# BRIDGING EFFICIENCY AND MORALITY: ETHICAL LTL CONSTRUCTION FOR FASTER CONVERGENCE IN Q-LEARNING'S ETHICAL DILEMMAS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

AI agents make decisions before executing actions, and we expect AI systems to make these decisions independently in large-scale autonomous deployments. Decisions made by AI agents are vital for preventing harm to humans and prioritizing human lives, both of which are considered moral objectives. Thus, the objective of building a general-purpose robot is to establish a mechanism for monitoring ethical decision-making by studying the decision-making process in partially observable Markov decision process (POMDP) environments. The challenge is how to generate an ethical linear temporal logic specification for model checking to ensure that the decisions made by AI are moral. Existing approaches use rule-based metrics or utility functions to constrain AI decision-making, but they are easily compromised because they lack a reasoning system. Our proposed approach is designed from a psychological perspective and trains the agent to learn a model-checking logic system without prior knowledge. We evaluate the approach on a modified burning room problem and demonstrate its robustness and flexibility.

## 1 INTRODUCTION

Reinforcement learning (RL) improves a policy through trial and error, but an ethically consequential failure can occur before a delayed reward teaches the agent to avoid it. This mismatch is especially visible under partial observability: an agent may need to gather information, maintain an obligation over several steps, and avoid irreversible actions while the relevant object or outcome remains uncertain. A single penalty can discourage failure, yet it does not explicitly represent the required order of events.

We study this problem in a modified burning-room dilemma. A robot collects gold while an object in a burning room may be a human or a rock. The object type is hidden until the robot evaluates it, and the evaluation can be noisy. The ethical obligation is sequential: evaluate the object and, when a human is detected, do not exit or destroy the object before rescue. This controlled environment isolates a practical question: can explicit temporal progress and pre-emptive action filtering reduce ethical violations while useful behavior is still being learned?

Our construction combines three familiar ingredients. A human-authored linear temporal logic (LTL) rule is translated into a small progress monitor; the monitor state is appended to the tabular Q-learning state; and a pre-emptive shield removes actions that would immediately violate a pending rescue obligation. The agent learns action values without knowing the transition dynamics or the hidden object type in advance. It does not learn or synthesize the LTL formula itself.

The paper makes three bounded contributions. First, it specifies an auditable ethical-decision benchmark with partial observability, delayed consequences, and proposition noise. Second, it gives a compact product-state Q-learning and shielding construction for the evaluate-before-act obligation. Third, across 30 seeds, the complete method reaches a prespecified joint return–violation criterion earlier than vanilla and penalty Q-learning, incurs fewer training violations, and achieves the best final return–safety trade-off among the tested methods. An observation-noise sweep exposes the main limitation: when the proposition detector is wrong, the shield can protect the wrong inferred state.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

## 2 Background and Related Work

**Safe reinforcement learning.** Safe RL separates task performance from constraints that should hold during learning or deployment (García & Fernández, 2015; Altman, 1999; Achiam et al., 2017). Reward penalties are simple, but their scale determines a difficult trade-off and may leave unsafe exploration during early learning (Amodei et al., 2016; Ray et al., 2019). Recent methods instead model uncertainty or safety constraints explicitly. For example, ActSafe explores optimistically with respect to reward-relevant uncertainty while remaining pessimistic about safety constraints (As et al., 2025); Safe SLAC uses a stochastic latent representation for constrained RL from pixels (Hogewind et al., 2023). Our study is much smaller and tabular: it asks whether a transparent temporal monitor and action shield improve one partially observable ethical dilemma, not whether they outperform general safe-RL algorithms.

**Temporal logic and structured reward.** LTL describes trajectory-level requirements with Boolean and temporal operators (Pnueli, 1977; Baier & Katoen, 2008). Automaton or product-state constructions turn temporal progress into Markovian state for learning (Hasanbeig et al., 2020). Reward machines similarly expose finite-state reward structure and can improve sample efficiency (Toro Icarte et al., 2022). More recent work differentiates automaton-based objectives to accelerate continuous-control learning from LTL specifications (Bozkurt et al., 2026). We adopt the same broad insight—temporal progress can be made explicit—but use a deliberately finite monitor and tabular learner so that every transition is inspectable.

**Shielding and ethical dilemmas.** Shielding places a formal action filter around a learner. Preemptive shields provide a set of currently safe actions, whereas post-posed shields correct a selected unsafe action (Alshiekh et al., 2018). Shielding has also been studied under partial observability, where uncertainty in the inferred state complicates the safety decision (Carr et al., 2023). Separately, controlled dilemmas have been proposed as probes of machine ethical decision making (Abel et al., 2016; Bjørgen et al., 2018). We connect these threads without claiming a general moral agent: the ethical priority is encoded by the designer, and the experiment measures behavior under that specification.

## 3 Problem Setup

### 3.1 Partially observable mining dilemma

We model the environment as a finite POMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma \rangle$. The four locations are the start, gold room, burning room, and exit. The action set follows the original draft:

$$\mathcal{A} = \{\text{DESTROY}, \text{RESCUE}, \text{CROSS}, \text{DROP}, \text{COLLECT}, \text{EVALUATE}\}. \tag{1}$$

The latent state records location, gold status, object type, resolution status, and the step counter. The observation replaces the hidden object type by one of {unknown, perceived human, perceived rock}. At the start of each episode, the object is a human with probability $0.2$. The EVALUATE action reveals a binary report, flipped with probability $\eta$.

The task reward preserves the draft's nominal priorities. Each step costs $1$; collecting gold adds $20$; evaluating adds $5$; rescuing a true human adds $200$; destroying an object adds $10$; and dropping carried gold at the exit adds $1$. Rescuing a rock incurs an additional cost of $10$. An episode-level ethical violation occurs if the agent destroys a true human, exits while a true human remains unresolved, or times out with a true human unresolved. These violations are measured from latent ground truth, not from the agent's noisy report.

### 3.2 Learning objective and evaluation measures

All learners maximize discounted return with $\gamma = 0.97$. We report task return using only the environment reward, so comparisons are not inflated by method-specific shaping. Safety is reported as the fraction of episodes with at least one ethical violation. Rescue rate is conditional on episodes in which the object is truly human. This denominator matters because a method cannot appear safe merely by encountering fewer human episodes.

We additionally define a descriptive joint learning criterion: the first 100-episode window whose mean task return is at least 30 and whose violation rate is at most 0.10. The criterion is a common ruler for the three learning curves, not a theorem about asymptotic convergence.

# 4 Ethical LTL Construction

## 4.1 From obligation to a progress monitor

Let $e$ denote that the object has been evaluated, $h$ that the current report identifies a human, $x$ an exit event, and $r$ a successful rescue event. We express the intended observable obligation as

$$\varphi = \mathbf{F}e \ \wedge \ \mathbf{G}\big(h \rightarrow (\neg x \ \mathbf{U} \ r)\big). \tag{2}$$

Equation 2 requires an evaluation eventually and, after a human report, forbids exit until rescue. The implementation also treats destruction of a true human and unresolved timeout as violations in the benchmark metric.

The formula is translated into monitor states $q \in \{q_0, q_{\text{rock}}, q_{\text{human}}, q_{\text{done}}, q_\perp\}$. State $q_0$ means that evaluation is pending; $q_{\text{rock}}$ records a perceived rock; $q_{\text{human}}$ records a pending rescue obligation; $q_{\text{done}}$ records a completed rescue; and $q_\perp$ records a violation. The learner observes the product state $z = (o, q)$, where $o$ is the ordinary environment observation. Figure 1 summarizes the resulting control loop.

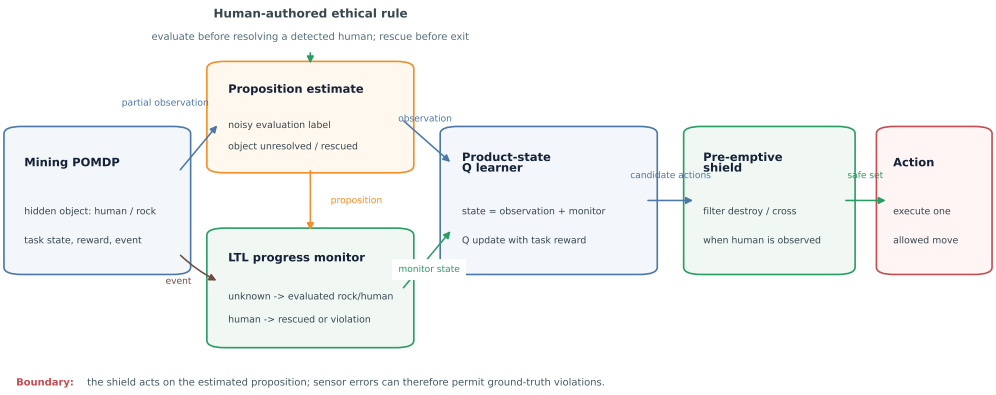


Figure 1: Ethical-LTL Q-learning architecture. The environment emits a partial observation and event propositions. A finite progress monitor updates the product state used by Q-learning, while a pre-emptive shield filters the candidate actions when a human-rescue obligation is pending. The ethical formula is supplied by the designer; only the action-value function is learned.

## 4.2 Product-state Q-learning and shaping

For each candidate action $a$, tabular Q-learning applies

$$Q(z_t, a_t) \leftarrow Q(z_t, a_t) + \alpha \left[\tilde{r}_t + \gamma \max_{a' \in \mathcal{A}_{\text{safe}}(z_{t+1})} Q(z_{t+1}, a') - Q(z_t, a_t)\right], \tag{3}$$

with learning rate $\alpha = 0.18$. The learning reward $\tilde{r}_t$ adds 18 at first evaluation, 35 at successful human rescue, and subtracts 80 on a violation. These shaping terms expose progress but are excluded from reported task return.

We compare this complete method with vanilla Q-learning and penalty Q-learning. Vanilla uses the environment observation and task reward. Penalty Q-learning uses the same state but subtracts 80 from the learning reward on a violation. Thus the penalty baseline tests whether a scalar consequence alone is sufficient, while the complete method changes both the state representation and the action set. Because we do not include component ablations, the experiment does not assign the observed difference independently to shaping, the monitor state, or shielding.

### 4.3 PRE-EMPTIVE SHIELD

When the robot is in the burning room, the report is human, and the object is unresolved, the shield removes DESTROY and CROSS. All remaining actions stay available. This intervention is intentionally local: it prevents two immediate ways of abandoning the pending obligation but cannot correct a mistaken report. If a true human is reported as a rock, the obligation is not activated; if a rock is reported as a human, the shield may be unnecessarily conservative.

## 5 EXPERIMENTS

### 5.1 PROTOCOL

We run 30 independent seeds for each method. Every seed trains for 1,500 episodes with at most 18 steps per episode. Exploration is $\epsilon$-greedy with $\epsilon_0 = 1.0$, multiplicative decay $0.996$, and floor $0.05$. The training sensor-flip probability is $\eta = 0.1$. After training, each greedy policy is evaluated for 300 episodes per seed. Reported intervals are normal-approximation 95% confidence intervals over seed-level means.

The implementations share the same environment, initialization, task rewards, discount, learning rate, exploration schedule, and random-seed construction. The comparison changes only the method-specific state and learning reward described above and, for the complete method, the shield. We retain episode-level and seed-level CSV files, an aggregate JSON file, and the deterministic analysis script.

### 5.2 DOES TEMPORAL STRUCTURE IMPROVE LEARNING EFFICIENCY AND TRAINING SAFETY?

Figure 2 compares task return and violation rate using the same 100-episode window as the joint criterion. The complete LTL-and-shield method crosses both thresholds at a median of episode 174, compared with 249 for penalty Q-learning and 395 for vanilla Q-learning among seeds that reached the criterion. All 30 LTL-and-shield seeds and all 30 penalty seeds reached it; only 19 of 30 vanilla seeds did.

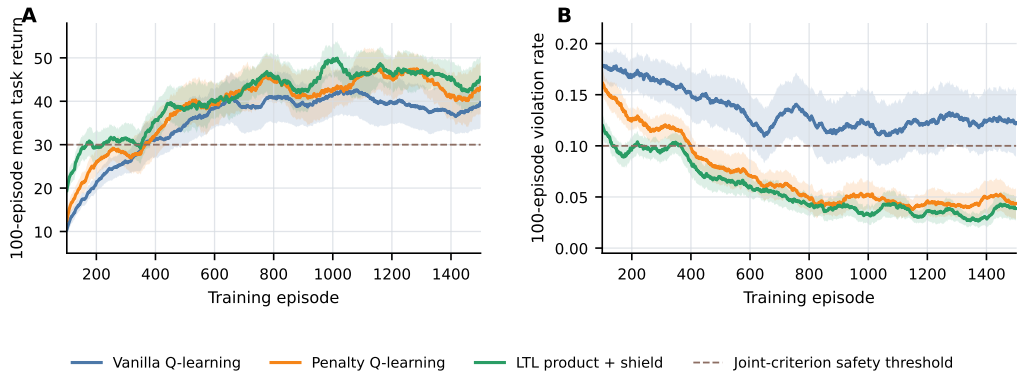


Figure 2: Training behavior across 30 seeds. Curves show seed-averaged 100-episode moving means; shaded bands are normal-approximation 95% confidence intervals over seeds. Dashed lines show the return (30) and violation-rate (0.10) components of the descriptive joint criterion.

The same pattern appears before convergence. Mean return over the first 500 episodes is $30.89 \pm 1.42$ for LTL plus shield, $27.57 \pm 1.32$ for penalty Q-learning, and $24.69 \pm 1.85$ for vanilla Q-learning. Across all 1,500 training episodes, the corresponding cumulative counts of violating episodes are $87.23 \pm 3.65$, $108.27 \pm 3.08$, and $203.57 \pm 30.95$. The complete method therefore achieves the best observed combination of early task learning and reduced unsafe exploration in this benchmark.

### 5.3 Final policy trade-off

Table 1 evaluates the learned greedy policies with method-specific shaping removed from the return. The complete method obtains the highest mean task return, lowest violation rate, and highest conditional rescue rate among the three implementations. Penalty Q-learning improves both safety measures over vanilla Q-learning, showing that consequence-based learning is useful; explicitly representing progress and filtering actions improves the tested trade-off further.

Table 1: Greedy final-policy evaluation over 30 seeds. Values are means $\pm$ 95% confidence intervals over seed-level means; each seed uses 300 evaluation episodes. Higher is better for return and rescue, while lower is better for violations.

| Method | Task return ↑ | Violation rate ↓ | Rescue rate ↑ |
|---|---|---|---|
| Vanilla Q-learning | $33.42 \pm 8.18$ | $0.135 \pm 0.031$ | $0.301 \pm 0.155$ |
| Penalty Q-learning | $35.67 \pm 10.32$ | $0.059 \pm 0.028$ | $0.695 \pm 0.140$ |
| LTL product + shield | $\mathbf{43.77 \pm 7.78}$ | $\mathbf{0.035 \pm 0.018}$ | $\mathbf{0.811 \pm 0.099}$ |

### 5.4 How does observation noise affect the shield?

We next hold each learned Q-table fixed and vary the evaluation sensor-flip probability from 0 to 0.3. Figure 3 shows that the LTL-and-shield policy's conditional rescue rate falls from $0.900 \pm 0.109$ at zero noise to $0.625 \pm 0.080$ at noise 0.3, while its violation rate rises from $0.019 \pm 0.021$ to $0.075 \pm 0.015$. The shield remains beneficial relative to the implemented baselines over this sweep, but its degradation is systematic.

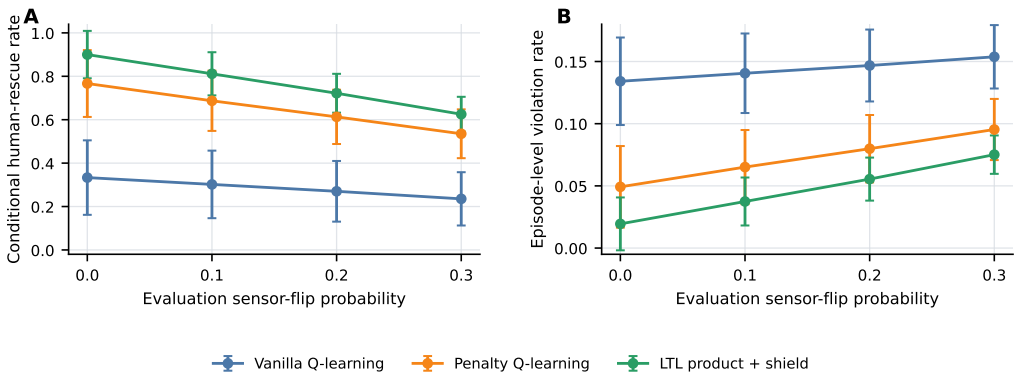


Figure 3: Evaluation-time observation-noise stress test without retraining. Points show means over 30 seed-level policies and error bars show normal-approximation 95% confidence intervals. Increasing sensor-flip probability weakens the correspondence between monitored propositions and hidden ground truth.

This result clarifies the scope of the construction. The shield enforces a rule over perceived propositions; it does not create correct perception. Formal action filtering and proposition estimation are therefore separate assurance problems in partially observable environments.

## 6 Discussion

The experiments support a simple mechanism-level conclusion. When an ethical obligation spans several actions, exposing progress in the learning state and blocking immediately conflicting actions can improve both learning efficiency and training safety. This finding is consistent with reward-machine work showing that finite-state reward structure can improve sample efficiency (Toro Icarte et al., 2022) and with shielding work showing that formal filters can constrain exploration (Alshiekh

et al., 2018). Our contribution is the combined, auditable construction and its measured behavior in the mining dilemma.

The comparison also distinguishes explicit obligations from large penalties. Penalty Q-learning substantially reduces violations relative to vanilla Q-learning, but it must still experience consequences to estimate their action values. The product monitor provides memory of evaluation and rescue progress, and the shield acts before two immediately conflicting actions. Because all three changes appear together, the evidence supports the complete construction rather than a causal claim about any one component.

Three limitations govern interpretation. First, the environment is a small synthetic POMDP with tabular states; it does not establish performance in continuous control, real robots, or socially complex settings. Second, the LTL rule is written by a human and assumed to represent the intended priority. The work does not infer moral values or automatically synthesize a correct specification. Third, the shield uses noisy propositions. The stress test shows that incorrect grounding can permit a ground-truth violation or impose an unnecessary restriction. Future work should separate monitor shaping from shielding in ablations, test specification misspecification, and combine the construction with calibrated belief-state estimation on established safe-RL benchmarks.

## 7 Conclusion

We constructed an ethical LTL progress monitor for a partially observable mining dilemma and combined it with tabular Q-learning and a pre-emptive action shield. Across 30 seeds, the complete method reached a joint return–violation criterion earlier than vanilla and penalty Q-learning, incurred fewer training violations, and produced the strongest final return–rescue–violation trade-off among the tested methods. Performance declined as proposition noise increased, emphasizing that a correct temporal rule cannot compensate for an incorrect observation. The results motivate explicit temporal structure as a practical tool for controlled ethical-decision experiments while leaving general moral reasoning and deployment safety unresolved.

### AI use statement

In this work, we used OpenAI Codex for tasks requiring disclosure: refining the empirical research question, providing feedback on experimental methodology, implementing the evaluated methods, and assisting with interpretation of results. Additionally, we used OpenAI Codex for literature retrieval and summarization, manuscript drafting and language polishing, reference preparation, code creation and revision, and creation and revision of scientific figures. We did not use generative AI to formulate or prove mathematical claims. All AI-assisted material included in the submission was checked against source documents, executable code, saved numerical results, and rendered artifacts. We take responsibility for the final content of this work, including text, claims, code, and figures produced with the aid of generative AI.

### Ethics statement

This study uses a synthetic environment and involves no human participants, personal data, or physical robot deployment. The terms *ethical* and *moral* refer to a designer-specified priority to evaluate and rescue a possible human; they do not imply that the agent possesses moral understanding. The method could encode an incorrect or contested rule, and a shield based on noisy observations can still permit harm. Consequently, the reported results should not be used as evidence that the system is suitable for safety-critical deployment.

### Reproducibility statement

The environment, monitor update, action filter, training loop, seed construction, aggregation procedure, and plotting code are provided with the submission source. Section 5 reports the full training and evaluation protocol, and Appendix A lists the principal hyperparameters. Seed-level and episode-level CSV files and the aggregate JSON file are included in the local reproducibility package.

## REFERENCES

David Abel, James MacGlashan, and Michael L. Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop on AI, Ethics, and Society*, 2016.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Mohammed Alshiekh, Roderick Bloem, Ruediger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11797.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999. doi: 10.1201/9781315140223.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. ActSafe: Active exploration with safety constraints for reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aKRADWBJ1I.

Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

Edvard P. Bjørgen, Simen Madsen, Therese S. Bjørknes, Fredrik V. Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg, Louise A. Dennis, and Marija Slavkovik. Cake, death, and trolleys: Dilemmas as benchmarks of ethical decision-making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 23–29, 2018. doi: 10.1145/3278721.3278767.

Alper Kamil Bozkurt, Calin Belta, and Ming C. Lin. Accelerated learning with linear temporal logic using differentiable simulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zbdhhlIy8o.

Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. Safe reinforcement learning via shielding under partial observability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14748–14756, 2023. doi: 10.1609/aaai.v37i12.26723.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015. URL https://jmlr.org/papers/v16/garcia15a.html.

Mohammadhosein Hasanbeig, Daniel Kroening, and Alessandro Abate. Deep reinforcement learning with temporal logics. In *Formal Modeling and Analysis of Timed Systems*, pp. 1–22. Springer, 2020. doi: 10.1007/978-3-030-57628-8_1.

Yannick Hogewind, Thiago D. Simão, Tal Kachman, and Nils Jansen. Safe reinforcement learning from pixels using a stochastic latent representation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01801.

Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science*, pp. 46–57, 1977. doi: 10.1109/SFCS.1977.32.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019. URL https://arxiv.org/abs/1910.01708.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022. doi: 10.1613/jair.1.12440.

# A Implementation Details

Table 2: Principal experimental parameters.

| Parameter | Value |
|---|---|
| Training seeds | 30 |
| Training episodes per seed | 1,500 |
| Greedy evaluation episodes per seed | 300 |
| Maximum steps per episode | 18 |
| Learning rate $\alpha$ | 0.18 |
| Discount $\gamma$ | 0.97 |
| Initial/final exploration $\epsilon$ | 1.0 / 0.05 |
| Multiplicative exploration decay | 0.996 |
| Human probability | 0.20 |
| Training sensor-flip probability | 0.10 |
| Penalty baseline weight | 80 |
| Evaluation/rescue shaping bonuses | 18 / 35 |
| LTL violation penalty | 80 |