CA2X2 FORUM 2026

AI-Enabled Digital Wargaming for Responsible Multi-Domain Decision Support in a NATO Simulation Ecosystem

Stian Osmundsen Chief Executive Officer, Levato AS

Luis E. Velazquez Chief Technology Officer, Levato AS

Espen Askvik President, Levato AS

Track: Experimentation, Exercises and Wargaming for Warfighting Readiness · AI Capability Development · Decision Support · Wargaming

Unclassified · Levato AS proprietary

THE PROBLEM

Awareness has outpaced decision

Data fusion and a shared operating picture describe the present with authority. Converting that picture into analysed, defensible options remains a manual staff process.

1

Combinatorial

The decision space branches across timing, sequencing, allocation, posture and risk. Staffs typically compare three COAs, qualitatively.

2

Accountable

A recommendation that cannot be explained in operational terms cannot support a decision under national caveats and the Law of Armed Conflict.

3

Time-bound

Analysis that arrives after the decision point has no operational value. A capability needing months of development per mission fails this test.

Two families of AI are currently applied to this problem. Each fails a different one of these three tests.

Neither method is sufficient alone

LARGE LANGUAGE MODELS

Generality without traceability

- Prone to hallucination. In a crisis, confidently wrong is worse than openly uncertain.
- Reasoning emerges from billions of general-text parameters — not auditable in operational terms.
- Handle time-series and positional data poorly — the sequencing and geometry that decide outcomes.
- Structurally dependent on data to retrieve, which at this level is scarce or classified.

REINFORCEMENT LEARNING

Traceability without tractability

- Learns from observable interaction. Behaviour can be inspected, reproduced, corrected.
- Well proven where the state space is bounded and the reward structure is clear.
- Cost grows faster than complexity: multiple domains, echelons, long horizons, adaptive adversaries.
- No independent means of learning what a commander treats as acceptable risk.

The failure modes are complementary, not shared — which is the observation everything that follows is built on.

Three methods, each used where its strengths lie

1

Large language models

INTERPRETATION & PREPARATION

Interpret strategic context, handle natural language interaction, generate simulation-ready scenarios, synthesise outputs. Grounded in doctrine and verified data through retrieval — not open internet sources.

Provides generality

2

Imitation learning

HUMAN INTENT & RISK

5–20 human-versus-human digital wargames, 20–120 minutes each. Decisions, sequencing, implicit risk acceptance and outcomes are captured automatically as guidance data.

Provides judgment

3

Reinforcement learning

SCALE & TRACEABILITY

Thousands of simulation iterations exploring alternative decisions, adversary reactions and outcomes across the parameter space, per role and per domain.

Provides traceability

The language model never makes the operational recommendation. It frames the problem and synthesises results — the simulation-trained agents make the decisions.

THE CONSTRAINT

The training data does not exist

Rare	High-level operational missions are infrequent by nature. There is no large corpus to learn from.
Classified	Where data exists it is frequently restricted, blocking conventional machine-learning pipelines.
Non-transferable	Ukraine drone lessons are specific to that terrain, EW intensity and adversary culture. They do not port to the High North.
Biased by era	Historical sets encode the force structures and threat perceptions of a particular time — biases invisible to those reusing them.

Consequence: a single universal agent forces excessive abstraction. Mission-specific agents are a strategic necessity, not a technical preference.

Current simulation cannot supply the data

Fidelity–scale gap

Tools optimise for precision or for force-level scale — rarely both in one framework.

Scalability limits

Physics-only environments do not scale to multi-day, multi-domain, branching missions.

Black-box models

Hidden assumptions transfer silently into agent behaviour and defeat explainability.

Training time

Generating and validating data exceeds planning timelines. Compute is not the constraint.

Fragmented data

Licensing, proprietary formats and closed ecosystems block joint and coalition integration.

Single-agent design

Most AI assumes one agent and one environment. Command is hierarchical and distributed.

Seventh limitation: scripted adversaries. Predictable OPFOR narrows the behaviour agents ever encounter, and weakens what they learn.

Five steps, running in parallel with planning

1

Historical base training

Doctrine, historical data, open-source operational knowledge.

2

Mission contextualisation

Calibration to theatre, adversary and constraints from C2, intelligence and repositories.

3

Human-led wargames

5–20 iterations capturing intent, sequencing and risk acceptance.

4

Synthetic mass training

Thousands of parallel runs across the defined parameter space.

5

Mission-ready agents

COA evaluation, planning and live decision support. All decisions traceable.

Human judgment sets intent, framing and acceptable risk. Machine scale explores the branches, edge cases and sensitivities that manual wargaming cannot reach.

ARCHITECTURE

Four modular, federated layers

1	**User interaction**	Secure web interface. Text, voice and structured input. Commander's intent expressed in operational language; products returned in NATO staff formats.
2	**AI orchestration & agents**	Coordinator, interpreter, ORBAT, scenario, response and orders generation. COA generation and test orchestration. RL agents for friendly and adversary roles across domains.
3	**Simulation & data generation**	Vectorized engine. Thousands of parallel or branched runs under time compression across land, maritime, air, cyber, space and EM. Transparent adjudication; structured capture of every run.
4	**Knowledge, memory & integration**	Doctrine, adversary profiles, orders formats. Persistent context across planning cycles. REST, MCP, DIS, HLA, SISO-aligned structures.

Operates standalone as an analytical environment, or as a federated decision-support node inside a wider AI-enabled warfighting architecture.

Fusion sees the present. Simulation explores the futures.

INGEST

COP state vectors, ISR-derived context, force posture and activity from fusion and C2 systems; doctrine and standing assumptions from repositories.

ANALYSE

Scenario and ORBAT generation, human wargaming, agent-driven COA generation and ensemble testing under explicit assumptions.

RETURN

Ranked COAs with trade-offs, risk and sensitivity analysis, escalation assessment, order-supporting artefacts — back into operational workflow.

Complementary, not competitive

Fusion and COP services describe the present state of the operational environment with authority. A wargaming and simulation capability explores the futures that follow from candidate decisions. Decision advantage requires both, connected through open standards rather than bespoke integration.

WORKED EXAMPLE — WHAT THE STAFF RECEIVES

High North reinforcement: two COAs, one disruption, one advisory

What we claim — and what we do not

WE CLAIM

Transparency
Open models. Variables, assumptions and adjudication logic exposed and adjustable.

Traceability
Input assumptions through simulation outcomes to recommendation — recorded and replayable.

Explainability
Decisions trace to operational cause and effect commanders recognise.

Reliability
Agents fail observably and correctably in simulation before deployment.

Human oversight
Human-in-the-loop and human-on-the-loop embedded in the workflow.

WE DO NOT CLAIM

Freedom from bias

All models embed assumptions. The defensible claim is that ours are transparent, inspectable and challengeable.

Shared accountability

Accountability is not shared with the machine. The system recommends; the commander decides; the commander is accountable.

Any autonomy in effect

This is decision support. It has no relationship to autonomous lethal systems.

Trustworthiness here is architectural, not procedural — a property of how the system is built, not an assurance layered over it afterwards.

EVIDENCE

Maturity, and what is not yet proven

TRL 8–9

Simulation and wargaming core. Sustained operational use in officer education, force training and force design experimentation.

TRL 6–7

AI pipeline. Agent training, automated scenario builder and closed RL loop demonstrated in controlled environments.

The evidence base for the simulation core does not transfer to the agent pipeline. Three limitations stated explicitly:

Inherited assumptions

Agents trained largely on synthetic data inherit the assumptions of the generating model. Transparency mitigates this; external validation against historical and exercise data is still required.

Generalisation of intent

Imitation learning from 5–20 wargames captures the risk posture of the staffs who played — which may not generalise across command cultures.

Institutional interfaces

Federated value depends on interface availability and data-sharing arrangements. These are institutional problems, not technical ones.

IN CLOSING

Not automated decision-making — better-informed commanders.

Equipped to compare COAs with greater confidence, transparency and speed — while retaining full responsibility and authority for operational decisions.

The open question for this community

What metrics and protocols should the NATO M&S community adopt to evaluate simulation-trained decision-support agents for operational credibility, before they inform live planning? That is a collective question, not a vendor one.

stian@levato.no · luis@levato.no