V&V of RL for Military Decision Support Systems

Two Novel Robustness Tests

Analikwu, B.O. (Brendan) | MESAS '26

Buis, C.F. (Carlijn) | Voogd, J.M. (Jeroen) | Schadd, M.P.D. (Maarten)

21 October 2026

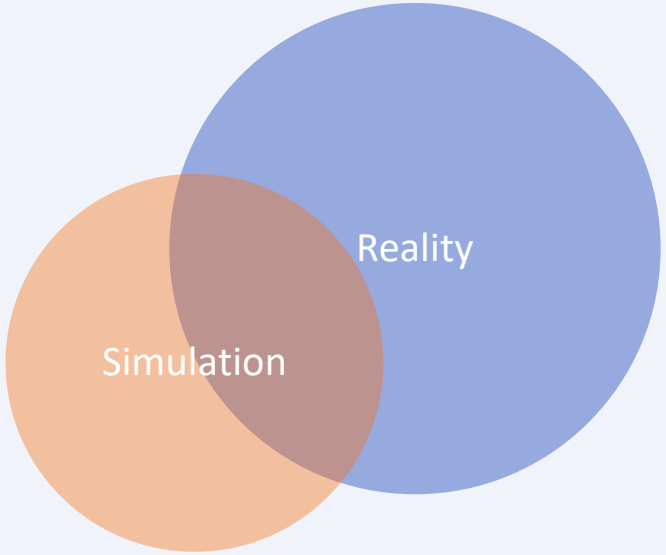


RL policies are trained in one world, deployed in another

Military reinforcement learning

- Limited real-world data
- Strong reliance on simulation

But: simulation ≠ reality → Simulation-to-reality gap

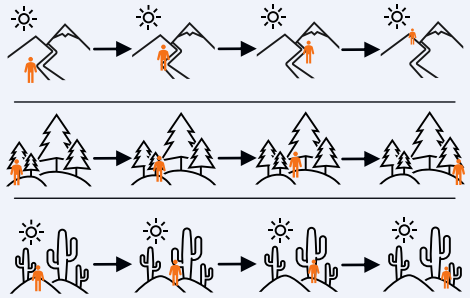


What happens when the deployment environment changes?

Why is robustness testing harder for RL?

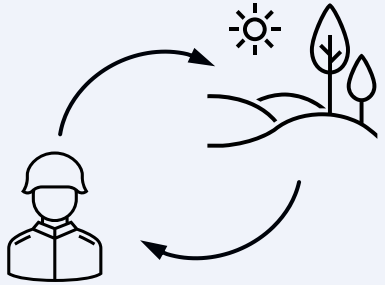
1. Scalability

Behaviour must be assessed over many time steps, scenarios and trajectories.



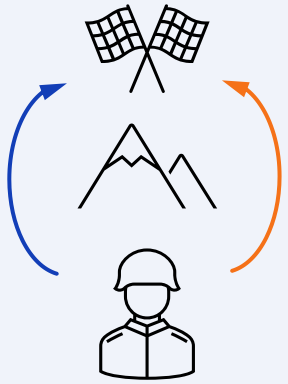
2. Sequentiality

Actions change the future states the policy encounters.



3. Non-unique behaviour

Different policies can achieve the same objective differently.



Performance alone does not tell us robustness

Problem 1

Environmental changes affect achievable performance.

E.g. introduce obstacles and performance drops



43.03s



45.80s

Performance drop may reflect harder environment, not less robust policy

Problem 2

Similar performance can arise from different behaviour.

Policy A

- Wait → Wait → Run → Run

Policy B

- Walk → Walk → Walk → Walk

Similar performance ≠ similar behaviour

We need performance evidence + behavioural evidence

Two complementary robustness tests

BNPR

Benchmark-Normalised Performance Ratio

“Does the policy still perform well relative to what is achievable?”

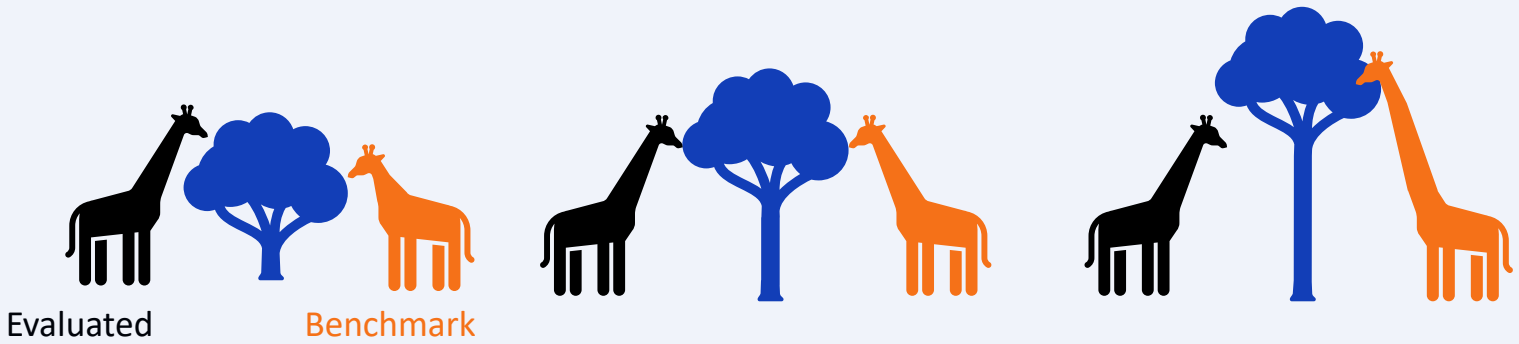
WATD

Weighted Action Transition Distance

“Does the policy make decisions differently?”

Same test principle:

Fixed evaluated policy + environmental variation + specialised benchmark



BNPR

Benchmark-Normalised Performance Ratio

$$\mathrm{BNPR}(v) = \frac{P_e(v)}{P_b(v)}$$

- $P_e(v)$ = performance of fixed evaluated policy in environment v
- $P_b(v)$ = performance of benchmark in environment v

BNPR separates changes in policy quality from changes in environmental difficulty

WATD

Weighted Action Transition Distribution

1. Record action sequences

Evaluated: Walk → Walk → Read → Walk → Hold → …

Benchmark: Walk → Read → Walk → Walk → Walk→ …

2. Count action transitions

From ↓ to →	Walk	Read	Hold
Walk	42	18	5
Read	31	2	1
Hold	12	1	7

			Hold
			3
Read	27	6	1
Hold	7	3	9

3. Convert each row to a probability distribution and compare

E.g. current action = Walk:

Evaluated: *[0.65, 0.27, 0.08]*

Benchmark: *[0.80. 0.15, 0.05]*

Then: compute total variation distance for each current action

4. Combine across current actions

Compute weighted sum of the distances

WATD measures differences in action sequences

Case study: Evaluating soldier information concepts

Goal: reach destination

- Cover a given distance
- As quickly as possible
- While maintaining situation awareness

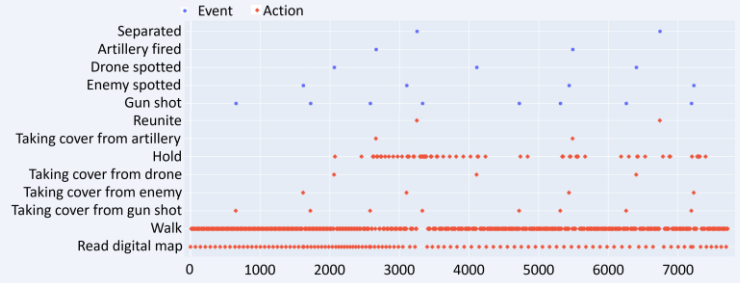
Interrupted by **events** that increase stress

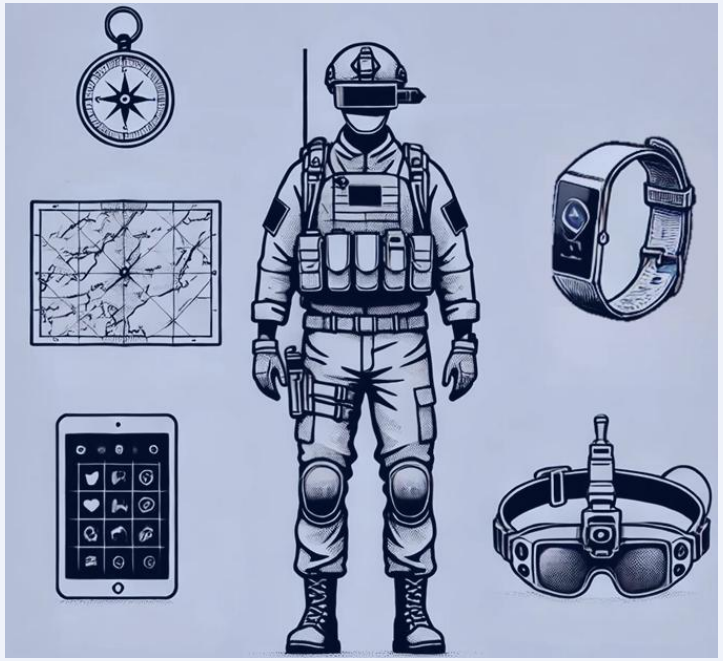
- Incoming fire
- Drones

Policy actions

- **Walk**
- **Read map** → improves speed
- **Hold** → reduces stress

Task: walk in the right direction, while managing information and stress





Experimental design

Three environmental parameters varied (one at a time)

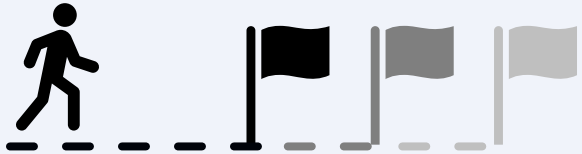
- **Travel efficiency** – Speed of soldier after obtaining information
- **Distance to goal** – Initial travel distance
- **Max. stress level** – Level of stress soldier can withstand

Benchmark construction

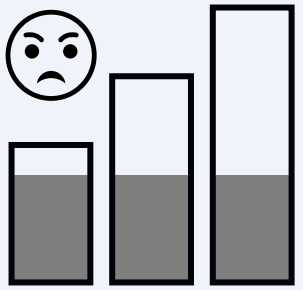
- Train 10 PPO agents per environment, select best
- 300,000 training steps
- 50 evaluation episodes



Travel efficiency



Distance to goal



Max stress level

Results

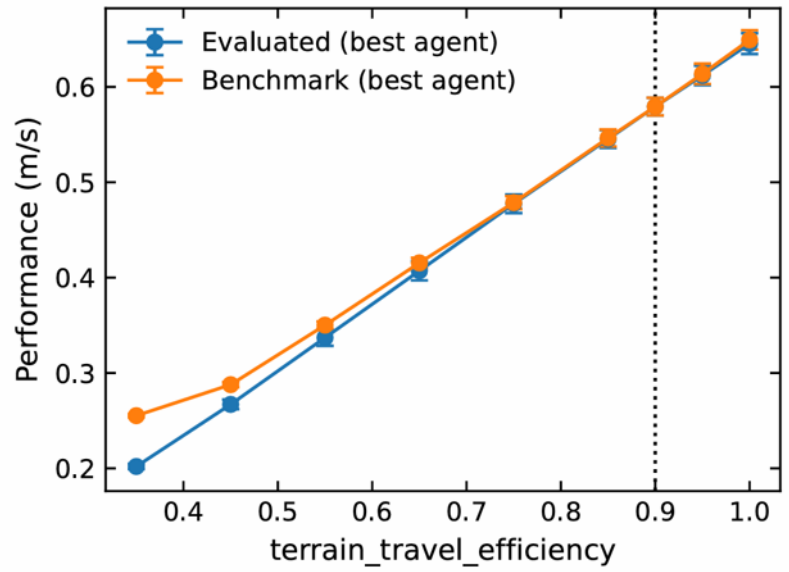
Training efficiency

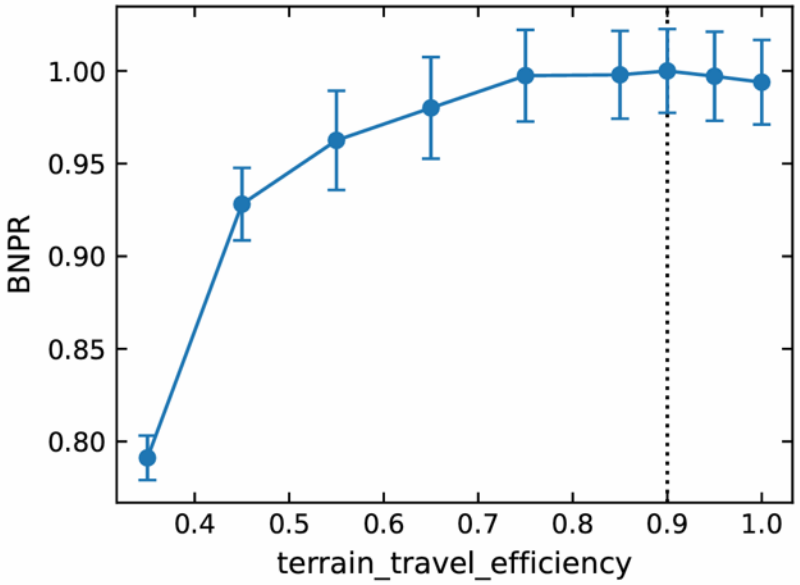
Nominal value: 0.9

- Absolute performance changes substantially
- But evaluated policy tracks benchmark

➢ Lower performance does not necessarily mean lower robustness.

BNPR reveals: much of performance change comes from the environment.



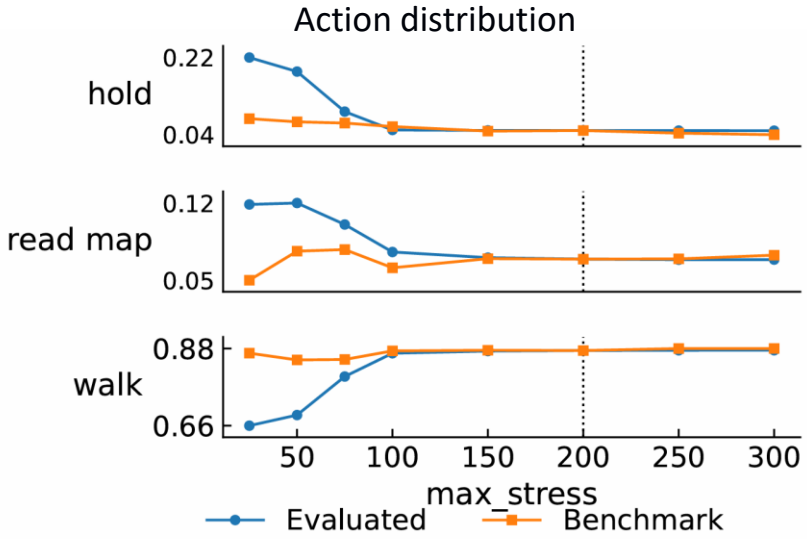


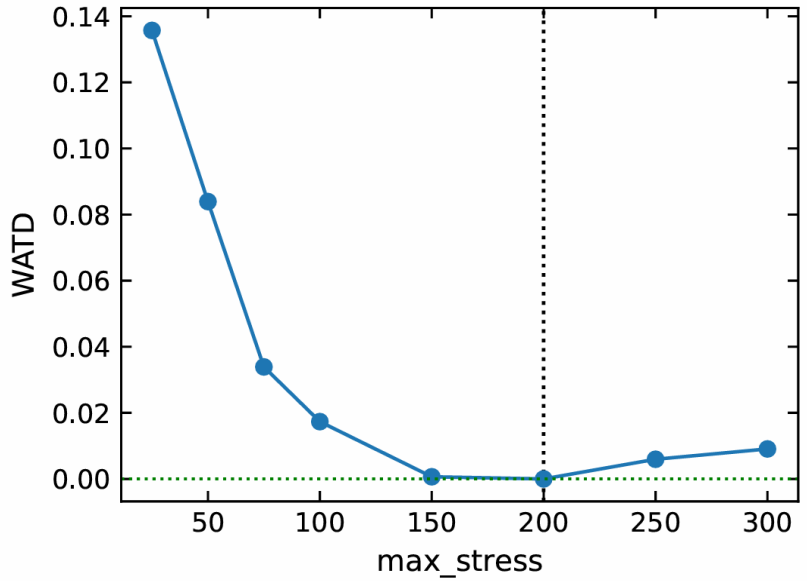
Max stress level

Nominal value: 200

- Low max stress → action distributions diverge
- Near nominal setting → distributions converge

➢ WATD increases as action distributions diverges.





Distance to goal

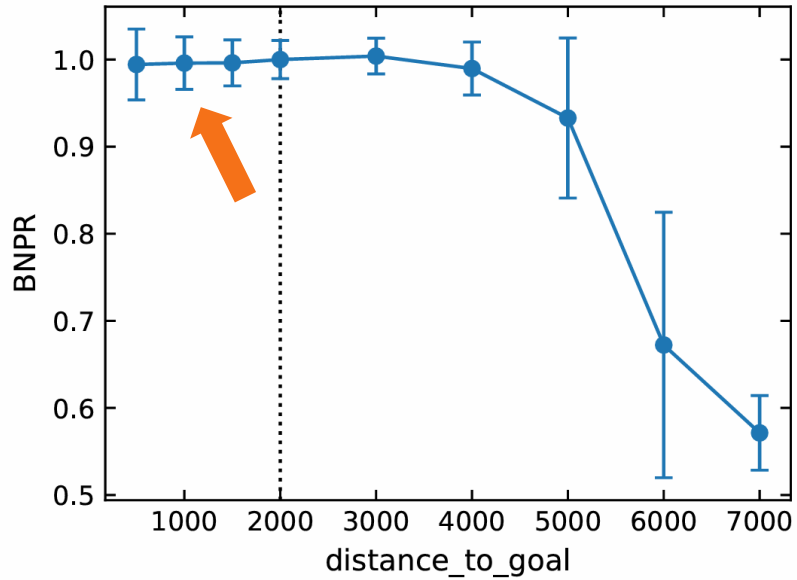
Nominal value: 2000m

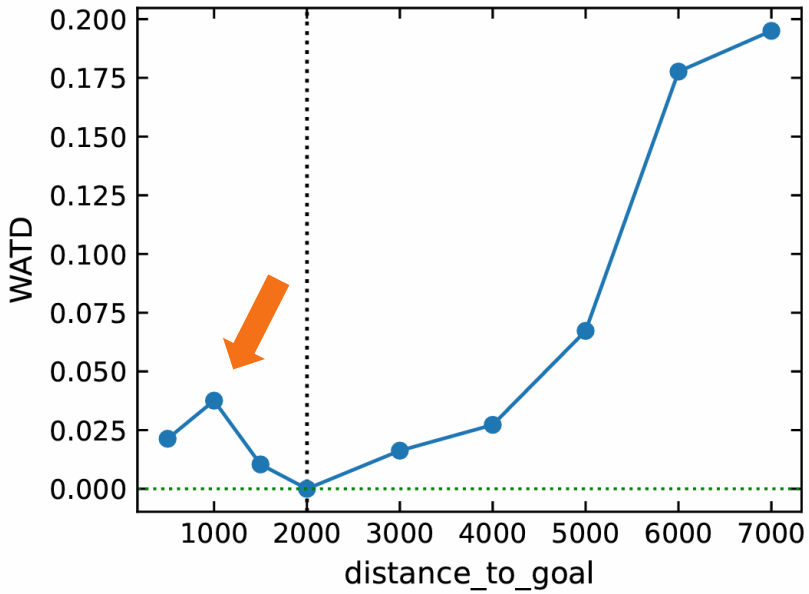
- Mostly: lower BNPR = higher WATD
- But: at 1000m, WATD ~1, but higher WATD

➢ High performance can hide behavioural changes.

➢ WATD can indicate different behaviour for similar performance.

BNPR and WATD give complementary information.





From scores to V&V evidence

- BNPR and WATD as evidence in verification and validation
- Neither metric defines acceptability by itself.
- Acceptance threshold must be mission- and system specific

	Low WATD	High WATD
High BNPR	✓ Close to benchmark	⚠ Inspect behaviour
Low BNPR	⚠ Inspect model	⚠ Clear issue



Takeaways

1. Absolute performance is not enough
2. Behaviour matters too
3. BNPR + WATD provide complementary V&V evidence

Future work

- State-conditioned WATD
- Operational acceptance threshold

Questions?