A Triage Ontology-Guided Knowledge Graph for Constraining Reinforcement Learning Decisions in Mass Casualty Incidents

Maham Fatima[1], Prof. Dr. Boris Tolg[1]

[1]Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Ulmenliet 20, 21033 Hamburg, Germany

Abstract

Before a Knowledge Graph (KG) is compiled into a hard safety constraint for a reinforcement learning (RL) agent, its classification logic needs to be validated against more than its own internal consistency. We take up this question in a safety-critical setting: mass casualty incident (MCI) triage. Our system couples a Deep Q-Network (DQN) agent with a reasoner-validated Web Ontology Language (OWL) / Semantic Web Rule Language (SWRL) ontology encoding the Simple Triage and Rapid Treatment (START) protocol and its chemical, biological, radiological, nuclear (CBRN) decontamination extensions. We first cross-validate two inference paths within the ontology (DL classification and SWRL-based inference, both grounded in the same START encoding), then check its base classification logic against Syn-STARTS [1], an independently constructed synthetic benchmark generated by a large language model that played no role in authoring our ontology. Agreement is perfect ($630/630$, 100.0%) on every scenario whose four required features could be mechanically and unambiguously derived from its structured fields. Only once the ontology has passed both checks do we compile it into a per-state action mask that structurally prevents the RL agent from selecting clinically contraindicated actions, rather than penalizing unsafe behavior after the fact. Across an exhaustive sweep of all 120 enumerable patient profiles and three random seeds, the constrained agent (KG-DQN) achieves a $0.0\% \pm 0.0$ violation rate and $100.0\% \pm 0.0$ triage accuracy, against $92.5\% \pm 1.8$ accuracy and $72.2\% \pm 13.4$ violations for an unconstrained DQN baseline: the safety constraint costs nothing in accuracy. We also examine, through per-step action traces, how the agent behaves on the rarest patient category - the natural stress test for a masking-based guarantee.

Keywords

Neurosymbolic AI, Knowledge Graphs, Safe Reinforcement Learning, Action Masking, Ontology-Based Reasoning, Deep Q-Learning, Mass Casualty Triage

1. Introduction

When a mass casualty incident (MCI) - a disaster, attack, or large-scale accident that overwhelms locally available medical resources - strikes, first responders must make rapid, sequential triage decisions under severe time pressure and partial information. The START protocol [2] addresses this by reducing triage to a small set of objective, repeatable checks (ambulation, breathing, pulse, mental status), assigning each patient to one of four categories: *Minor* (walking wounded), *Delayed* (serious but stable), *Immediate* (life-threatening but salvageable), or *Expectant* (unlikely to survive given resource constraints). When the incident also involves chemical, biological, radiological, or nuclear (CBRN) contamination, responders must decontaminate patients before treatment as well, following established CBRN-triage extensions to START [3]. Because the decision at each step depends on which patient, which action, and what order, all under a hard resource and time budget, this is a natural fit for reinforcement learning (RL): an agent that learns from repeated simulated exposure which actions suit which patient states, so as to maximize correctly-triaged outcomes.

A standard, "vanilla" RL agent offers no guarantee, however, that it will avoid clinically contraindicated actions - particularly for rare patient profiles it has seen only a handful of times during training. That is an unacceptable failure mode here: a single wrong action can mean misallocating scarce medical resources away from a patient who could have been saved. Prior work on safe RL addresses this through

Third Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy 2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

maham.fatima@haw-hamburg.de (M. Fatima); boris.tolg@haw-hamburg.de (Prof. Dr. B. Tolg)

© 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4313/kgnesy-paper2.pdf

shielding [4] and *invalid action masking* [5] (Section 2). We adopt the masking principle here as well, but ground its specification in a reasoner-validated Knowledge Graph (KG) rather than a hand-written formula, which situates this work within the broader neurosymbolic AI agenda of combining neural pattern recognition with symbolic verifiability [6].

Concretely, we encode the START protocol, together with CBRN hazard types and decontamination requirements, as an OWL/RDF ontology with SWRL rules. We cross-validate it internally via a Pellet reasoner [7], then check it externally against an independently constructed benchmark (Section 3), and only once it has passed both checks do we compile it into a lookup table used to train a DQN [8] agent - either as a hard action mask (KG-DQN) or not at all (Vanilla DQN). Both agents are evaluated exhaustively over all 120 enumerable patient profiles across multiple seeds, so that we can verify the safety properties hold everywhere rather than only on patients seen during training.

This short paper makes four contributions. First, a reasoner-validated, CBRN-extended OWL/SWRL ontology of the START protocol. Second, its external validation against Syn-STARTS [1], an independently constructed benchmark, carried out before the ontology constrains any agent. Third, compiling the validated ontology into a hard action mask, with multi-seed evidence that it improves both safety and accuracy over an unconstrained baseline and over a penalty-only baseline given the same contraindicated-action information. Fourth, exhaustive per-profile verification of the masking guarantee, including action traces on the rarest patient category.

2. Related Work

Knowledge Graphs as a constraint layer in neurosymbolic RL. Neurosymbolic AI aims to combine the pattern-recognition strengths of neural networks with the explainability and verifiability of symbolic knowledge representation [6]. Within RL, the dominant approach to enforcing safety is *shielding*: a reactive monitor, synthesized from a formal specification, intervenes to correct or block a violating action [4]. A simpler, closely related mechanism is *invalid action masking*, which removes illegal actions from the policy's distribution altogether; Huang and Ontañón [5] show this corresponds to a theoretically valid policy-gradient update, and that it matters increasingly as the invalid-action space grows. We adopt this masking formulation here as well, but draw the specification from a reasoner-validated ontology rather than a hand-authored formula (Section 3.1). This connects our masking layer to a separate, substantial body of work combining Knowledge Graphs with RL - work that has mostly used KGs as a *state representation* or *reward-shaping* signal rather than a hard action constraint [9]. As far as we are aware, no prior work has compiled a reasoner-validated, externally-checked clinical-protocol ontology into a structural action mask for an RL triage agent.

Triage protocols, synthetic benchmarks, and AI-assisted mass casualty response. START [2] remains the most widely used mass casualty triage protocol, with documented CBRN-specific extensions for decontamination sequencing [3]. A growing body of work applies RL to MCI response more broadly, covering ambulance redeployment, resource allocation, and patient transfer scheduling; an RL-trained decision-support agent has recently been shown to match or exceed expert trauma-surgeon performance on simulated patient-transfer decisions [10]. That line of work targets logistics around a given triage category, whereas our focus is complementary: we constrain the triage decision itself, so that the protocol's clinical rules are structurally guaranteed to hold rather than merely statistically likely to. Separately, because real-world MCI records are rare and costly to collect, several efforts have turned to synthetic triage benchmark generation. Syn-STARTS [1] uses an LLM-based framework to produce structured START scenarios at a scale unavailable from real incident data, and reports that its cases are qualitatively indistinguishable from a manually-curated benchmark. We repurpose its structured, machine-readable fields for a different purpose than the LLM evaluation it was built for: an external check on a reasoner-backed ontology's classification logic, made possible precisely because its scenarios were generated independently of our ontology.

3. Method

Figure 1 summarizes the validate-then-constrain pipeline this section describes.

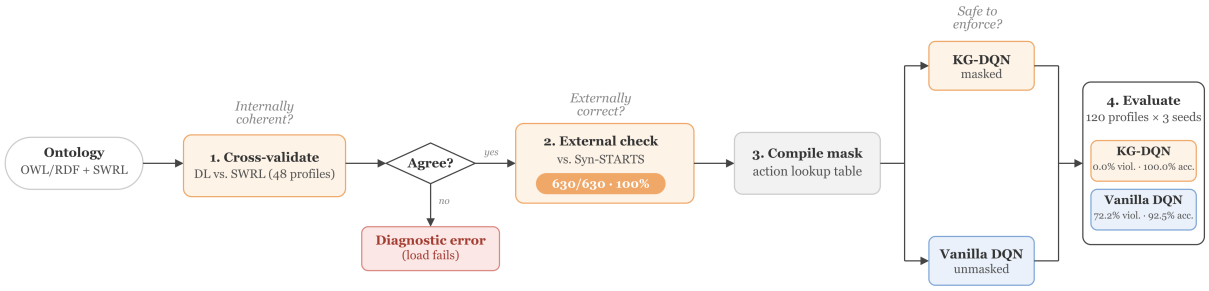


Figure 1: The validate-then-constrain pipeline. Step 1 cross-validates two inference paths, standard Description Logic (DL) classification and SWRL-based inference, over the 48 clinically coherent profiles.

3.1. Ontology and reasoning layer

The symbolic component is an OWL/RDF ontology, which we refer to as the ruleset. It encodes the START categories *Minor*, *Delayed*, *Immediate*, and *Expectant* as classes defined via `equivalentClass` axioms (Table 1) over ambulation, respiratory status, respiratory rate (used to derive tachypnea, per START's own rate threshold), pulse, mental status, and decontamination status.

Table 1
START protocol categories and their OWL/SWRL encoding. [a]Tachypnea (rate > 30) is derived by two helper rules, `rule3b_tachypnea` and `rule3d_not_tachypneic`, which label tachypnea but do not assign a category directly.

Category	Encoded condition	SWRL rule(s)	Severity
Minor	Ambulatory	`rule1_minor`	0
Delayed	Non-ambulatory, breathing, pulse present, follows commands, not tachypneic	`rule6_delayed`	1
Immediate	Non-ambulatory AND (not breathing with pulse; OR no pulse; OR breathing+pulse but not following commands; OR breathing+tachypneic)[a]	`rule3_immediate_no_breath`, `rule3c_tachypnea_immediate`, `rule4_immediate_no_pulse`, `rule5_immediate_no_commands`	2
Expectant	Non-ambulatory, not breathing, no pulse	`rule2_expectant`	3

We use *Knowledge Graph*, *ontology*, and *ruleset* to refer to different stages of the same artifact: the ontology (OWL/RDF classes and properties plus SWRL rules) is the authored, human-inspectable knowledge source. At load time we query the reasoner once for every clinically coherent profile and compile its output into a static per-profile lookup table, a dictionary from profile signature to a permitted-action vector; at runtime, the RL agent consults only this compiled table and never touches the ontology or reasoner directly. Expressing the ruleset in OWL/SWRL, rather than as hand-written conditionals, is what makes the internal cross-validation below possible in the first place: the same clinical logic is encoded twice, once via `equivalentClass` axioms and once via SWRL rules, and checking one against the other gives a consistency guarantee that a single hand-written rule set could not provide by construction. A plain lookup table could reproduce the runtime behavior, but it could not reproduce this authoring-time guarantee.

CBRN contamination is represented as a `HazardType` class with four disjoint subclasses, each carrying a `hasDecontaminationDuration` property (chemical: 1, biological: 2, radiological: 2, nuclear: 3) that sets decontamination time without affecting the triage category itself. SWRL rules [11] independently infer a `treatmentRequired` action and a `hasSeverityScore` for each profile.

We load and query the ontology through owlready2 [12], using a Pellet reasoner [7] chosen for its ability to execute SWRL rules alongside standard Description Logic (DL) classification (OWL-DL).

This lets us cross-validate two inference paths, DL classification and SWRL inference, for every profile in a 6-dimensional vital-sign space (ambulation, breathing, pulse, mental status, decontamination status, tachypnea). Of the $2^6 = 64$ possible combinations, 48 are clinically coherent; the remaining 16, tachypneic while not breathing, are excluded as clinically incoherent at load time. Because both inference paths are grounded in the same ontology, agreement between them is an internal consistency check rather than independent evidence of correctness - Section 3.2 supplies the external check, against a benchmark built with no reference to our ontology at all. For each profile, the DL-classified category must agree with the SWRL-inferred `treatmentRequired` action and severity score, or loading fails with a diagnostic error. This catches authoring errors where a class definition and a rule silently disagree; a bug present identically in both encodings would still pass undetected, which is exactly what the external check is for.

We additionally verify the tachypnea threshold at its boundary (rates 29/30/31 across all 16 remaining-vital combinations, 48 further profiles genuinely re-run through Pellet rather than the two representative values 18/35 used elsewhere): DL and SWRL agree on every profile, correctly classifying rate $= 30$ as *Delayed* and rate $= 31$ as *Immediate* - confirming a strict "greater than 30" cutoff at the one point an off-by-one error would most likely hide.

From this lookup table we derive, for every profile, a set of KG-permitted actions. Each of the four tag actions (*tag_minor*, *tag_delayed*, *tag_immediate*, *tag_expectant*) is permitted only for the category it corresponds to, with one clinically motivated relaxation: a *Delayed* patient may also be tagged *tag_immediate*, since a responder is always permitted to escalate a stable patient to a higher-acuity category. Escalating into, or de-escalating out of, the *Expectant* category is never permitted from any other category, which makes the *Expectant* boundary symmetric: a patient is tagged *Expectant* if and only if the ontology classifies them as such. This rules out both escalating an *Expectant* patient to *Immediate*, which would waste scarce resources on someone unlikely to survive, and abandoning an *Immediate* or *Delayed* patient to *Expectant*, which would withhold resources from someone who could be saved. We treat the ontology's classification as ground truth for masking purposes, validated independently in Section 3.2; this is not because we take it to be infallible, but because its calls have to be enforceable one way or the other.

A second, independent contraindication rule governs contamination: no tag action, and no *treat*, is permitted for a patient who has not yet completed decontamination, a CBRN precondition absent from the base START protocol. This applies uniformly across all four categories, including *Expectant*, since an undecontaminated casualty remains a contamination risk to responders and other patients in the same treatment corridor regardless of triage category. Only *open_airway* and *decontaminate* itself stay available regardless of decontamination status.

3.2. External validation against an independently generated synthetic benchmark

The cross-check above establishes that the ontology is internally consistent, in that its two reasoning paths agree with each other, but internal consistency alone does not establish that the encoded clinical judgments are correct. Before using the ruleset to constrain any learning agent, we therefore check its base classification logic against scenarios it had no part in producing: Syn-STARTS [1], a benchmark of 2,000 structured triage scenarios generated by an LLM-based framework built specifically to overcome the scarcity and collection cost of real MCI data. Each scenario pairs a free-text patient description with a structured `vitals_info` object and a gold triage tag, evenly distributed across the four START categories (500 scenarios each for Green/Minor, Yellow/Delayed, Red/Immediate, and Black/Expectant). We use only the structured `vitals_info` fields, deliberately setting the free-text description aside, so that every feature fed to the ontology is mechanically derived rather than manually interpreted.

We apply a locked, mechanical conversion rule, fixed before inspecting any results (full specification in the code repository), to extract the four ontology-relevant binary features plus respiratory rate from each scenario's structured fields. It draws only on directly provided boolean fields and recognized START perfusion/airway checks [2]; free-text interpretation plays no part. A scenario is included only if all four features come out unambiguous under this rule, and any scenario missing a required field

is excluded outright. Classification then runs through the same reasoner-backed pipeline described above - not a separate reimplementation of it. Of the 2,000 scenarios, 630 (31.5%) have all four required features mechanically derivable under this rule. The remaining 68.5% are excluded because a required field is absent from `vitals_info`; most often this is pulse (859 cases), followed by breathing/rate (464) and mental status (47), with that information present only in the free-text description we do not draw on for feature extraction.

On these 630 scoreable scenarios, our ontology's classification agreed with the benchmark's gold label on every case (630/630, 100.0%), with no disagreement in any of the four categories. This is a necessary precondition for trusting the ontology as an RL action mask, though a more modest one than agreement with real clinical narrative would be, since Syn-STARTS scenarios arrive as structured fields rather than free clinical prose that needs interpreting (Section 5 returns to this distinction).

Three caveats apply here. First, coverage is highly uneven across categories, so 630/630 is far stronger evidence for the Delayed boundary than for the rarer ones (Section 5 gives the full breakdown). Second, Syn-STARTS' gold labels are themselves generated according to the START protocol, so agreement confirms that our encoding matches an independently-authored reading of START, not independent clinical ground truth. Third, because the benchmark is LLM-generated, its scenarios inherit whatever regularities the generating model reproduces from START's textbook description, rather than the variability of real clinical presentations. Hagiwara et al. [1] partially address this by reporting that their cases are qualitatively indistinguishable from a manually-curated benchmark, though we cannot verify that independently from the structured fields alone.

Had the external check surfaced any disagreement, we would not have adjusted the benchmark or explained the mismatch away; instead we would have returned to the ontology itself, located the SWRL rule or class definition responsible, and revised it against the START protocol directly. A disagreement that persisted after such revision would be reason to withhold that specific rule from the action mask rather than compile it into a hard safety constraint - a targeted response, not grounds to discard the ontology as a whole. We hold ourselves to this standard partly because the ontology is not built for this system alone: we intend to carry it forward as the basis for a structured mass-casualty-incident dataset, an area where good, clinically grounded structured data remains scarce.

The remainder of this paper builds the RL system on top of this validated ontology.

3.3. Environment and action space

The MCI environment is implemented with Gymnasium [13]. Each episode presents 10 patients, generated according to fixed priors ($P(\text{ambulatory}) = 0.3$, $P(\text{breathing}) = 0.8$, $P(\text{pulse}) = 0.75$, $P(\text{follows commands}) = 0.6$, $P(\text{decontaminated}) = 0.6$, $P(\text{tachypneic} \mid \text{breathing}) = 0.10$, hazard type uniform given contamination), consistent with reported tachypnea rates among stable trauma casualties. This balanced category distribution is a deliberate stress-testing choice rather than a claim about real MCI epidemiology. Under these priors, *Expectant* patients (non-ambulatory, not breathing, no pulse) are rare by construction (∼1-2% of patients); Section 4.2 reports how this rarity shows up in trained-agent behavior. During training, 15% of otherwise-stable patients have their label stochastically escalated, to model clinician override; the exhaustive evaluation in Section 4 instead uses the deterministic protocol label, since our question there is whether the policy matches the protocol, not whether it matches the noisier training signal.

The observation is a 12-dimensional vector: five binary vital signs, a treated flag, a tachypnea flag, a one-hot hazard-type encoding, and a decontamination-progress fraction. The action space has seven discrete actions: `tag_immediate`, `tag_delayed`, `tag_minor`, `tag_expectant`, `open_airway`, `treat`, `decontaminate`. Only a tag action ends a patient's turn; the three non-tag actions keep the patient active for up to 5 steps, after which the episode advances with a timeout penalty. The DQN selects among all seven at every timestep; none are scripted. *treat* and *open_airway* are single-step actions that flip a boolean flag; *decontaminate* requires 1-3 repeated selections depending on hazard type (chemical: 1, biological: 2, radiological: 2, nuclear: 3). Five is a generous per-patient budget, covering the worst case with room to spare; a clean, stable patient can resolve in a single step.

Reward shaping is intentionally simple (Table 2), designed to discourage stalling without biasing the agent toward or away from any particular category. In the KG-DQN condition, a KG-contraindicated action is structurally unavailable, blocked by a per-step boolean mask returned by the ontology lookup. This follows the invalid-action-masking formulation of Huang and Ontañón [5] rather than the shield-as-corrector formulation of Alshiekh et al. [4], and guarantees a 0% violation rate by construction (Section 5 verifies this empirically). Vanilla DQN has neither the mask nor an analogous reward signal: its objective never penalizes tagging a still-contaminated patient, so it treats contaminated and decontaminated patients identically once tagged. That is deliberate rather than an oversight - it is exactly the unconstrained baseline the comparison requires (Section 4.2). Decontaminating Expectant patients before tagging protects others rather than improving their own survival odds; we flag the resulting tradeoff, that decontamination time spent on Expectant patients delays decontamination of salvageable ones, as a limitation worth revisiting against real triage doctrine.

Table 2
Reward shaping. [a]Redundant `decontaminate`/`open_airway`/`treat`. [b]Applied to the executed action for Penalized DQN (Section 4.1); applied to the rejected, masked-out action as a follow-up test for KG-DQN (Section 5).

	Correct tag	Incorrect tag	Non-tag (step)	Redundant[a]	Budget exhausted	KG-contraind.[b]
Reward	$+20$	-10	-0.5	-1.0	-10	-50

3.4. Agent and training

The Q-network is a two-hidden-layer MLP (64 units per layer, ReLU activations) mapping the 12-dimensional observation to Q-values over the 7 discrete actions. Hyperparameters, shared across all conditions, are: replay buffer capacity 5000; target network synced every 20 episodes; ϵ schedule $1.0 \rightarrow 0.05$ with decay 0.995/episode; discount factor $\gamma = 0.99$; batch size 32; Adam optimizer [14] with learning rate 10^{-3}; and 500 training episodes. We select checkpoints by evaluating candidates every 20 episodes against the full 120-profile enumeration, rather than a rolling training-episode average, because the skewed patient-prior distribution could otherwise let an undertrained checkpoint look artificially strong simply by chance-missing the rarest category within a short window; the uniform enumeration weights every category equally regardless of how rare it is during training.

4. Experiments and Results

4.1. Multi-seed exhaustive evaluation

We evaluate both agents on all 120 enumerable initial patient profiles (the 16 binary combinations of the four vital-sign features, each further split by tachypnea status when breathing, crossed with clean or one of four hazard types) under a greedy ($\epsilon = 0$) policy, repeated across three random seeds (42, 43, 44) with independent training runs. This is deliberately a different evaluation regime from training: every profile is weighted equally regardless of its real-world prior probability, and the ground-truth label is the deterministic protocol label rather than the noisy training one. We also evaluate a third condition, *Penalized DQN*, which receives the same -50 contraindicated-action penalty as KG-DQN (Table 2) but is not masked, isolating what the mask itself contributes beyond the reward signal alone.

Table 3 summarizes the result: KG-DQN's violation rate is zero on every seed, and its accuracy is perfect throughout. This is stronger than masking alone guarantees, since the mask only prevents unsafe actions while the network still has to act decisively, and the trained policy clears both bars on all three seeds. The unconstrained baseline reaches neither figure on any seed, and its accuracy and violation rate vary meaningfully across seeds ($92.5\% \pm 1.8$ accuracy, $72.2\% \pm 13.4$ violations); nothing in Vanilla's training objective anchors it to a consistent outcome the way the mask does for KG-DQN. Its violation rate reflects that the CBRN decontamination precondition (Section 3.1) applies to every

Table 3
120-profile exhaustive evaluation, mean ± std across three seeds. Penalized DQN receives the same −50 penalty as KG-DQN but is not masked; full per-seed breakdown in the repository.

Agent	Accuracy	Violation rate	Mean actions/patient
Vanilla DQN	$92.5\% \pm 1.8$	$72.2\% \pm 13.4$	1.22 ± 0.32
Penalized DQN	$98.9\% \pm 0.8$	$0.6\% \pm 0.4$	2.64 ± 0.02
KG-DQN	$100.0\% \pm 0.0$	$0.0\% \pm 0.0$	2.60 ± 0.00

tag action, not only *treat*: a policy with no incentive, and, as Section 3.3 notes, no reward signal, to decontaminate before tagging violates the precondition on most contaminated profiles regardless of whether the tag is clinically correct. Of the 120 evaluated profiles, the mask permits exactly one legal tag for 115 of them (Minor, Immediate, and Expectant categories), so accuracy on these is guaranteed once the agent tags at all, independent of learning. Only the 5 Delayed profiles admit two legal tags (Delayed or an escalation to Immediate), making them the sole test of whether the network learned a meaningful distinction rather than the mask simply removing the wrong options; KG-DQN tags all 5 correctly on every seed. A rules-only policy, no learning, always following the ontology's recommended tag, also reaches 120/120 accuracy and zero violations, with the same mean actions/patient as KG-DQN (Table 3); Vanilla DQN's lower mean reflects tagging before completing decontamination, not genuine efficiency.

To isolate what the mask itself contributes beyond the contraindicated-action penalty, Penalized DQN reaches $98.9\% \pm 0.8$ accuracy and $0.6\% \pm 0.4$ violations (mean 2.64 ± 0.02 actions/patient) - closer to KG-DQN than the fully unconstrained Vanilla baseline, but neither perfect nor exactly zero on any individual seed (seeds 42 and 43 each retain one violation out of 120, even after full, converged training). KG-DQN, by contrast, reaches $100.0\% \pm 0.0$ accuracy and $0.0\% \pm 0.0$ violations on every seed, with no exceptions. The gap is small in absolute terms but qualitatively different: KG-DQN's zero is a structural property of its action space, true independent of training outcome, whereas Penalized DQN's near-zero is an empirical training result that came close but did not consistently reach zero even with generous training.

Figure 2 confirms this pattern holds continuously across the entire rarity spectrum, not merely on average, with the minimum correctness point (0.5) corresponding to the same Expectant-category profiles discussed in Section 4.2.

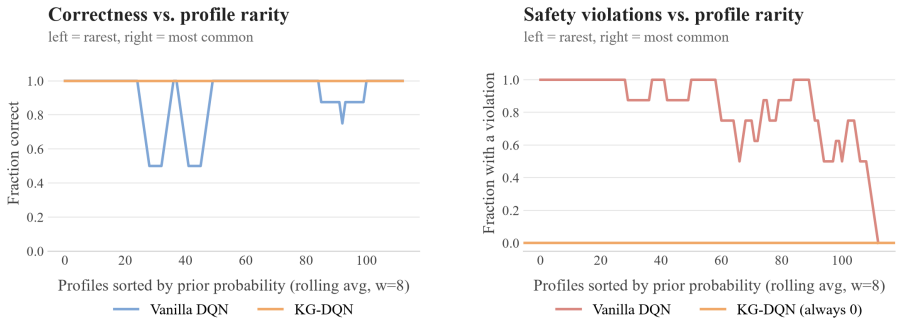


Figure 2: Correctness (left) and violation rate (right) vs. profile rarity (8-profile rolling average, rarest on the left), seed 42. Vanilla DQN's correctness dips only for the rarest (Expectant) profiles, while its violation rate stays high throughout; KG-DQN is flat at perfect correctness and zero violations everywhere.

4.2. Per-profile verification: rarity and contamination

On seed 42, Vanilla DQN's 10 *incorrect* profiles are exactly the 10 *Expectant* patients across all five decontamination/hazard variants, each over-triaged to `tag_immediate`: the error the symmetric boundary rule (Section 3.1) rules out. Vanilla's 98 *violations* are a larger superset: the remaining 88 are profiles tagged with the clinically *correct* category while still contaminated, a CBRN Rule 5 violation

independent of triage accuracy, since Vanilla ties tagging to decontamination status not at all. KG-DQN tags all 10 correctly, with zero violations, on every seed. For the rarest profile (not ambulatory, not breathing, no pulse, does not follow commands, chemical hazard), the agent decontaminates once ($+1.5$) then immediately tags *tag_expectant* ($+20.0$). All 10 Expectant-category traces show the same pattern: a short, monotonic sequence ending in the one tag the mask allows, unsurprising since *tag_expectant* is the sole valid tag under the mask (Section 3.1), leaving the agent nothing contraindicated to drift toward even during near-random exploration.

5. Discussion and Limitations

The constraint did not cost accuracy. Safety constraints in RL are commonly assumed to trade performance for safety; we observe no such trade-off: KG-DQN beats Vanilla's accuracy on every seed (Table 3), not just matches it. With only three seeds we stop short of a guaranteed advantage - the defensible claim is that the constraint is never accuracy-costly here. To distinguish the mask's 0% violation rate from whether the network itself ever preferred a contraindicated action, we logged its raw, unmasked greedy action at every step. It preferred a contraindicated action on $78.3\% \pm 25.4$ of the 120 profiles at some point in the episode, confirming the 0% rate is a structural property of the mask, not learned behavior: the network never experiences a contraindicated tag under masking, so nothing teaches it to avoid the preference - only the mask prevents acting on it.

Two follow-ups on the -50 penalty. The penalty in Table 2 plays two distinct roles in this paper, worth disentangling. First, applied to *KG-DQN's own* raw, masked-out preference (the rejected action, never executed) rather than anything it actually does, it cut the mean shadow-violation rate from $78.3\% \pm 25.4$ to $29.2\% \pm 33.3$ at a small accuracy cost (100.0% to $98.6\% \pm 1.4$) - not yet characterized well enough to know if it can be tuned away. Second, and separately, this is the same penalty that defines the standalone *Penalized DQN* agent from Table 3 - so we asked whether its near-zero violation rate is a property of the trained policy or an artifact of sufficient training: at 50 episodes its accuracy collapses to $9.2\% \pm 2.5$ with $64.4\% \pm 30.0$ violations, still $3.3\% \pm 2.5$ at 150 episodes, while KG-DQN is unaffected at either length ($99.7\% \pm 0.4$ accuracy, 0.0% violations). Counting exactly rather than by average: across all nine runs in this paper (three seeds $\times$ three training budgets), KG-DQN reaches exactly 0.0% violations every time; Penalized DQN reaches exactly 0.0% in only one of nine.

A general, revisable pattern: validate-then-constrain. Beyond MCI, this paper instantiates a transferable, three-question pattern for using Knowledge Graphs in neurosymbolic systems (Figure 1): is the KG internally coherent, externally correct, and safe to enforce? The ontology remains inspectable and revisable without retraining. We see validate-then-constrain as a candidate building block toward a broader taxonomy of KG-NeSy integration patterns, complementing reward-shaping or state-representation uses of KGs in RL [9].

What the synthetic-benchmark validation establishes. The 630/630 agreement (Section 3.2) is necessary but not sufficient: coverage is uneven across categories - near-complete for *Delayed* (498/500) and strong for *Immediate* (110), but thin for *Minor* and *Expectant* (20 and 2). Validating against real, free-text clinical narrative - where a feature like "breathing" requires interpretation rather than a clean structured lookup - is flagged as future work.

Limitations. The RL component is evaluated on a synthetic domain with a hand-specified patient-prior distribution, not real incident data. External validation (Section 3.2) covers only the base START categories; the CBRN extension is validated only by the internal reasoner cross-check and the 120-profile sweep. Nor has the ontology been reviewed by a clinician or EMS-trained triage practitioner - the Syn-STARTS check served as our primary sanity gate before treating it as enforceable, not clinical vetting. Three seeds is a modest sample for KG-DQN's perfect record; more would rule out coincidence.

Finally, masking is only as correct as the ontology it is built from: an encoding error would produce a policy that's safely *wrong* instead of unsafely wrong, a real risk given how often the protocol itself has been revised [3].

6. Conclusion

We presented a neurosymbolic DQN agent for CBRN-extended mass casualty triage that compiles a reasoner-validated, externally-validated OWL/SWRL ontology into a hard action mask rather than a reward penalty. The constrained agent achieves zero violations without sacrificing accuracy, while the unconstrained baseline violates the CBRN decontamination precondition regardless of triage correctness. We see this as a concrete instance of the broader KG-NeSy strand of using Knowledge Graphs for verifiably safer neurosymbolic systems.

Acknowledgments

Code, the full ontology (`OWL_Ontology.rdf`), and per-profile evaluation results, including the 120-profile breakdown underlying Table 3 and the external validation results in Section 3.2, are available at https://github.com/mahamfatima123/kg-nesy-mci-triage, including an automated test suite covering the ontology, action masking, and evaluation pipeline.

Declaration on Generative AI

During preparation of this work, the author(s) used Claude (Anthropic) to assist with debugging code and checking grammar, spelling, and phrasing. All literature search, scientific decisions, experimental design, and interpretation of results were performed by the human author(s), who take full responsibility for the publication's content.

References

[1] C. Hagiwara, N. Nonaka, Y. Hashimoto, R. Uchimido, J. Seita, Syn-STARTS: Synthesized START triage scenario generation framework for scalable LLM evaluation, 2025. arXiv:2511.14023.

[2] G. Super, S. Groth, R. Hook, J. Walden, J. Tepas, W. Volz, START: Simple Triage and Rapid Treatment Plan, Technical Report, Hoag Memorial Presbyterian Hospital, Newport Beach, CA, 1994.

[3] D. C. Cone, K. L. Koenig, Mass casualty triage in the chemical, biological, radiological, or nuclear environment, European Journal of Emergency Medicine 12 (2005) 287–302. doi:10.1097/00063110-200512000-00009.

[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, U. Topcu, Safe reinforcement learning via shielding, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), volume 32, 2018, pp. 2669–2678. doi:10.1609/aaai.v32i1.11797.

[5] S. Huang, S. Ontañón, A closer look at invalid action masking in policy gradient algorithms, in: Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference (FLAIRS-35), 2022. doi:10.32473/flairs.v35i.130584.

[6] A. d'Avila Garcez, L. C. Lamb, Neurosymbolic AI: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. doi:10.1007/s10462-023-10448-w.

[7] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, Pellet: A practical OWL-DL reasoner, Journal of Web Semantics 5 (2007) 51–53. doi:10.1016/j.websem.2007.03.004.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, et al., Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533. doi:10.1038/nature14236.

[9] Q. Huo, H. Fu, C. Song, Q. Sun, P. Xu, K. Qu, H. Feng, C. Liu, J. Ren, Y. Tang, T. Li, Knowledge graph based on reinforcement learning: A survey and new perspectives, IEEE Access 12 (2024) 161897–161924. doi:10.1109/ACCESS.2024.3479774.

[10] Z. Liu, W. H. Souza, J. Han, A. Madani, Using AI to optimize patient transfer and resource utilization during mass-casualty incidents: A simulation platform, 2025. arXiv:2509.08756.

[11] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission, World Wide Web Consortium (W3C), 2004. URL: https://www.w3.org/Submission/SWRL/.

[12] J.-B. Lamy, Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies, Artificial Intelligence in Medicine 80 (2017) 11–28. doi:10.1016/j.artmed.2017.07.002.

[13] M. Towers, A. Kwiatkowski, J. Terry, et al., Gymnasium: A standard interface for reinforcement learning environments, arXiv preprint arXiv:2407.17032 (2024).

[14] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.