisions, not elapsed time or vendor certification. This model, applied consistently, makes the governance framework legible to the people operating inside it — which is a prerequisite for holding them accountable when governance fails.

The fifth decision is resolving data access governance before edge deployment, not after. Any AI deployment that runs inference at the edge — on devices, at sites, inside operational technology — requires a clear, legally documented account of who controls data access at the last-mile layer, what the connectivity provider's rights are with respect to data transmitted across their infrastructure, and what happens to inference outputs when connectivity is degraded. This is not a standard connectivity contract question. Most standard connectivity contracts were written before edge inference was a realistic deployment scenario and do not address it. The organizations that discover this after deployment are in a significantly weaker negotiating position than the ones that resolve it before go-live.

The sixth decision is building the governance review cadence around the AI system's operational life cycle rather than the organization's standard IT review schedule. The MIT Sloan research on adaptive governance is direct on this: governance must adapt across system life cycles. An AI system's risk profile changes as its deployment context changes — as new data flows through it, as usage patterns shift, as edge deployments multiply. A governance framework reviewed annually against a static risk assessment is not adaptive governance. It is compliance theater. The review cadence should be tied to operational triggers — a defined volume of decisions, a defined pattern of corrections, a defined change in deployment scope — rather than the calendar.

The Human System That AI Governance Must Not Forget

There is a dimension of cognitive governance that the preceding sections have addressed structurally but not organizationally: the human system through which AI decisions are reviewed, approved, and acted upon. The governance architecture for AI is not only the architecture surrounding the model. It is the architecture surrounding the people who interact with the model's outputs. These are not the same design problem, and conflating them produces governance frameworks that are technically sound and operationally unworkable.

The core issue is what behavioral research on choice architecture established decades ago: the way options are presented determines which options get selected, before any deliberate reasoning occurs. A human reviewer presented with an AI output that is formatted as a conclusion, expressed with confidence, and placed at the end of a long recommendation

chain will, under operational pressure, approve it. Not because the reviewer is incompetent or negligent. Because the presentation architecture has already shaped the cognitive frame. The conclusion is the default. Challenging it requires energy the operational context does not reliably supply.

This is why cognitive governance must address the human behavioral layer, not just the AI technical layer. The review interfaces through which humans interact with AI outputs are behavioral architecture. They shape cognition before the reviewer has read a word of the output. A review interface designed to surface assumptions alongside conclusions, to present uncertainty as information rather than weakness, and to require explicit acknowledgment of the cost of being wrong — that interface is doing governance work that no policy document can replicate. The organizations that understand this design their review interfaces as carefully as they design their models.

The authority-accountability mismatch surfaces here with particular clarity. Research from IMD on AI and human capital leadership identifies the authority-accountability gap as a central governance challenge — the CHRO, or any senior leader nominally accountable for AI-related people decisions, often has neither the technical visibility nor the operational authority to regulate what AI systems are actually doing inside the workflows they nominally own. This is not an HR problem. It is a governance design problem. The accountability map must be aligned to the actual decision flow, not the organizational chart. If the AI system is making the substantive decision and the human is reviewing the conclusion, the human's accountability must be defined in terms of the review quality, not the outcome — because the human's actual leverage is on the review, not the underlying decision.

The concentration of AI risk among power users — identified in research on enterprise AI usage patterns — adds another human-layer complication. In most organizations, a small group of sophisticated users drives the majority of AI interactions and influences the majority of consequential AI outputs. These users typically operate with greater autonomy, use more advanced prompting techniques, and interact with higher-stakes workflows than the average user. They are also the least likely to be covered by governance frameworks designed for average-case deployments. The cognitive governance architecture must be designed for the actual risk distribution, not the median user. This means that the most consequential oversight mechanisms should be applied where the most consequential decisions are being made — among the power users, in the high-stakes workflows, at the edge of organizational authority.

The behavioral architecture of cognitive governance is ultimately about making it structurally easier to do the right thing than to default to the convenient thing. That is what CBT does for a person operating under cognitive distortion. That is what well-designed choice architecture does for a decision-maker under time pressure. And that is what cognitive governance must do for the human-AI system operating under the compounding pressures of scale, speed, and probabilistic uncertainty.

Conclusion

The argument across this paper reduces to a single structural observation: organizations that have not built cognitive governance infrastructure are not deploying AI at scale — they are accumulating AI-shaped operational risk without the structural capacity to detect, contain, or learn from it. The two look identical from the outside, and for a while they produce similar outputs. They diverge when something goes wrong, when scale amplifies a small systematic error into a large operational failure, or when a competitor that built the governance layer uses it as an engine for compounding improvement rather than just risk reduction.

The competing technology arguments — edge versus central, telco versus cloud, domain-specific model versus mega-LLM — are real and consequential. The battle for data access at the last mile will shape enterprise AI architecture for the next decade. The migration of inference to the device and site layer is restructuring what organizational decision-making means at a physical level. These are not background conditions. They are the structural context within which cognitive governance must be designed.

But the organizations that get this right will not have succeeded because they won the infrastructure argument. They will have succeeded because they built a system that learns. Not a system that generates more answers, faster — any vendor can provide that. A system that challenges its own assumptions, retains its own errors, calibrates its own confidence, and expands its own authority only as fast as governance can absorb it. That is the compound asset. It is not replicable by an organization that treated governance as an afterthought, because the learning requires the architecture to have been in place when the errors occurred, not after they were discovered.

The organizations that cannot quickly replicate this are not the ones that lack AI technology. They are the ones that deployed AI at scale without building the surrounding system — the ones whose defaults were inherited, whose memory architecture was never built, whose confidence was never structurally challenged, and whose authority maps never caught up to where the decisions were actually being made.

The specific capability that the organizations that get this right will have built is a closed feedback loop between AI operational performance and AI governance calibration — a mechanism through which every error informs the system's next configuration, every demonstrated reliability earns expanded authority, and every assumption surfaced becomes a test against evidence. That is not a feature. It is an organizational capacity built over time, on decisions that were made deliberately, before the failures made them unavoidable.

BDG Advisory | bdgadvisory.com/perspectives/cognitive-governance