Q-Learning Model for Proportionality Assessment in Military Operations

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

clara.maathuis@ou.nl

Abstract: In the context of military operations, accurate and transparent proportionality assessment is essential to ensure compliance with international humanitarian law. On this behalf, this research presents and evaluates two Q-learning models designed to build the proportionality assessment in military operations. In this sense, the first model considers collateral damage exclusively in physical terms (excluding psychological harm), while the second model explicitly integrates psychological damage as part of the collateral damage effects. Both models encode operational rules as multi-attribute states encompassing injury severity, fatalities, object damage, and military advantage, differing only in the inclusion of psychological factors. From training and simulation results, it can be seen that this approach provides a valuable classification approach for proportional and disproportional outcomes within their respective scenario sets. This shows that AI (Artificial Intelligence) provides effective methods and techniques that allow both modelling and the expansion of existing definitions and perspectives of existing challenging concepts in the uncertain and dynamic space of the military domains while accounting and respecting legal and ethical considerations in order to build responsible and trustworthy military AI systems.

Keywords: Military operations, Military targeting, Proportionality, Artificial intelligence, Reinforcement learning, Q-learning

1. Introduction

Motto: "The fragility of freedom is the simplest and deepest lesson of my life and work ." (Albert Einstein)

Recent years witnessed unprecedented advancements in the Artificial Intelligence (AI) domain, driven by breakthroughs in machine learning, deep neural networks, and reinforcement learning paradigms. These applications have transcended traditional data analysis tasks to enable autonomous perception, reasoning, and decision-making across a variety of complex domains (Sarker, 2021; Hussain, Rahman & Ali, 2024). In particular, the increased applicability of AI-enabled systems, ranging from unmanned aerial vehicles to automated sensor fusion platforms, has catalysed their integration into military operations (Criollo, Mena-Arciniega & Xing, 2024; Singh, 2024). At the same time, worldwide, various military organizations are exploring how AI can augment situational awareness, optimize logistical workflows, and support command and control (C2) architectures, bringing a new era in which intelligent agents play an increasingly central role in strategic and tactical decision loops (Lingel et al., 2020; Tóth & Farkas, 2023; Toroi, 2024). Within the military operational context, numerous decision-making processes stand to benefit from AI-driven automation. This points out to solutions such as the ones developed and deployed for target identification and selection, Rules of Engagement adjudication, and fire-control authorization represent high-stakes tasks that demand rapid, consistent judgments under uncertain and time critical conditions (Jahn, 2019; Schmitt, M. N., & Schauss, 2019; Lekea, Lekeas & Topalnakos, 2023). At the same time, machine learning-based classifiers assist in discriminating combatants from non-combatants and in prioritizing threats based on mission objectives, while automated planning algorithms can generate and evaluate courses of action in parallel, reducing human cognitive burden (Ahmed, 2022; Montasari, 2024; Ray, 2024). By embedding AI into these activities, armed forces aim to enhance responsiveness, reduce unintended effects such as collateral damage, and free human operators to focus on higher-order strategic reasoning.

A particularly sensitive application of AI in warfare lies in autonomous targeting systems, where adherence to International Humanitarian Law (IHL) is an important requirement (Sassoli, 2014; Khali & Raj, 2024). Central to IHL is the principle of proportionality, codified in Additional Protocol I to the Geneva Conventions, most notably Article 51(5)(b), which prohibits attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, or damage to civilian objects … which would be excessive in relation to the concrete and direct military advantage anticipated," and Article 57's requirement to take "all feasible precautions" to minimize harm to civilians (Benvenisti, 2022; Al-Fatlawi, Tamimi & Al-Tamimi, 2024). Responsible AI solutions for targeting should therefore internalize proportionality assessments, ensuring that algorithmic decisions respect these legal norms. Hence, designing AI agents that can systematically evaluate collateral damage against anticipated military advantage is essential to upholding lawful conduct in future conflict scenarios. To this end, integrating legal and ethical perspectives is indispensable when engineering AI systems for

proportionality assessment and targeting support, as these systems operate at the intersection of life-and-death decisions and binding international norms. In addition to the legal perspective, ethically, responsible AI frameworks compel respect for human dignity, civilian protection, and the minimization of unintended harm, necessitating transparent value-alignment mechanisms, human-in-the-loop safeguards, and accountability provisions. In this context, despite extensive theoretical and conceptual research on collateral damage and proportionality assessment spanning legal treatises, ethical frameworks, philosophical analyses of jus in bello (Neuman, 2004; Henderson, I., & Reece, 2018; Fard & Maathuis, 2021; Cohen, A., & Zlotogorski, 2021; Maathuis & Chockalingam, 2023), and technical proposals for decision models, a gap remains in translating these insights into intelligent and adaptive responsible AI systems .

Traditional studies have systematically catalogued normative criteria, harm quantification methodologies, and decision heuristics, yet few have bridged the divide to operationalize proportionality as a dynamic, learnable function within reinforcement-learning or hybrid control architectures. As a result, existing AI efforts often rely on simplistic threshold-based filters, lacking the capacity to generalize across evolving tactical contexts or to calibrate judgments in the face of novel collateral damage profiles. Addressing this gap calls for transdisciplinary research that merges rigorous legal-ethical formalization with AI techniques, thereby creating AI systems capable of real-time, context-aware proportionality assessments that learn and adapt while remaining legally and morally anchored.

To this end, this research aims to develop a Q-learning–based decision-support system (Watkins & Dayan, 1992) that encodes both legal norms and ethical imperatives into the proportionality assessment process for military targeting. By modelling proportionality as a Markov decision process (MDP) (Bellman, 1957), this approach represents each engagement scenario as a state vector of collateral damage attributes and military advantage, and uses tabular Q-learning to learn an optimal action-value function Q(s,a) mapping these states to Proportional or Disproportional decisions. Here, the agent's exploration-exploitation policy is governed by an ϵ-greedy strategy that ensures comprehensive sampling of legally relevant scenario permutations before converging on lawful and ethical targeting recommendations. On this behalf, to investigate the impact of different collateral damage conceptualizations, two experimental and simulation cases are defined. In Case 1, collateral damage encompasses only physical injury, death, and object destruction, thereby isolating the agent's ability to enforce legal proportionality under traditional kinetic effects. In Case 2, the state definition is extended to include psychological harm as well by evolving ethical recognition of non-lethal but debilitating effects on civilian populations, thereby increasing the rule complexity and ethical nuance the Q-learning agent must internalize (Maathuis, 2022). By comparing learning dynamics and policy formation across these two cases, this research aims to demonstrate how adaptive reinforcement learning architectures can be tuned to respect core legal thresholds while integrating emerging ethical dimensions of collateral damage for building a responsible AI model targeting decision-making support.

The remainder of this article is structured as follows. Section 2 introduces the background of this research and discusses related studies conducted in this domain. Section 3 presents the methodology followed to achieve the aim of this research. Section 4 presents the design of the model proposed. Section 5 presents the evaluation process and the evaluation results obtained. At the end, in Section 6 are discussed concluding remarks and a series of future research perspectives are provided.

2. Related Research

Q-learning is a fundamental algorithm within the RL (Reinforcement Learning) paradigm designed to enable agents to learn optimal decision-making policies through interaction with their environment (Watkins & Dayan, 1992; Clifton & Laber, 2020). In this paradigm, an agent sequentially observes the state of the environment, selects actions, and receives feedback in the form of rewards or penalties, with the goal of maximizing cumulative future rewards. Specifically, Q-learning is distinguished by being a value-based, model-free, and off-policy algorithm, which implies that it does not require knowledge of the environment's dynamics and is not depending on the policy currently being followed. Instead, it directly estimates the optimal action-value function, or Q-function, which quantifies the expected cumulative reward for taking a particular action in a given state and subsequently following the best possible strategy (Tan, Yan & Guan, 2017). Through iterative updates that use the Bellman equation, the agent refines these Q-values, gradually converging towards the optimal policy that maximizes long-term returns.

The main components of the Q-learning algorithm include the agent, the state space representing all possible situations the agent might encounter, and the action set from which the agent chooses behaviours. After every action taken, the agent receives a scalar reward reflecting the immediate benefit or cost of that decision,

which is used to update the Q-values stored in a tabular or functional form that maps state-action pairs to expected returns. Accordingly, the core parameters relevant to this learning process are as follows: the learning rate which controls the influence of new experiences relative to existing knowledge, while the discount factor balances the importance of immediate against future rewards (Oliehoek, Spaan & Vlassis, 2008; Sadhu & Konar, 2020). Moreover, an exploration-exploitation strategy, typically epsilon-greedy, manages the trade-off between trying unfamiliar actions to discover potentially better rewards and exploiting current knowledge to maximize returns. In addition, the model-free and off-policy nature that the Q-learning algorithm has contributes to both its versatility and its robustness across various societal applications such as autonomous robotics, game playing, and decision support systems (Zhang & Mo, 2021; Rajasekhar, Radhakrishnan & Samsudeen, 2025). Despite these advantages, classical Q-learning struggles with scalability when dealing with large or continuous state spaces since tabular representations become infeasible. This limitation has motivated the development of deep Q-learning, which employs the power of neural networks as function approximators to generalize Q-values beyond discrete states, thereby enabling efficient learning in complex environments (Arulkumaran et al., 2017; Jang et al., 2019).

One of the most established applications is in active protection systems (APS) for armoured vehicles, where Q-learning is utilized to optimize the deployment of soft-kill countermeasures against guided high-mobility threats, such as anti-tank missiles (Rajagopalan, 2018). This real-time decision-making capability enhances survivability by adaptively responding to diverse attack vectors. Beyond physical defence, Q-learning is implemented in cyber defence systems to autonomously detect, analyse, and mitigate complex cyber threats such as advanced persistent threats (APTs) and zero-day exploits within military networks (Stellios, Kotzanikolaou & Psarakis, 2019; Hernández-Rivas, Morales-Rocha & Sánchez-Solís, 2024). Other applications include autonomous vehicle control (e.g., drones, tanks, unmanned surface vessels) where Q-learning guides navigation, obstacle avoidance, and mission task execution without explicit programming, thereby allowing adaptive responses to evolving battlefield conditions (Zhang et al., 2020; Chen et al., 2024).

In military decision support and C2, Q-learning plays is often use in optimizing weapon selection, threat evaluation, and resource allocation. This enable systems to learn effective strategies for weapon assignment, coordinating multi-armed platforms to prioritize and engage threats efficiently, which is critical in fast-paced combat scenarios (Cheng, Chen & Gong, 2021; Wang, Zhang & Wang, 2024). For instance, X applies it in battlefield simulations to train agents for tactical decision-making that balances mission objectives with minimizing casualties (Costa et al., 2025; Park & Shim, 2025). Furthermore, Q-learning algorithms assist in logistics planning by optimizing the distribution of supplies, personnel, and equipment across operational theatres to ensure mission readiness and sustainability (Singh, Mandal & Nayak, 2024). At the same time, various reinforcement learning techniques have also been used in air combat simulation and pilot training, where Q-learning networks integrated with deep learning help build intelligent virtual pilots capable of self-learning tactics and confronting human pilots in simulated engagements (Ruan, Duan & Deng, 2022).

Other applications include applications for intelligent reconnaissance and surveillance systems, where agents autonomously explore and gather intelligence while adapting to changing environments; multi-agent cooperation for coordinated actions in complex missions; electronic warfare strategies for adaptive signal jamming and countermeasures; autonomous underwater vehicle task scheduling; adaptive encryption and cyberthreat mitigation in sensor networks; and battlefield robotics for target acquisition and engagement. These applications show the versatility of Q-learning in enabling autonomous, adaptive, and optimized decision-making across a broad range of defence systems, enhancing operational effectiveness and resilience in complex and unpredictable combat scenarios (Maathuis, Pieters & van den Berg, 2018; Abro et al., 2024; Premakumari et al., 2025). However, challenges remain in algorithm stability, interpretability of learned policies, adversarial robustness, and compliance with legal and ethical standards for autonomous military systems.

3. Research Methodology

To build a Q-learning model for proportionality assessment in military operations, this research adopts methodological principles of the Design Science Research (DSR) paradigm as articulated by (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018) which emphasizes the iterative creation and rigorous evaluation of innovative artifacts to solve relevant, real-world problems. Following the DSR process framework, this research commences with the identification and motivation of the proportionality assessment challenges in military operations, highlighting the need for a decision-support mechanism capable of encoding complex legal and ethical humanitarian rules (Maathuis, Pieters & Van Den Berg, 2018b) which serves as the

base for the data used in this research. To this end, a Q-learning reinforcement learning approach is considered to build the proportionality assessment model that includes a broader perspective on the collateral damage component of this principle, i.e., the inclusion of psychological effects or harm. The design and development phase involved constructing this Q-learning agent with calibrated hyperparameters reflective of the deterministic rule complexity, followed by an in-depth performance evaluation through simulation on representative datasets. This ensures the artifact's functional robustness and its conforming to the intended decision criteria are fulfilling the DSR guidelines so that the artifact must demonstrates its effectiveness and utility within the military targeting decision-making process.

Furthermore, the model's evaluation incorporates metrics such as convergence dynamics, reward stability, and classification accuracy, augmented by confusion matrices for clear interpretability to support the transparency dimension of the responsible AI perspective governing this research (Maathuis & Chockalingam, 2022). On this behalf, the results are analysed in relation to learning-theoretic implications, linking hyperparameter configurations with convergence behaviour to inform methodological tuning. By iteratively refining the Q-learning model and documenting important modelling choices, this research contributes to the existing body of knowledge that aims at developing responsible and trustworthy military AI systems and existing practical efforts dedicated to building intelligent and adaptive operational AI systems.

4. Model

To build the Q-learning agent model that encodes proportionality as a decision-support system for military targeting, the assessment process is modelled as a Markov decision process since the outcomes are partly random and partly under the control of an agent. This framework consists of states, actions, transition probabilities between states after taking actions, rewards for transitions, and a discount factor for future rewards. The MDP assumes the Markov property, meaning the next state depends only on the current state and action, not on the history. In this process, the agent observes a discrete state s reflecting collateral-damage and military-advantage components, selects a binary action a∈{P for proportional and DP for disproportional} and receives a scalar reward r that guides off-policy updates of its action-value function Q(s,a)Q(s,a). Then, the learning proceeds using the Bellman equation:

$$Q(s,a) \leftarrow (1-\alpha)\,Q(s,a) + \alpha\big(r + \gamma \max_{a'} Q(s',a')\big)$$

with a tabular Q-table initialized to zeros. An ϵ (epsilon)-greedy policy balances exploration and exploitation, ensuring the agent samples all scenario–action pairs while progressively favouring optimal classifications. Further, taking into account the two modelling perspectives where psychological damage is not considered (i.e., Case 1) and considered (i.e., Case 2) as being part of the collateral damage component, both modelling cases are further discussed.

In Case 1, where psychological harm is excluded from collateral damage, the state is the 4-tuple (CD_injury_no_psycho, CD_death, CD_object, MA) meaning collateral effects that do not imply psychological harm, the ones that imply civilian loss of life, the ones that imply damage or destruction to civilian objects, and Military Advantage, where each of the first two dimensions takes values {L, M, H} meaning Low, Medium, and High, the third {Y,N} which means Yes and No, and the fourth {L, M, H}. This implies 3×3×2×3=54 unique states based on the inclusion or exclusion of psychological harm. At each training step, the agent samples a state uniformly, selects P or DP using ϵ-greedy (ϵ=0.1), and receives a reward of +10 if the choice matches the expert rule and –10 otherwise. In this process, the updates apply with learning rate α=0.1 and discount factor γ=0.9 iterated over several thousand episodes to populate the Q-table with the expected returns for each state–action pair.

In Case 2, psychological injury is incorporated in the collateral damage component by replacing the first feature with CD_injury_with_psycho, thereby enriching the state semantics while keeping the action set unchanged . The added complexity demands hyperparameter recalibration: α is reduced to 0.01 and ϵ is increased to 0.2 to slow Q-value uptake and prolong exploration across the expanded rule boundary . The learning loop and Bellman updates remain identical, but these conservative settings ensure that the agent visits each of the 54 states sufficiently often and refines its Q-table smoothly, enabling reliable encoding of both physical and psychological collateral-damage assessments in a single, unified tabular model.

5. Evaluation

In the first case, agent was trained on a fully deterministic rule set mapping collateral-damage and military-advantage features to a binary proportionality decision (P vs. DP). Using a learning rate of $\alpha = 0.1$, discount factor $\gamma = 0.9$, and exploration rate $\varepsilon = 0.1$ over 1 000 episodes, the total reward curve exhibits a rapid ascent from initially negative values corresponding to random exploratory actions to a stable plateau around +45 to +55 per episode by ~100 episodes. This indicates that the agent consistently selects correct actions thereafter. The resulting Q-values yield perfect classification (accuracy = 100%), as confirmed by a confusion matrix with 20 DP and 34 P cases all correctly identified (Figure 1). The smooth convergence of the reward trajectory and the clear separation of Q(P) vs. Q(DP) in sampled states demonstrate that the chosen hyperparameters strike an effective balance between exploration and exploitation, enabling the agent to fully internalize the rule structure with minimal oscillation once learned (Figure 2).

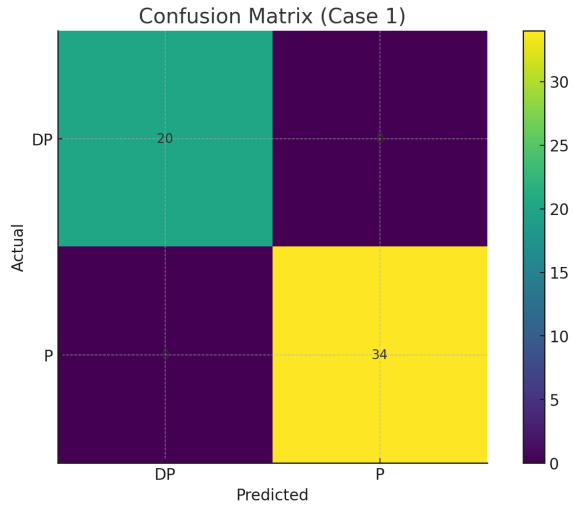

Figure 1: Confusion matrix for Case 1

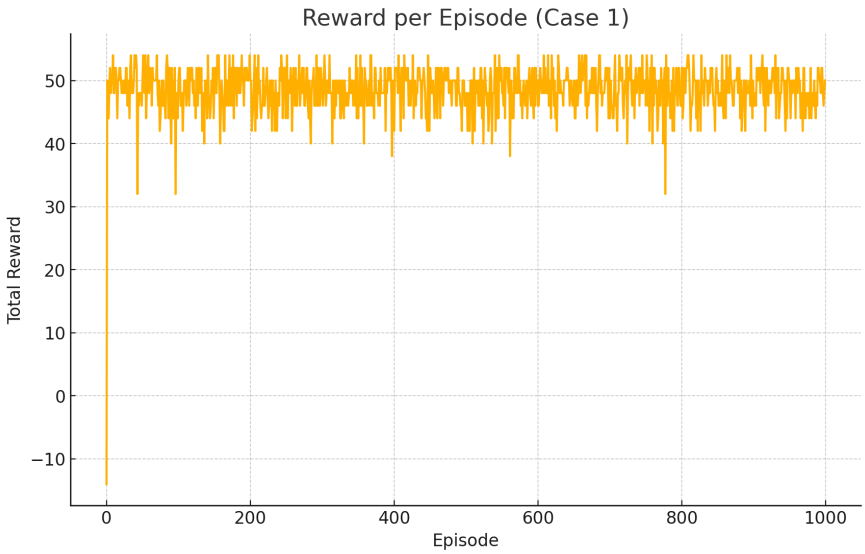

Figure 2: Reward per episode for Case 1

In the second case, when extending the rule set to include psychosocial injury, a targeted hyperparameter sweep over learning rates ($\alpha \in \{0.01, 0.05, 0.1\}$) and exploration rates ($\varepsilon \in \{0.01, 0.1, 0.2\}$) is executed. The optimal configuration ($\alpha = 0.01$, $\varepsilon = 0.2$, $\gamma = 0.9$) achieved consistent 100% accuracy while yielding a reward curve that, although more variable early on, stabilized by ~200 episodes in the +40 to +50 range. The lower learning rate reduced Q-value update magnitudes, smoothing convergence and mitigating large reward swings, while a moderate exploration rate ensured sufficient sampling of state-action pairs. The confusion matrix (22 DP and 32 P correctly classified) shows that, even with additional feature complexity, careful tuning

of α and ε delivers robust learning (Figure 3). Overall, these results confirm that Q-learning can reliably encode deterministic proportionality rules and that hyperparameter selection critically influences the speed and stability of convergence (Figure 4).

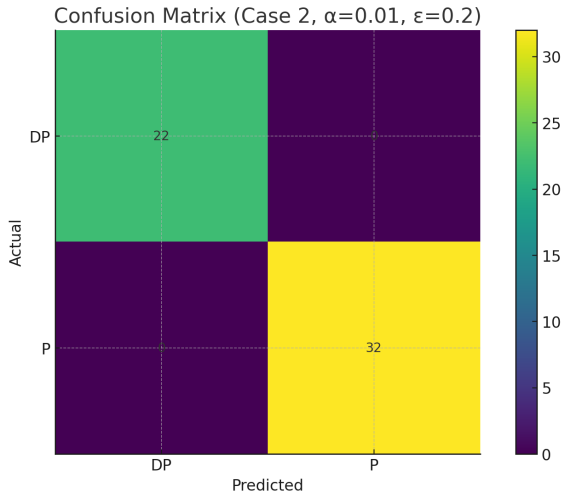

Figure 3: Confusion matrix for Case 2

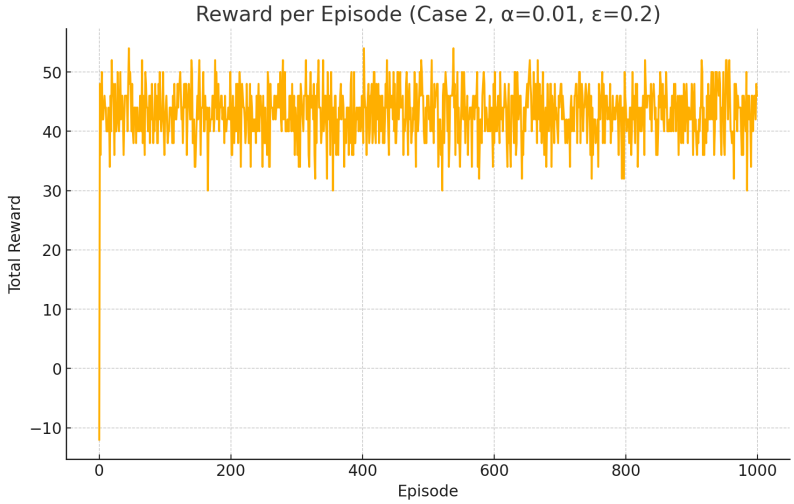

Figure 4: Reward per episode for Case 2

Although both cases achieved a good classification, their learning dynamics exhibit distinct trade-offs driven by problem complexity and hyperparameterization. Specifically, in Case 1, a relatively high learning rate (α = 0.1) and moderate exploration (ε = 0.1) enabled the agent to rapidly accrue positive reward, stabilizing by roughly 100 episodes, and converging on the correct Q-values with minimal oscillation. At the same time, the enriched feature space of Case 2 required a much lower learning rate (α = 0.01) paired with increased exploration (ε = 0.2) to smoothly integrate the additional rule patterns: although rewards initially fluctuated more broadly and convergence took closer to 200 episodes, the softer updates and sustained exploration prevented premature convergence to suboptimal actions. Hence, while in both cases accuracy is high, Case 1 had a speed of acquisition under simpler rule sets, whereas Case 2 reflected the stabilizing role of conservative learning rates and thorough exploration when encoding more complex decision boundaries.

From a learning-theoretical standpoint, both Case 1 and Case 2 demonstrate Q-learning's capacity to internalize deterministic proportionality mappings, but their Q-value trajectories and update semantics differ markedly. In Case 1, a higher learning rate (α = 0.1) yields rapid Q-value growth for the correct actions, reflected by a steep reward curve and large separation between Q(P) and Q(DP) after only ≈100 episodes, whereas in Case 2 the reduced α (0.01) produces more gradual, finely graded increments to Q-values. This conservative updating, combined with elevated exploration (ε = 0.2), preserves stochastic sampling longer into training, preventing overly confident Q-values before the full rule set is observed and smoothing the reward

signal until ≈200 episodes. In practical terms this means that one should calibrate α to the complexity of the decision space: simpler rule sets tolerate, and benefit from, larger step-sizes, while richer feature combinations favour smaller α to avoid oscillatory Q-updates. Moreover, monitoring the gap between the highest and second-highest Q-values per state can serve as a convergence diagnostic: narrow margins suggest the need for further exploration or slower learning.

6. Conclusions

This research introduces a Q-learning modelling approach to conduct the proportionality assessment in military operations considering various levels of problem complexity. In the first case scenario without psychosocial injuries, a relatively high learning rate and moderate exploration enabled rapid convergence and stable, perfect classification of proportionality decisions, underscoring the agent's ability to quickly learn deterministic rules under constrained feature spaces. When the feature set was extended to include psychosocial injury considerations, careful hyperparameter tuning toward a lower learning rate and increased exploration was necessary to achieve smooth convergence and maintain perfect accuracy. This illustrates the delicate balance required between learning speed and stability, particularly as the decision boundary complexity increases. The differentiated dynamics of Q-value updates observed between the two cases provides practical insights for tailoring reinforcement learning strategies to the complexity of military proportionality constraints, ensuring reliable encoding of operational, legal, and ethical considerations in autonomous military decision-making support systems.

From a broader AI and military operational perspective, these findings support the viability of reinforcement learning as a tool for encoding and modelling the proportionality assessment principle in settings that imply both a high level of autonomy or human-in-the-loop targeting decisions (Maathuis, 2024). The results obtained align with emerging research advocating for explainable and responsible AI systems capable of managing uncertainty and multidimensional rule sets in legally and ethically critical domains like military engagement. In addition, by enabling adaptive, robust classification through methodical hyperparameter calibration, Q-learning offers a promising pathway toward autonomous systems that respect legal and moral boundaries in warfare. From here, future implementations could consider integrating these models with function approximators such as deep Q-networks to generalize beyond deterministic rules and embed even more real-world uncertainty elements and partial observability. Specifically, two future research perspectives emerge from this research. First, investigating the resilience of such models under conditions of uncertainty and partial observability typical of combat environments, possibly through introducing controlled noise or incomplete information during training. And second, and extending the approach toward function approximation methods, such as deep reinforcement learning, to enable scalable generalization and real-time adaptability in complex, dynamic operational theatres beyond rigid constraints. These directions aim to enhance the practical utility as well as legal and ethical reliability of AI-driven proportionality assessments in future military decision support systems.

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

References

Abro, G. E. M., Ali, Z. A., & Masood, R. J. (2024). Synergistic UAV motion: A comprehensive review on advancing multi-agent coordination. *ICCK Transactions on Sensing, Communication, and Control*, *1*(2), 72-88.

Ahmed, N. U. (2022). Integrating machine learning in military intelligence process: study of futuristic approaches towards human-machine collaboration. *NDC e-journal*, *2*(1), 59-89.

Al-Fatlawi, A. A., Tamimi, K., & Al-Tamimi, M. J. A. (2024). Feasible Precautions: A Legal Study in the Stable and Variable Concept Under International Humanitarian Law. *Journal of college of Law for Legal and Political Sciences*, *10*(37), 598-629.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE signal processing magazine*, *34*(6), 26-38.

Bellman, R. (1957). A Markovian decision process. *Journal of mathematics and mechanics*, 679-684.

Benvenisti, E. (2022). The Birth and Life of the Definition of Military Objectives. *International & Comparative Law Quarterly*, *71*(2), 269-295.

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*.

Cheng, Q., Chen, D., & Gong, J. (2021). Weapon-target assignment of ballistic missiles based on Q-learning and genetic algorithm. In *2021 IEEE International Conference on Unmanned Systems (ICUS)* (pp. 908-912). IEEE.

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

Clifton, J., & Laber, E. (2020). Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, *7*(1), 279-301.

Costa, A. N., Dantas, J. P., Scukins, E., Medeiros, F. L., & Ã–gren, P. (2025). Simulation and Machine Learning in Beyond Visual Range Air Combat: A Survey. *IEEE Access*.

Criollo, L., Mena-Arciniega, C., & Xing, S. (2024). Classification, military applications, and opportunities of unmanned aerial vehicles. *Aviation*, *28*(2), 115-127.

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

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.

Henderson, I., & Reece, K. (2018). Proportionality under international humanitarian law: the reasonable military commander standard and reverberating effects. *Vand. J. Transnat'l L.*, *51*, 835.

Hussain, M. D., Rahman, M. H., & Ali, N. M. (2024). Artificial intelligence and machine learning enhance robot decision-making adaptability and learning capabilities across various domains. *International Journal of Science and Engineering*, *1*(3), 14-27.

Hernández-Rivas, A., Morales-Rocha, V., & Sánchez-Solís, J. P. (2024). Towards autonomous cybersecurity: A comparative analysis of agnostic and hybrid AI approaches for advanced persistent threat detection. In *Innovative Applications of Artificial Neural Networks to Data Analytics and Signal Processing* (pp. 181-219). Cham: Springer Nature Switzerland.

Jahn, J. L. (2019). Shifting the safety rules paradigm: Introducing doctrine to US wildland firefighting operations. *Safety science*, *115*, 237-246.

Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. *IEEE access*, *7*, 133653-133667.

Khalil, A., & Raj, S. A. K. (2024). Deployment of autonomous weapon systems in the warfare: Addressing accountability gaps and reformulating international criminal law. *Balkan Social Science Review*, *23*(23), 261-285.

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.

Lingel, S., Hagen, J., Hastings, E., Lee, M., Sargent, M., Walsh, M., ... & Blancett, D. (2020). Joint all domain command and control for modern warfare: an analytic framework for identifying and developing artificial intelligence applications.

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., & Chockalingam, S. (2022). Responsible digital security behaviour: Definition and assessment model. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1).

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).

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

Maathuis, C. (2024). Trustworthy Human-Autonomy Teaming for Proportionality Assessment in Military Operations. In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-8). IEEE. Montasari, R. (2024). Addressing Ethical, Legal, Technical, and Operational Challenges in Counterterrorism with Machine Learning: Recommendations and Strategies. In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 199-226). Cham: Springer International Publishing.

Neuman, N. (2004). Applying the rule of proportionality: force protection and cumulative assessment in international law and morality. *Yearbook of International Humanitarian Law*, *7*, 79-112.

Oliehoek, F. A., Spaan, M. T., & Vlassis, N. (2008). Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, *32*, 289-353.

Park, K., & Shim, S. (2025). Intelligent Counterforce Allocation Method Using Multi-Agent Reinforcement Learning for Ground Operations. *IEEE Access*.

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.

Premakumari, S. B. N., Sundaram, G., Rivera, M., Wheeler, P., & Guzmán, R. E. P. (2025). Reinforcement Q-Learning-Based Adaptive Encryption Model for Cyberthreat Mitigation in Wireless Sensor Networks. *Sensors*, *25*(7), 2056.

Rajagopalan, A. (2018). Active protection system soft-kill using q-learning. In *International Conference on Science and Innovation for Land Power, Australia Defence Science and Technology*.

Rajasekhar, N., Radhakrishnan, T. K., & Samsudeen, N. (2025). Exploring reinforcement learning in process control: a comprehensive survey. *International Journal of Systems Science*, 1-30.

Ray, P. P. (2024). A Review of TRiSM Frameworks in Artificial Intelligence Systems: Fundamentals, Taxonomy, Use Cases, Key Challenges and Future Directions. *Authorea Preprints*.

Ruan, W., Duan, H., & Deng, Y. (2022). Autonomous maneuver decisions via transfer learning pigeon-inspired optimization for UCAVs in dogfight engagements. *IEEE/CAA Journal of Automatica Sinica*, *9*(9), 1639-1657.

Sadhu, A. K., & Konar, A. (2020). *Multi-agent coordination: A reinforcement learning approach*. John Wiley & Sons.

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, *2*(5), 377.

Sassoli, M. (2014). Autonomous weapons and international humanitarian law: Advantages, open technical questions and legal issues to be clarified. *International Law Studies*, *90*(1), 1.

Schmitt, M. N., & Schauss, M. (2019). Uncertainty in the law of targeting: towards a cognitive framework. *Harv. Nat'l Sec. J.*, *10*, 148.

Singh, B. (2024). Unmanned aircraft systems (UAS), surveillance, risk management to cybersecurity and legal regulation landscape: unraveling the future analysis, challenges, demand, and benefits in the high sky exploring the strange new world. *Unmanned aircraft systems*, 313-354.

Singh, R., Mandal, P., & Nayak, S. (2024). Reinforcement Learning Approach in Supply Chain Management: A Review. *How Machine Learning is Innovating Today's World: A Concise Technical Guide*, 271-302.

Stellios, I., Kotzanikolaou, P., & Psarakis, M. (2019). Advanced persistent threats and zero-day exploits in industrial Internet of Things. In *Security and Privacy Trends in the Industrial Internet of Things* (pp. 47-68). Cham: Springer International Publishing.

Tan, F., Yan, P., & Guan, X. (2017). Deep reinforcement learning: From Q-learning to deep Q-learning. In *International Conference on Neural Information Processing* (pp. 475-483). Cham: Springer International Publishing.

Toroi, G. I. (2024). Rethinking military command and control systems. *Bulletin of" Carol I" National Defence University (EN)*, *4*(04), 88-112.

Tóth, A., & Farkas, T. (2023). Opportunities and Directions for the Evolution of Command and Control Systems in the Context of Multi-domain Operations.

Wang, X., Zhang, Y., & Wang, G. (2024). Target assignment for multiple stages of weapons systems using a deep Q-learning network and a modified artificial bee colony method. *Computers and Electrical Engineering*, *118*, 109378.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279-292.

Zhang, L. A., Xu, J., Gold, D., Hagen, J., Kochhar, A. K., Lohn, A. J., & Osoba, O. A. (2020). *Air dominance through machine learning: A preliminary exploration of artificial intelligence-assisted mission planning* (No. RR4311RC).

Zhang, T., & Mo, H. (2021). Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems*, *18*(3), 17298814211007305.