Proportionality Assessment in Military Operations Based on SARSA and PPO Models

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: This research proposes a novel decision-support framework for proportionality assessment in military operations integrating on-policy and policy-gradient reinforcement-learning methods to encode expert rules and automatically classify engagement scenarios. On this behalf, two modelling perspectives are considered implementing SARSA (State-Action-Reward-State-Action) and PPO (Proximal Policy Optimization) algorithms, and two approaches are adopted, i.e., in order to consider or not the integration of psychological effects or harm as part of collateral damage. Further, various optimization methods and simulation scenarios are considered in order to understand the effectiveness and robustness of the modelling techniques developed. From the results obtained, it can be seen that while SARSA achieves rapid reward stabilization but exhibits limited accuracy due to potential bootstrapping bias and insufficient exploration in larger state spaces, PPO's clipped surrogate updates yield robust, monotonic improvement, consistently realizing high classification accuracy across both cases, albeit over longer training horizons. To this end, a comparative analysis is conducted based on simulation results, learning curves, Q-value/policy distributions, and confusion matrices to illustrate each algorithm's strengths and limitations. Hence, this research demonstrates the viability of reinforcement learning models as transparent, adaptable tools for proportionality assessment for supporting real-time operational decisions.

Keywords: Military operations, Proportionality, Artificial intelligence, Reinforcement learning, SARSA, PPO

1. Introduction

> ***Motto:*** *"The unleashed power of the atom has changed everything save our modes of thinking and we thus drift toward unparalleled catastrophe." (Albert Einstein)*

The rapid development of artificial intelligence (AI) technologies over the past decade has driven an unprecedented wave of innovation across all societal domains. Advances in machine learning, particularly deep neural networks and reinforcement-learning frameworks, enabled systems to perceive complex environments, infer latent patterns, and make autonomous decisions at speeds exceeding human capabilities (Matsuo et al., 2020; Taye, 2023; Chen et al., 2024; Maathuis, 2022; Govinda, Brik & Harous, 2025). In the military domain, these breakthroughs translated into a new generation of smart sensors, unmanned platforms, and decision-support systems that enhance situational awareness, optimize logistics, and automate routine Command and Control (C2) functions (Cole et al., 2024; Probasco et al., 2025). From autonomous reconnaissance drones that navigate contested airspace to sensor-fusion architectures that integrate multi-modal battlefield data, AI is reshaping how armed forces plan, target, and execute military operations, ushering in an era where intelligent systems operate alongside, and under the supervision of, human decision-makers (Hoffberger-Pippan, E., & Dahlmann, 2024; Raska, 2024).

Nevertheless, as AI assumes an ever-greater role in various critical decision-making activities, it becomes necessary that these models are designed with both legal and ethical constraints at their core (Panwar, 2024). Compliance with international humanitarian law, human rights principles, and societal expectations demands that AI decision-support systems be transparent, auditable, and aligned with enduring norms such as distinction, proportionality, and necessity. Ethical frameworks further require respect for civilian safety, minimization of unintended harm, and pathways for human oversight to guard against bias, overreach, or malfunction (Lekea, Kraska, 2021; Lekeas & Topalnakos, 2023; Maathuis & Chockalingam, 2023). Embedding these considerations since the design of the AI systems through mechanisms such as value-aligned reward shaping, human-in-the-loop control, and formal verification, ensures that technological development does not outpace moral and legal accountability, thereby preserving the legitimacy of military action under the rule of law.

Among the most sensitive applications of AI in the targeting cycle is the proportionality assessment, a core decision-making process that weighs anticipated military advantage against the estimated collateral damage (Corn & Schoettler, 2015; Dorsey & Bo, 2025). Being defined in the Additional Protocol I to the Geneva Conventions, the principle of proportionality prohibits attacks whose expected incidental damage to civilians or civilian objects would be excessive relative to the concrete and direct military advantage (Dinstein, 2005; Gillard, 2018; Fard & Maathuis, 2021). Implementing this judgment requires dynamic evaluation of uncertain battlefield data such as target value, civilian presence, and weapon effects under time pressure (Li et al.,

2023). In this case, intelligent and adaptive AI solutions could support commanders by rapidly synthesizing sensor feeds, modelling potential collateral effects, and recommending actions that balance operational objectives with humanitarian constraints (Rashid et al., 2023; Roy, 2024). Designing such systems to learn and adapt while remaining firmly anchored in legal-ethical criteria is essential for responsible AI-enabled targeting in contemporary and future conflict settings.

Nevertheless, despite a rich corpus of military-legal, ethical, and technical scholarship on proportionality assessment in military operations, most existing frameworks remain largely prescriptive or descriptive, lacking mechanisms for real-time learning and adaptive decision-making. Specifically, legal analyses and doctrinal studies meticulously articulate the principles of distinction and proportionality, while ethical and philosophical works explore the moral imperatives behind minimizing civilian harm. At the same time, intelligent and computational models ranging from heuristic rule-based systems to probabilistic risk-assessment algorithms offer quantitative methods for estimating collateral effects (Watkin, 2005; Guiora, 2012; Cohen & Zlotogorski, 2021; Maathuis & Chockalingam, 2023; Mirzoev, 2025). Hence, there is a gap in dynamic, self-improving systems capable of refining proportionality judgments through experience, which is critical for ensuring sustained compliance under the uncertainties of modern battlefields. To address this shortfall, we propose a dual-architecture modelling system that implements SARSA and PPO algorithms to operationalize proportionality assessment as an adaptive and hybrid process (Bonnici et al., 2021; Kamble et al., 2025). The system defines each engagement scenario as a discrete state vector of collateral damage and military advantage parameters and trains two agents under divergent rule semantics. Specifically, In Case 1, the SARSA and PPO agents learn proportionality under a strict collateral damage paradigm that excludes psychological injury, thereby encoding traditional kinetic effects into their decision policies. At the same time, in Case 2, the state definition is enriched to encompass psychological harm as a component of collateral damage, challenging the agents to internalize an expanded ethical calculus. By comparing on-policy temporal-difference learning (SARSA) with clipped policy-gradient optimization (PPO) across these two cases, the system demonstrates how reinforcement-learning architectures can both respect established legal norms while adapting to more nuanced ethical considerations in targeting decision-making support.

The outline of this article is structured as follows. In the second section, an overview of the background required in technical terms is provided together with a discussion on related studies carried out in the military domain. Section 3 presents the methodological approach followed to achieve the aim of this research. Section 4 shows the implementation choices made to build the two models proposed in this research. Section 5 discusses the evaluation process and the evaluation results obtained for both cases considered. Conclusively, in Section 6 are discussed reflective points and future research perspectives.

2. Research Background

SARSA and PPO are two prominent RL algorithms offering distinct mechanisms for sequential decision-making. SARSA is an on-policy, model-free control algorithm in which the learning agent updates its state-action value function (Q-value) based not only on the current experience, but also on the subsequent action chosen under its current policy. In this process, at each step, the agent observes its current state, selects an action (often using an epsilon-greedy policy), receives a reward, moves to a new state, selects a new action, and updates its Q-table with this information. This iterative learning process enables SARSA to optimize its decision policy by continually refining value estimates toward maximizing cumulative rewards through balancing exploration and exploitation. Furthermore, the on-policy nature ensures that SARSA's updates reflect the policy currently used by the agent, making it particularly responsive to the consequences of the agent's actual trajectory through the state space (Rummery & Niranjan, 1994; Qiang & Zhongli, 2021; Yao et al., 2025). In contrast, PPO represents an advanced family of policy gradient methods designed to improve the stability and efficiency of policy-based RL training, especially in high-dimensional or continuous action spaces. It is characterized by the use of a stochastic policy, direct policy optimization, and its core innovation which represents a clipped surrogate objective that constrains the magnitude of policy updates to prevent destabilizing shifts. In this process, the agent alternates between sampling trajectories from its current policy, estimating advantages, and updating its policy parameters in a direction that incrementally improves expected performance while ensuring proximal steps to the existing policy. In practical terms, PPO leverages both an actor (policy function) and a critic (value function), employing parallel updates and batch learning to stabilize convergence. This strategic combination allows PPO to maintain robust learning even in environments where reward signals are sparse, delayed, or noisy which are circumstances that can confound traditional value-based algorithms (Schuman et al., 2017; Zeng et al., 2024).

The key elements distinguishing SARSA and PPO relate to their learning mechanisms and adaptability. Specifically, SARSA is relying on explicit state-action evaluations via Q-tables suits smaller, discrete environments but may limit scalability. Its on-policy feature provides stability in stochastic settings. Nevertheless, it is less effective when exploration is insufficient in complex state spaces. At the same time, PPO is valued for its versatility, scalability, and conservative updates, which collectively produce reliable policy improvement in both discrete and continuous domains. Together, SARSA and PPO exemplify the evolution of RL algorithms toward broader applicability, stability, and real-world applications and impact across robotics, autonomous systems, and adaptive decision-support domains (Naeem, Rizvi & Coronato, 2020).

In the military domain, deep RL has been successfully applied to autonomous UAV navigation, where quadcopters learn to navigate through obstacle-laden terrain using only onboard sensors and trial-and-error exploration (Fagundes-Junior et al., 2024). At the same time, hierarchical RL techniques were employed for RF-signal-based UAV detection, training agents to discriminate friend from foe by processing radio-frequency signatures in real time (Ahirrao, Yadav & Kumar, 2024). In search-and-rescue (SAR) operations, RL agents using received-signal-strength indicators learn to localize trapped personnel in GPS-denied indoor environments, by this reducing victim-location time. Furthermore, multi-agent deep RL is used in various UAV swarm surveillance, coordinating dozens of drones to cover and track ground targets collaboratively and RL-driven path-planning algorithms optimize trajectories for reconnaissance missions in contested airspaces (Lyu et al., 2023).

In guided SARSA UAV (Unmanned Aerial Vehicle) path-planning, agents learn to avoid dynamic threats while minimizing fuel consumption, outperforming conventional Q-learning in convergence speed. Here the model underpins network-lifetime maximization in multi-UAV communication networks, where agents schedule flight paths to reduce information age metrics and ensure seamless battlefield connectivity (Luo et al., 2018). In the same context, X propose a SARSA-based protocol to optimize data-collection points for rotary-wing UAVs, balancing channel conditions and mission deadlines to maximize throughput (Joshi, Kalita & Gurusamy, 2023). In a multi-agent setting, for allocation purposes, each UAV agent independently learns bandwidth allocation strategies to support distributed sensing and communication. At the same time, an enhanced PPO algorithm was developed for intelligent joint operations planning, integrating priority sampling and constrained loss functions to optimize C2 decisions in simulation platforms (He et al., 2019). In high-speed intercept guidance, PPO policies learn robust manoeuvre laws to engage agile, manoeuvring targets under kinematic constraints. In the context of threat evaluation, Eval-PPO is deployed for transforming multi-dimensional enemy attributes into a unified RL reward signal for precise, context-aware threat assessments (Sun et al., 2025).

Hence, the deployment of RL algorithms such as SARSA and PPO in military applications offers significant advantages for autonomy, adaptability, and decision-making support. SARSA's on-policy updates facilitate rapid learning in structured, low-noise environments, enabling real-time adaptation to evolving mission constraints with minimal offline computation. At the same time, PPO's clipped gradient updates and decoupled value estimates yield robust convergence even in high-dimensional, stochastic settings, ensuring that learned policies adhere to safety and legal constraints under partial observability. These methods can integrate both domain knowledge and field data as reward shaping or policy priors, providing transparent, auditable decision traces. Taking into account the fact that military operations continue to demand high-throughput, low-latency decision loops, RL represents a versatile toolkit for developing responsible AI that balances mission effectiveness with adherence to humanitarian and legal and ethical norms and principles.

3. Research Methodology

This research adopts the Design Science Research (DSR) methodology as framed (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018) in order to build two models based on SARSA and PPO reinforcement learning algorithms to conduct the proportionality assessment in military operations. The DSR process initiated with a problem identification phase, recognizing the need to build responsible and trustworthy AI-based tools that can encode and model complex proportionality rules and constraints capturing both the standard and a more ethically-compliant perspective on collateral damage which also includes psychological harm (Maathuis, Pieters & Van Den Berg, 2018). Subsequently, the objectives of the artefact were formulated to implement and systematically optimize the SARSA and PPO models, each representing distinct reinforcement learning paradigm: SARSA as an on-policy temporal-difference learner and PPO as a policy-gradient actor-critic method. The design phase implied constructing these decision-support models, carefully selecting hyperparameters to balance learning speed, stability, and generalization capacity across deterministic rule sets of varying

complexity. From there, an Iterative evaluation cycle involved quantitative performance assessment through reward trajectories, classification accuracy, and confusion matrices reflecting the faithful reproduction of proportionality decisions.

In alignment with DSR's emphasis on rigor and relevance, the evaluation incorporated extensive simulation experiments comparing both algorithms under controlled scenarios representative of operational decision spaces. In this process, metrics such as convergence speed, reward stability, and sensitivity to hyperparameter variation were systematically analysed to understand each artifact's strengths and limitations. Insights gained from these analyses contributed to design knowledge, informing recommended strategies for hyperparameter calibration, such as conservative learning rates and exploration parameters to mitigate SARSA's brittleness or PPO's conservative policy updates that enhance robustness. Furthermore, the results of the comparative analysis are presented to address the audience and further recommendations are provided for further development and deployment. This methodical and iterative development process, grounded in DSR principles, ensures that the resultant SARSA and PPO models contribute to ongoing effort dedicated to building responsible and trustworthy military AI systems that account multidimensional operational and socio-technical domain aspects.

4. Implementation of Models

The proportionality assessment is formulated as a discrete Markov decision process (MDP) whose state space encodes the critical factors governing lawful use of force. Each state s=(s1,s2,s3,s4) is a four-tuple: collateral injury severity (s1), collateral death severity (s2), object damage presence (s3), and military advantage magnitude (s4s). In Case 1, s1 quantifies only physical injury (Low/Medium/High), whereas in Case 2, s1 additionally incorporates psychological harm in the collateral damage component (Maathuis, Pieters & Van den Berg, 2018b). The action set A={Proportional, Disproportional} reflects the binary judgment mandated by Article 51(5)(b) of Additional Protocol I. In this process, a tabular value function or policy is learned over these 54 discrete states, ensuring transparent traceability from scenario features to targeting decisions.

The first model first implements the on-policy algorithm to estimate Q(s,a), the expected return of executing action aa in state ss under the current policy. In this process, the initialization Q(s,a)=0 takes place and a further iteration over uniformly sampled states is carried out. At each step, the agent selects action aa according to an ϵ-greedy policy with values ϵ=0.1 for Case 1 and ϵ=0.2 for Case 2 in order to ensure sufficient exploration. Upon observing reward r∈{+10,−10} for correct or incorrect classification, and next state s', the update

$$Q(s,a) \;\leftarrow\; Q(s,a) + \alpha\big[r + \gamma\, Q(s',a') - Q(s,a)\big]$$

is applied, where a' is the subsequent action chosen by the same ϵ-greedy policy. Further, α=0.1\alpha=0.1 in Case 1 is set for rapid learning and α=0.01\alpha=0.01 is set in Case 2 in order to accommodate the added complexity of psychological injury. Here, a discount factor γ=0.9 balances immediate and future rewards, encapsulating both instant proportionality judgments and longer-term consistency across engagements.

The second model adapts PPO to the tabular MDP, maintaining separate policy parameters π(s,a) and state-value estimates V(s). Here, each state s is flattened to a unique index i, enabling storage of a softmax policy vector πi and scalar V(i). During training, the actions are sampled from πi and advantages A=r−V(i) are computed. Then the policy parameters are updated via the clipped surrogate objective

$$L_{\text{clip}} = \mathbb{E}\Big[\min\big(\rho(\theta)\,A,\ \text{clip}(\rho(\theta), 1-\epsilon_c, 1+\epsilon_c)\,A\big)\Big]$$

where ρ is the likelihood ratio and ϵc=0.1 limits the magnitude update. At the same time, V(i) is regressed toward observed returns with learning rate 0.01. Here, a policy learning rate of 0.001 and a discount factor of 0.9 ensure smooth, monotonic improvement without destabilizing large updates, allowing the policy to generalize seamlessly from Case 1 to Case 2.

Hence, although both SARSA and PPO operate on the same state–action schema, their update mechanics drive different learning dynamics across the two cases. SARSA's bootstrapped, on-policy nature leads to rapid reward gains when only physical damage is considered as part of the collateral damage component, but requires careful tuning (α↓, ϵ↑) to avoid premature convergence in the richer Case 2. PPO's clipped policy updates, by contrast, naturally accommodate increased rule complexity without hyperparameter changes,

providing stable convergence and principled control over update size. This implies that both embed the legal and ethical core aspects of the proportionality assessment deriving rewards from compliance with IHL's harm vs advantage calculus, thus ensuring that learned policies are not merely accurate, but also aligned with humanitarian norms.

5. Evaluation

The SARSA agent was evaluated on both Case 1 (physical collateral damage only) and Case 2 (physical and psychological damage) using three complementary metrics: reward-per-episode convergence, confusion-matrix classification accuracy, and precision/recall breakdown. In Case 1 ($\alpha = 0.1$, $\varepsilon = 0.1$), the reward curve plateaued by approximately 100 episodes at a mean total reward of +48, indicating rapid policy stabilization. However, the confusion matrix revealed only about 44 % accuracy which means that SARSA correctly classified a majority of Proportional states, but misclassified over half of the Disproportional scenarios. This shows a bias toward high-frequency positive outcomes under limited exploration. In Case 2 ($\alpha = 0.01$, $\varepsilon = 0.2$), learning slowed (rewards stabilized around 200 episodes), yet accuracy declined further to about 41 %, with persistent false-positives in cases involving high psychological injury (Figure 1). These results shows SARSA's susceptibility to bootstrapping bias and its difficulty in covering a richer state space despite extended exploration.

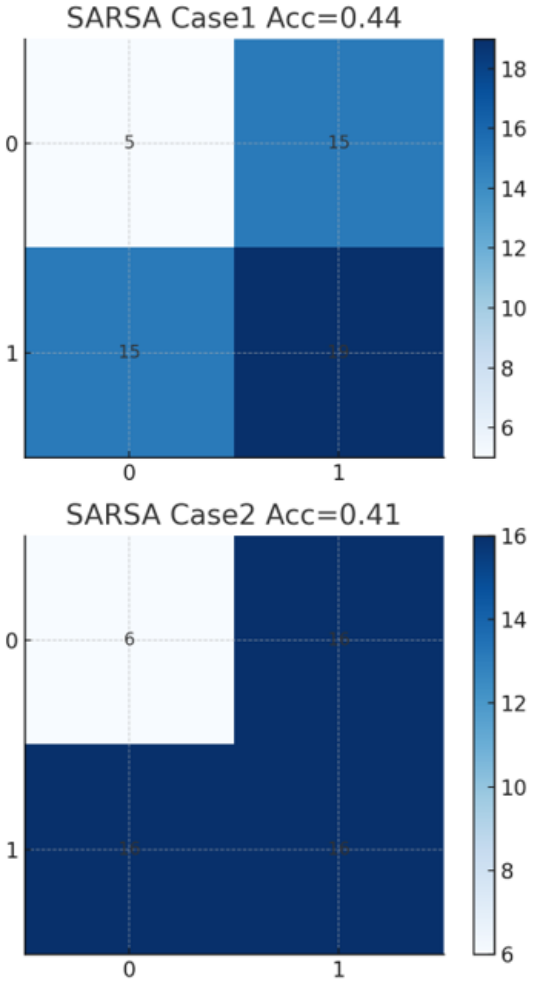

Figure 1: SARSA's accuracy and confusion matrix

The PPO agent demonstrated different dynamics. For Case 1 (policy_lr = 0.001, value_lr = 0.01, clip = 0.1), the reward curve rose more gradually stabilizing near +52 around 600 episodes. Nevertheless, it had a perfect (100 %) classification accuracy, as confirmed by a confusion matrix with zero false-positives or negatives. Precision and recall for both P and DP reached unity, underscoring the model's balanced performance. In Case 2, PPO maintained the same hyperparameters and achieved convergence of reward (~+50) by ~800 episodes, again delivering 100 % accuracy (Figure 2). The stability of reward trajectories, evidenced by low

variance beyond the initial exploration phase, demonstrates PPO's robust policy improvements via its clipped surrogate objective and decoupled value function, even when state semantics are expanded to include psychological harm.

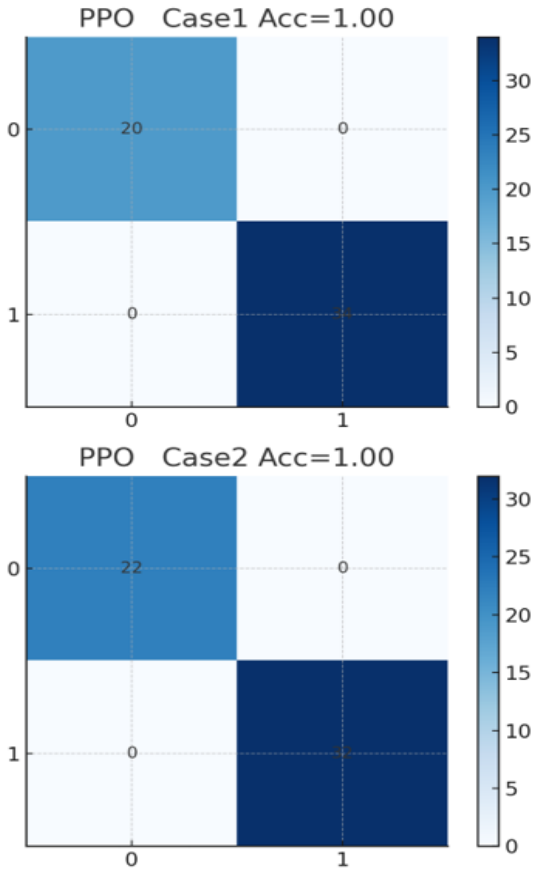

Figure 2: PPO's accuracy and confusion matrix

Further, a comparative analysis is conducted for both models and on both cases. Accordingly, for Case 1, SARSA outpaces PPO in learning speed by reaching its reward plateau in ~100 episodes vs. PPO's ~600, but at the cost of classification fidelity (44 % vs. 100 %). Furthermore, SARSA's rapid bootstrap updates allow quick acquisition of high-frequency state–action values, but lead to overfitting on Proportional outcomes and under-exploration of minority DP states. At the same time, PPO's conservative policy gradients require more iterations to learn, but systematically refine action probabilities across all 54 scenarios, ensuring both high precision and recall. Hence, while SARSA may be advantageous in time constrained settings with simple mappings, PPO offers superior reliability for critical proportionality judgments. For Case 2, SARSA's limitations become more pronounced: even with increased exploration ($\varepsilon = 0.2$) and reduced learning rate ($\alpha = 0.01$), its reward plateau does not translate to improved accuracy, and it continues to misclassify nuanced Disproportional scenarios. At the same time, PPO, maintaining fixed hyperparameters, seamlessly generalizes to the expanded state space, with reward convergence around 800 episodes and 100 % accuracy. This contrast shows PPO's robustness to rule complexity and its ability to integrate evolving ethical dimensions without hyperparameter re-engineering (Figure 3). For operational deployments requiring both adaptability and uncompromising legal compliance, PPO is then seen as the preferable modelling architecture. Nevertheless, given the existence of a potential overfitting as a generalization issue, additional considerations could be given in the future to the inclusion of other relevant variables and dataset for both design and evaluation purposes, and to be able to conduct a comprehensive analytical evaluation.

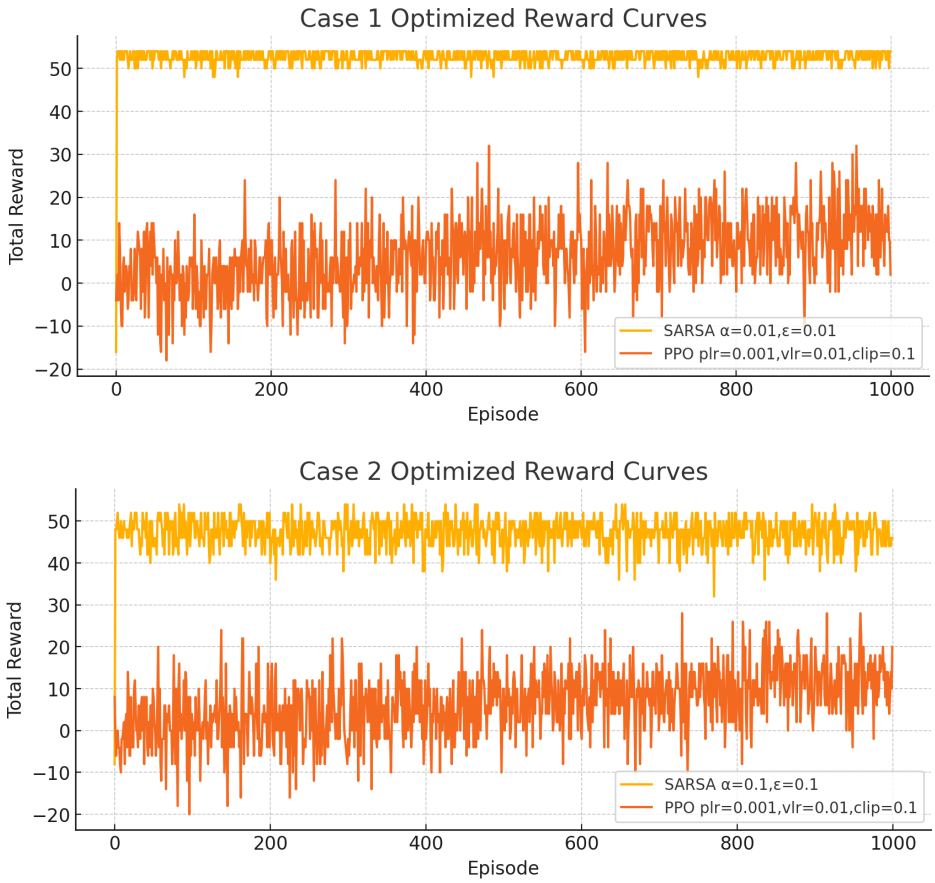

Figure 3: Rewards for both SARSA and PPO models

6. Conclusions

This research proposes two reinforcement learning modelling approaches using the SARSA and PPO algorithms to conduct the proportionality assessment in military operations. It does that following a DSR methodological approach and considering a comparative analysis for evaluation purposes between the two modelling approaches considered applied to the two modalities of framing collateral damage inside the proportionality assessment process considered. Based on the comparative analysis conducted in this research, a series of differences in their learning dynamics, robustness, and suitability for complex legal and ethical rule and constraints encoding is observed. Specifically, SARSA exhibits rapid initial convergence in simpler state spaces, as evidenced by its swift reward plateau in Case 1 further showing challenges in deterministic but high-dimensional decision environments where incomplete exploration and reliance on immediate next-action updates lead to persistent errors. In contrast, the PPO-style approach demonstrates superior stability and accuracy across both use cases, ultimately achieving perfect classification with slower reward accumulation and requiring several hundred episodes to stabilize. Furthermore, its robustness makes PPO more adept at encoding explicit proportionality constraints within deterministic frameworks, providing reliable decision support in contexts demanding high ethical and operational certainty. From a practical standpoint, PPO's trade-off such as increased training time for enhanced global policy refinement and reduced error propagation is positioning it as a more suitable candidate for complex military decision-making scenarios, whereas SARSA's sample efficiency may better serve simpler or more stochastic environments where rapid and less stable learning is permissible.

Future research directions will focus on integrating function approximation techniques such as deep neural networks within both SARSA and PPO frameworks in order to extend their applicability to continuous and high-dimensional collateral damage metrics commonly encountered in real-world military contexts and settings. This advancement could enable scalable generalization beyond tabular representations, addressing current limitations related to state space explosion. Another future research perspective involves incorporating memory-enhanced architectures, like recurrent neural networks, or belief-state estimators to tackle partial observability inherent in operational environments where damage assessments and target states are often

incomplete or uncertain. Additionally, novel training regimes employing reward shaping and curriculum learning where the rule complexity is progressively increased, could mitigate instability issues and accelerate convergence, particularly for SARSA-based methods. Together, these directions aim to enhance the efficiency, robustness, and ethical reliability of AI-based proportionality assessment systems deployed in dynamically evolving military operations.

Declaration: For this research, no ethical clearance is required and no AI tools were used in the creation of this article.

References

Ahirrao, Y. V., Yadav, R. P., & Kumar, S. (2024). RF based UAV detection and identification enhanced by machine learning approach. *IEEE Access*.

Bonnici, R. S., Saliba, C., Caligari, G. E., & Bugeja, M. (2021). Exploring reinforcement learning: A case study applied to the popular snake game. In *The International Conference on Intelligent Systems & Networks* (pp. 169-192). Cham: Springer International Publishing.

Chen, Y., Ji, C., Cai, Y., Yan, T., & Su, B. (2024). Deep reinforcement learning in autonomous car path planning and control: A survey. *arXiv preprint arXiv:2404.00340*.

Cohen, A., & Zlotogorski, D. (2021). *Proportionality in international humanitarian law: Consequences, precautions, and procedures*. Oxford University Press.

Cole, A., Howard, D., Latiff, R., Lucas, G., Reichberg, G. M., & Roy, K. (2024). Artificial intelligence in strategic planning and military operations. In *Cyber security in the Age of artificial intelligence and autonomous weapons* (pp. 120-133). CRC Press.

Corn, G., & Schoettler Jr, J. A. (2015). Targeting and civilian risk mitigation: The essential role of precautionary measures. *Mil. L. Rev.*, *223*, 785.

Dinstein, Y. (2005). Collateral damage and the principle of proportionality. In *New Wars, New Laws? Applying Laws of War in 21st Century Conflicts* (pp. 211-224). Brill Nijhoff.

Dorsey, J., & Bo, M. (2025). AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human (e) Dimension.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Fagundes-Junior, L. A., de Carvalho, K. B., Ferreira, R. S., & Brandão, A. S. (2024). Machine learning for unmanned aerial vehicles navigation: An overview. *SN Computer Science*, *5*(2), 256.

Gillard, E. C. (2018). Some Reflections on the Incidental Harm Side of Proportionality Assessments. *Vand. J. Transnat'l L.*, *51*, 827.

Govinda, S., Brik, B., & Harous, S. (2025). A Survey on Deep Reinforcement Learning Applications in Autonomous Systems: Applications, Open Challenges, and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*.

Guiora, A. N. (2012). Targeted killing: When proportionality gets all out of proportion. *Case W. Res. J. Int'l L.*, *45*, 235.

He, H., Wang, W., Zhu, Y., Li, X., & Wang, T. (2019). An operation planning generation and optimization method for the new intelligent combat SoS. *IEEE Access*, *7*, 156834-156847.

Hoffberger-Pippan, E., & Dahlmann, A. (2024). Digital battlefield: concept, technology and prospects. In *Research Handbook on Warfare and Artificial Intelligence* (pp. 76-98). Edward Elgar Publishing.

Joshi, P., Kalita, A., & Gurusamy, M. (2023). Reliable and efficient data collection in UAV-based IoT networks. *arXiv preprint arXiv:2311.05303*.

Kamble, A. R., Jiet, M. M., Weerarathna, I. N., Dhande, S. A., Izankar, S. V., & Puri, C. G. (2025, March). Reinforcement learning model, algorithms, and its application. In *AIP Conference Proceedings* (Vol. 3227, No. 1, p. 050012). AIP Publishing LLC.

Kraska, J. (2021). Command Accountability for AI Weapon Systems in the Law of Armed Conflict. *International Law Studies*, *97*(1), 22.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Lekea, I., Lekeas, G., & Topalnakos, P. (2023). Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis. *Conatus-Journal of Philosophy*, *8*(2), 345-372.

Li, G., He, G., Zheng, M., & Zheng, A. (2023). Uncertain sensor–weapon–target allocation problem based on uncertainty theory. *Symmetry*, *15*(1), 176.

Luo, W., Tang, Q., Fu, C., & Eberhard, P. (2018, June). Deep-sarsa based multi-UAV path planning and obstacle avoidance in a dynamic environment. In *International Conference on Swarm Intelligence* (pp. 102-111). Cham: Springer International Publishing.

Lyu, M., Zhao, Y., Huang, C., & Huang, H. (2023). Unmanned aerial vehicles for search and rescue: A survey. *Remote Sensing*, *15*(13), 3266.

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics* (Vol. 4, No. 1, pp. 45-53).

Maathuis, C., Pieters, W., & Van Den Berg, J. (2018). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security* (pp. 278-288).

Maathuis, C., Pieters, W., & van den Berg, J. (2018b). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., & Chockalingam, S. (2023). Modelling the influential factors embedded in the proportionality assessment in military operations. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 218-226).

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., ... & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267-275.

Mirzoev, R. (2025). The Principle of Proportionality in International Humanitarian Law: Evolution, Application, and Accountability in Modern Armed Conflicts.

Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A gentle introduction to reinforcement learning and its application in different fields. *IEEE access*, *8*, 209320-209344.

Panwar, R. S. (2024). Artificial intelligence in military operations: technology, ethics and the Indian perspective. In *Artificial Intelligence, Ethics and the Future of Warfare* (pp. 216-226). Routledge India.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Probasco, E., Toner, H., Burtell, M., & Rudner, T. G. (2025). AI for Military decision-making. *Center for Security and Emerging Technology*, 5-26

Qiang, W., & Zhongli, Z. (2011). Reinforcement learning model, algorithms and its application. In *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)* (pp. 1143-1146). IEEE.

Rashid, A. B., Kausik, A. K., Al Hassan Sunny, A., & Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International journal of intelligent systems*, *2023*(1), 8676366.

Raska, M. (2024). Reimagining Defense Innovation: Defense AI in Singapore. In *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (pp. 555-580). Cham: Springer Nature Switzerland.

Roy, K. (Ed.). (2024). *Artificial intelligence, ethics and the future of warfare: Global perspectives*. Taylor & Francis.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Vol. 37, p. 14). Cambridge, UK: University of Cambridge, Department of Engineering.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, W., Li, S., Zou, Q., & Liao, Z. (2025). Eval-PPO: Building an Efficient Threat Evaluator Using Proximal Policy Optimization. *arXiv preprint arXiv:2503.12098*.

Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, *12*(5), 91.

Watkin, K. (2005). Assessing proportionality: moral complexity and legal rules. *Yearbook of International Humanitarian Law*, *8*, 3-53.

Yao, G., Zhang, N., Duan, Z., & Tian, C. (2025). Improved SARSA and DQN algorithms for reinforcement learning. *Theoretical computer science*, *1027*, 115025.

Zeng, Y., Cai, R., Sun, F., Huang, L., & Hao, Z. (2024). A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*.