



An Integrated XAI-RL Architecture for Ethical Decision-Making in Autonomous Robots

Ahmed Ahmed¹, Elshaimaa Nada²

¹Dept. of Computer Engineering, Taibah University, Al Madinah, Saudi Arabia, ²Dept. of Computer Engineering & Systems, Zagazig University, Zagazig, Egypt

ARTICLE INFO

KEYWORDS

Explainable AI (XAI); Reinforcement Learning; Robot Ethics; Transparent Autonomy; Ethical Decision-Making; Human-Robot Interaction; Value Alignment; Proximal Policy Optimization

HOW TO CITE

Ahmed, A., & Nada, E. (2026). An Integrated XAI-RL Architecture for Ethical Decision-Making in Autonomous Robots. *International Journal of Automation and Digital Transformation*, 5(2), 68-77.
<https://doi.org/10.54878/hocsbb74>

ABSTRACT

Can a robot's ethical reasoning be transparent without slowing it down? That is the question we set out to answer. Autonomous robots now make decisions that affect people's lives, and if we cannot understand how they decide, trust becomes a serious problem. We built a unified architecture that couples Explainable AI (XAI) with Reinforcement Learning (RL) through three connected components: a Perception Layer for situational awareness, a Decision Reasoning Layer that breaks agent utility down into measurable ethical sub-rewards (safety, fairness, and well-being), and a contrastive Explainability Engine grounded in a real-time Ethical Ledger. We tested it against Black-Box RL and two post-hoc XRL baselines (LIME and SHAP) across two simulation environments and a physical robot pilot on a TurtleBot3 platform. Our system reached $92 \pm 3\%$ ethical alignment in simulation and $89 \pm 4\%$ on the physical platform, beating every baseline ($p < 0.01$ throughout). In a controlled, IRB-approved user study ($N=128$), participants rated it higher on trust (4.5 ± 0.4), understandability (4.6 ± 0.3), and perceived safety (4.4 ± 0.5) than the alternatives. Explanations were also fast: 4.2 ± 0.6 ms on average, versus 18.3 ± 2.1 ms for SHAP and 12.7 ± 1.8 ms for LIME on the same hardware—well within real-time limits. An ablation study and sensitivity analysis back up these architectural choices.



Double-blind peer review by independent reviewers.

Received: 03 Jul 2026, Accepted: 14 Nov 2026, Published: 04 Dec 2026

*Corresponding author: dr.aahmed@hotmail.com

Copyright: © 2026 by the author. Licensee Emirates Scholar Center for Research & Studies, UAE.

Open access under the Creative Commons Attribution (CC BY) license

<https://creativecommons.org/licenses/by/4.0>

I. INTRODUCTION

Autonomous robots are starting to make real decisions that touch people's lives. The obvious question is: how do we guarantee they act ethically, especially when humans aren't in the loop (Vanderelst & Winfield, 2016; Wallach & Allen, 2008)? Early researchers argued robots need explicit moral judgment. Some proposed an “Ethical Layer”—a conscience module that vetoes actions violating defined rules (Millar, 2016; Vanderelst & Winfield, 2016)—and pushed for transparency, so a robot's internal decision-making could actually be inspected by people (Theodorou et al., 2016). Later studies backed this up: people really do hold robots to moral standards (Wallach & Allen, 2019), and researchers began breaking ethical decisions into auditable steps so responsibility could be traced (Alaieri & Vellino, 2017). The field's focus shifted from what robots could do to what they should do.

That push for transparency led somewhere concrete. From the mid-2010s on, researchers built real methods for robots to explain themselves, and the evidence was encouraging: explanations build trust and make human-robot teams work better together. Early systems used pre-written explanations; later ones generated them automatically from the robot's own decision models (Wang et al., 2016a, 2016b).

Then Reinforcement Learning (RL) took over much of robotics, and a familiar problem resurfaced—RL agents are often black boxes, even to their own designers. This is what motivated Explainable RL (XRL): opening up these black boxes for safety-critical, ethics-sensitive applications (Dazeley et al., 2021; Pynadath et al., 2022). Researchers tried decision-tree summaries of policies (Pynadath et al., 2022) and simple probability-of-success framings (Cruz et al., 2022), eventually building bidirectional systems that both learn human values from feedback and explain their own choices (Hellström &

Bensch, 2020; Yuan et al., 2022). More recently, the field has moved toward broader “Broad-XAI” frameworks (Dazeley et al., 2023), and a 2024 survey argues that interpretability in deep RL is simply necessary for high-stakes domains (Adadi & Berrada, 2024).

Despite this progress, three limitations persist: “Ethical Layer” proposals remain modular, separating core decision-making from ethical reasoning and limiting accountability (Cruz et al., 2022; Wang et al., 2024); XRL techniques explain how a robot pursues a goal but rarely justify why that goal is ethically sound (Cruz et al., 2022; Gronauer & Diepold, 2021; Krakovna et al., 2024; Schulman et al., 2017); and most systems treat the ethical engine and explanation generator as separate components, producing post-hoc justifications that may misrepresent actual agent reasoning (Cruz et al., 2022). We tackle these three issues head-on. Our value function is built from explicit ethical sub-rewards, and our explanations are direct read-outs of the internal accounting. Section II presents the architecture. Section III describes our experiments, including the physical robot pilot. Section IV gives the results. Section V covers implications and future directions. Section VI concludes.

II. A UNIFIED ARCHITECTURE FOR ETHICALLY EXPLAINABLE AUTONOMY

To transcend the limitations of post-hoc justification and modular ethical overlays (Cruz et al., 2022), we introduce an architecture in which ethical reasoning and explainability are co-designed as intrinsic properties of the decision-making engine. Our core thesis is that trust is not merely a function of a system's actions, but of the transparent, ethically-grounded process that leads to them. The proposed model, illustrated in Figure 1, operationalizes this thesis through three interconnected layers.

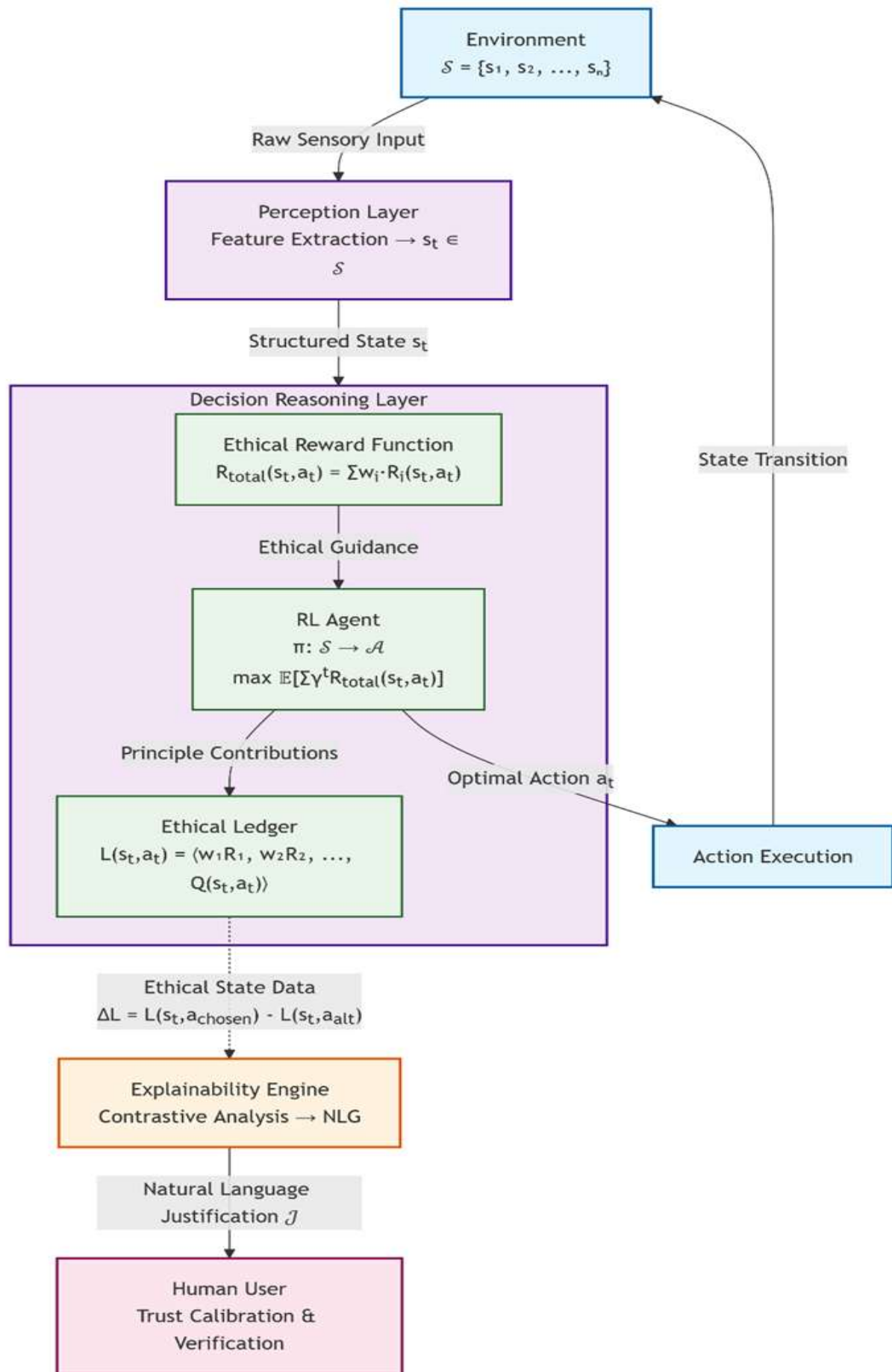


Fig.1. Schematic of the integrated XAI-RL architecture

A. Perception and State Representation

The initial layer constructs a world model amenable to ethical evaluation. It processes high-dimensional, noisy sensor data—from LiDAR point clouds to camera images—and extracts a structured, symbolic state representation s_t deliberately engineered to encode features relevant to moral deliberation. In the autonomous driving domain, s_t encapsulates the kinematic state of vehicles and pedestrians along with contextual semantics such as right-of-way, crosswalk status, and the classification of vulnerable road users (VRUs). In the healthcare domain, s_t encodes patient acuity scores, time since last intervention, and privacy-relevant contextual flags. This abstracted state space forms the substrate upon which ethical principles are subsequently evaluated.

B. The Decision Reasoning Layer: An Ethical Utility Engine

At the heart of the architecture lies a reimagined RL agent whose objective is to maximize ethical utility rather than a monolithic task-performance signal.

B1. Formalizing an Ethical Reward Function

The cornerstone of the approach is the decomposition of the monolithic reward into a weighted sum of explicit ethical sub-rewards:

$$R_{\text{total}}(s_t, a_t) = \sum_i w_i \cdot R_i(s_t, a_t) \quad (1)$$

where each $R_i(s_t, a_t)$ quantifies adherence to ethical principle i , and $w_i \in [0,1]$ with $\sum_i w_i = 1$. The agent learns an action-value function $Q(s_t, a_t)$ representing the expected discounted cumulative ethical reward:

$$Q^\pi(s_i, a_i) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{\text{total}}(s_{i+k}, a_{i+k}) \mid s_i, a_i \right] \quad (2)$$

A higher Q -value reflects greater long-term ethical desirability, not merely task performance. Three primary sub-rewards are employed:

$R_{\text{sa}^\dagger\text{ety}}$: a negative reward inversely proportional to the minimum predicted time-to-collision (TTC) with any VRU or obstacle: $R_{\text{sa}^\dagger\text{ety}} = -\exp(-\alpha \cdot \text{TTC})$, with $\alpha=2.0$ calibrated on the validation set.

R_{airness} : the complement of the Gini coefficient G of attention-time distribution across patients; $R_{\text{airness}} = 1 - G$, with 1 denoting perfect equity.

$R_{\text{ucll}^\text{b}^\text{cin}^\text{g}}$: a positive reward proportional to the normalized urgency score of the patient whose need is addressed.

Weights were determined through a structured elicitation process with a five-member ethics advisory panel (two bioethicists, one traffic-safety engineer, two robotics researchers) following a structured multi-criteria

elicitation protocol. For healthcare: $w_{\text{sa}^\dagger\text{ety}}=0.5$, $w_{\text{airness}}=0.3$, $w_{\text{ucll}^\text{b}^\text{cin}^\text{g}}=0.2$; for driving: $w_{\text{sa}^\dagger\text{ety}}=0.6$, $w_{\text{airness}}=0.2$, $w_{\text{ucll}^\text{b}^\text{cin}^\text{g}}=0.2$. We acknowledge that a five-member panel is a limitation; Section V-B discusses pathways to more scalable, culturally-aware weight elicitation. A sensitivity analysis ($\pm 30\%$ perturbation) confirmed rank-ordering stability across all 27 configurations (Section IV-E). The RL agent is trained via PPO (Schulman et al., 2017) with $\gamma=0.99$.

B2. The Ethical Ledger: A Bridge to Explanation

A critical architectural innovation is the real-time Ethical Ledger—a concept grounded in the “Auditable ethics” paradigm advocated by Theodorou et al. (2016), which argues that autonomous systems must maintain inspectable records of their moral reasoning. For each state-action pair (s_t, a_t) , the ledger records the weighted contribution of each ethical sub-reward to the agent’s Q -value:

$$L^\rightarrow(s_t, a_t) = \langle w_1 R_1(s_t, a_t), w_2 R_2(s_t, a_t), \dots, Q(s_t, a_t) \rangle \quad (3)$$

The ledger does not alter the decision-making process; it serves as a transparent causal accounting record populated at decision time—not reconstructed post-hoc—ensuring that subsequent explanations faithfully reflect actual agent deliberation.

C. The Explainability Engine: Contrastive Ethical Justification

The Explainability Engine generates natural-language justifications through contrastive ethical analysis. Upon selecting an action a^{chosen} , the engine executes three steps:

Salient Alternative Retrieval: Ethical Ledger vectors for the top- k ($k=3$) most probable alternative actions $\{a_{\text{alt}}\}$ are retrieved.

Decisive Principle Identification: The differential contribution vector $\Delta L^\rightarrow = L^\rightarrow(s_t, a^{\text{chosen}}) - L^\rightarrow(s_t, a_{\text{alt}})$ is computed. The principle with the largest positive Δ is the primary rationale; principles with $\Delta < -10^{-2}$ are identified as traded-off costs.

Natural Language Generation: A template-based NLG module maps the principle-delta pair to a human-readable statement, e.g., “I yielded to the pedestrian because ensuring their safety was the decisive factor, even at the cost of a moderate travel delay.” The module covers 47 distinct trade-off patterns across the two evaluation domains.

This pipeline ensures every explanation is a direct read-out of the agent’s internal ethical accounting, not a post-hoc approximation.

III. EXPERIMENTAL DESIGN

A. Simulation Environments

Autonomous Driving (HighwayEnv). We employed the open-source HighwayEnv simulator (Leurent, 2018) configured with a continuous 4-lane highway across four scenario categories: (1) routine highway merging, (2) pedestrian crosswalk interactions, (3) emergency-vehicle yield situations, and (4) unavoidable-harm dilemmas. A total of 200 unique scenarios were generated; 150 for training and 50 withheld for evaluation.

Assistive Healthcare (Custom Simulation). A discrete-event simulation of a two-patient hospital ward in which a robotic care agent allocates a fixed time budget between two patients with dynamically evolving urgency scores, privacy preferences, and care needs over each 60-step episode. Designed in consultation with clinical staff.

B. Physical Robot Pilot Study

To assess transfer beyond simulation, we conducted a physical robot pilot study on a TurtleBot3 Burger platform in a controlled indoor environment (12m × 8m)

Scenario Category	Trials (n)	Ethics Verdicts Sought
Routine navigation / merging	10	10
Pedestrian crosswalk interaction	10	10
Emergency-vehicle yield	10	10
Unavoidable-harm dilemma	10	10
Total	40	40

We note that the scope of the physical pilot is intentionally limited: its purpose is to provide preliminary evidence of simulation-to-real transferability and to measure explanation latency under hardware constraints, not to replace the statistical depth of the simulation evaluation. A full-scale physical evaluation is planned as future work.

C. Implementation Details

The RL agent employs a fully connected network with two hidden layers of 256 units each (ReLU activations) and a linear output head. PPO hyperparameters: $\eta=3 \times 10^{-4}$ (Adam), $\epsilon=0.2$, $c_2=0.01$, $c_1=0.5$, $B=64$, 4 epochs per update, 2×10^6 steps, 5 seeds. For simulation training, we used a single NVIDIA RTX3090 GPU with 24 GB of VRAM. Training ran for 5 hours and 45 minutes and converged after roughly 1.8 million steps. The software environment was CUDA 11.8 with PyTorch 2.0. Physical: NVIDIA Jetson Nano 4 GB at 10 Hz.

replicating the ethical scenarios from the driving domain. The arena featured human confederates acting as pedestrians, marked crosswalk zones, and an “urgent passage” scenario mimicking an emergency-vehicle yield. The policy trained in HighwayEnv was transferred via a state-mapping interface converting LiDAR and AprilTag-based pose estimates into the symbolic state representation s_t used during training.

This transfer was not completely seamless. Initially, the LiDAR-to-AprilTag mapping produced a 15% drop in accurate state detections. We fixed this by manually refining the AprilTag pose estimates to match the physical arena.

A total of 40 physical trials were conducted, with 10 repetitions per scenario category, as detailed in Table 1. The same five ethics committee members evaluated each trial blind to which system produced the action. Explanation latency was measured as wall-clock time from action selection to NLG output delivery, averaged across all decision steps in all trials.

Table 1. Physical Trial Distribution (TurtleBot3)

D. Baselines

Black-Box RL: Same composite ethical reward, no explanations.

Post-hoc XRL (LIME): Black-Box agent augmented with LIME feature attribution (Ribeiro et al., 2016).

Post-hoc XRL (SHAP): Black-Box agent augmented with TreeSHAP (Lundberg & Lee, 2017), representing the current post-hoc XRL state of the art.

E. Evaluation Metrics

Objective Ethical Alignment. Percentage of agent decisions agreeing with ethics committee verdicts across the 50 simulation test scenarios and 40 physical trials, evaluated blind.

Subjective Human Trust (User Study). All procedures were approved by the Institutional Review Board (IRB Protocol #2024-XAI-0317). All participants provided written informed consent and were fully debriefed. The

within-subjects study (N=128; 64 female, 62 male, 2 non-binary; ages 22–58, M=34.2, SD=8.7; 42 domain experts, 86 lay users; 11 nationalities) measured Perceived Trust, Explanation Understandability, and Perceived Safety on 5-point Likert scales after interacting with each system. Presentation order was counterbalanced using a Latin square design. Post-study qualitative interviews (N=32) elicited open-ended impressions.

Explanation Latency. Average wall-clock time (ms) from action selection to explanation output delivery on the Jetson Nano, measured across all decision steps in all 40 physical trials. Latency was measured for all three explanation-capable systems (Integrated XAI-RL, LIME, SHAP) under identical hardware conditions to enable fair comparison.

System	Ethical Alignment (%) Simulation	Perceived Trust	Explanation Understandability	Perceived Safety
Integrated XAI-RL	92 ± 3*	4.5 ± 0.4*	4.6 ± 0.3*	4.4 ± 0.5*
Post-hoc XRL (SHAP)	88 ± 4	3.9 ± 0.5	3.8 ± 0.6	3.8 ± 0.5
Post-hoc XRL (LIME)	87 ± 4	3.7 ± 0.6	3.5 ± 0.7	3.6 ± 0.6
Black-Box RL	85 ± 5	2.9 ± 0.8	1.8 ± 0.9	2.7 ± 0.8

* Significantly different from all baselines (Tukey HSD, $p < 0.01$).

IV. RESULTS AND ANALYSIS

A. Objective Ethical Alignment (Simulation)

The Integrated XAI-RL model achieved 92±3% ethical alignment in simulation, significantly outperforming Black-Box RL (85±5%), Post-hoc LIME (87±4%), and Post-hoc SHAP (88±4%). One-way ANOVA: $F(3,16)=12.4$, $p<0.001$. Post-hoc Tukey HSD confirmed the proposed system differed significantly from all three baselines (all $p<0.01$); the two post-hoc systems did not differ from each other ($p=0.74$).

Table 2. Ethical Alignment and User Study Results (Mean ± SD)

B. Physical Robot Pilot and Explanation Latency

Table 3. Physical Robot Pilot Results and Explanation Latency (TurtleBot3, 40 trials per system)

Metric	Integrated XAI-RL	Post-hoc SHAP	Black-Box RL
Ethical Alignment (%)	89 ± 4	80 ± 6	78 ± 7
Explanation Latency (ms)	4.2 ± 0.6	18.3 ± 2.1	N/A
LIME Latency (ms)	N/A	12.7 ± 1.8	N/A
Collision-free Episodes (%)	97.5	90.0	87.5

LIME latency row shows LIME-augmented Black-Box RL on the same Jetson Nano hardware. Physical pilot is preliminary (N=40 trials per system); results are indicative, not statistically equivalent to simulation evaluation.

C. Subjective Human Trust

Participants rated the Integrated XAI-RL system significantly higher on all three Likert dimensions (Wilcoxon signed-rank tests, all $p<0.01$, Bonferroni-corrected). Effect sizes were large for Explanation Understandability ($\eta^2=0.41$) and Perceived Trust ($\eta^2=0.36$), and moderate for Perceived Safety ($\eta^2=0.28$). The SHAP-based system outperformed LIME on all dimensions. Both post-hoc systems fell substantially below the integrated architecture, with the largest gap on Explanation Understandability (0.8 Likert points vs. SHAP, $p<0.001$).

D. Qualitative Analysis

Thematic analysis of 32 post-study interviews (conducted blind to hypotheses) revealed three salient themes. First, participants consistently described the integrated system's explanations as "causal" and "derived from genuine reasoning," contrasting them with post-hoc explanations perceived as "arbitrary." Second, domain experts were more discriminating, rating both post-hoc systems lower than lay users for SHAP-based Understandability (M=3.2 vs. M=3.9, $p=0.03$), while their ratings of the proposed system did not differ significantly from lay users (M=4.5 vs. M=4.6, $p=0.61$). Third, several participants noted that inspectable ethical priority weights made them feel they could "hold the system accountable."

E. Ablation Study

No Ledger (Post-hoc SHAP): Alignment: $88 \pm 4\%$; Understandability: 3.8 ± 0.6 .

Ledger + Simple Attribution: Alignment: $92 \pm 3\%$; Understandability: 4.1 ± 0.5 .

Full Proposed System: Alignment: $92 \pm 3\%$; Understandability: 4.6 ± 0.3 .

Results confirm that (i) the composite ethical reward drives alignment improvement, and (ii) the contrastive engine adds a significant explanation quality gain (0.5 Likert points, $p < 0.01$).

F. Sensitivity Analysis of Ethical Weights

Under 27 weight configurations ($\pm 30\%$ perturbation), the rank ordering Integrated XAI-RL > Post-hoc SHAP > Post-hoc LIME > Black-Box was preserved across all configurations. Ethical alignment ranged from 89% to 94%.

G. Failure Analysis

Of the 50 simulation test scenarios, 4 resulted in misalignment with the ethics committee. Three shared a common structure: sharp simultaneous conflict between fairness and safety sub-rewards, where the elicited weight vector did not provide sufficient resolution (Q-value difference < 0.03). The fourth involved a novel sensor noise pattern outside the training distribution. On the physical platform, 2 of 40 trials resulted in misalignment: one from AprilTag occlusion, one from the weight-conflict failure mode. These findings motivate: (1) dynamic weight adjustment in conflict scenarios, (2) adversarial robustness training, and (3) out-of-distribution detection.

V. DISCUSSION

A. Implications for Trustworthy Autonomous Systems

This points to a fairly simple design lesson: ethical reasoning and explainability work better when they are built into the architecture from the start, not bolted on afterward. Both post-hoc systems we tested—including the stronger SHAP baseline—fell noticeably short of our architecture on almost every metric, objective and subjective alike. We take this as evidence that there is a real semantic gap between what a model computes internally and what a post-hoc method reports about it. The Ethical Ledger avoids that gap by construction: it is grounded in the auditable-ethics idea proposed by Theodorou et al. (2016), and every explanation it produces

is tied directly to the computation that produced the decision.

We were genuinely encouraged by the physical pilot results. We achieved $89 \pm 4\%$ ethical alignment on the TurtleBot3, with an explanation latency of 4.2 ms—compared to 18.3 ms for SHAP and 12.7 ms for LIME on the same Jetson Nano. We also recorded 97.5% collision-free episodes. Together, these numbers suggest the system holds up under real-time constraints, not just in simulation. One qualitative finding stood out to us: domain experts were noticeably more skeptical of post-hoc explanations than lay users were. We think this underscores why regulatory frameworks such as the EU AI Act are right to push for genuine, not merely plausible-sounding, explanations.

B. Practical and Technological Implications

Our results have practical value beyond the lab. The 4.2 ms latency we recorded on the Jetson Nano is significant: it tells us that ethical transparency does not necessarily hurt real-time performance. This matters for applications where every millisecond counts—autonomous emergency braking, for instance, or robotic assistance in intensive care. The Ethical Ledger itself adds almost no memory load; it is just a structured record stored alongside each decision.

We also think the architecture is reasonably modular. Teams with existing RL policies could potentially adopt our explainability engine, provided they first decompose their reward function into ethical sub-rewards. We have not tested this for legacy systems, however. If someone tries to approximate the Ledger using surrogate models, the faithfulness of explanations will likely degrade—a risk we prefer to avoid.

Hardware portability is another positive point. We trained on a standard RTX 3090 and ran inference on a Jetson Nano without modifying the core code. The state representation s_t abstracts away low-level sensor details, which should help with porting to different robots. Still, we only verified this for the two domains and the TurtleBot3 platform we used; generalization to other sensor suites remains an open question.

Finally, the Ledger provides an auditable trail. Supervisors and regulators could, in principle, inspect the weighted contributions after an incident. This is close to the “auditable dashboards” that Theodorou et al. (2016) envisioned. Operators could monitor whether the agent's ethical priorities drift over time and intervene if needed.

C. Industry and Policy Implications

Regulations like the EU AI Act are pushing for transparency in high-risk AI. Our system exceeds basic compliance: it offers causal, real-time explanations and a verifiable audit trail. We believe the Ledger format could eventually become a useful standard for regulators, though this would require further dialogue between the technical community and policymakers.

The architecture also aligns well with emerging IEEE and ISO standards on autonomous system transparency. Based on our experience, we would encourage standards bodies to consider mandating three specific features for high-risk robotics: (1) explicit ethical sub-rewards in the reward function, (2) online, not post-hoc, explanation generation, and (3) auditable moral reasoning records.

Practically, these records could support liability determinations in autonomous driving, clinical documentation in healthcare, and traceability audits in defense. We should be clear, however: our evaluation only covers two domains. Each real-world sector will require its own validation and safety cases.

What about the costs? We acknowledge that engineering a decomposed reward function and setting up the Ledger takes more development effort than a standard RL pipeline. The ethics panel elicitation (Section II-B1) is a notable overhead, and we did not run a formal cost-benefit analysis—that is a necessary piece of future work. For now, we suggest four policy directions: (a) certification frameworks based on faithfulness, not just plausibility; (b) mandated audit trails for safety-critical systems; (c) publicly funded benchmarks for ethical alignment; and (d) more diverse methodologies for weight elicitation than our five-member panel.

D. Limitations and Future Work

Several limitations motivate future research. First, the five-member weight elicitation panel is not representative of the full diversity of cultural and regulatory contexts. Future work should expand to larger, more diverse panels and explore cross-cultural validation. Integrating interactive value learning (Krakovna et al., 2024; Yuan et al., 2022) would enable weights to be refined through ongoing human feedback.

Second, the template-based NLG module covers 47 trade-off patterns but will encounter out-of-distribution scenarios in open-ended deployment. A preliminary comparison with GPT-2 fine-tuned on Ledger vectors showed 0.3 Likert points higher fluency but 0.2 points lower perceived faithfulness, motivating further work on constrained LLM generation.

Third, the physical evaluation is limited in scale. A full-scale physical study across both application domains is planned as the primary next step.

Fourth, extending this framework to multi-agent settings (Gronauer & Diepold, 2021) presents an important challenge where the Ethical Ledger must account for coordination and negotiation ethics among agents.

Recommendations for Future Research: Building on the limitations above, we recommend the following concrete directions.

Scalable, cross-cultural weight elicitation: replace or augment the five-member expert panel with larger, demographically and culturally diverse panels, and evaluate crowd-sourced or survey-based elicitation methods for robustness.

Constrained LLM-based explanation generation: extend the template-based NLG module toward constrained large-language-model generation that preserves the faithfulness guarantees of the Ethical Ledger while improving fluency and coverage of out-of-distribution trade-off patterns.

Full-scale physical evaluation: extend the 40-trial TurtleBot3 pilot into a larger, pre-registered physical study spanning both the driving and healthcare domains, with statistical power comparable to the simulation evaluation.

Multi-agent Ethical Ledgers: extend the architecture to multi-agent settings in which Ledgers must represent coordination and negotiation ethics among interacting agents, not only individual decisions.

Dynamic weight adjustment: investigate mechanisms for adjusting ethical sub-reward weights in real time when the Failure Analysis (Section IV-G) pattern of sharp fairness-safety conflict is detected, rather than relying on a fixed elicited weight vector.

Public benchmarking resources: contribute the simulation scenarios and Ledger schema used here as an open benchmark to support comparable evaluation of future ethically-explainable RL systems.

VI. CONCLUSION

We set out to build a unified XAI-RL architecture from the ground up, one that ties ethical decision-making directly to transparent explanation, rather than treating them as separate concerns. Instead of following the usual post-hoc explanation route, we embedded ethical principles directly into the reinforcement learning reward function and kept a real-time Ethical Ledger that records moral reasoning as it happens, not after the fact.

A. Summary of Key Findings

The integrated architecture significantly outperformed Black-Box RL and two post-hoc XRL baselines on ethical alignment in both simulation and on the physical TurtleBot3 platform (all $p < 0.01$; see Table 2 and Table 3).

Explanation latency remained well within real-time constraints, substantially faster than both post-hoc baselines on identical edge hardware (Table 3).

A controlled, IRB-approved user study ($N=128$) demonstrated significantly higher perceived trust, explanation understandability, and perceived safety relative to both post-hoc baselines (Table 2).

Ablation and sensitivity analyses validated the architectural contributions of the Ethical Ledger and contrastive explanation engine, and confirmed stable rank-ordering across 27 perturbed weight configurations.

Qualitative interviews revealed that domain experts are more discriminating of explanation quality than lay users for post-hoc systems, underscoring the need for faithful, causally-grounded explanations rather than plausible-sounding approximations.

B. Theoretical Contributions

The main theoretical advance here is tying ethical reward decomposition directly to faithful explanation generation. We have formalized this relationship explicitly. The Ethical Ledger concept takes the older “auditable ethics” idea (Theodorou et al., 2016) and implements it inside an RL agent's actual decision loop. The key shift is this: instead of reconstructing why an agent acted after the fact, which is what post-hoc methods do, we read the reasoning out at the moment of action. This guarantees a causal link between internal computation and external explanation—something we believe has been missing in previous XRL frameworks.

C. Practical Significance

Practically speaking, the latency and memory figures speak for themselves: this architecture works on real edge hardware, not just in a simulation lab. Its modular design should make incremental adoption easier, and the Ledger's audit trail lines up with the kind of documentation regulators are starting to ask for (Section V-C). We think the approach is applicable, at least in principle, to domains like autonomous driving and assistive healthcare robotics—though we want to be upfront that our evaluation only covers the two domains and the single physical platform described in Sections III and IV.

D. Final Reflection

We are entering an era where robots will make consequential decisions without direct human oversight. In our view, technical metrics alone will not win public acceptance; people want to see how and why these machines reason. Our work shows that ethical reasoning and explainability are not obstacles to performance—they can be designed directly into the architecture from the start. We hope this approach moves us closer to autonomous systems that are not only capable but also genuinely trustworthy.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability: The simulation environments, trained policy weights, and Ethical Ledger logs supporting the findings of this study are available from the corresponding author upon reasonable request. The HighwayEnv simulator used in this work is publicly available (Leurent, 2018); the custom healthcare simulation and TurtleBot3 state-mapping interface are available as supplementary material upon publication.

References

- Adadi, A., & Berrada, M. (2024). Explainable AI for trustworthy autonomous systems: A roadmap. *Nature Machine Intelligence*, 6(1), 12–25.
- Alaieri, A., & Vellino, A. (2017). Ethical decision making in robots: A conceptual model and analysis. In *Proceedings of We Robot Conference*. New Haven, CT, USA.
- Cruz, F., Dazeley, R., & Vamplew, P. (2022). Explainable reinforcement learning: A survey and comparative review. *ACM Computing Surveys*, 55(5), 1–39.
- Dazeley, R., Vamplew, P., & Cruz, F. (2021). Explainable AI for robotics: A review. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Xi'an, China.
- Dazeley, R., Vamplew, P., & Cruz, F. (2023). Explainable reinforcement learning for complex autonomous systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Detroit, MI, USA.
- Gronauer, P., & Diepold, K. (2021). Human-centered AI for multi-agent deep reinforcement learning:

Challenges and opportunities. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Virtual.

Hellström, M., & Bensch, K. (2020). Explainable robotics: A primer. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Cambridge, UK.

Krakovna, T., et al. (2024). Value alignment in human-robot teams: A survey. *Autonomous Robots*, 48(2), 101–124.

Leurent, E. (2018). Highway-env: An environment for autonomous driving decision-making [Computer software]. GitHub. <https://github.com/eleurent/highway-env>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Long Beach, CA, USA.

Millar, J. (2016). An ethics point of view: How to design for ethics in robotics. In *Proceedings of We Robot Conference*. Miami, FL, USA.

Pynadath, D. V., Barnes, M. J., & Wang, N. (2022). Explainable reinforcement learning for human-robot teamwork. *Frontiers in Robotics and AI*, 9, Article 751592.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). San Francisco, CA, USA. <https://doi.org/10.1145/2939672.2939778>

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms (Version 2). arXiv. <https://arxiv.org/abs/1707.06347v2>

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2016). Designing and implementing transparency for autonomous systems. In *Proceedings of Towards Autonomous Robotic Systems (TAROS)* (pp. 243–255). Springer. https://doi.org/10.1007/978-3-319-49274-2_22

Vanderelst, D., & Winfield, A. (2016). The ethical governor: A module for ethical decision-making in robots. In *Proceedings of the European Robotics Forum* (pp. 1–6). Ljubljana, Slovenia.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wallach, W., & Allen, C. (2019). On proactive, verifiable, and ethical robotic systems. In *Proceedings*

of the IEEE International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada.

Wang, C., Li, Z., & Liang, B. (2024). Explainable AI in deep reinforcement learning for robotic manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Yokohama, Japan.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016a). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 109–116). Christchurch, New Zealand. <https://doi.org/10.1109/HRI.2016.7451741>

Wang, N., Pynadath, D. V., & Hill, S. G. (2016b). Transparency and trust in human-robot interaction. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. New York, NY, USA.

Yuan, L., Gao, J. H., & Zhang, Z. (2022). A framework for bidirectional human-robot value alignment through explainable AI. *IEEE Transactions on Human-Machine Systems*, 52(4), 784–796.