From Sensors to Decisions

MixCache.com

Table of Contents

Introduction

Real-time decision-making under uncertainty has always been the heartbeat of combat systems. Today, that heartbeat is increasingly synchronized by software. From Sensors to Decisions: End-to-End Architectures for Real-Time AI in Combat Systems is a technical primer for engineers and architects who must turn heterogeneous, high-rate sensor data into timely, trustworthy actions. Rather than treating sensing, computation, modeling, and decision logic as separate crafts, this book takes an end-to-end perspective: what matters is the latency, reliability, and fidelity of the entire loop from first photon or packet to an informed human decision.

End-to-end thinking starts with mission context and constraints. In contested environments, assumptions we enjoy in enterprise AI—abundant compute, stable networks, clean labels, and generous latency budgets—do not hold. Power, size, weight, and cooling are tight. Links are intermittent or denied. Data is noisy, incomplete, or deceptive. Against this backdrop, the engineering challenge is to design architectures that degrade gracefully, prioritize the right computations at the right time, and surface explanations that enable operators to act with confidence.

At the front of the pipeline, modern platforms host a tapestry of modalities—RF and radar, EO/IR imagery, acoustic and seismic cues, navigation sources, and more. Each brings distinct phenomenology and uncertainty models. Turning these streams into coherent situational awareness requires precise time synchronization, careful calibration, and schema discipline so that downstream components can reason over shared coordinates and semantics. This book outlines practical patterns for time distribution, metadata handling, and uncertainty propagation that keep fusion tractable without erasing the nuance that makes multi-modal sensing powerful.

Pushing intelligence to the edge is essential when milliseconds matter. We will examine compute substrates—from CPUs and GPUs to FPGAs and specialized accelerators—through the lens of SWaP-C and thermal realities, and discuss scheduling strategies on real-time operating systems that guarantee deadlines. On the model side, the emphasis is on latency-aware design: architectures that are modest in footprint but rich in signal, compression techniques that preserve mission-critical accuracy, and streaming inference patterns that align with sensor cadence rather than batch-minded habits from the cloud.

Between sensors and models lies the connective tissue: streaming pipelines, middleware, and communications. In practice, success hinges on disciplined publish–subscribe topologies, backpressure and queueing strategies, and serialization choices that respect both bandwidth and evolution of schemas. Because

communications are often degraded, we focus on prioritization and summarization at the source, graceful fallback modes, and observability that works even when connectivity does not—health telemetry, embedded profiling, and lightweight traces that inform both operators and maintainers.

Decisions are not made by models alone. They emerge from policy logic that fuses estimates, confidence, and context with human judgment. We will explore decision engines that combine rule-based guards with probabilistic reasoning and learned policies, always emphasizing transparency, override paths, and human–machine teaming. Robustness is a cross-cutting concern: systems must withstand adversarial conditions, deception, faults, and drift. Security is likewise foundational; we treat zero-trust principles and secure update pathways as non-negotiable features of any fielded edge AI.

Finally, building the right system includes proving it is right. The later chapters present verification, validation, and test methods tuned for real-time AI: from simulation and digital twins to hardware-in-the-loop and instrumented field trials. We round out the journey with interoperability and standards, deployment and CI/CD at the tactical edge, and governance—ethical frameworks, accountability, and compliance with the law of armed conflict. Throughout, case studies highlight architecture trade-offs and practical lessons. The goal is not to prescribe a single blueprint, but to equip you with patterns, checklists, and mental models to design solutions that are fast, resilient, and responsible—truly end-to-end from sensors to decisions.

CHAPTER ONE: Mission Requirements and Operational Context

Before an engineer even thinks about a convolutional neural network or a publish–subscribe bus, the first questions—the absolutely foundational ones—must revolve around the mission. What are we trying to achieve? Who is the operator, and what information do they truly need to act? And perhaps most critically, under what gnarly, unpredictable, and often hostile conditions will this carefully crafted system have to perform? Without a clear-eyed understanding of the operational context, even the most elegant technical solution risks becoming a magnificent white elephant, impressive in theory but useless in the field.

Combat systems, by their very nature, operate in environments designed to confound and destroy. This isn't a data center with redundant power and air conditioning, nor is it a self-driving car navigating pristine suburban streets. We're talking about platforms that might be rocking violently on a turbulent sea, vibrating furiously in the back of a tactical vehicle, or screaming through the sky at Mach speed, all while experiencing extreme temperatures and constant electromagnetic interference. The sensors themselves might be coated in dust, obscured by fog, or targeted by lasers. This isn't just an inconvenience; it's the baseline reality.

Consider the classic intelligence, surveillance, and reconnaissance (ISR) mission. A commander needs to know what's happening over the horizon, right now. This isn't a request for a batch report that arrives next Tuesday. This is a demand for actionable insight within seconds, or even milliseconds, to cue a defensive maneuver or prosecute a fleeting target. The "decisions" in our book title are often time-critical, irreversible, and carry immense consequences. An AI that takes five seconds to identify a threat might as well take five hours if the threat is already upon you.

The concept of "real-time" itself demands careful definition within this context. In enterprise IT, real-time might mean a few seconds, or even sub-second responses for interactive applications. In combat systems, real-time often translates to microseconds or tens of milliseconds, driven by the physics of engagements. A missile traveling at hypersonic speeds closes distances terrifyingly quickly. A radar needs to process returns fast enough to track multiple objects simultaneously and predict their trajectories before they become an immediate danger. The definition of "real-time" is always dictated by the mission and the inherent timelines of the physical world the system interacts with.

Another critical facet of the operational context is the contested environment. Unlike

commercial applications where ubiquitous connectivity is assumed, tactical networks are often intermittent, low-bandwidth, and subject to active jamming or denial-of-service attacks. This fundamentally changes how we think about data flow and processing. Cloud-based AI, reliant on constant, high-speed data links to massive compute clusters, simply isn't an option for many edge applications. The intelligence, therefore, must reside *at the edge*, close to the sensors and the decision-makers. This necessitates significant ingenuity in model design, data compression, and decentralized architectures.

The adversary is not a passive observer. They are actively trying to deceive, disrupt, and degrade our systems. This introduces the concept of robustness to adversarial conditions and deception, which we will delve into later in the book. But from a mission requirements perspective, it means the AI must not only be accurate but also resilient. It must be able to operate effectively even when confronted with spoofed signals, camouflaged targets, or deliberate attempts to overwhelm its sensors with noise. The system needs to be smart enough to recognize when it's being fooled, or at least to flag uncertainty in a way that allows a human operator to intervene.

Moreover, the "fog of war" isn't just a metaphor; it's a constant factor. Information will be incomplete, ambiguous, and contradictory. Sensors will fail. Operators will be stressed and fatigued. The AI, therefore, cannot be a black box that demands blind trust. It must be able to provide context, confidence levels, and ideally, some form of explainability or provenance for its decisions. This isn't about satisfying academic curiosity; it's about enabling a human operator to understand *why* the system is recommending a particular action, especially when the stakes are incredibly high. The interface between machine intelligence and human cognition becomes paramount.

The sheer diversity of combat platforms also drives architectural requirements. An AI system designed for a small, unmanned aerial system (UAS) with strict power and weight constraints will look vastly different from one deployed on a large naval vessel with significantly more compute and power available. Similarly, the specific threats and operational profiles for air defense differ dramatically from those for ground-based reconnaissance or undersea warfare. Each domain brings its own unique set of sensor modalities, environmental challenges, and latency requirements. There is no one-size-fits-all solution; context is king.

When defining mission requirements, engineers often categorize them into functional and non-functional requirements. Functional requirements describe what the system *does*: "The system shall identify enemy vehicles," or "The system shall track targets within a 100km radius." Non-functional requirements, however, are equally, if not more, important for combat AI. These define *how well* the system performs its functions, often under duress. Think of metrics like latency, accuracy, precision, recall, robustness, availability, maintainability, and security. These "ilities" are where the rubber meets the road in tactical AI.

Latency, for example, isn't just a single number. It's an end-to-end measure from the moment a photon hits a sensor to the instant an actionable decision is presented to an operator. This encompasses sensor acquisition time, signal processing, data transmission, AI inference, and the rendering of information on a display. Each stage contributes to the overall latency budget, and optimizing one stage in isolation without considering the others is a fool's errand. We will explore how to dissect and minimize these latencies throughout the pipeline.

Accuracy and precision are also more nuanced than they might appear in a textbook. What constitutes "acceptable" accuracy depends entirely on the mission. For identifying a high-value target, precision might be paramount, even if it means sacrificing some recall. For early warning of a potential threat, high recall might be prioritized, even if it comes with a higher false alarm rate that a human can filter. These trade-offs are not technical decisions alone; they are driven by operational doctrine and risk tolerance.

The concept of "graceful degradation" is another non-functional requirement that looms large in contested environments. Unlike a commercial application that might simply crash when a network connection is lost, a combat system needs to continue operating, albeit perhaps with reduced capabilities. This might mean prioritizing critical functions, reverting to simpler models, or relying more heavily on human input. Designing for graceful degradation requires architectural foresight, anticipating failure modes, and building in resilience from the ground up. It's about ensuring the system doesn't just stop working; it adapts to adversity.

Maintainability and deployability are also crucial considerations that often get overlooked in the initial excitement of building cutting-edge AI. Combat systems have long operational lifespans, often spanning decades. This means the AI models and software must be updateable, debuggable, and extensible in the field, sometimes by personnel with limited technical expertise. Over-the-air updates, secure software delivery, and robust health monitoring systems are not optional luxuries; they are mission imperatives. The ability to deploy a patched model to a remote platform in a contested zone is as vital as the model's accuracy itself.

Security, of course, is foundational. In combat systems, the consequences of a compromised AI can be catastrophic. Zero-trust architectures, secure boot, encrypted communications, and tamper-resistant hardware are not buzzwords; they are essential safeguards against malicious actors attempting to inject false data, alter model weights, or disable critical functions. The entire chain, from sensor to decision, must be secured, recognizing that every component is a potential vulnerability.

Understanding the specific platform constraints is also paramount. Power, size, weight, and cooling (SWaP-C) are often the most severe limitations at the edge. A drone might

have only a few watts of power available for its entire compute payload, demanding incredibly energy-efficient processors and highly compressed models. A fighter jet has strict weight limits and needs to dissipate significant heat generated by its electronics. These physical realities directly dictate the types of sensors, processors, and AI architectures that are even feasible. Ignoring SWaP-C constraints early in the design phase is a surefire way to engineer a solution that will never leave the lab.

The human element remains central. Even the most advanced AI in combat systems serves to augment human decision-makers, not replace them entirely. Therefore, understanding the cognitive load on operators, their training levels, and their existing workflows is critical. The AI's outputs must be intuitive, timely, and presented in a way that minimizes confusion and maximizes comprehension. This requires a deep understanding of human factors engineering and careful design of human-machine interfaces. The goal is a seamless partnership, where the AI handles the data deluge and presents distilled insights, allowing the human to focus on higher-level judgment and strategy.

Finally, the regulatory and ethical landscape adds another layer of complexity. The use of AI in combat systems raises profound questions about accountability, bias, and the Law of Armed Conflict (LOAC). While these are subjects for a later chapter, they are fundamentally mission requirements. Any AI system deployed in a combat role must be designed and verified to operate within these legal and ethical frameworks. This means building in safeguards against unintended consequences, ensuring transparency in decision-making processes, and rigorously testing for biases that could lead to discriminatory outcomes. These aren't just "soft" requirements; they are hard constraints that shape the entire architectural design.

In essence, Chapter 1 serves as a stark reminder that technology for technology's sake has no place in combat systems. Every technical choice, from the type of sensor to the neural network architecture, must be traceable back to a clear operational need and evaluated against the harsh realities of the contested environment. The foundation of a successful real-time AI system begins not in a silicon foundry or a data science lab, but in a deep, empathetic understanding of the warfighter's mission and the unforgiving world they inhabit.

This is a sample preview. Purchase the book to read the full content.

Visit MixCache.com to purchase the complete book.