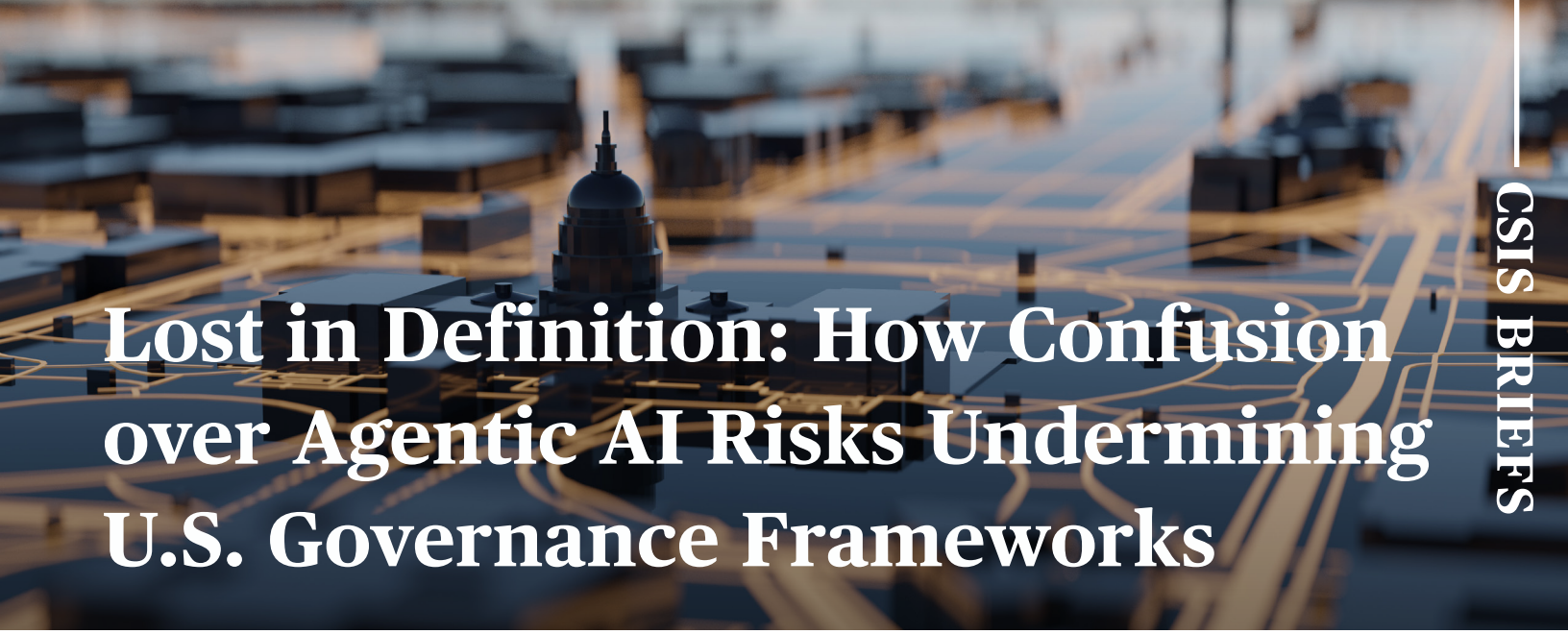

CSIS BRIEFS

Lost in Definition: How Confusion over Agentic AI Risks Undermining U.S. Governance Frameworks

By Yasir Atalan, Ian Reynolds, and Ben Jensen

JANUARY 2026

THE ISSUE

- "Agentic AI" is an umbrella term that covers a wide range of systems from basic chat assistants to complex autonomous workflows. This ambiguity risks undermining U.S. governance frameworks and creating procurement vulnerabilities that expose organizations to mismatched capabilities and unaccounted risks. If the same vague word is applied to a helpful chatbot and a combat-ready swarm, the United States could accidentally deploy a system with the power to start an operation before that system understands the context or risks involved.
- The danger is not that the AI lacks intelligence, but that it lacks judgment; a system might be smart enough to execute a task perfectly yet fail to realize that a sudden change in the local situation makes that task a catastrophic mistake. This is particularly critical in national security environments where decisionmaking stakes are high.
- There is a need for clearer context charts for AI workflows so that commanders know exactly which tasks a machine can handle alone and exactly which who is responsible if the machine makes a mistake. The U.S. government must establish a relational, capability-based taxonomy that specifies where systems sit in workflows, what authorities they exercise, and how accountability is distributed. This approach shifts acquisition and oversight from a narrow focus on technical features to understanding broad organizational impact and ensuring evaluation matches real operational risk.

INTRODUCTION

Artificial intelligence systems that can plan, act, and operate across multiple steps with limited human supervision are increasingly described as "agentic AI." Across defense policy speeches, industry white papers, and technical briefings, the term has quickly become shorthand for what many view as the next phase of **AI capability**. According to a recent survey, **35 percent** of a wide range of organizations in 116 countries have begun to use agentic AI systems. Governments and militaries are **already experimenting** with such systems in planning, logistics, intelligence analysis, and decision support, while private sector investment in agentic technologies continues to **grow** rapidly.

Yet, despite its widespread use, there is no shared **understanding** of what qualifies as an agentic AI system. Industry actors, consultancies, and technology providers use the term to describe a wide range of tools, from simple chat-based assistants to autonomous systems that can initiate actions, coordinate with other systems, and operate over extended periods without direct supervision. In the **national security domain**, this definitional ambiguity has important consequences. First, it undermines test and evaluation. When everything from a script (a simple preprogramed sequence

of instructions) to a fully autonomous decision support system can be branded as an agent, establishing standardized evaluation regimes becomes a challenge. Second, when procurement documents request "agentic capabilities" without operational specifications, vendors can satisfy requirements in name while delivering vastly different systems. Finally, when governance frameworks are designed for targeted—narrow—applications, organizations risk missing how agentic systems reshape broader organizational workflows, authority, and accountability structures.

At present, the first draft of what counts as an agent is largely being written by industry. This brief, however, argues that governments must take a more active role in defining agentic AI to establish a governance foundation that matches how agentic systems may change organizational dynamics. The paper shows how inconsistent definitions undermine acquisition, testing, and oversight. Moreover, drawing on insights from industry, philosophy, and national security practice, the brief proposes a capability-based taxonomy grounded in a relational understanding of AI agency, one that shifts evaluation from asking what a system can do to how a system reshapes organizational decisionmaking, delegation of authority, and accountability structures.

This approach operationalizes a relational perspective on AI agency. In doing so, it recognizes that agency emerges not from isolated technical systems but from how those systems reshape organizational decisionmaking when embedded in planning processes and institutional authorities. To implement this framework, the brief recommends that defense agencies (1) require relational classification for all agentic AI programs, (2) update procurement processes to mandate capability mapping that captures workflow position and authority delegation, and (3) adopt governance frameworks that prioritize organizational impact over narrow technical benchmarks. By grounding conceptual understandings of agentic systems in operational context rather than marketing language, this brief argues that the United States can better align procurement, evaluation, and governance with the emerging realities of human-AI decisionmaking in the context of national security and other domains.

INDUSTRY'S DIVERGING DEFINITIONS

Industry provides the clearest example of definitional inconsistency in agentic AI. For instance, some firms focus on autonomous orchestration of tools and workflows while others stress natural language interfaces powered by large language models (LLMs). Additional definitions require the system to learn and adapt based on feedback or may define agents as simply rules-based automation behind a front end, user-facing chatbot. In fact, a simple internet search leads to a range of definitions offered by Anthropic, Amazon, Google, IBM, and others. The table below lists a set of definitions from leading firms.

Table 1: Industry Definitions of Agentic AI

Firm	Definitions
OpenAI (2023)	■ Systems that can take actions over extended periods that consistently advance specified goals without every behavior being pre-scripted ■ A system's "agenticness" as the degree to which it can achieve complex goals in complex environments with limited direct human supervision
Anthropic (2024)	■ Acknowledge customers use the term "agent" for a wide range of systems, from long-running autonomous tool-using processes to prescriptive workflows ■ These are grouped as "agentic systems" but the term "agent" is reserved for LLM-centered loops in which the model itself decides which tools to call, in what order and when to stop
IBM (2024)	■ AI systems that independently perform tasks by assembling workflows that use available tools ■ Agents go beyond simple language interfaces to include decisionmaking, problem-solving, and interaction with external environments
AWS (2025)	■ Software that interacts with its environment, gathers data, and uses that data to carry out self-directed tasks aimed at predefined human goals ■ Humans specify objectives while the agent chooses appropriate actions without continuous supervision

Google Cloud (2024)	■ Software systems that use AI methods to pursue goals and complete tasks on behalf of users ■ Display reasoning, planning and memory and have enough autonomy to make decisions, learn, and adapt over time
McKinsey (2025)	■ The tools people use to interact with AI ■ Can automate and execute complex tasks that would normally require humans, such as natural language processing, analysis, and decision support
Deloitte (2025)	■ Autonomous components that extend what generative AI can do inside organizations ■ Multiagent systems are made up of role-specific agents that understand requests, plan workflows, coordinate actions, collaborate with humans and validate outputs
Mistral (2025)	■ LLM enabled autonomous systems that can plan, use tools, and complete tasks to complete complicated goals based on user guidance

Note: Definitions are paraphrased.

These industry definitions can be classified according to a range of features emphasized by each of the different organizational conceptualizations of AI agents. These factors include:

- **System Autonomy:** Can the system act autonomously based on guidance?
- **Need for an LLM:** Does the system explicitly rely on an LLM?
- **Learning/Feedback:** Does the system need the capacity to learn and update behavior?
- **Memory:** Does the system need either a short- or long-term memory to tailor behavior to user needs?
- **Planning/Reasoning:** Can the system break down user guidance into tasks and subtasks for completion?
- **Collaboration:** Does the system collaborate either with other computational systems or other humans?
- **Perception/Observation:** Does the system have a perception mechanism to take in additional data and react to external environment?

Table 2: Summary of Definitions of AI Agents by Key Industry Players

Company	Autonomy	Explicit about LLMs	Learning/ feedback	Memory	Reasoning and planning	Collaboration	Perception/ observation
Anthropic	Yes	Yes	Yes	Qualified	Yes	No	No
IBM	Yes	Yes	Yes	Yes	Yes	Qualified	Qualified
AWS	Yes	Yes	Yes	Yes	Yes	Yes	Yes
McKinsey	Yes	Qualified	Qualified	Yes	Yes	Yes	No
Google	Yes	Qualified	Yes	Qualified	Yes	Yes	Yes
Deloitte	Yes	Qualified	Yes	Qualified	Yes	Qualified	No
OpenAI	Yes	Yes	No	No	Yes	No	No
Mistral	Yes	Yes	No	No	Qualified	Qualified	No

Note: green = robustly discussed/required; yellow = briefly mentioned/conditional or limited scope; red = not mentioned/required.

Source: CSIS analysis of company statements.

As demonstrated in Table 2, industry discussions of AI agents tend to vary across the seven identified features. Other than with respect to the requirement of autonomy, there is conceptual variation in all categories, either in the degree to which these characteristics are addressed, or with respect to if they are mentioned at all. This indicates that, at least in terms of the material reviewed for this brief, there are a range of ways in which industry players discuss AI agents.

Yet while requirement of autonomy is shared across industry players at a high level, conceptual variation still exists within this category. For instance, organizations such as **IBM** and **AWS** explicitly lay out a spectrum of autonomy in agents ranging from simple reactive agents, which act based on strict rules, to learning agents that can automatically update behavior based on past actions. For organizations including **IBM** and **OpenAI**, this spectrum explicitly excludes chatbots from counting as agents. Others including **McKinsey** appear to suggest that LLM chatbots could be considered agents, stating, "If you've ever interacted with a customer service chatbot or asked a gen AI model to write you a sonnet, then you're likely already familiar with a rudimentary version of AI agents." Moreover, some industry-based researchers have **noted** that there is no clear "binary distinction" between AI agents and generative AI models such as OpenAI's GPT-4. The upshot is, even when industry organizations agree that a feature such as autonomy is a key factor in AI agents, inconsistency on what *counts* as autonomous behavior remains.

PHILOSOPHICAL ROOTS OF AI AGENCY

Philosophical approaches toward agency and agentic systems have shaped AI research for decades. This brief identifies three main philosophical categories. First, **internalist** approaches treat agency as a characteristic of a bounded technical system. With respect to system behavior, questions of agency therefore focus on the model's representation of the world and how it maps inputs to outputs in a fashion that determines system action. For instance, a language model that plans and executes tasks based on trained model **weights** reflects this tradition. Moreover, this intellectual orientation influences many ongoing discussions of tool-using LLM agents, because planning and task selection still result from the basic structure of perception of inputs leading to the agentic system's eventual action (i.e., which tool to call based on a user query). In a defense context, this is akin to an autonomous surveillance drone that identifies a target and selects a sensor for tracking based solely on its preprogrammed recognition software, without factoring in external variables like shifting diplomatic rules of engagement.

Second, **embodied/phenomenological** approaches treat agency as inseparable from dynamic interactions with an environment, emphasizing what the system can reliably do in context through tight sensing and action loops. A robotic rover navigating terrain using sensorimotor coupling rather than internal representations exemplifies this view. Governance tools rooted in an embodied approach would therefore emphasize safety envelopes, progressive field trials, and certification in realistic scenarios. For the military, this approach is mirrored in the testing of autonomous undersea vehicles, where the focus is not on the software's logic but on its physical ability to safely navigate unpredictable ocean currents and avoid collisions in real time.

Finally, **relational approaches** treat agency as emerging from relationships among configurations of technology, people, and institutions. Thus, the key question is not whether an AI model has agency, but how agency is produced when a system is embedded in workflows, doctrine, interface design, and organizational routines. As an example, a workflow tool reorganizing how a **military staff** generates options, delegates authorities, assesses risks, and assesses the process alongside ongoing operations reflects this orientation.

Each philosophical lens implies a different approach to governance, structuring what policymakers see as the object of regulation, where they locate risk, and which governance tools are perceived as most impactful. For instance, an **internalist framing** concentrates risk in the model, directing governance toward benchmark performance, robustness tests, and technical guardrails. Seen in this light, an AI decision support system is primarily a technical artifact to be evaluated and aligned. Risk concentrates in the LLM's inference behavior and governance focuses on technical safeguards, benchmarks, and maintaining "**meaningful human control**" at specific decision points. Accordingly, the primary object of governance is the AI system itself. In the case of a military decision support system, evaluation might focus on factors such as the capacity to generate targets at a desired rate, in line with predetermined rules of engagement. Therefore, such evaluation metrics are narrow and targeted, highlighting the

specific system's performance characteristics related to task completion, rather than how the integration of the AI enabled tool might shift responsibility and control away from human analysts.

An **embodied perspective** concentrates risk in the platform's physical interaction with environments, directing governance toward safety certification, operating envelopes, rules of deployment, and scenario-driven field testing. A system such as an autonomous intelligence drone is therefore a platform to be tested and certified in physical context. Risk lies in physical interaction with complex environments. As such, the primary object of governance is the reliability of the platform operating autonomously at physical distance from human oversight. In this case, an embodied approach to agency results in a narrow focus with respect to military use cases and evaluation procedures. For instance, in the example of an unmanned rover, governance metrics would focus on validating the embodied system's capacity to navigate select environments rather than how the system's deployment might reshape traditional military roles and practices of intelligence, surveillance, and reconnaissance.

Finally, a **relational conceptualization** concentrates risk in sociotechnical configurations, directing governance toward procurement incentives, data curation, training pipelines, doctrine, workflow design, and distributed accountability structures. Accordingly, an AI agent embedded into a joint planning cell may represent a workflow and organizational reconfiguration. Risk is best analyzed as distributed throughout a sociotechnical system via misaligned incentives, overreliance, erosion of expertise, and opaque chains of responsibility. Governance, therefore, must address doctrine, procurement, training, and legal frameworks shaping human-AI teams. The primary focus of governance is the sociotechnical system. In military applications, this means that evaluating an AI-enabled planning system requires examining not just the algorithm's optimization capabilities, but how its recommendations influence staff workflows, whether junior officers feel empowered to question machine-generated action recommendations, and how accountability is distributed when decision outputs contribute to mission failure.

For national security applications, a relational perspective offers the most robust analytical framework for understanding effective governance and implementation. Defense organizations' use of agentic AI will rarely involve isolated technical systems making autonomous decisions. Instead, AI agents will be embedded in command structures, planning processes, and organizational workflows where agency is distributed across human operators, technical systems, doctrinal constraints, and institutional authorities. Consider an AI-enabled decision support system generating operational courses of action. The system's outputs emerge not just from a model's technical characteristics, but from engineering decisions about which databases it can access, interface design structuring how commanders interact with recommendations, training that shapes operator trust and skepticism, and organizational rules defining delegation boundaries. When things go wrong, responsibility cannot be traced to the technical system alone but is distributed across designers who specified access permissions, commanders who defined use parameters, and institutions that established oversight procedures.

While internalist and embodied lenses remain useful for specific governance tasks, such as model alignment or platform safety certification, they risk misplacing responsibility and obscuring where things go wrong when AI systems reshape decisionmaking processes. A relational perspective keeps organizational and systemic factors as primary sites of both risk assessment and governance intervention. Without this relational framing, a military organization might deploy an AI system marketed as a simple planning assistant only to discover it has altered their operational planning cycle, reduced debate about the proper course of action, and created ambiguity about whether staff officers or the AI system bear responsibility for flawed decisions. With a relational approach there is a higher likelihood that such risks become visible prior to operational integration.

THE CURRENT GOVERNANCE GAP

The absence of clear definitions creates three specific policy problems that a relational approach to agentic systems is uniquely positioned to address. First, corporate definitions may influence norms of adoption due to their public prevalence. When vendors call basic automation an "agent," they upsell modest tools as transformational. Moreover, when they market powerful tool-using systems as simple assistants, they sidestep scrutiny. This ambiguity undermines test and evaluation. If everything is branded as an agent, standardized evaluation regimes matching real risk become impossible. Organizations may overinvest in

model-level evaluations while underinvesting in evaluating how systems reshape organizational decisionmaking.

Second, acquisition practices lack conceptual foundations. To remain at the cutting edge, defense procurement documents may increasingly request "agentic capabilities" without properly defining operational specifications. Vendors then can satisfy agentic requirements in name, with vastly different systems, making comparison difficult. Government buyers may believe they are procuring agents that reason, plan, and act across systems when they are actually receiving glorified chatbots or, conversely, systems with more autonomy than realized. Without a relational framework asking what authorities are delegated to which system components under what conditions, acquisition becomes checking technical feature boxes rather than understanding broad organizational impact.

Third, in the absence of definitional and conceptual coherence, governance frameworks face the threat of being mismatched to actual risks. As a result, evaluation teams could struggle to decide which safety cases or legal reviews apply when determining whether systems resemble traditional decision support tools, autonomous platforms, or new members of human-machine teams. Without clear conceptual taxonomies, government agencies risk over-focusing on narrow task-specific benchmarks while ignoring how agents reshape organizational workflows, authority, and accountability.

A relational approach addresses these problems by shifting the governance question from "What can this system do?" to "How does this system reshape organizational decisionmaking, and what authorities are delegated under what constraints?".

A CAPABILITY-BASED TAXONOMY FOR AGENTIC AI

Currently, industry firms are dominating the first draft of conceptualizations of agentic AI. Yet the impact of these emerging conceptualizations on factors such as role delegation, rules of engagement, legal responsibility, and broader accountability is not adequately discussed within the policy community. A practical way forward is establishing a capability-based classification for AI systems exhibiting agent-like behavior. This taxonomy operationalizes a relational approach to agentic AI governance. Rather than asking internalist-driven questions (e.g., is this system an agent because it has x, y, or z characteristic?) or questions rooted in an embodied approach (e.g., can this platform safely operate autonomously via its sensor systems?), a capability taxonomy forces organizations to answer relationally oriented considerations which recognize that agency in defense contexts emerges from the interaction of technical capabilities, organizational workflows, and human practice. A relationally oriented taxonomy should capture the following questions:

- **Positionality in Workflows:** Where does the system sit in organizational decisionmaking (e.g., does it generate options, route information, and/or execute decisions)?
- **Authority Delegation:** What can the system autonomously initiate versus what requires human approval?
- **Teaming Structure:** Does the system operate in human-AI teams or multi-agent configurations?
- **Accountability Mapping:** Who is responsible when the system acts (e.g., developers, operators, commanders, or institutions)?
- **Temporal Scope:** How are system outputs shaped by temporally distributed design and workflow processes?
- **Human Practice:** Which human routines are disrupted and how?

This taxonomy makes explicit the organizational and relational dimensions of agentic AI. When acquisition officers specify workflow position and authority delegation, they should think about how the system reshapes decision processes. When commanders define collaboration structures, they should think about human-AI teaming dynamics. When evaluators map accountability, they should think about distributed responsibility chains. These are inherently relational questions that cannot be answered by examining technical systems in isolation. Moreover, such questions better align with emerging conceptualizations of **agentic warfare** and how military practices and organizations, such as **command staffs**, must be reconfigured.

RECOMMENDATIONS

This paper offers three recommendations for the U.S. government related to how to conceptualize and integrate agentic systems. First, as argued above, defense agencies should adopt a relationally influenced, capability-based classification of agentic AI for all new programs involving

planning, acting, or tool use. The National Institute of Standards and Technology, the Department of Defense Chief Digital and Artificial Intelligence Office, and the Office of Management and Budget can jointly define a simple schema and require its use in strategies, inventories, and risk assessments. Every system labeled agentic or deployed with agent-like features should be tagged according to this scheme, explicitly capturing both technical capabilities and organizational contexts such as which workflows the system enters, what authorities are delegated, where human oversight sits, and how broader system outputs will be evaluated.

Second, acquisition and test and evaluation processes should be updated to reflect this classification. Procurement documents should require vendors to supply agent capability sheets that map proposed systems onto agreed categories. As such, test plans should be conditioned on these capabilities. High-autonomy, tool-using agents touching critical systems must trigger rigorous scenario-based testing, red teaming, and organizational exercises that examine human-machine teaming and accountability. Low-autonomy advisory tools may be subject to less comprehensive evaluation. However, in either case, evaluation should extend beyond narrow technical metrics to include organizational uplift studies measuring actual impact on decision quality, tempo and accountability structures, metrics best visible through a relational lens.

Third, governance frameworks should be explicitly multi-lens. While the relational approach provides the best overarching framework for defense applications, internalist and embodied perspectives remain useful for specific tasks. Model-level alignment and benchmarking can be used to ensure important technical baselines, and platform certification and safety envelopes will still be important for managing the risks of any physical system operating autonomously. But these must be understood as components within a larger relational governance structure addressing how systems reshape organizational practices. Policymakers should be cognizant about which framework they apply to which governance task and ensure relational considerations including workflow design, delegation boundaries, and distributed accountability receive adequate attention.

Over time, this structure would help allies and partners align their own standards and avoid incompatible definitions. By grounding definitions in operational capabilities and organizational context rather than marketing language, the United States can establish a governance framework that travels across defense institutions and international partnerships, demonstrating a commitment to being a global leader in responsible AI implementation and governance innovation. ■

Yasir Atalan *is a deputy director and data fellow in the Futures Lab at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Ian Reynolds*** *is the postdoctoral fellow for the Futures Lab in the International Security Program at CSIS.* ***Benjamin Jensen*** *is director of the Futures Lab and a senior fellow for the Defense and Security Department at CSIS.*

This report is made possible by general support to CSIS. No direct sponsorship contributed to this report.