Microsoft

Edge capability in the era of AI

The evolving Microsoft tactical cloud platform

Operating at the Tactical Edge

In our previous paper, we introduced the Microsoft Tactical Cloud Platform concept. We highlighted the ever-evolving landscape of global defense and intelligence, and the demand for real-time data processing, analysis, and decision-making has never been a more critical focus. Cloud computing continues to emerge as a transformative technology for information-driven operations, offering unparalleled innovation, scalability, agility, security, and accessibility. The rapid emergence of artificial intelligence (AI) and language models is pushing digital advancement to the next level.

Defense organizations operating at the tactical edge are increasingly facing major challenges, broadly categorized as:

- Constraints imposed by an austere and remote operating environment
- Increased cognitive load on individuals conducting operations due to exponential growth in the volume, veracity, and velocity of data
- Survivability and the need for distributed nodal command and control

In remote or hostile environments, the ability to digitally transform is particularly complex. Significant challenges are encountered when utilizing cloud services, primarily centered around connectivity, sprawling networks, and legacy constraints. The mission environment is becoming increasingly dependent on sensors for Intelligence, Surveillance, and Reconnaissance (ISR) capability in tactical situational awareness, and developing common operating pictures. However, the ability to fuse data from multiple sources and from legacy investments in diverse technologies requires breaking down stovepipes and presenting data in a common format for AI inference.

The surge in digital technology, particularly within remote and hostile environments, has led to a substantial increase in frontline operators' complexity and cognitive load. The increasing preference for distributed nodal command and control, coupled with the necessity for rapid data analysis, has placed an unprecedented strain on operators. This is especially evident in situations requiring electromagnetic spectrum analysis, where access to upstream operations centers is often restricted. Furthermore, the integration of advanced systems like Robotics and Autonomous Systems (RAS) demands dedicated operators to focus on manual system operation, which impacts their ability to maintain situational awareness or focus on other tasks. The dilemma posed here is whether technological advancements inadvertently compromise decision-making abilities due to the heightened cognitive burden on end-users.

Microsoft is extremely well-placed to support defense and intelligence organizations and their Defense Industrial Base partners in tackling these challenges head-on. Over the last 30 years, Microsoft Research has steadily advanced vision, speech, and language technologies. Including being the first organization to reach human parity in object recognition, speech, and machine reading comprehension.

Making sense of the complex

As the battlefield operating environment becomes increasingly digital, commanders are balancing the needs of survivability against traditional HQ posture, making dispersed C2 nodes vital for survivability. In maritime, land, air, or space domains, deploying low-latency computing at mission edge is essential to support this new C2 paradigm. It will enhance decision support by ingesting, storing, processing, and exploiting data at the point of need. This shift will also see a surge in the volume of intelligence data across increasingly numerous sensor feeds, with all outputs needing to be fused and processed.

The ability to optimize data processing at the edge is constrained not only by the limitations in size, weight, and power (and hence limits on compute power) but also by the proliferation of legacy systems generating multiple types of non-common data. The solutions that we deliver must be flexible enough to operate under these constraints, and we must consider the need to provide a range of hardware types that exist or are emerging. For example, Leonardo DRS currently provides hardware to more than 160,000 military vehicles worldwide. Hence, it's important to ensure that future solutions can be exploited for these types of hardware and Generic Vehicle Architectures (GVA).

We increasingly see modern IoT architecture and Kubernetes processes in manufacturing and smart-factory developments. By integrating proprietary protocols into a common data format accessed via an assets data distribution service (DDS), we can enable seamless access to data from various defense systems such as vehicles, weapons, and sensors.

Making sense of the complex *continued*

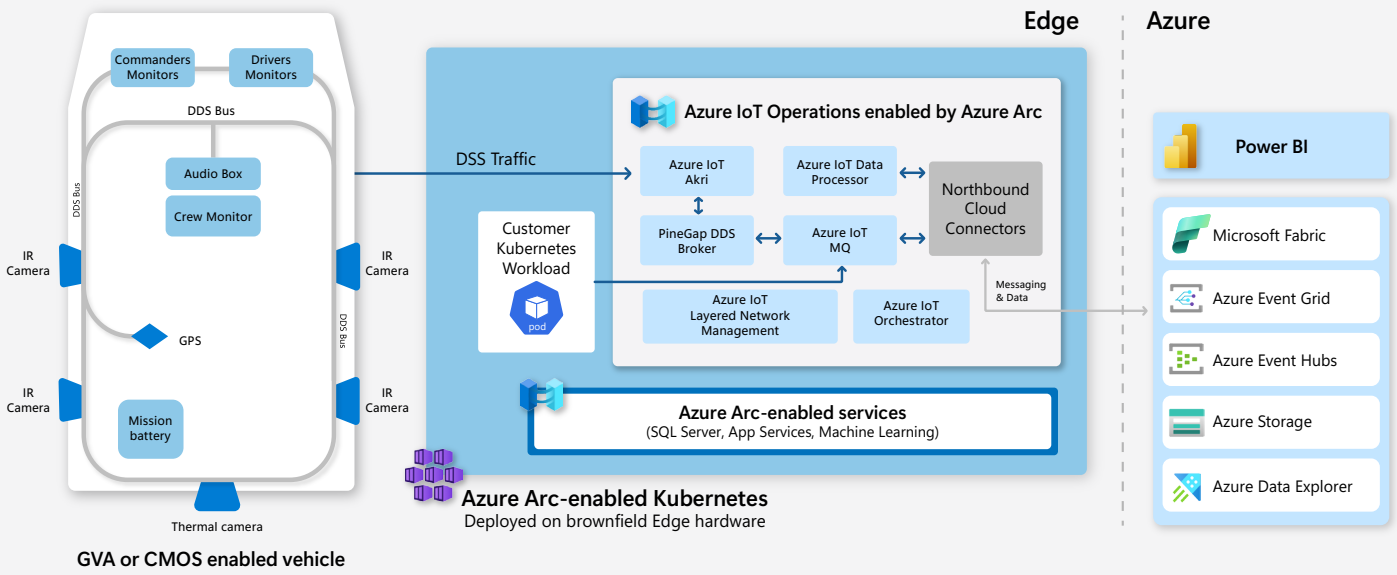

Diagram 1

This is clearly pertinent to vehicle health systems and predictive maintenance scenarios, but the application goes much further than this. The standardized data processing pipeline allows for efficient AI inference and, potentially, visualization at the source (via the Battle Management System (BMS)). As well as the ability to move data back to a headquarters cloud environment for in-depth analysis and model training. Through the breakdown of traditional stovepipes, promoting interoperability and enabling the ability to fuse multiple data sources, this approach ensures that critical information is readily available for tactical decision-making.

Moreover, by deploying AI models directly onto edge compute, the architecture supports instantaneous inference from local sensor data, significantly reducing latency and improving the accuracy of situational awareness.

Decisive action powered by AI

We have surmised that using Artificial Intelligence (AI) and Machine Learning (ML) during missions will become critical to effective C2 environments. Central to this rise in importance is the advent of language models.

In 2022, language models were limited to natural language understanding and basic image processing within their training sets. By 2023, the AI ecosystem evolved with retrieval-augmented generation (RAG), enabling models to integrate external data for complex interactions. Today, agentic AI enhances language comprehension and reasoning, decomposing instructions, executing tasks, and working alongside humans to improve operational efficiency.

This advancement has significant implications for mission capabilities, and we see early application of this technology featuring in a range of use cases, such as:

Voice Transcription/Translation – We have already seen that when paired with Push To Talk (PTT) voice radios, digital audio voice streams can be captured for real-time transcription, translations, and augmentation with other sources of data. Translation services at the edge, provided over tactical communications, have significantly interested defense organizations globally. However, the potential for capture of voice goes far beyond this. Agentic AI applications capture events for plotting in BMS and auto-associating with callsigns. Downstream, an AI agent can take a call for re-supply and automatically suggest to a human operator the tasking of a robotic/UAV asset to deliver the necessary material directly to the reported source. Provision of this tipping and cueing effect via natural voice will likely have a marked impact in reducing cognitive load for operators, increasing the effectiveness of operations for mission success. Microsoft is leading the way with extensive innovation and the use of tactical radio and AI at the edge, ensuring this is an operational reality.

Robotic Command & Control Orchestration – Microsoft Defense and Intelligence is already starting to push the boundaries of the use of GPT models to control and orchestrate Remote Autonomous Systems (RAS). With an intent to release operators from the need to manually operate these systems, we can free human resources to concentrate on the specifics of their mission and reduce the force protection overhead required to keep operators safe. This is achieved, again, by providing voice commands for the control and direction of the RAS system, which can intelligently create workflows, carry out the task (with feedback notifications as necessary), and periodically report updates or new interactions.

For example, an operator could deliver a voice command to a UAV: “take off and fly to 50m and proceed to grid reference [123456]. Once on location carry out a spiral search pattern and report when you observe any vehicles or humans within the sector.”

Decisive action powered by AI *continued*

The system automatically plans the evolution, provides verbal requests to execute the next step (as necessary), and initiates a verbal cue when there is movement in the sector.

Furthermore, the operator can query what the asset sees, request specific details, and place further commands such as "track the red vehicle." This can be achieved at the far edge, relieving operators of the need for constant focused control of assets and instead allow them to concentrate on other priority tasks.

ISR analysis for Mission – Working with multiple agents and multi-modal sensors for defense use cases, we can increase the accuracy and range of surveillance and provide a multi-layered approach to detection and action. Additionally, we can insert updated tracks into the Battle Management System to provide automated handover across the agent network, informing the human operators efficiently.

For example, in a copilot in action case, the tactical operator 'commands' (using voice) a mobile or static sensor in the battlespace to monitor a given field of view, reporting (via voice cue) when an object of interest enters that sector. If an individual is identified, the agent automatically describes a human, and the operator asks the AI agent to calculate their location in two minutes based on current movement. The agent forecasts the individual's position, and the operator then directs the agent to transfer 'overwatch' to a UAV operator, while other data, such as OSINT and ELINT, is automatically fused to provide cross-pollination, enrich the intelligence picture, and provide audio cues to the operator as the picture develops.

Query of Battle Management Systems – Across the use cases captured above, we have demonstrated several examples of how the outputs received can be automatically fed into the BMS. This capability can be further enhanced by using AI agents to query mission data that is both internal and external to the BMS. Commanders and staff can articulate verbal queries to interrogate the operational overview (the Common Operating Picture (COP)) to support decision-making and achieve decision advantage over an adversary. For example, a J4 logistics planner might ask, "What is the ammunition status for Charlie Company?" or "When will ammunition resupply be required, given a medium rate of engagement?" This is a simple example that typically requires significant resources and time to address, but it also holds the potential for much greater complexity as AI use matures.

By providing verbal query options, the system allows information to be accessed more humanly and on-demand, reducing the staff effort required for briefing and data analysis. AI agents can handle the manual workload, easing the human cognitive load to enable better and faster decision-making.

Agentic AI explained

So, what do we understand about the advancement and application of Agentic AI? When discussing Agentic AI, it's crucial to highlight the characteristics that distinguish an agent from tools like ChatGPT or traditional co-pilot assistants seen in office settings. There are five key nonlinear elements that define agentic capabilities:

Planning – Instead of diving right into a task, an AI agent can step back and plan. This structured approach prevents errors, as we often see in traditional language model implementations with robots. For instance, when instructing a robot to walk ten meters, sit down, and take a photo, it executes all the commands at once, resulting in failure. Proper planning by the agent means that a series of commands can be broken down into individual steps. The AI agent can reason out the planning and ensure that each step is placed in a logical order. In addition, the planning element can also call on the 'reflection' capability to determine that each step has been completed successfully and that it is appropriate therefore to move to the next step.

Reflection – Current models like ChatGPT provide answers but don't validate them, as they lack a built-in 'reflection' capability. Sure, you can tell the model that the response isn't correct, and it will return a different answer, but this is sub-optimal and creates issues for safety critical operations that we see in military use cases.

GitHub Copilot, another pertinent example, is an excellent aid to writing code snippets but it falls short of verifying if the code it produces actually works. In an optimal scenario an agent would not only generate code but also run it, fix errors, and revalidate until everything functions correctly. This ability to 'reflect' and ensure completeness is crucial to confirm that tasks are executed properly and is relevant to each subsequent step in the Agentic AI lifecycle.

Use of tools – The third element refers to functions where a task surpasses what an AI knows, and it therefore needs external support to execute. For instance, in our robotic example, OpenAI does not inherently know how to control robots. To solve this dilemma, we can create a tool. For example, by writing code that controls the robot, and then providing OpenAI with a manual that explains each function, including how to move the robot forward, the necessary parameters, and the expected results. When the AI encounters a step it can't perform, it checks its manual for a corresponding tool, gathers the needed information, executes the task, and processes the response.

Today's ecosystem includes tools like search engines, math tools, image generators, and analytical tools, which enhance extensibility. This is crucial for proprietary industry capabilities, allowing handoffs to external sources.

Agentic AI explained *continued*

Collaboration – Next we have multi-agent collaboration. This is important for two reasons: Creating clear boundaries and ensuring agents are task specific. Agents should focus on specific tasks to avoid confusion from handling too many things, much like microservice APIs that handle tightly coupled domain models. For example, one agent manages the camera, another oversees movement, and a third handles voice interaction, all managed by a planner agent. This ensures logic and that validation occurs within each specific domain.

This concept can be expanded, allowing different types of agents to collaborate. For example, an agent might work with a UAV, handing off tasks like following an object. A key point is that all interactions can be conducted through natural language, offering flexible workflows and enhancing autonomous capabilities.

Memory – This cycle is further powered by memory, where the agent retains and can recall prior inputs, actions, and outcomes. With this capability, the agent learns from past decisions, improving future actions and refining its planning and reflection.

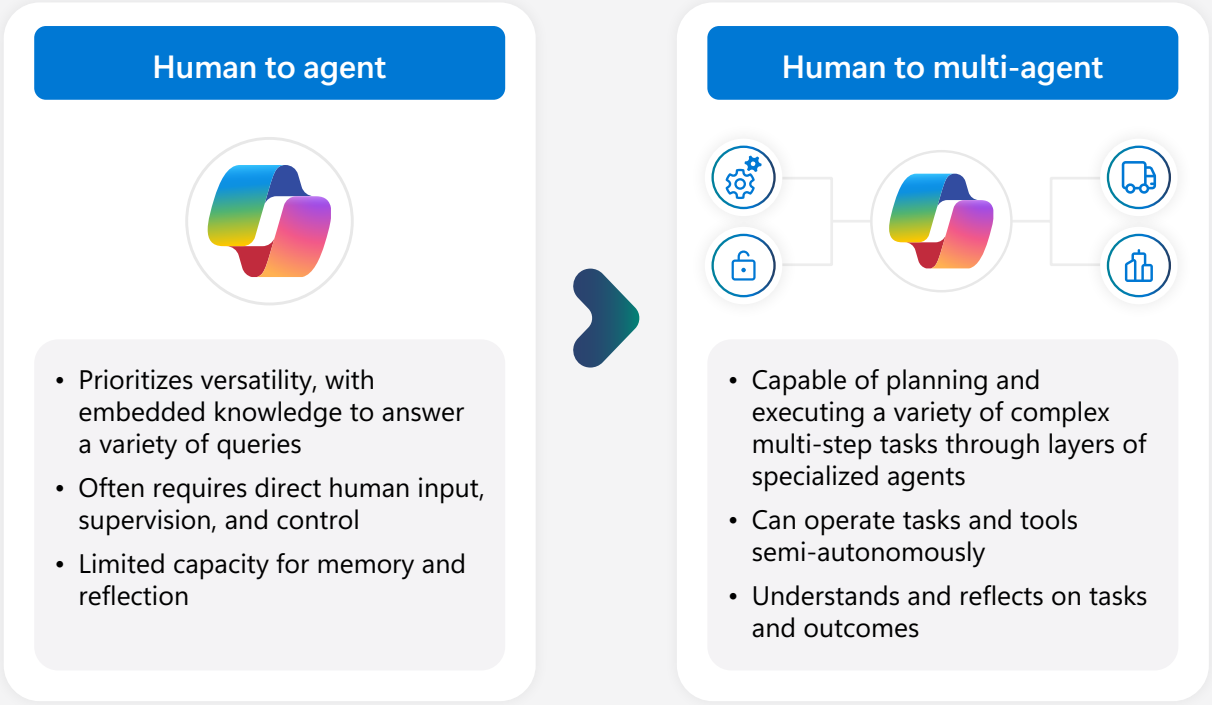

Diagram 2

Agentic AI explained *continued*

REACT framework – Collectively, these five characteristics form a framework known as REACT, which is shorthand for reasoning and action. Reasoning involves planning and reflection, while action is about the execution. Within the process, using tools and delegating to agents is crucial, as is the concept of memory. For instance, ChatGPT builds a memory of operations and conversations. This is crucial in agentic cycles, as it helps track previous actions, successes, failures, and the given instructions throughout each process step. Together, this ensures a solid understanding of what has previously transpired.

The key difference between traditional non-agentic AI workflows, often seen in zero-shot prompts, and the more advanced, agentic workflows we've been discussing can be seen in diagram 2. On the left of the image, we see the non-agentic workflow. This is typical of earlier language models, where you give the AI a task—like writing an essay—and it just dives in, starting from point A and pushing through until the task is complete. In this case, it's expected to write an entire essay on a given topic without checkpoints or reflection. There's no opportunity to revise, correct, or adapt during the process—it's a one-way execution. The AI doesn't stop to think about whether it's heading in the right direction or whether any adjustments are needed. This approach is functional but has clear limitations, especially when higher accuracy or more complex reasoning is required.

Now, compare this to the agentic workflow on the right. Here, the agent doesn't simply start and finish the task in one go. Instead, it begins by creating an outline for the essay, considering what research might be needed. It then moves to writing a first draft but remains aware that this draft is just an initial step. The agent will reflect on the draft, determining if parts need revision or if additional research is required. This cycle of reflection and revision is where the real power of agentic AI comes into play. The model doesn't just execute a task; it actively thinks about what it's doing, adapts, and seeks improvement throughout the process. Furthermore, the agentic cycle also allows for collaboration between agents; in this example, it could call on another agent that has particular research skills or knowledge, etc. It may also call on other tools that are not generally available, perhaps an instruction set for the formatting of the essay that is very specific to the organization concerned.

This comparison highlights the shift from rigid execution to dynamic, reflective task managements—making agentic AI far more suitable for real-world, complex scenarios where tasks are not always straightforward and often benefit from adaptation and collaboration.

Agentic AI for mission

As we move from discussing the theoretical aspects of agentic AI, let's dive into its real-world applications, especially in the defense sector. Speed, precision, and data are critical on the modern digital battlefield. We've seen major advancements in command-and-control environments in places like Eastern Europe and the Middle East. The key is to process information in near real-time while keeping humans central to decision-making.

Enter human-machine teaming. Picture modern soldiers working with AI as their digital co-pilots. Instead of dealing with keyboards or complicated interfaces, they can use natural language, or voice commands through military radios. Imagine an operator talking to AI systems just like they would with a colleague: "Identify threats at these coordinates" or "What's the status of drone operations?" This hands-free, seamless interaction improves situational awareness and enhances decision-making by combining AI's analytical power with human intuition and judgment.

The result? More efficient missions, quicker responses, and a trusted pairing of humans and machines. This allows warfighters to focus on tactical operations while AI handles data processing and situational analysis in the background.

Agentic AI for mission *continued*

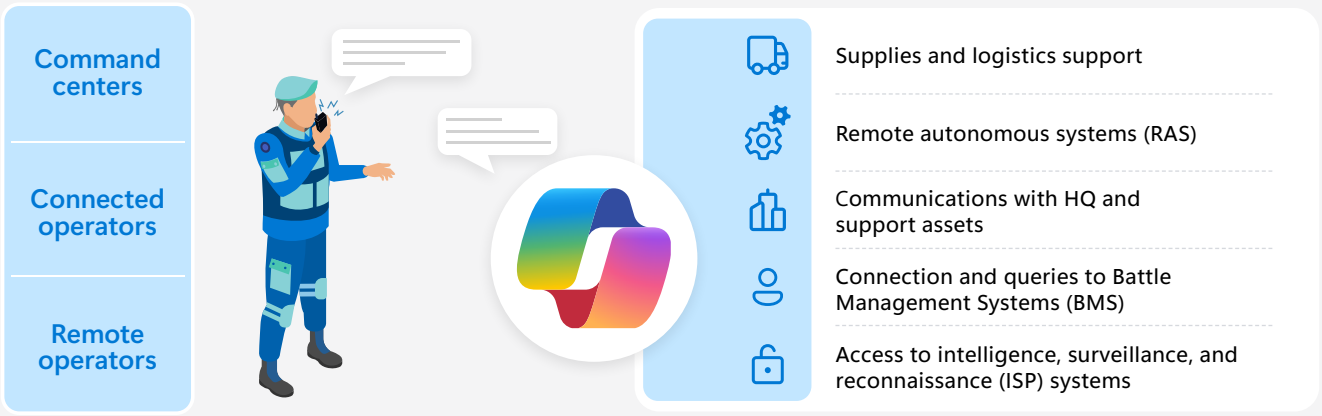

Diagram 3

As we delve further into the concept of agentic AI, it's crucial to understand how these capabilities can be seamlessly integrated into an organization's workflow, especially for field operators. This is where tactical copilots come into play.

Copilot is an AI-supported assistant that allows operators, particularly those in forward positions, to delegate specific tasks using natural language. These AI copilots are designed for real-world conditions. They are multimodal, meaning they can handle different data types such as video, sensor feeds, or command-and-control tasks like issuing instructions to autonomous systems.

One of the standout features of copilots is their user interface design. Generally operated via voice communication over existing radio networks or lightweight chat clients, these systems offer a seamless interaction experience. Commands as simple as "Scan this sector for movement" or "Deploy UAV support to grid coordinates" highlight how easy these tools are to use. Moreover, these AI agents are built to function even when communication channels are compromised or unavailable. This resilience is essential for mission-critical operations, ensuring that copilots continue to operate based on pre-collected data and seek human confirmation when necessary.

Copilots can also autonomously manage long-term tasks, periodically checking in with operators. For example, once an operator initiates a reconnaissance mission, the copilot can take over, providing updates and requesting feedback as needed. This ensures that humans stay in control while offloading repetitive or intensive work to the AI.

Incorporating tactical copilots into your workflow can revolutionize the way your organization handles complex operations. By offering an intuitive interface, robust performance under duress, and the ability to manage tedious tasks, these AI assistants ensure that operators can focus on what really matters—making critical decisions in dynamic environments.

Microsoft AI principles

Responsible AI solutions that reflect principles rooted in timeless values

Microsoft is committed to advancing AI through principles that put people first.

We put our responsible AI principles into practice through the AI, Ethics, and Effects in Engineering and Research (Aether) Committee, as well as our Office of Responsible AI (ORA). The Aether Committee advises our leadership on the challenges and opportunities presented by AI innovations. ORA sets our rules and governance processes, working closely with teams across the company to support the effort.

Microsoft AI serves to enhance human capabilities, not replace them. It's designed to embody principles such as fairness, inclusivity, reliability and safety, transparency, privacy and security, and accountability. By using AI to optimize administrative functions and services, stakeholders can focus on what matters most: human-centered design, decision-making, and empathy.

In public justice and safety agencies, decision-makers should consider not just what AI can do, but what it should do to innovate in a reliable and trusted way. It's critical to understand the components of a holistic approach to AI that will help agencies turn meaningful innovation into actionable results for their communities. The AI business school learning path offers a blueprint and roadmap designed for decision-makers to realize the true transformational impact of AI in the public sector.