AI Support Systems in the Military

Division of Science and Mathematics
University of Minnesota, Morris

John Gulon

April 9, 2026

Road Map

- Why AI Decision Support Matters
- What is ReLeGSim?
- How the Model Architecture works
- How Reward & Training shape Behavior
- Experimental Results

Why is AI in the Military?

- Data rich environment
- Human limits (OODA: Observe, Orient, Decide, Act)
- Faster, Adaptive Decision Support
- Decision Support, **not** Replacement

What is ReLeGSim?

- Reinforcement Learning-based Ground Simulation
- Battalion-level Simulation (BLUE vs RED)
- Grid-based Battlefield
- AI Generates Tactical Commands
- Reinforcement Learning
- Designed by Möbius et al.

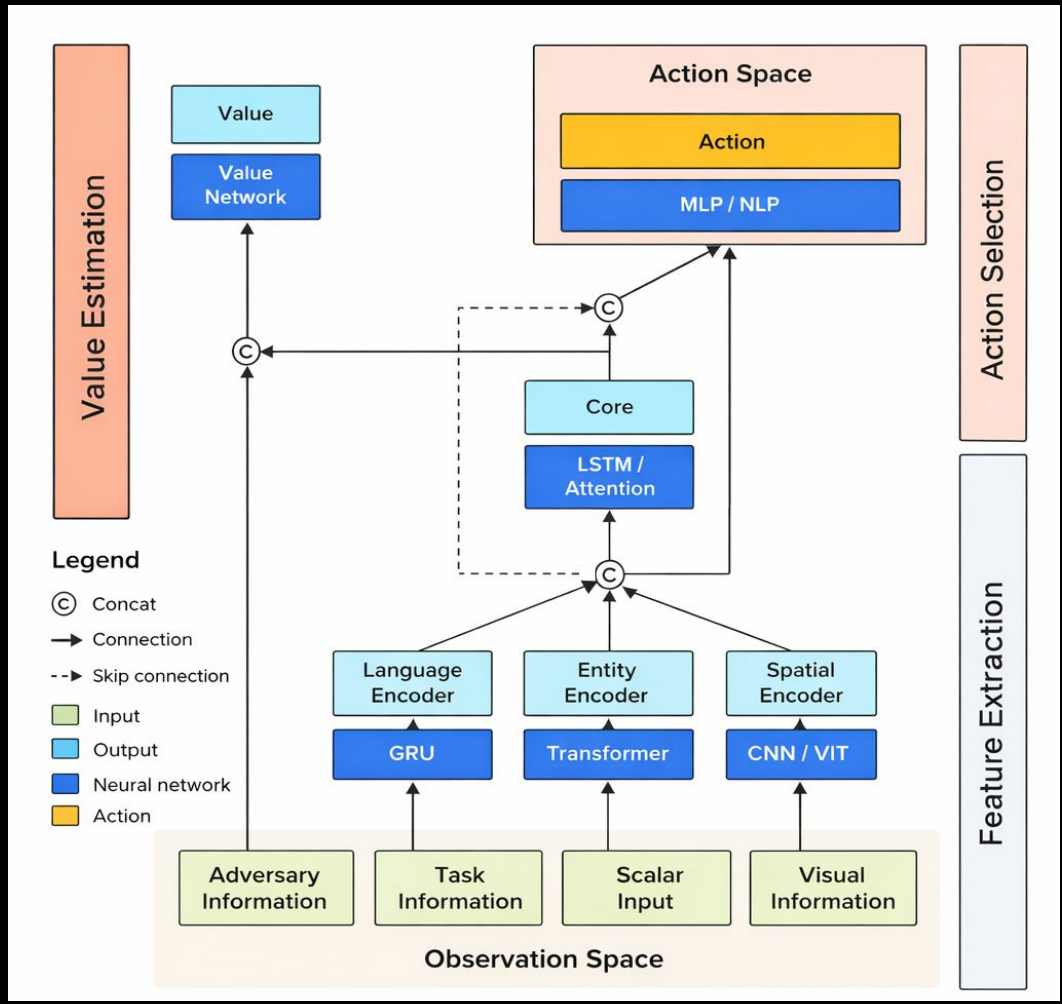
Map adapted from Möbius et al. 2024

Model Architecture

- Multi-modal input
- Feature Extraction
- Decision Network
- Action & Value Outputs

Disclaimer: Equations used are theoretical, they may or may not be used by sources.

Architecture adapted from Möbius et al. 2024



Reward Function

Measuring Long-term Success

$$R(s_t, a_t) = w_1 r_{mission} + w_2 r_{enemy} + w_3 r_{friendly}$$

Define what the AI considers “good behavior” (Policy)

s_t = state, time step. a_t = action, time step.

Why Reward Design Matters

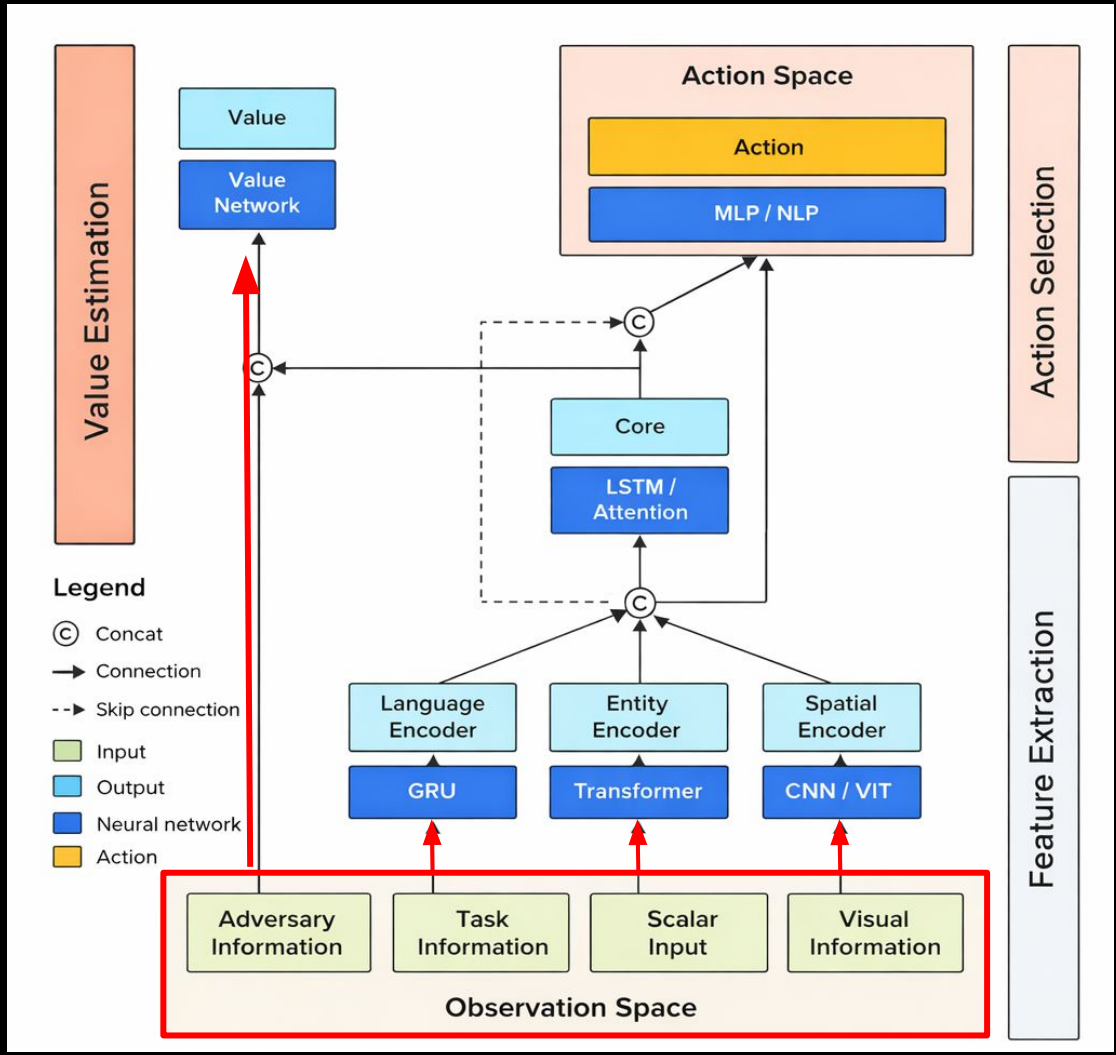
- Reward defines Behavior

  Simple Rewards = Narrow Strategies

  Multi-objective Rewards = Realistic Tactics

- Doctrine Alignment

Observation Space



Observation Space

- Adversary Information
- Task Information
- Scalar Input
- Visual Information

S_t = f(enemy, task, friendly, terrain)

Equations Road Map

1. Input State

- Enemy armored units near B4
- Task: Capture Objective, Minimize Losses
- Friendly companies w/Artillery Support
- Forest, road, open terrain (Map)

S_t = f(enemy, task, friendly, terrain)

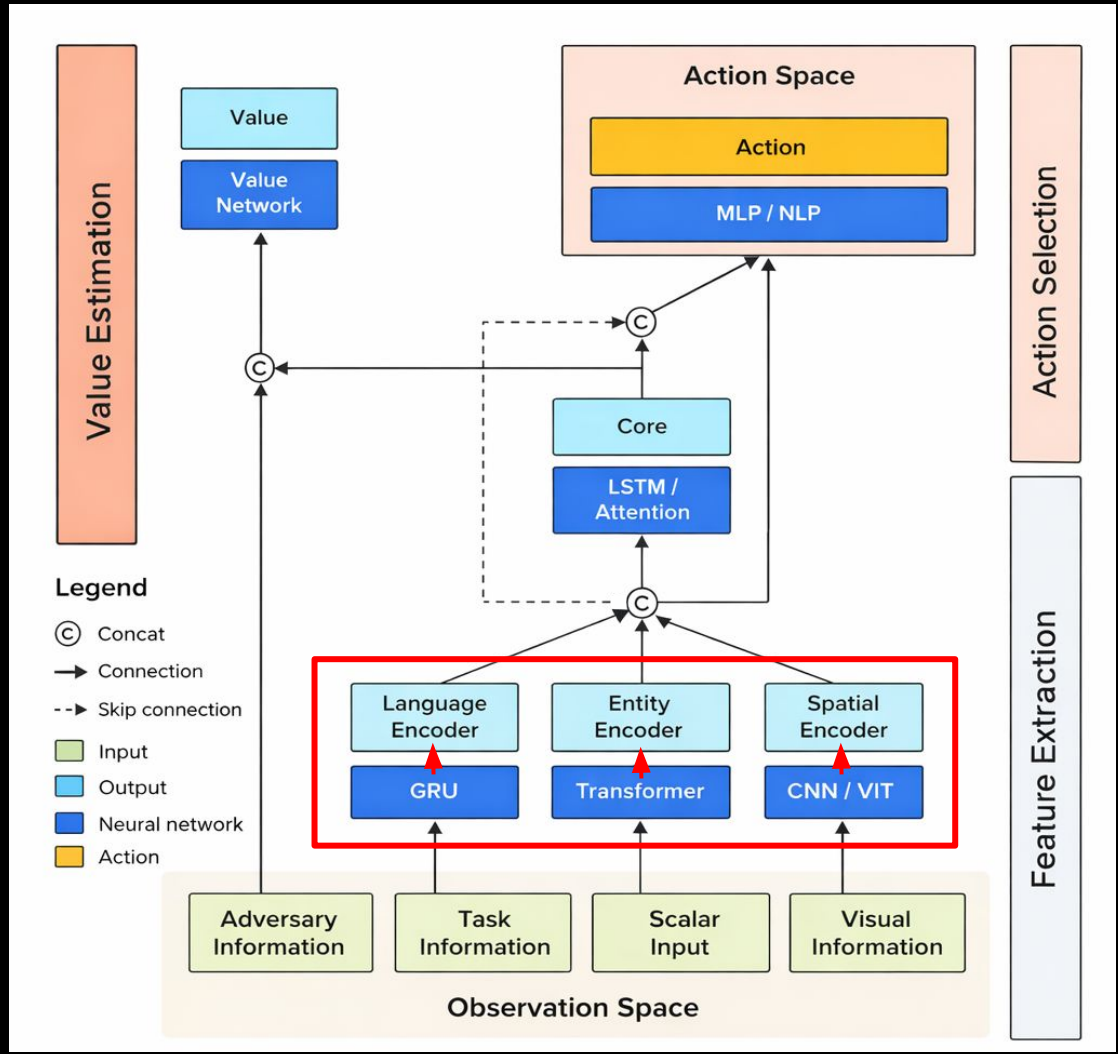
Observation (Model View)

- enemy_forces_table
- own_forces_table
- combat_support_table
- objective
- image
- prev_action
- action_mask (valid actions)

Feature Extraction

- GRU -> Command Intent
- Transformer -> Unit Relationship
- CNN/ViT -> Spatial Features

$h_t = [\, h^{lang}, h^{entity}, h^{visual}, h^{scalar} \,]$



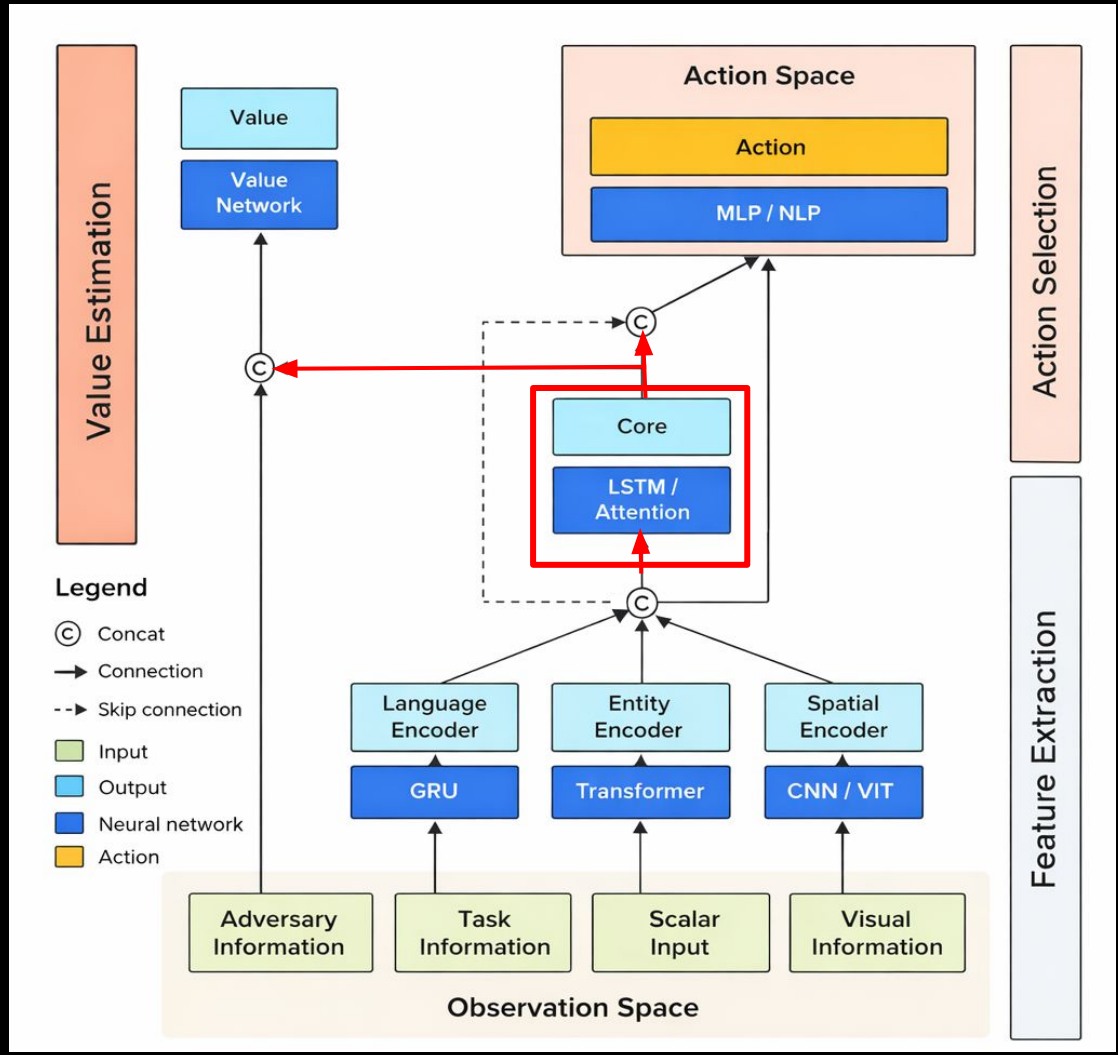
Core Network

- LSTM -> Memory Over Time

$$h_t = LSTM(h_t, h_{t-1})$$

- Attention -> Focus on Important Info

$$c_t = \sum \alpha_i x_i$$



Equations Road Map

2. Core

$$h_t = [\, h^{lang}, h^{entity}, h^{visual}, h^{scalar} \,]$$

Core (Recurrent + Attention)

$$h_t = LSTM(h_t, h_{t-1})$$

$$c_t = \sum \alpha_i x_i$$

x_i = individual elements/features inside h_t ($h_t = [x_1, x_2, x_3, \ldots]$)

α_i = learned importance weights

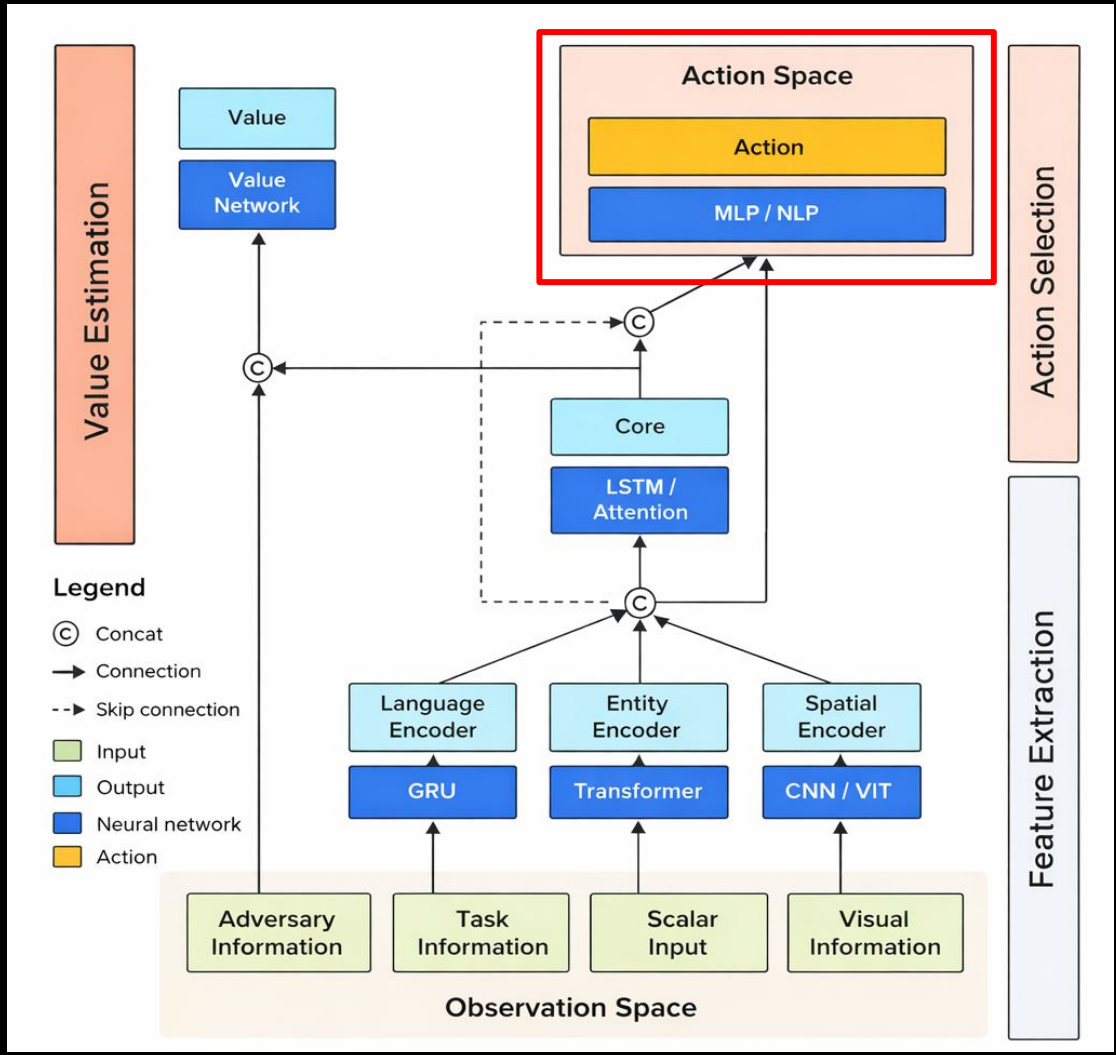
Action Selection

Outputting Commands

- Move
- Attack
- Observe

$\pi(a \mid s) = P(a_t \mid s_t)$

$a_t = \text{argmax}\ \pi(a \mid s)$



Equations Road Map

3. Action Selection

$$h_t^{core}$$

Policy (Probability Distribution)

$$\pi(a \mid s) = P(a_t \mid s_t)$$

- Maps State -> Probabilities over actions
- Constrained by:
  - Action mask
  - Available resources

Action Choice

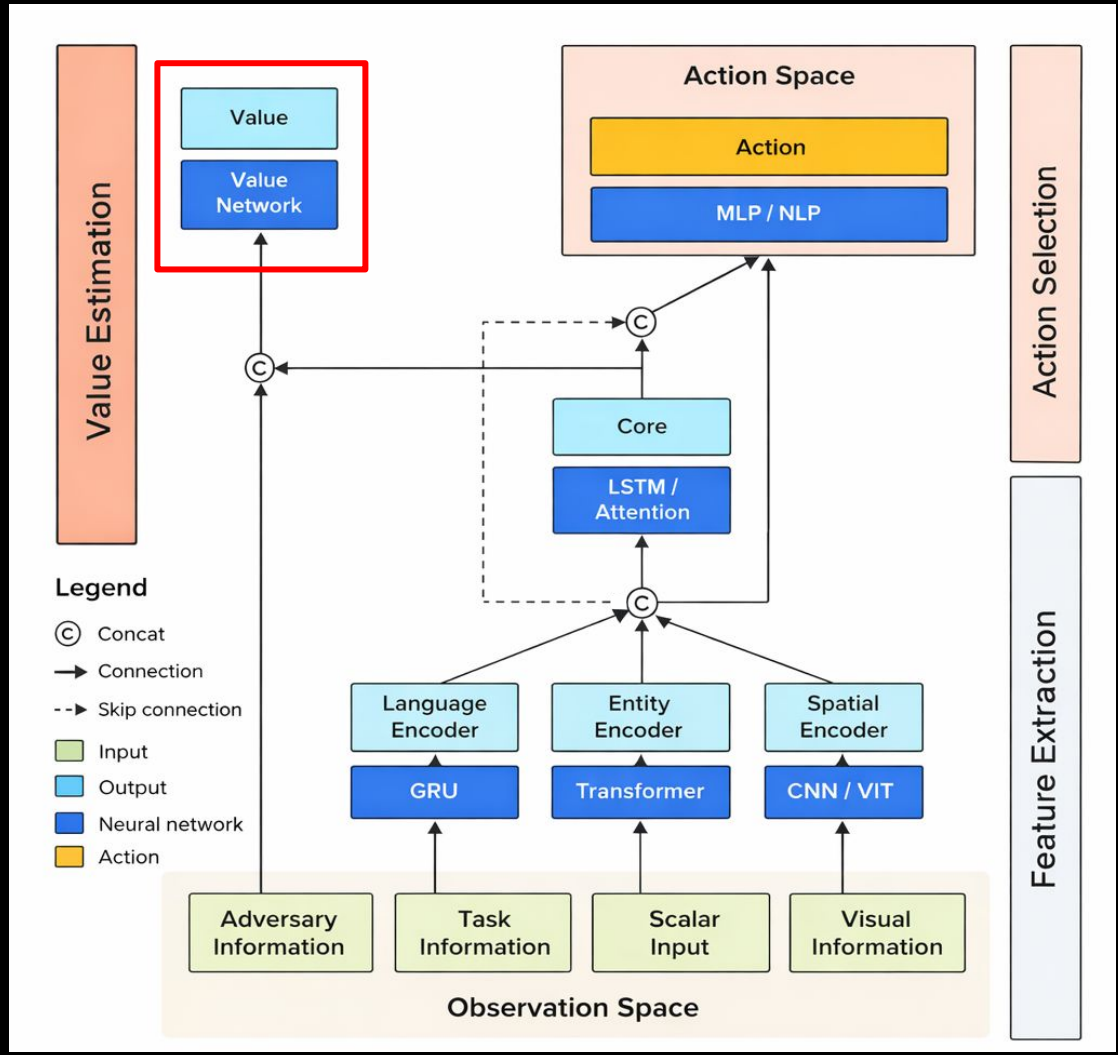
$$a_t = \text{argmax}\ \pi(a \mid s)$$

- Selects highest probability action
- Outputs command
  - Move
  - Attack
  - Observe

Value Estimation

Measuring Long-term Success

$$V(s_t) = E\left[\sum \gamma^k R\right]$$



Equations Road Map

4. Value Estimation

$$h_t^{core}$$

Value Function

$$V(s_t) = E\left[\sum \gamma^k R\right]$$

- Estimates total future reward
- Evaluates current state quality

Output: $V(s_t) \in R$, a single scalar value for state quality.

Full Learning Objective

Maximizing **Long-term** reward while following mission **priorities**.

$$J(\pi) = E\left[\sum \gamma^t R(s_t, a_t)\right]$$

Combining action, value, reward into a single objective.

Training Loop

- observe state $\mathbf{s}_t$
- compute features and core representation
- output action and value
- execute action in simulator
- receive reward and next state
- update policy/value network
- repeat over many time steps and episodes

Reward Function Reminder

Measuring Long-term Success

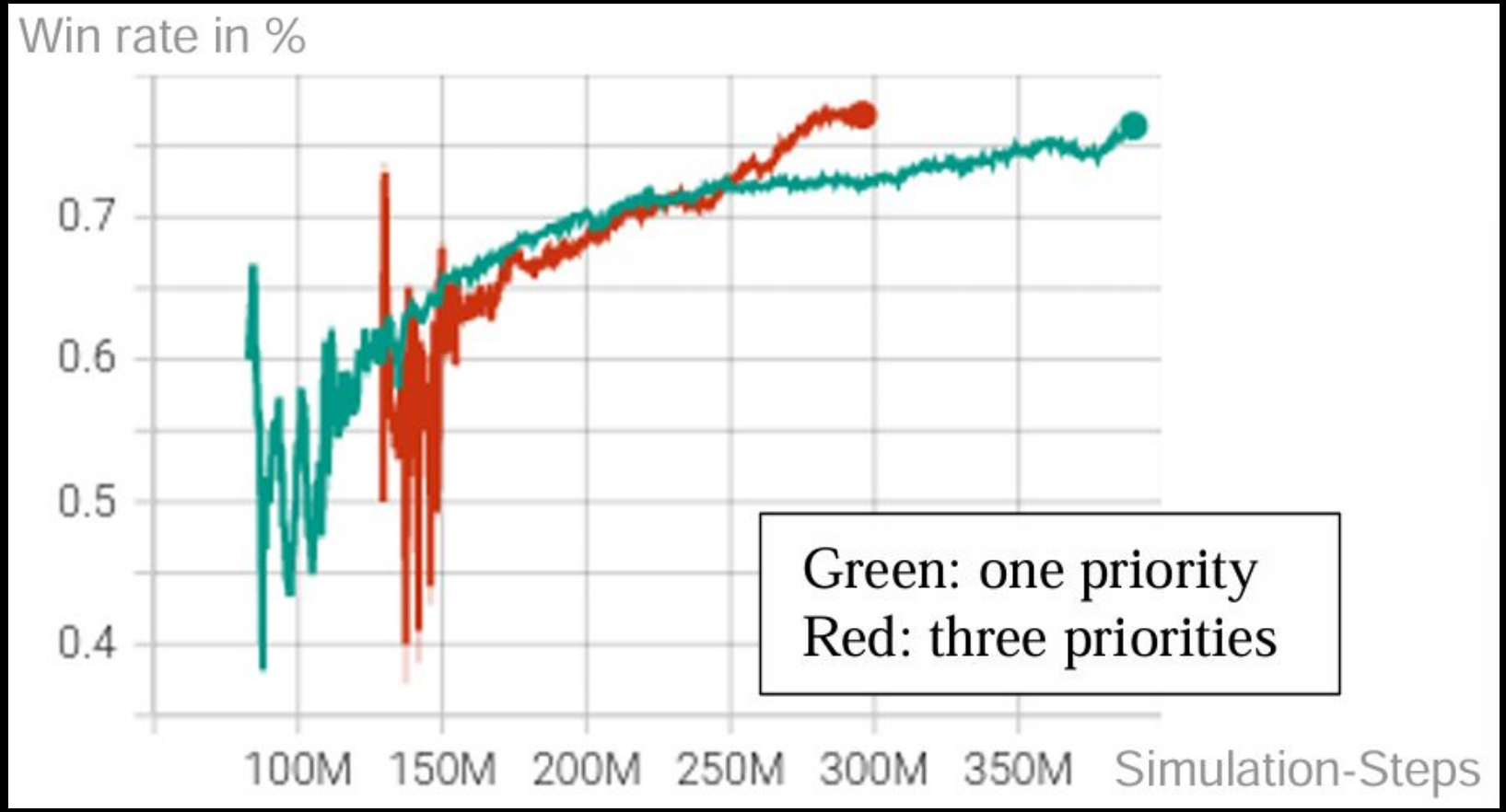
$$R(s_t, a_t) = w_1 r_{mission} + w_2 r_{enemy} + w_3 r_{friendly}$$

Define what the AI considers “good behavior” (Policy)

s_t = state, time step. a_t = action, time step.

Experimental Results

Training Results from Möbius et al. 2024



Experimental Results

Early Training

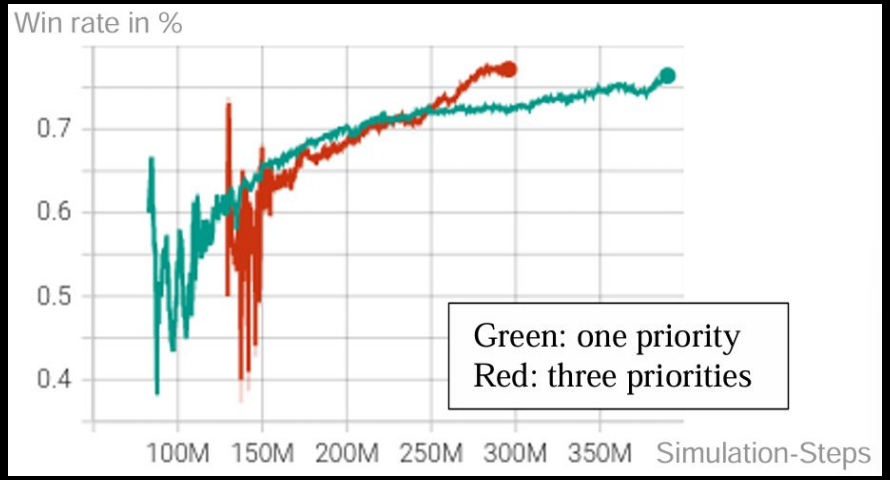
- Multi-objective = slower

Later Training

- Multi-objective = better performance

Reason

- More exploration
- More realistic strategies



Connection to Doctrinal Alignment

Results Support the Doctrine

- Real Military Decisions = Multi-Objective
- Reward = Doctrine Encoding
- Better Performance = Better Alignment

Human-in-the-Loop

- AI Suggests
- Human Decides
- Prevents Misalignment

Key Takeaways

- Multi-modal AI = Battlefield Awareness
- Reward = Doctrine
- Architecture = Decision Pipeline
- Risk = Misalignment
- Human Control = Priority

Questions?

Citations

Main Source

Michael Möbius, Daniel Kallfass, Matthias Flock, Thomas Doll, and Dietmar Kunde. 2024. Incorporation of Military Doctrines and Objectives into an AI Agent via Natural Language and Reward in Reinforcement Learning. In Proceedings of the Winter Simulation Conference (WSC '23). IEEE Press, 2357–2378.

Supporting Source

Michael Möbius, Daniel Kallfass, Thomas Doll, and Dietmar Kunde. 2023. AI-Based Military Decision Support Using Natural Language. In Proceedings of the Winter Simulation Conference (WSC '22). IEEE Press, 2082–2093.

Supporting Source