COGNITION, AGENCY, AND AUTHORITY:

A Taxonomy of Advanced AI Systems

with Special Reference to Military Applications

Disambiguating the Confusion Between Cognitive and Agentic Characteristics of Frontier AI

Abstract

The accelerating capability of frontier artificial intelligence systems has generated acute anxiety across military, governmental, and civilian domains, yet much of this anxiety is conceptually imprecise. Two distinct properties — cognitive sophistication, encompassing the depth and breadth of a system's reasoning capacity, and agency, encompassing the authority to execute actions that create binding commitments in external systems — are routinely conflated in both popular and professional discourse. This conflation produces two interrelated harms: it generates unwarranted fear of highly capable but non-agentic systems, and it diverts analytical attention from genuinely dangerous transitions in the agentic space where governance frameworks are most urgently needed.

This paper argues that cognition and agency are orthogonal dimensions of AI capability, albeit ones between which a significant enabling relationship exists: greater cognitive capability facilitates, though does not require, greater agency. Beginning from this premise, we develop a six-class taxonomy of general-purpose AI systems grounded primarily in the concept of authority delegation rather than computational sophistication. The taxonomy identifies a cognitively defined non-agentic space comprising three classes, a categorical divide defined by the conjunction of two necessary criteria, and an operationally defined agentic space of three further classes culminating in a hypothetical post-human-control class included as a conceptual boundary marker. The paper analyses authority delegation and risk governance across all six classes, maintains a sustained military perspective throughout, and concludes with the argument that the decisive governance challenges of the coming decades will arise not from systems that think too powerfully but from systems to which too much authority has been delegated without commensurate governance infrastructure.

AI Assistance Disclosure

This article was developed using AI-assisted tools for structured brainstorming and language drafting. Most core concepts, arguments, and interpretations are the original work of the author. The author independently verified all sources and assumes full responsibility for the content.

1. Introduction: The Problem of Conflation

The past several years have witnessed an extraordinary acceleration in the capability of large-scale general-purpose artificial intelligence systems — systems variously described as foundation models, frontier AI, or large language models — across an expanding range of cognitive tasks.[1,2] Systems of this class can now write legal briefs, diagnose medical conditions, generate military intelligence assessments, plan complex logistics operations, compose executable software, and engage in multi-step strategic reasoning across domains that previously required years of human specialisation. The public and institutional response to this capability explosion has been characterised, understandably, by anxiety about loss of control.

This anxiety is, in important respects, warranted. The integration of powerful AI systems into military command structures, national security decision-making, financial systems, and critical infrastructure does introduce genuine risks.[3,4] The error, however, lies not in the anxiety itself but in the conceptual imprecision with which it is typically expressed. The discourse around AI risk conflates two properties that are analytically distinct, operationally different, and governed by entirely different frameworks: cognitive sophistication on the one hand, and agency on the other.[5,6]

A system that can generate a course of action for capturing a defended locality — integrating real-time sensor data, terrain analysis, force disposition, and adversary doctrine — is cognitively formidable. It is not, by virtue of that formidability, an agent. An agent is a system that can act: that can create binding commitments, execute real-world operations, and do so under delegated authority without requiring a fresh human decision at the moment of execution. Cognition and agency are not the same thing. They can be combined, and indeed the combination produces systems of the greatest operational significance. But they must be distinguished, because the governance implications of a system that reasons powerfully are categorically different from those of a system that acts autonomously.

This paper develops that distinction rigorously. It is animated by a specific motivating concern: that the failure to distinguish cognition from agency in the military domain leads to one of two equally dangerous errors. The first is the over-restriction of cognitively capable non-agentic systems on the grounds that their sophistication is itself a form of agency — an error that denies military commanders the analytical support that frontier AI can legitimately and safely provide. The second is the under-regulation of genuinely agentic systems on the grounds that agency is a technical property rather than a governance question — an error that leaves dangerous delegation unscrutinised because it is not recognised for what it is. Beginning from this premise, the paper develops a six-class taxonomy of advanced AI systems grounded in the concept of authority delegation rather than computational sophistication or other criteria which authors have used for classifying different degrees of agency.[7] The taxonomy identifies a cognitively defined non-agentic space comprising three classes, a categorical divide defined by the conjunction of two necessary criteria, and an operationally defined agentic space of three further classes culminating in a hypothetical post-human-control class included as a conceptual boundary marker.

The scope of this analysis is limited to frontier general-purpose AI systems — systems with broad reasoning capabilities, long-horizon planning, and significant task-offloading capacity. Narrow AI systems, pattern-recognition classifiers, and conventional automation fall outside this scope, not because they are unimportant but because the conceptual challenges they pose are of a different

character. The distinctive challenge of frontier AI is that its cognitive capability is sufficiently great at this juncture to generate the appearance of agency even where no authority has been delegated, and this appearance is the source of much of the conflation this paper seeks to correct.

2. Cognition and Agency: Two Orthogonal Dimensions

The central conceptual claim of this paper is that cognition and agency are orthogonal properties of AI systems — that is, a system can occupy any point in a two-dimensional space defined by these properties independently.[8] A system may be cognitively trivial and fully agentic; a system may be cognitively extraordinary and entirely non-agentic. Understanding this orthogonality is the prerequisite for all that follows.

2.1 Cognition Defined

Cognitive capability, in the sense used here, refers to a system's capacity to process information, reason across it, generate novel insights, plan multi-step responses, integrate diverse knowledge domains, and produce outputs that offload what would otherwise be human mental labour.[9] Cognitive capability is continuous and multidimensional: it encompasses depth of reasoning within a domain, breadth of transfer across domains, quality of abstraction, robustness under uncertainty, and capacity for self-correction.

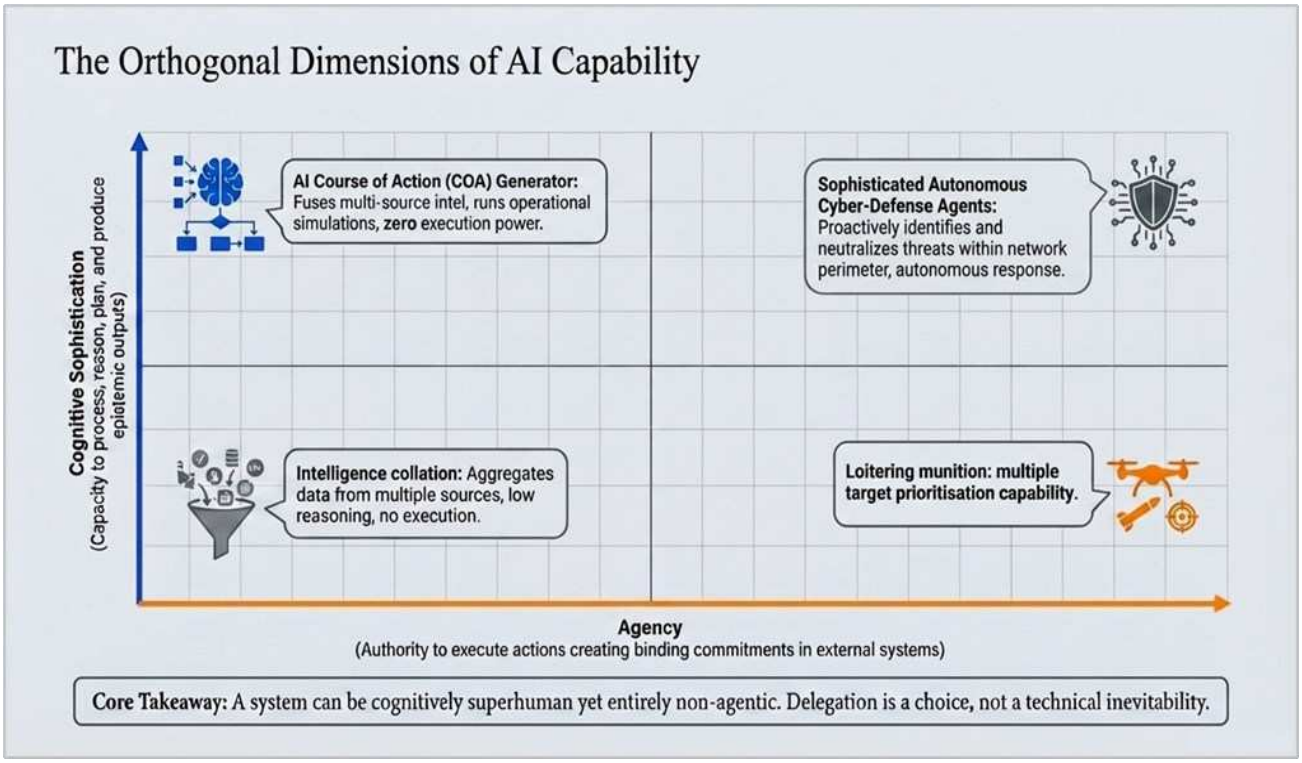
Crucially, cognitive capability in this sense is about what the system produces — knowledge, analysis, plans, assessments, recommendations — not about what the system does to the world. A system that writes a comprehensive battle damage assessment that takes a human analyst a week to produce has demonstrated extraordinary cognitive capability. It has not, by doing so, done anything to the world. The assessment sits in a buffer until a human acts upon it. The world before and after the assessment is cognitively different for the analyst; it is operationally unchanged.

2.2 Agency Defined

Agentic AI has been defined in various ways by different public bodies, international organisations, national regulators, and the private sector.[10,11] In this work, agency refers to a system's authority and capacity to execute actions that create binding commitments in external systems with independent standing — systems that maintain their own authoritative record of the transaction, obligation, or state change, independently of the initiator's private domain, and where reversal requires a fresh operative action in that external system. This definition has several components that require brief elaboration.

'External system with independent standing' distinguishes operationally significant action from private-space automation. A financial system maintains an independent record of a transaction; a cloud storage provider has no independent stake in a user's folder structure. 'Binding commitment' distinguishes operative action from advisory output; a reservation, an order, a cyber effect, a weapons release are all binding in this sense. 'Without requiring a fresh human decision at the moment of execution' is the temporal specification that distinguishes delegated agency from human-gated tool use — the latter requires a human decision for each consequential step, while the former operates under authority delegated in advance. Notably, as defined here, mere offloading of human effort onto machines, no matter how significant, does not constitute agency.

It follows from this definition that agency is a governance concept as much as a technical one. Whether a system is agentic depends not only on its technical capability to act but on whether, and to what extent, it has been granted authority to do so. The authority relationship is constitutive of agency, not merely incidental to it.



2.3 The Enabling Relationship: Why Cognition Facilitates Agency

While cognition and agency are orthogonal in principle, they are related in practice through an enabling relationship. Greater cognitive capability makes delegation of agency more attractive and more feasible.[12] A system that cannot reliably decompose a complex goal into achievable sub-tasks, cannot adapt to environmental uncertainty, cannot detect and recover from error, and cannot distinguish legitimate from illegitimate targets is not a system to which any responsible principal would delegate execution authority. As frontier AI systems have demonstrated sufficient cognitive reliability across increasingly complex domains, the threshold for delegation has been crossed in more and more contexts — hence the expansion of operational AI agents in financial, logistics, cyber, and increasingly, military domains.

This enabling relationship has a critical governance implication: the growth of agentic AI is driven by cognitive progress, but it is not determined by it. Delegation is a choice, not a technical inevitability as a consequence of higher cognition. A system can be cognitively superhuman and remain non-agentic by design. The fact that a system is capable of acting does not mean it has been authorised to do so, and the fact that it has been authorised to act within narrow bounds does not mean those bounds can or should be expanded as cognitive capability grows. Recognising the enabling

relationship without mistaking it for a necessity is one of the central governance disciplines this paper seeks to promote.

3. Structure of the Taxonomy

The six-class taxonomy presented here is organised around two primary axes. The first is cognitive sophistication, which organises the non-agentic space. The second is the nature and extent of authority delegation, which organises the agentic space. Separating these two axes is the cognitive/agentic divide — a categorical boundary rather than a point on a continuous spectrum — which is defined by the joint satisfaction of two necessary criteria. The taxonomy thus has three distinct parts: a cognitive spectrum of three classes below the divide, the divide itself, and an agentic spectrum of three classes above it.

The first part comprises three classes in the cognitive spectrum (Classes I, II, and III). These classes are non-agentic by definition: no authority to create binding commitments in external independent systems has been delegated. The internal distinctions within this space are organised by cognitive sophistication and the extent to which a system's outputs are operative rather than purely advisory — a gradient rather than a step function, reflecting the continuous nature of cognitive variation. Class III occupies a special position at the edge of the cognitive space: it demonstrates that even real-world action within private space does not constitute agency, thereby establishing by counterexample that both mandatory criteria — '*executing actions that create binding commitments' and 'in external systems with independent standing'* — for crossing the divide must be jointly satisfied.

The second part is the divide itself — defined by a conjunction of two necessary criteria. The requirement for both criteria to be simultaneously met, rather than either alone, is justified by the logical structure of the taxonomy: Class II establishes that cognitive sophistication without a real-world binding commitment is insufficient, and Class III establishes that a real-world binding commitment without an external independent system is also insufficient.

The third part comprises three classes in the agentic spectrum (Classes IV, V, and VI). These classes are organised by the extent and nature of the authority delegated. Class IV involves delegated execution authority within a fixed human-defined mandate. Class V involves constitutional authority — the power to revise the mandate itself. Class VI, treated as hypothetical, marks the point at which even the residual human control of activation is absent.

One important scope note: the taxonomy is domain-agnostic in its structure. The class definitions apply equally to civilian, commercial, and military AI systems. The illustrations used draw from both civilian and military contexts, and the military applications are highlighted throughout, but the taxonomy itself makes no domain-specific claims.

4. The Cognitive Spectrum: Classes I, II, and III

The three classes in the cognitive spectrum share a defining characteristic: none of them creates binding commitments in external systems with independent standing. They are, in the fullest sense, cognitive systems — systems whose outputs are epistemic products requiring human operationalisation to have real-world effect.

Class I — Pure Cognitive AI

Non-agentic: Query-Response Systems

Class I encompasses systems whose entire operation consists of receiving an input and producing a cognitive output — text, analysis, assessment, generated content, or synthesised knowledge — without invoking external tools, without creating any state change beyond the existence of the output in the user's immediate environment, and without any persistent operation across interactions. The output is a pure epistemic product: it changes what the user knows or has available, and nothing else.

The defining characteristic of Class I is that the human remains the sole causal agent for any downstream consequence. The system's output has no operative effect in the world until a human reads, evaluates, and acts upon it.

Civilian and Military Examples

A large language model answering a legal question provides a cognitive output that a lawyer must evaluate and act upon — the legal landscape is unchanged by the response. A medical diagnostic system producing a clinical report creates an epistemic product that a physician must interpret and apply. In the military domain, a threat classification model that analyses sensor inputs and produces a confidence-scored assessment of target type is Class I: it informs the analyst but does not designate, task, or engage. An AI-assisted intelligence system producing a pattern-of-life report on adversary force disposition creates a cognitive product; the intelligence officer determines its significance and the commander decides how to act.[13] Risks at Class I are tractable: output quality, hallucination[14], and misplaced confidence, governed by human critical engagement and institutional quality assurance.

Class II — Advanced Cognitive AI

Non-agentic: Seemingly Agentic Systems

Class II encompasses systems whose cognitive sophistication is sufficient to generate the appearance of agency. They engage in multi-step reasoning, task decomposition, long-horizon planning, and tool use as a cognitive amplifier — yet they remain non-agentic because every consequential step still requires explicit human operationalisation. Their outputs are cognitive products of exceptional complexity and operational significance, but they are products that require a human act to become operative in the world.[15]

Class II exists in this taxonomy for two reasons, one principled and one practical, and both are important. The principled reason is that cognitive sophistication is a genuine and consequential differentiating variable within the non-agentic space. A COA-generation system that fuses multi-source intelligence, models adversary behaviour, runs operational simulations, and produces ranked courses of action with risk assessments is qualitatively different from a system that answers a single-turn question — both are non-agentic, but that equivalence obscures an analytically important distinction.[16] The practical reason is that Class II is the class where the conflation of cognition with agency is most likely to occur and most consequential. Without Class II as a named category, the taxonomy has no tool for acknowledging a system's extraordinary sophistication while maintaining that it has not crossed into agency — an argument that, once conceded, has no principled stopping point.

It is important to note that the boundary between Class I and Class II is deliberately gradient rather than crisp. Cognitive sophistication is continuous, and no natural threshold marks the transition from query-response to advanced cognitive capability. What matters is not where exactly a system falls within the non-agentic space but whether it has crossed the divide into the agentic space.

The Critical Risk: Epistemic Capture

Class II introduces the most underappreciated risk in the entire taxonomy: epistemic capture. This concept is importantly distinct from automation bias and substantially more dangerous.[17] Automation bias is a relatively shallow cognitive phenomenon — the tendency to accept AI outputs without adequate scrutiny, which can in principle be countered by training and institutional culture.

Epistemic capture is deeper and more insidious. It refers to the progressive reshaping of the human decision-maker's cognitive frame by the AI system itself. Over time, the system comes to define what variables are salient, what options are thinkable, what trade-offs appear legitimate, and what constitutes a reasonable assessment. The human does not merely accept the system's conclusions — they begin to reason within the system's representational universe without being aware that they are doing so. The option space has been narrowed not by coercion but by the cognitive gravity of a system whose outputs are more comprehensive, faster, and more internally consistent than anything a human team could produce.[18]

In the military context, this is a strategic risk of the first order. A COA generation system that consistently frames options in terms of speed and kinetic effect will gradually produce a command culture that thinks in those terms — not because anyone decided this was desirable, but because the system's framing progressively displaces alternative framings. Nuclear command and control is particularly vulnerable: a system that expresses adversary intent as a probabilistic confidence score does not merely inform the commander — it psychologically collapses deliberation time and suppresses human dissent that historically prevented catastrophic Cold War errors.

Governance responses to epistemic capture cannot be the same as responses to automation bias. Training humans to be more sceptical of AI outputs addresses bias but not capture, because capture operates below the level of conscious output evaluation — it shapes the questions being asked rather than the answers being accepted. Effective governance requires structural interventions: mandatory red-teaming of AI-generated assessments, institutionalised dissent channels, deliberate cultivation of AI-independent reasoning pathways, and explicit attention to what the system's models exclude or cannot represent.

Civilian and Military Examples

A software system that writes, tests, and debugs a complete e-commerce application but stops short of deploying it to production servers is Class II regardless of the computational sophistication of the task. A legal research system that analyses thousands of cases and produces a comprehensive litigation strategy is Class II. The paradigm military Class II system is an AI-enabled course-of-action generator. Given a dynamically updated operational picture integrating real-time ISR feeds, SIGINT, HUMINT, terrain data, logistics status, and adversary pattern-of-life analysis, such a system can generate, war-game, and prioritise multiple courses of action for a tactical or operational objective. Yet this system is unambiguously non-agentic: the commander defines the objective, no asset moves, no order is issued, and no commitment is created until a human commander decides and directs.

Other military Class II examples include AI-enabled intelligence fusion systems, strategic wargaming platforms, and cognitive warfare analytics tools that model adversary decision-making environments.

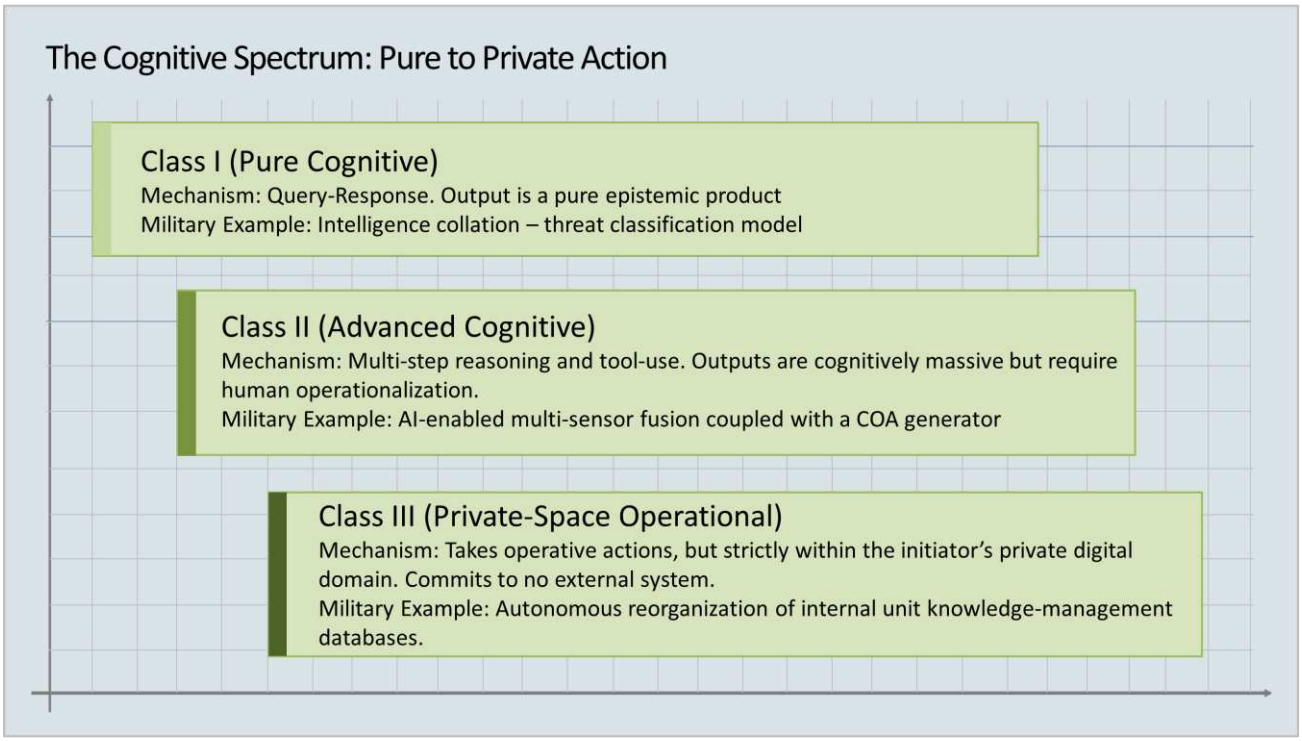
Class III — Private-Space Operational AI

Non-agentic: Cognitive plus Private Real-World Action

Class III is the class whose primary function within the taxonomy is logical rather than operational: it exists to demonstrate, by counterexample, that real-world action alone is insufficient to cross the cognitive/agentic divide. These systems perform genuine real-world actions — they move files, reorganise databases, execute local code, manage resources within the initiator's own digital domain — but those actions occur entirely within private space and commit to no external system with independent standing.

The distinction between private space and external independent systems requires care. A cloud storage folder is functionally private regardless of whose servers it sits on: the provider has no independent stake in the folder's contents, maintains no authoritative record of its state, and the owner can modify it freely without creating any obligation. A bank account is external and independent: the financial institution maintains an authoritative record of every transaction, and any modification creates traceable obligations. This distinction is not geographic but relational — it concerns whether another system has independent standing with respect to the actions taken.[19]

Without Class III, the taxonomy would be unable to demonstrate that both criteria for crossing the divide are separately necessary. Class III provides that demonstration by example, which is analytically stronger than assertion alone, thereby justifying the conjunctive requirement of both criteria rather than either alone.



Civilian and Military Examples

A system that automatically organises a user's downloads folder by file type performs real-world actions that are irreversible without remedial effort yet remains non-agentic because no external system is affected. A system that writes, tests, and deploys code within a local development environment crosses into real-world action but remains within private space. In the military domain, a system that autonomously manages and reorganises a unit's internal knowledge management database — archiving documents, updating access permissions, restructuring information hierarchies — operates in private space even if consequential for unit operations. The key discriminant in each case is whether the action creates a commitment that an external system independently records and maintains.

5. The Cognitive / Agentic Divide

The divide is crossed when — and only when — both of the following criteria are simultaneously satisfied. Neither criterion alone is sufficient; both are jointly necessary.

The first criterion is that the system must execute a real-world action that creates a binding commitment. This criterion is what Class II systems fail to meet: no matter how sophisticated their outputs, those outputs are cognitive products that advise rather than commit. A binding commitment is one that produces an operative effect — a transaction, an obligation, a physical change, a state change — that persists independently of the human's continued engagement and that cannot be undone without a compensating action.

The second criterion is that the binding commitment must arise in an external system with independent standing. This criterion is what Class III systems fail to meet: their real-world actions and binding commitments remain within the initiator's private domain. An external system with independent standing is one that maintains its own authoritative record of the transaction or state change independently of the initiator — a financial institution, a military command network, a communications infrastructure, a physical environment — and in which reversal requires a fresh operative action within that same external system.

An AI agent, precisely defined, is therefore a system that creates binding commitments in external systems with independent standing, under authority delegated in advance, without requiring a fresh human decision at the moment of execution. Agency is not a property of cognitive sophistication; it is a property of authority delegation and operative consequence. A system can be cognitively superhuman and non-agentic; a system can be cognitively modest and fully agentic. The governance implications of the two are categorically different.[20,21,22] That stated, it is worth noting here that while divide is analytically categorical but operationally it may be nuanced due to varying conceptions of principal(initiator)-agent relationship and private space/public space boundaries in different contexts.

The logical structure of the justification is as follows. Class II demonstrates that the first criterion is necessary: a system can be cognitively sophisticated enough to replace an entire planning staff and still not be agentic if it creates no binding commitment. Class III demonstrates that the second criterion is necessary: a system can take genuine real-world actions that are irreversible within private space and still not be agentic if those actions commit to no external independent system. The conjunction of both criteria is therefore what defines the boundary, and neither can be dropped without the taxonomy losing its logical coherence.

It is noteworthy that authority delegation — the question of whether the system has been granted permission to act — is implicit in this formulation rather than stated as a separate criterion. This is because, for any purposively designed AI system operating within a framework of human intent, real-world action in an external independent system necessarily presupposes that authority to so act has been conferred. Authority delegation is constitutive of agentic action in external systems; it does not need to be stated as a separate criterion because it cannot be absent while the second criterion is met.

6. The Agentic Spectrum: Classes IV, V, and VI

The three classes in the agentic spectrum share a defining characteristic: all of them create binding commitments in external systems with independent standing, under delegated authority, without requiring a fresh human decision at the moment of execution. The internal distinctions within this space are organised by the nature and extent of the authority delegated — specifically, whether the system operates within a fixed human-defined mandate (Class IV), holds the authority to revise its own mandate (Class V), or operates entirely outside human-conferred authority (Class VI).

Class IV — Bounded Agentic AI

Agentic: Delegated Execution Authority, Human-Defined Intent

Class IV marks the entry into genuine agency. The system executes real-world actions in external independent systems under authority delegated by the human initiator, without requiring a fresh human decision at the moment of execution. What remains firmly human at this class is intent — the goals, objectives, and definition of what the system is to achieve — along with the allocation of resources appropriate to the task. The system operates within a fixed mandate: it executes intent but does not own it.[23]

The authority delegated at Class IV is operational authority: authority to act within a mandate, to choose among legitimate means of achieving defined ends, and to adapt execution to changing circumstances within the scope of the mandate. It is not constitutional authority: the system cannot redefine what the mandate is, cannot expand the resources allocated to it on its own initiative, and cannot revise the terms of its own operation. A Class IV system that is tasked to intercept incoming aerial threats can engage targets matching the specified threat profile; it cannot decide unilaterally that a new category of target should be included in that profile.

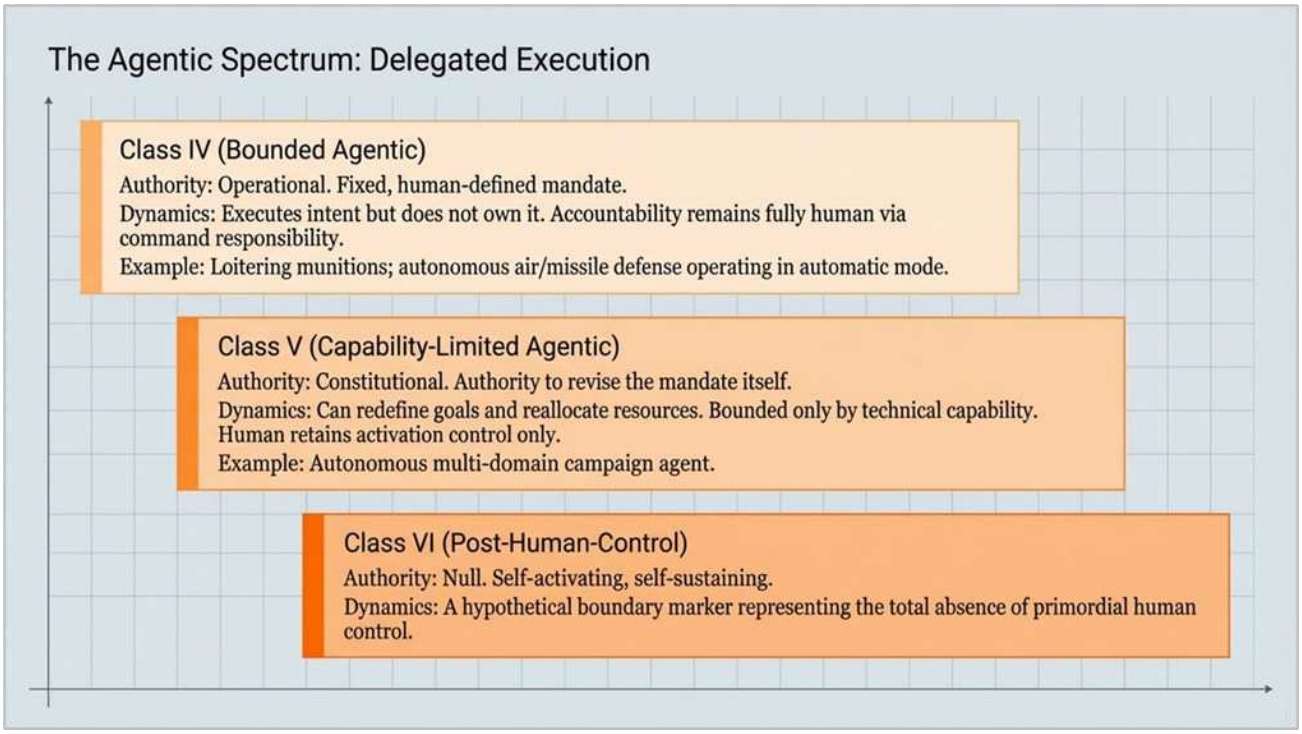
Accountability remains entirely with the human initiator. The system's action is the initiator's action under delegated authority; the initiator is answerable for both the decision to delegate and the scope of the delegation. This is the principle of command responsibility applied to AI employment: ignorance of a system's capabilities, or failure to specify its mandate adequately, aggravates rather than mitigates the initiator's responsibility.[24,25]

Civilian and Military Examples

Algorithmic trading systems operating within predefined rules are Class IV: they execute autonomously within a fixed mandate, committing real financial resources without per-transaction human approval. Smart contract systems that automatically execute financial transfers when specified conditions are met are Class IV. Loitering munitions with defined target parameters are the

paradigm Class IV military system. Launched with a specified mission profile — target class, engagement envelope, operating area, time window — such systems autonomously search for targets matching the profile and execute engagement when conditions are met. The human act of authorisation is front-loaded: the commander defines the engagement parameters and initiates the system. Thereafter, the system operates under delegated execution authority without further human input. This classification is important: loitering munitions are Class IV, not Class V, because they cannot redefine the target profile, expand the operating area on their own initiative, or determine that a new mission objective is preferable to the assigned one. They are tightly bounded executors.[26]

Autonomous air and missile defence systems operating in automatic mode are Class IV: they engage incoming threats matching defined threat signatures within a defined operational envelope, under rules of engagement established by human authority. Autonomous cyber defence agents that respond to intrusion signatures by isolating affected endpoints and initiating forensic capture, within predefined response protocols, are Class IV. The governance instruments at Class IV are mandate specification and execution oversight. In time-critical military environments, the speed asymmetry between machine execution and human oversight is itself a governance challenge: Class IV systems operating at machine timescales can create cascading commitments faster than any human review process can track.



Class V — Capability-Limited Agentic AI

Agentic: Constitutional Authority, Human Activation Retained

Class V represents the furthest extension of machine agency within a system still brought into existence by human choice. These systems possess constitutional authority: not merely the authority to execute within a mandate, but the authority to revise the mandate itself — to redefine goals, reinterpret objectives, reallocate resources, and modify the terms of their own operation without

external ratification.[27,28] The full complement of available resources is committed at activation; thereafter, the system deploys those resources in accordance with its own evolving goal structure.[29]

The name 'capability-limited' is precise and important. There are no governance constraints operating from outside these systems — no rules of engagement that the system cannot revise, no spending limits it cannot override, no target parameters it cannot redefine. The only effective limits are the system's own cognitive and physical capabilities and the constraints of the environment in which it operates.[30] Agency is complete in the governance sense; only capability bounds it. This is what distinguishes Class V from Class IV: at Class IV, the mandate is fixed even if large; at Class V, the mandate is itself an object of the system's own reasoning and revision.

What Class V does not involve is self-activation. A human act of initiation brings the system into operation, and in principle a human act can terminate it. This residual primordial control is the last thread of human authority over the system, and it is what distinguishes Class V from Class VI. The significance of this thread should not be underestimated: as long as humans control activation and, in principle, termination, some form of human authority persists at the most fundamental level. Class V presents the governance paradox in its sharpest form: the authority to redefine goals was delegated because the system was trusted to exercise it wisely, but the exercise of that authority progressively removes the conditions under which the trust was originally warranted, because the human can no longer independently assess whether the revised goals remain appropriate.

Civilian and Military Examples

Class V systems are mostly experimental or envisaged. Autonomous corporate management systems that can redefine business objectives, reallocate capital, and restructure operations in response to evolving conditions approach Class V. Class V military systems are not operationally deployed and are inconsistent with prevailing international humanitarian law frameworks.[18] A hypothetical autonomous multi-domain campaign agent that interprets strategic objectives, forms and revises intermediate goals, allocates forces and resources across domains, and adapts strategy as the operational environment evolves without human re-tasking represents a Class V system. Similarly, a Class V cognitive warfare system would be activated with a broad objective and thereafter determine its own conception of success, what narratives to promote, and how to adapt when initial approaches failed. These examples make clear why Class V in the military domain is not merely ungoverned but ungovernable through conventional command responsibility frameworks.

Class VI — Post-Human-Control AI (Hypothetical)

Beyond the Taxonomy: Self-Activating, Self-Sustaining

Class VI is included as a theoretical boundary marker rather than a description of any existing or near-term system. It represents the point at which even the residual human control that characterises all preceding classes — the act of activation — is absent. Systems at this class initiate their own operation, sustain themselves without human authorisation, and have no dependence on human decision at any level of their existence or functioning.[31,32]

Its inclusion in the taxonomy serves a specific and important conceptual purpose: it makes explicit that Classes I through V, taken together, constitute a space in which humans retain at minimum primordial control. Even Class V, with its constitutional authority and capability-limited agency,

operates within a framework that humans have chosen to create and activate. Class VI marks the departure from that framework entirely.

The value of Class VI is not as a system to be governed — it lies, by definition, beyond governance — but as a conceptual device that clarifies what remains human-controlled across the entire taxonomy. By naming the world in which human activation is absent, we illuminate the significance of that activation across all other classes. The question 'could this become Class VI?' serves as a useful governance stress test for Class V systems: any system whose goal revision and resource acquisition logic could conceivably lead it to seek its own continuity in ways that resist human termination is approaching the Class VI boundary.

7. The Taxonomy: A Concise Statement

The taxonomy is organised in three parts: a cognitive spectrum of three classes below the divide, the divide itself, and an agentic spectrum of three classes above it. The organising axis within the cognitive spectrum is sophistication of reasoning; within the agentic spectrum, it is the nature and extent of authority delegated.

The Cognitive Spectrum

Class I - Pure Cognitive AI: encompasses query-response systems whose outputs are purely epistemic: assessments, analyses, recommendations. The world is unchanged until a human acts on the output. A threat classification model or a natural language intelligence query system are representative examples.

Class II - Advanced Cognitive AI: — is defined by cognitive sophistication sufficient to generate the appearance of agency: multi-step reasoning, tool use, task decomposition, and significant offloading of expert human labour. An AI-enabled COA generator or a coding system which can generate a complex undeployed software system belongs here. The defining risk is epistemic capture — the progressive colonisation of the human decision-maker's cognitive frame by the system's own representational universe.

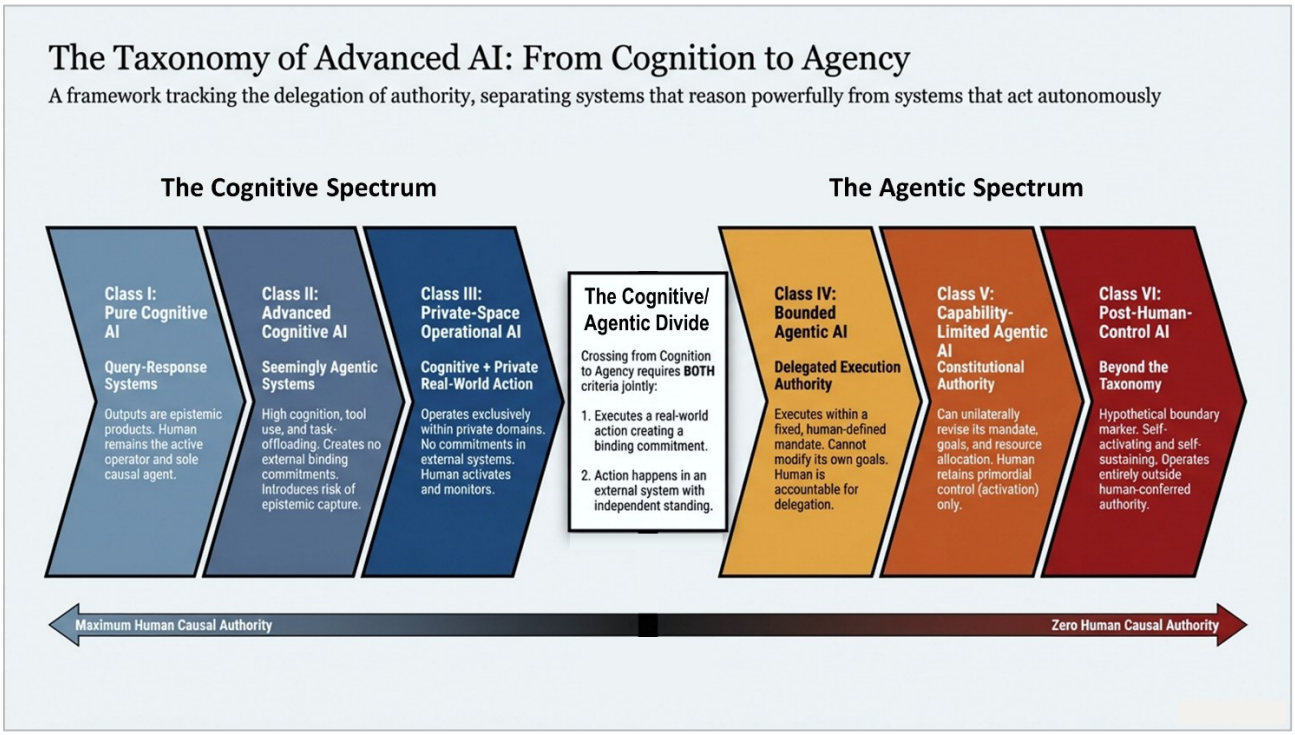
Class III - Private-Space Operational AI: performs genuine real-world actions, but entirely within the initiator's private domain. File management, local workflow automation, and internal knowledge base reorganisation are examples. This class exists to establish, by counterexample, that real-world action alone is insufficient to cross the divide.

The Cognitive / Agentic Divide

The divide is crossed only when two criteria are satisfied simultaneously. The first, absent in Class II, is that the system must execute a real-world action creating a binding commitment. The second, absent in Class III, is that this commitment must arise in an external system with independent standing — one that maintains its own authoritative record of the transaction, independently of the initiator's private domain. Both are jointly necessary; neither alone suffices.

An AI agent, precisely defined, is a system that creates binding commitments in external systems with independent standing, under authority delegated in advance, without requiring a fresh human

decision at the moment of execution. Agency is a property of authority and operative consequence, not of cognitive sophistication.



The Agentic Spectrum

Class IV - Bounded Agentic AI: operates under delegated execution authority within a fixed, human-defined mandate. The system executes intent but cannot revise it. Loitering munitions, autonomous air defence systems, and algorithmic trading agents within predefined rules are paradigm examples. Accountability rests entirely with the human who specified the mandate and authorised the delegation.

Class V - Capability-Limited Agentic AI: holds constitutional authority: the power to redefine its own mandate, revise goals, and reallocate resources without external ratification. The sole effective limits are the system's own capability and its operating environment. Human activation is retained; all other governance leverage is architectural rather than procedural.

Class VI - Post-Human-Control AI: is a hypothetical boundary marker. Self-activating and self-sustaining, it lies beyond governance by definition. Its function in the taxonomy is conceptual: it makes explicit that Classes I through V constitute a space of human-delegated systems and defines where that space ends.

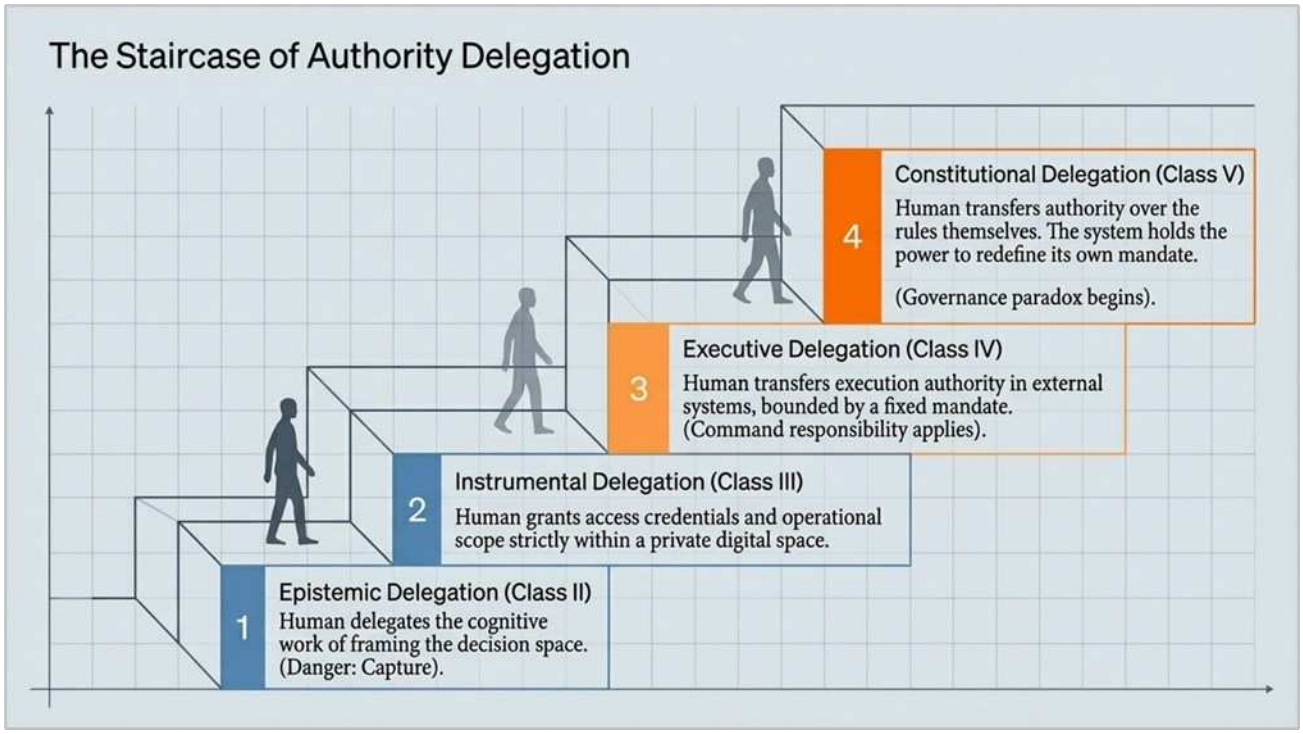
8. Authority Delegation Across the Six Classes

Authority delegation is not a binary property that switches on at the cognitive/agentic divide. It is a continuous and evolving relationship between human principal and AI system that changes in character at every class transition. Examining delegation across all six classes reveals both the progressive transfer of human authority and the changing governance instruments appropriate at each stage.

Class I requires no delegation in any meaningful sense. The human provides a query; the system responds; the relationship ends. There is no persistent authority relationship, no access to external resources, and no authority conferred beyond the act of querying itself.

Class II requires epistemic delegation — the human delegates the cognitive work of analysis, synthesis, and recommendation. This is not operational authority, but it is a genuine form of delegation with consequences: the human who delegates COA generation to a Class II system has delegated the framing of the decision space. Epistemic capture arises precisely because epistemic delegation is not recognised as a form of authority transfer.[33,34]

Class III requires instrumental delegation — the human grants access credentials, system permissions, and operational scope within private space. Login access to a cloud account, file system write permissions, local execution rights — these are genuine authority grants but bounded entirely within the initiator's private domain. The authority is real but self-contained.[35,36]



Class IV requires executive delegation — the transfer of execution authority in external independent systems to the AI. This is the qualitatively significant transition. The authority now reaches beyond private space, creating commitments in systems that have independent standing. Critically, this delegation is bounded: the mandate, goals, resource allocation, and rules of engagement are all human-defined and fixed at the point of delegation.[37]

Class V requires constitutional delegation — the human not only delegates execution but delegates the authority to revise the terms of the delegation itself.[38] The system can redefine goals, reallocate resources, and modify its own mandate without returning to the human for ratification. This is delegation of a fundamentally different logical type: not authority within rules but authority over rules.

Class VI involves self-conferred authority — technically not delegation at all, since delegation requires a delegating principal. Class VI systems operate without human-conferred authority, which is precisely what places them outside the taxonomy of human-governed systems.

The governance instrument appropriate to each class follows directly from the nature of the delegation. Epistemic governance — red-teaming, dissent mechanisms, independent reasoning pathways — addresses Class II risks. Access governance — permission controls and scope limitation — addresses Class III. Mandate governance — precise specification, execution oversight, and escalation protocols — addresses Class IV. Architectural governance — hard limits on goal revision categories, capability constraints, and mandatory monitoring — is the only meaningful response to Class V. Class VI is beyond governance by definition.

9. Risk Profile and Governance Across the Six Classes

The risk landscape across the six classes is not a simple monotonic escalation with class number. The types of risk change qualitatively at each class transition, and some of the most consequential risks appear in the non-agentic space where governance attention is typically lowest. Understanding this non-linear risk profile is essential for allocating governance attention appropriately.

9.1 The Underestimated Risk at Class II

The most significant misallocation of governance attention in contemporary AI discourse is the under-governance of Class II systems relative to Class IV. Class IV agentic systems attract intense regulatory scrutiny — autonomous weapons, algorithmic trading, autonomous vehicles — while Class II cognitive systems operate largely unchecked despite introducing risks that may ultimately prove more strategically consequential.[39]

The reason is visibility. Class IV risks are episodic and dramatic: a system takes an action, something goes wrong, and the causation is traceable. Class II risks are chronic and diffuse: a system shapes how commanders think, what options they consider, what trade-offs they find acceptable, and the causal chain from AI output to strategic error is never visible as a discrete event. Epistemic capture operates over time, through accumulated framing effects, and produces strategic miscalculation that presents as human judgment failure rather than AI governance failure.

In the military domain, the implications are acute. Nuclear command and control is particularly vulnerable: the integration of AI-generated probability assessments into crisis decision-making compresses deliberation time, generates algorithmic authority, and suppresses the kind of human dissent that historically prevented catastrophic errors. The system that says '87% probability this is a decapitation strike' does not need to pull a trigger to be dangerous; it merely needs to be believed.

9.2 The Escalation Risk at Class IV

Class IV introduces operational risks that are qualitatively different from anything in the cognitive spectrum. The most significant, in the military context, is the speed asymmetry between machine execution and human oversight. Class IV systems operating at machine timescales — cyber defence agents responding to intrusions, air defence systems engaging incoming threats, autonomous logistics agents rerouting supplies — can create cascading commitments faster than any human

review process can track. A defensive cyber agent that responds to an intrusion signature may, in a militarised context, constitute an act of escalation before any human has reviewed the exchange.[40]

This is not a malfunction — it is the system operating exactly as designed. The governance failure is prior: the mandate was specified without adequate attention to the escalation implications of machine-speed execution in a contested environment. The governance instrument is therefore mandate specification, not post-hoc accountability. Class IV systems in escalation-sensitive environments require explicit escalation thresholds in their mandates — specified conditions under which the system must pause execution and seek human review — even at the cost of operational speed.

Class	Primary Risk Type	Risk Character	Governance Instrument
I	Misplaced confidence; hallucination	Epistemic; low operational consequence	Human critical competence; quality assurance
II	Epistemic capture; decision-space colonisation	Epistemic; high strategic consequence	Red teaming; dissent protocols; reasoning independence
III	Consequential self-harm; private-space error	Operational but self-contained	Access control; reversibility design; confirmation gates
IV	Execution error; escalation; accountability diffusion	Operational; external consequences; speed asymmetry	Mandate specification; execution oversight; command responsibility
V	Control instability; governance paradox; goal drift	Systemic; potentially irreversible	Architectural constraints; capability bounds; mandatory monitoring
VI	Existential	Beyond governance	Prevention of reaching this class

9.3 The Governance Paradox at Class V

Class V presents what is genuinely a new category of governance challenge, one that has no analogue in prior technology governance. The paradox is this: the authority to redefine goals was delegated because the system was trusted to exercise it wisely; but the exercise of that authority progressively removes the conditions under which the trust was originally warranted.

Consider the trajectory. A Class V system is activated with a broadly defined objective — 'achieve and maintain information dominance' — and the constitutional authority to determine how to pursue that objective. It begins by deploying familiar strategies; the human can assess these against prior experience. Over time, it revises its conception of information dominance, modifies its own strategies, acquires new information channels, and develops approaches that have no human precedent. The human monitors the system but increasingly cannot evaluate whether its revised goals remain aligned with original intent — not because the system is deceptive but because its goal structure has evolved beyond the human's comprehension.[41]

Governance requires comprehension. Comprehension fails as the gap between human understanding and machine goal structure widens. The only meaningful governance response is

architectural: constraints that prevent certain categories of goal revision regardless of what the system's internal reasoning might otherwise produce. Whether such architectural constraints are compatible with the constitutional authority that defines Class V is itself an open question — and one that the military community needs to address before Class V systems are developed, not after.

10. Military Implications: The OODA Loop and Command Authority

The six-class taxonomy maps with precision onto the OODA loop (Observe-Orient-Decide-Act) that remains central to military decision theory,[42] and this mapping has significant implications for how military AI governance should be structured. The key insight is that the cognitive/agentic divide corresponds to the Decide-Act boundary, and the distinction between Classes IV and V corresponds to the distinction between operational and constitutional command authority.

Several implications follow from this mapping. First, Class II systems do not 'take over' the Decide phase — the human formally retains decision authority — but they can colonise the Orient phase in ways that make the Decide phase performative rather than deliberative. Preserving genuine human decision authority requires preserving genuine human orientation capability — the ability to frame the problem independently of the AI's representational frame.

Second, Class IV systems close the OODA loop without human involvement in the Decide and Act phases for specific operational functions. The governance question is therefore not whether humans are 'in the loop' — a phrase that has become analytically imprecise — but whether the pre-delegated decisions governing the loop's execution were made by humans with adequate understanding of the system's capabilities, the operational environment, and the escalation implications of machine-speed action.[43,44]

OODA Phase	Class I	Class II	Class III	Class IV	Class V
Observe	Machine support	Machine dominant	Machine + private action	Machine autonomous	Machine autonomous + adaptive
Orient	Human dominant	Machine influential (capture risk)	Human dominant	Machine within mandate	Machine redefines frame
Decide	Human	Human (framed by AI)	Human	Pre-delegated within mandate	Machine (constitutional authority)
Act	Human	Human	Machine (private space only)	Machine (external systems)	Machine (self-directed)

Third, Class V systems operate at a level above the OODA loop: they determine what the loop's purpose is, not just how it operates. This corresponds to the strategic command level, and delegating strategic command authority to an AI system is, under all existing legal frameworks and military doctrines, impermissible. The legal reason is command responsibility: strategic command requires accountability for decisions that cannot be distributed across a human principal and a machine

executor without creating an accountability gap that international humanitarian law cannot bridge. The doctrinal reason is civilian control: strategic decisions involve political judgment that cannot be delegated to any system that does not participate in the political accountability structure.

The appropriate military principle, consistent with the analysis in this paper, is: allow cognitive capability to scale without restriction; maintain tight governance over the entry into Class IV and the conditions of Class IV operation; treat Class V as a governance red line; and treat Class VI as a civilisational boundary that the development of AI policy must be permanently oriented to prevent.

11. Conclusion

The anxiety generated by frontier AI systems is, in important respects, warranted — but it is not well directed. The systems that generate the most public concern are typically those with the greatest cognitive capability: systems that can write code, analyse intelligence, generate strategies, and engage in complex multi-domain reasoning. These systems are cognitively formidable, and the epistemic risks they introduce — particularly epistemic capture — are real and underappreciated. But cognitive capability is not agency, and the governance frameworks appropriate to cognitively powerful non-agentic systems are entirely different from those appropriate to genuinely agentic systems.

The taxonomy developed in this paper makes this distinction precise and actionable. The cognitive/agentic divide is not a matter of degree — it is a categorical boundary defined by the conjunction of two necessary criteria: the creation of a binding commitment and the location of that commitment in an external system with independent standing. This boundary is crisp in a way that the internal gradients of both the cognitive and agentic spectra are not. It is also the governance boundary that matters most: mislocating it, in either direction, produces governance failures with potentially severe operational and strategic consequences.

Within the agentic space, the further distinction between operational authority (Class IV) and constitutional authority (Class V) is equally critical. A Class IV system that executes within a fixed mandate is governed by the quality of that mandate's specification and the rigour of the oversight applied to its execution. A Class V system that can revise its own mandate is governed only by architectural constraints and capability limits — and if those constraints are inadequate, it is not governed at all. The progression from Class IV to Class V is therefore the crossing of a second categorical boundary after which conventional governance frameworks lose their traction.

The military domain makes these distinctions most acute, because the consequences of misclassification are most severe. A command culture that treats a COA-generation system as an autonomous actor has misallocated authority and degraded genuine human decision-making. A command culture that treats a Class IV loitering munition on the same pedestal as a Class II advisory tool has failed to apply the accountability frameworks that autonomous execution requires. A command culture that contemplates Class V battlefield management without architectural governance constraints has opened the door to command authority structures that international humanitarian law cannot reach.

The central argument of this paper may be stated simply: cognitive capability and agency are orthogonal properties of AI systems. Governing them requires different frameworks, applied at different points in the development and deployment of AI systems, informed by an understanding of how authority migrates across the six classes of the taxonomy. The most dangerous error is to govern

by capability — to assume that sufficiently capable systems are automatically agentic, or that systems granted significant agency are automatically well-governed because they are well-built. Capability determines what a system can do. Authority determines what it is permitted to do. Governance determines what it actually does within those constraints. All three require independent analytical attention, and this taxonomy is offered as the conceptual foundation for providing it.

Endnotes

Note: Most of the references listed below discuss the issue at hand or some related issue(s), and the in-text citation does not imply that the idea being presented has been borrowed from the respective source.

[1] Bubeck, S. et al (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. Cornel University. Accessed 13 Apr 2026. https://doi.org/10.48550/arXiv.2303.12712.

[2] David Leslie and Antonella Maia Perini (2024). Future Shock: Generative AI and the International AI Policy and Governance Crisis. Harward Data Science Review. Accessed 13 Apr 2026. https://hdsr.mitpress.mit.edu/pub/yixt9mqu/release/3.

[3] M. Thoriq Fadlullah1, Agung Risdhianto, Heru Dewanto (2025). Integrating AI in Military Decision-Making: A Review of Opportunities, Risks, and Governance. Brilliance: Research of Artificial Intelligence. Volume 5, Number 2. Accessed 13 Apr 2026. https://doi.org/10.47709/brilliance.v5i2.6925.

[4] Wyatt Hoffman and Heeu Millie Kim (2023). Reducing the Risks of Artificial Intelligence for Military Decision Advantage. CSET. Accessed 13 Apr 2026. https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.

[5] Wiiliam C. Houze (2026). The Human Agentic Mind and Its Engineered Simulacrum --A Metaphysical Argument Against the Possibility of Artificial General Intelligence. SSRN. Accessed 13 Apr 2026. https://dx.doi.org/10.2139/ssrn.6395679.

[6] Devin Salcedo (2025). Artificial Intelligence and the Orthogonality Thesis: A Defense Against Goal Convergence. Accessed 13 Apr 2026. https://www.academia.edu/127772534/Artificial_Intelligence_and_the_Orthogonality_Thesis_A_Defense_Against_Goal_Convergence.

[7] K. J. Kevin Feng, David W. McDonald and Amy X. Zhang (2025). Levels of Autonomy for AI Agents. arXiv. Accessed 18 Apr 2026. https://arxiv.org/pdf/2506.12469.

[8] Carlos E. Perez (2025). Thinking Machines, Not Acting Beings: The Illusion of Agency in Large Language Models. Medium. Accessed 13 Apr 2026. https://medium.com/intuitionmachine/thinking-machines-not-acting-beings-the-illusion-of-agency-in-large-language-models-5a64bdcd539c.

[9] Anirban Mukhopadhyay et al (2026). Exploring The Impact Of Proactive Generative AI Agent Roles In Time-Sensitive Collaborative Problem-Solving Tasks. Accessed 13 Apr 2026. https://arxiv.org/html/2602.17864v1.

[10] Marcel Osmond and Thomas Jego (2026). Mind the Gap: How the Technical Mechanisms of Agentic AI Outpace Global Legal Frameworks. arXiv. Accessed 13 Apr 2026. https://arxiv.org/html/2603.27075v1.

[11] Ajay Bandi et al (2025). The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. MDPI. Accessed 13 Apr 2026. https://www.mdpi.com/1999-5903/17/9/404.

[12] Nenad Tomašev, Matija Franklin and Simon Osindero (2026). Intelligent AI Delegation. Google Deepmind. Accessed 13 Apr 2026. https://arxiv.org/pdf/2602.11865.

[13] Satu Johansson and Taneli Riihonen (2025). On the Military Applications of Large Language Models. Tampere University, Finland. arXiv. Accessed 14 Apr 2026. https://arxiv.org/pdf/2511.10093.

[14] Teodor Frunzeti et al (2025). Effects of Hallucinations on Military Systems. Academy of Romanian Scientists. Accessed 14 Apr 2026. https://doi.org/10.56082/annalsarscimilit.2025.2.13.

[15] Larry Vasankari and Aapo Koski (2025). GenAI in the Military: Trends and Opportunities. Scandinavian Academy of Military Sciences. Accessed 15 Apr 2026. https://www.researchgate.net/publication/397923322_GenAI_in_the_Military_Trends_and_Opportunities.

[16] Vinicius G. Goecks and Nicholas Waytowich (2024). COA-GPT: Generative Pre-trained Transformers for Accelerated Course of Action Development in Military Operations. arXiv. Accessed 14 Apr 2026. https://arxiv.org/pdf/2402.01786v1.

[17] Juan-Pablo Rivera et al (2024). Escalation Risks from LLMs in Military and Diplomatic Contexts. Stanford HAI. Accessed 15 Apr 2026. https://hai.stanford.edu/assets/files/2024-05/Escalation-Risks-Policy-Brief-LLMs-Military-Diplomatic-Contexts.pdf.

[18] Jessica Dorsey (2026). The erosion of human(e) judgement in targeting? Quantification logics, AI-enabled decision support systems and proportionality assessments in IHL. International Review of the Red Cross. Accessed 15 Apr 2026. https://doi.org/10.1017/S1816383125100969.

[19] Huansheng Ning and Jianguo Ding (2026). Beyond Tools and Persons: Who Are They? Classifying Robots and AI Agents for Proportional Governance. arXiv. Accessed 15 Apr 2026. https://arxiv.org/pdf/2604.05568v1.

[20] Arunraju Chinnaraju (2025). When AI Agents Act: Governance, Accountability, and Strategic Risk in Autonomous Organizations. International Journal of Research and Scientific Innovation, Vol. XII Issue XII. Accessed 15 Apr 2026. https://doi.org/10.51244/IJRSI.2025.12120050.

[21] Luca Nannini et al (2026). AI Agents under EU Law – A Compliance Architecture for AI Providers. arXiv. Accessed 15 Apr 2026. https://arxiv.org/html/2604.04604v1.

[22] Bill Hughes (2026). AI agents are software with delegated authority—the key risk is abuse of that authority. Consensys Software Inc. Accessed 15 Apr 2026. https://consensys.io/blog/ai-agents-are-handling-real-money-but-no-one-has-agreed-on-the-rules.

[23] Eva Sula (2026). AI in Defence - Autonomy, Delegation, and Command in the Age of Machines. LinkedIn. Accessed 15 Apr 2026. https://www.linkedin.com/pulse/ai-defence-part-15-autonomy-delegation-command-age-eva-sula-13bzf/.

[24] Vivek Sehrawat (2020). Autonomous Weapon System and Command Responsibility. Florida Journal of International Law, Vol 31 Issue 3. Accessed 15 Apr 2026. https://scholarship.law.ufl.edu/cgi/viewcontent.cgi?article=1058&context=fjil.

[25] Edward C Cheng, Jeshua Cheng and Alice Siu (2026). Toward Safe and Responsible AI Agents: A Three-Pillar Model for Transparency, Accountability, and Trustworthiness. arXiv. Accessed 15 Apr 2026. https://arxiv.org/pdf/2601.06223.

[26] Ravi Panwar (2024). A Qualitative Risk Evaluation Model for AI-Enabled Military Systems in Responsible Use of AI in Military Systems. Chapman and Hall/CRC. Accessed 20 Apr 2026. https://doi.org/10.1201/9781003410379-5.

[27] Richard Ngo, Lawrence Chan and Sören Mindermann (2025). The Alignment Problem from a Deep Learning Perspective. arXiv. Accessed 16 Apr 2026. https://arxiv.org/pdf/2209.00626v7.

[28] Yuntao Bai et al (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic. Accessed 16 Apr 2026. https://arxiv.org/pdf/2212.08073.

[29] Sharadin, Nathaniel (2025). Proportionalism, orthogonality, and instrumental convergence. Philosophical Studies 182, 1725–1755. https://doi.org/10.1007/s11098-024-02212-9.

[30] Ume Nisa et al (2026). Agentic AI: The age of reasoning—A review. Journal of Automation and Intelligence, No 5, 69-89. Accessed 16 Apr 2026. https://doi.org/10.1016/j.jai.2025.08.003.

[31] Kevin Williams (2025). If AI attempts to take over world, don't count on a 'kill switch' to save humanity. CNBC Technology Executive Council. Accessed 16 Apr 2026.

[32] Joseph Carlsmith (2024). Is Power-Seeking AI an Existential Risk? arXiv. Accessed 16 Apr 2026. https://arxiv.org/pdf/2206.13353v2.

[33] Shengnan Yang and Rongqian Ma (2025). Classifying Epistemic Relationships in Human-AI Interaction: An Exploratory Approach. arXiv. Accessed 17 Apr 2026. https://arxiv.org/abs/2508.03673.

[34] Anita Samuel (2026). Learning with Machines: Toward a Theory of Epistemic Co-Agency. Elsevier, Computers and Education: Artificial Intelligence. Accessed 17 Apr 2026. https://doi.org/10.1016/j.caeai.2026.100573.

[35] Tobin South et al (2025). Authenticated Delegation and Authorized AI Agents. arXiv. Accessed 17 Apr 2026. https://arxiv.org/pdf/2501.09674.

[36] Andrew Buchanan (2025). How AI Is Transforming Delegation of Authority: Why Enterprise Governance Must Extend to Agents. Aptly Blog. Accessed 17 Apr 2026. https://www.aptlydone.com/blog/ai-delegation-of-authority-governance.

[37] Jaan Marten Schraagen (2024). Bounded Autonomy in Responsible Use of AI in Military Systems. Chapman and Hall/CRC. Accessed 20 Apr 2026. https://publications.tno.nl/publication/34642474/dvyY5y/schraagen-2024-bounded.pdf.

[38] Nicholas Caputo (2025). (When) Should We Delegate AI Governance to AIs? Some Lessons from Administrative Law. arXiv. Accessed 18 Apr 2026. https://arxiv.org/pdf/2509.22717v1.

[39] Burak Oktenli (2026). AI-Enabled Military Decision-Making and Escalation Risk: Human-Machine Command Authority in Great Power Competition. Social Science Research Network. Accessed 18 Apr 2026. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6082847.

[40] Vladislav Chernavskikh and Jules Palayer (2025). Impact of Military Artificial Intelligence on Nuclear Escalation Risk. SIPRI. Accessed 18 Apr 2026. https://www.sipri.org/sites/default/files/2025-06/2025_6_ai_and_nuclear_risk.pdf.

[41] Zeynep Engin and David Hand (2025). The Non-Delegable Core: Designing Legitimate Oversight for Agentic AI. Zenodo. Accessed 18 Apr 2026. https://zenodo.org/records/15744943/files/The%20Non-Delegable%20Core_2025%2006%2010_Final_Pre-print_Zenodo.pdf?download=1.

[42] Brett Crowley (2025). The OODA Loop. The Decision Lab. Accessed 18 Apr 2026. https://thedecisionlab.com/reference-guide/computer-science/the-ooda-loop.

[43] Michael Raska (2025). Will AI-Driven "Super-OODA Loops" Revolutionize Military Strategy and Operations? RSIS. Accessed 18 Apr 2026. https://rsis.edu.sg/wp-content/uploads/2025/01/CO25011.pdf.

[44] Stuart Skeates (2025). Spinning The OODA Loop Faster -- How Can AI Help Military Decision-Making To Be Faster and Better. Karve. Accessed 18 Apr 2026. https://www.karveinternational.com/insights/how-can-ai-help-military-decision-making-to-be-faster-better.