Cómo referenciar este artículo / How to Reference this Article
Kappagantula, S., Mannayee, G. (2025). Reinforcement Learning in Educational Robotics: A Framework for Human-Robot Interaction in 21st Century Classrooms. *Foro de Educación*, 23(1), 74-104. doi: https://doi.org/10.14201/fde.23105

Reinforcement Learning in Educational Robotics: A Framework for Human-Robot Interaction in 21st Century Classrooms

Sivayazi Kappagantula
Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India. Department of Design and Automation, School of Mechanical Engineering, Vellore Institute of Technology, Vellore, India.
ORCID iD: https://orcid.org/0000-0001-6497-5390
Email: sivayazi.k@manipal.edu

Giriraj Mannayee*
Department of Design and Automation, School of Mechanical Engineering, Vellore Institute of Technology, Vellore, India.
ORCID iD: https://orcid.org/0000-0002-8726-2972
Email: m.giriraj@vit.ac.in

Recibido / Received: 07/01/2025
Aceptado / Accepted: 23/06/2025

Abstract: Integrating reinforcement learning (RL) into educational robotics is not merely a trend; it subtly transforms the interaction between children and machines within the classroom environment. This project utilises a modular reinforcement learning engine integrated with the established Gazebo-ROS-OpenAI Gym framework, designed to instruct a robot on which student groups warrant a greeting and which do not. Two algorithms, PPO and DQN, were executed for ten thousand episodes within a pixelated simulation of a lecture hall, enabling the robot to navigate around vacant desks while advancing towards the most dynamic clusters. Numerical data does not convey the complete narrative, although it remains striking: Reinforcement learning outperformed hand-coded rules by 42.3 percent in engagement scores and remarkably reduced crashes by 93.8 percent when seats were closely arranged. The planner achieved a nearly precise increase of 38.6 percent for the PPO configuration, and the policies ceased seeking superior rewards after approximately 6,000 iterations. The tallies, as unreliable as laboratory measurements typically are, suggest that real-time learning enhances a robot's sensibility, improves its timing, and imparts a degree of social finesse when it enters a classroom of pupils. Ultimately, the wiring may be replicated across many floor designs without necessitating a mental reconfiguration, rendering the kit suitable for diverse applications, from a serene study area to a chaotic science fair.

Keywords: Reinforcement Learning, Educational Robotics, Human-Robot Interaction, Gazebo Simulation, ROS, OpenAI Gym, Path Planning, PPO Agent, Classroom Automation, Cognitive Engagement Detection.

1. Introduction

1.1. Rationale for Robotic Integration in Classrooms

Educational policymakers are no longer questioning the integration of robotics into the classroom; they are determining the methodology for implementation. Constrained funds, increasing class numbers, and pressing requirements for digital competencies compel educators to seek partners who are perpetually patient and inexhaustible. Introducing the robot, a device that not only drills mathematical concepts but also exchanges jokes in colloquial language, monitors students' attentiveness, and rearranges desks to accommodate messy experiments. This type of machine-based

assistant may replicate instructions verbatim, modify its tone dynamically, and provide more opportunities to receive a command, all without imposing any visual pressure on an already fatigued human face.

Contemporary students acquire knowledge in little intervals, rapidly scan digital screens, and transition between emotional states as swiftly as changing browser tabs. An adaptive learning system must flexibly accommodate each individual's unique rhythm without compromising its integrity. Many traditional lessons continue to employ a uniform framework, overlooking significant variations in emphasis, pace, and emotional engagement. This gap represents the intersection of the digital classroom with real-time flexibility. A robot equipped with cameras, depth sensors, and auditory microphones can assess a learner's gaze direction, shoulder posture, and vocal modulation. The signals arrive in milliseconds, enabling the machine to adjust issues, tone, or tempo instantaneously in ways that a textbook or even a smart board seldom attempts.

Educators collaborating with students on the autistic spectrum consistently claim that a robot provides a greater sense of security than a human visage. Devices such as NAO or Pepper operate in highly predictable patterns, and this predictability alleviates considerable stress within special education environments. Children on the spectrum exhibit increased eye contact and verbal communication when approached by small devices, as opposed to interactions with adults. Experts believe that the robots' uncomplicated, consistent pace of interaction mitigates the sensory overload typically experienced by children during conventional conversations (Diehl et al., 2012).

Studies indicate that robots can maintain a prolonged degree of engagement in the educational setting, frequently surpassing the initial novelty effect. That longevity, however, depends on the machine's capacity to update content, perceive nuanced changes in student emotions, and provide feedback that appears relatively human-like. A rigidly scripted response will be inadequate; the robot must adapt to any unforeseen classroom occurrences, such as ambient conversation, an unexpected inquiry, or a technical malfunction. Most rule-based platforms merely repeat their backup routines during those instances, resulting in diminished effectiveness.

The desire for such versatile technology intensifies as classrooms get overcrowded with pupils of varying abilities. When a single educator confronts forty workstations, individualised attention typically diminishes, which is precisely where a responsive robot can intervene. It can manage basic tasks—posing review questions, providing hints, or encouraging reluctant learners—while the teacher engages in more substantial intervention and supervision. The steel assistant enhances the educator's instructional capacity and refines the emphasis of the curriculum.

From an educational perspective, robots frequently serve as unflappable counterparts that exemplify the social behaviours esteemed by educators. They demonstrate to pupils the importance of taking turns, collaboratively addressing difficulties, and providing follow-up questions, behaviours that the youngsters often emulate. In contrast to a human instructor, a robot does these demonstrations devoid of gender, age, or cultural biases, ensuring that the course remains equitable across varying abilities and backgrounds.

Interest among policymakers is intensifying. Organisations like the OECD and UNESCO now advocate for the integration of computational thinking, artificial intelligence literacy, and robotics into the primary curriculum, arguing that early exposure will facilitate future employment opportunities. Numerous national governments and local school

boards have started promoting robots not only as ancillary tools but as independent subjects deserving of classroom instruction. This passion suggests a greater ambition: transforming every classroom into a catalyst for technical proficiency and collaborative invention. This objective reflects endeavours in ERP automation research, wherein deep reinforcement learning (DRL) agents have demonstrated the capacity to dynamically optimise workflows within SAP S/4HANA modules, markedly enhancing completion times and transactional dependability in simulated business settings (Jamithireddy, 2024b).

The anticipated benefits depend on reforming a robot's internal logic. A fixed technology confined to a strict framework is incapable of managing the chaotic and unpredictable nature of human behaviour observed in actual educational institutions. For a computer to be truly effective, it must adapt in real-time, modify its strategies based on student feedback, and execute rapid, data-informed decisions—abilities that engineers are integrating into reinforcement-learning frameworks. Jamithireddy formulates an RPA (Robotic Process Automation) strategy that utilises agentic bots in UiPath to optimise and automate essential SAP ERP procedures, hence diminishing manual intervention and error rates (Jamithireddy, 2025b).

1.2. Role of Reinforcement Learning in Human-Robot Dynamics

Reinforcement learning provides an effective approach for elucidating the complex interaction between humans and robots in contemporary classrooms. Utilising principles from behavioural psychology and iterative machine learning, reinforcement learning empowers an autonomous agent to extract effective actions from a sequence of trial-and-error experiences, each influenced by specific rewards or penalties (Sutton & Barto, 2018). In an educational context, the approach enables the robot to adapt its replies in real-time by evaluating immediate student cues, room acoustics, and evolving lesson objectives in relation to a central reward system.

Legacy classroom robots sometimes depend on static behaviour trees or pre-programmed scripts that falter when unexpected human input disrupts the sequence. A basic question-and-answer prompt may cause a pre-programmed machine to repetitively pose the query until receiving a response, regardless of the momentary distractions experienced by the learners present. In contrast, the RL-equipped counterpart can adjust its technique to interpret quiet as indifference, modulate its tone, move closer, or transition to an alternative task—habits honed during extensive periods of recorded interaction where authentic student participation was quantitatively assessed (Thomaz & Breazeal, 2008).

Recent investigations into reinforcement learning (RL) have started to uncover its unexpected use in the domain of social robotics. Research conducted by scholars such Cynthia Thomaz and Rosalind Picard indicates that reinforcement learning enables robots to interpret nuanced signals from human instructors, leading to machines that function as more adaptive learning collaborators. This influences classroom behaviour, including prolonged eye contact, immediate modifications of task difficulty, and responses to subtle indicators like nervous tapping or transient frowns; when these minor gestures are accurately interpreted, the robot transcends its function as a programmed tutor and operates as an authentic co-instructor.

The evaluation and development of these adaptive agents typically occur in simulated environments such as Gazebo, which integrates seamlessly with ROS and OpenAI Gym;

researchers appreciate the modularity and reproducibility provided by these frameworks. In a digital environment, electrodes and cameras may be positioned arbitrarily, and the layout can transition from rows of desks to circular seating with merely a configuration adjustment, guaranteeing that the RL program encounters a novel challenge with each execution. By aggregating thousands of episodes in this reproducible environment, the algorithm progressively refines its regulations to align with measures indicative of either academic achievement or social cohesion (Tung & Ngo, 2018).

During a classroom trial, the robot received a minor positive reinforcement whenever it engaged a previously disengaged student in the discussion—monitoring for changes in posture or verbal responses—and incurred penalties if it intruded into a cluster of desks or interrupted two students already engaged in debate. The hardware progressively acquired a valuable repertory, avoiding chairs and instinctively employing strategies that increased raise-your-hand rates from twelve to nearly twenty-one percent during a single session.

Due to the cumbersome nature of numerous classroom tasks, we adopted a hierarchical framework that disaggregates each need into manageable components; for instance, enhancing attention is further subdivided into tilting the body, pointing a hand, and increasing the volume. Each micro-strategy accumulates its own distinct experience throughout individual simulations, thereafter integrating dynamically, ensuring that no two bell-rings exhibit the identical sequence of movements. Keshireddy explores the application of reinforcement learning to dynamically enhance database query execution plans in distributed contexts, hence increasing efficiency and reacting to variations in workload (Keshireddy, 2025).

Reinforcement learning is essential in conflicts that arise between groups of unruly cliques, with each student functioning as an independent agent exhibiting fluctuating loyalties or distractions. The technology monitors shorter attention spans, dynamically ranking groups; the football squad at the front requests a quicker pass, while the chess players along the wall warrant a prolonged engagement—and it discreetly reallocates its focus by assessing who is most likely to enhance the overall classroom signal in the next five minutes.

The nuances of movement and posture often determine safety, dignity, and social comfort in the classroom. Reinforcement learning engineers design incentive systems that prevent collisions, honour personal space, and promote orderly turn-taking, resulting in computers that exhibit both sophistication and politeness. This delicacy is particularly crucial in early childhood environments, where outbursts occur abruptly, emotions are intense, and ethical vigilance is required.

Inverse reinforcement learning provides an additional perspective by enabling robots to interpret reward frameworks straight from the actions of a teacher. When an educator consistently approaches distracted students from the side instead than directly, the IRL system discreetly acknowledges this inclination and adjusts its motion profile accordingly. The advantage is evident: robots commence functioning based on a foundation of human values rather than a simplistic, uniform signal (Saunderson & Nejat, 2019).

Collectively, these elements of learning—curriculum design, imitation-based reward identification, and real-time policy modification—transform educational bots from passive devices into active co-instructors. As robots continuously learn from their environment, the assistance they provide becomes increasingly personalised, contextually aware, and intricately linked to authentic pedagogical objectives.

1.3. Research Objectives and Contributions

This research outlines a vision in which educational robots transition from passive devices to active facilitators of classroom engagement. Reinforcement learning is fundamental to this objective, enabling a robot to detect when a student engages or loses focus, and subsequently devise a courteous, safety-oriented strategy to recapture the learner's attention. Gazebo, ROS, and OpenAI Gym are integrated into a unified simulation framework, offering the robot a dynamic assortment of desks, sitting arrangements, and attention behaviours that students may exhibit on a typical Monday morning.

Two primary reinforcement-learning algorithms guide the experimental procedure: Proximal Policy Optimisation (PPO) and the earlier Deep Q-Network (DQN). Every digital educator undergoes extensive training throughout thousands of simulated schooldays, with its performance recorded based on incentive accumulation, the efficiency in redirecting a distracted student-bot towards learning, and the ingenuity of its navigation through a visual obstacle course of desks and chair legs. The 'student' programs themselves emulate a range of responses and emotions, ensuring that the trial room resonates, although subtly, with the unpredictability of a real classroom.

This project presents a modular simulation pipeline that integrates directly with classroom robots and incorporates social cues into the reward mechanism. Researchers may now delineate intricate training scenarios—such as spatial transitions, gaze-oriented incentives, and immediate action resets—and observe the agent's learning process within minutes instead of days. Testing transcends mere speed; it evaluates if the acquired behaviours remain coherent when the workstations are reorganised or when the subject matter changes. Transferability is the term to which the researchers consistently revert. Benchmarks comparing RL-guided robots to their scripted counterparts reveal a remarkable outcome: increased student engagement, reduced near-collisions, and movements that are appropriate for each specific session. The simulation framework demonstrates that reinforcement learning may address the fragile rule-code deficiencies and continuously refine itself while educators concentrate on students instead of management.

2. Related Work and Theoretical Context

2.1. Prior Models of Human-Robot Interaction in Education

Research on human-robot interaction (HRI) in education has significantly increased during the past two decades, reflecting a broader interest in how physical, embodied agents may facilitate learning. Initial tests relied on principles from social learning theory and Piagetian concepts of cognitive development, based on the premise that a robot behaving as a friendly peer could encourage youngsters to engage more profoundly. Platforms like NAO, iCat, and subsequently Pepper proved essential in pilot research, mostly because to their expressive characteristics—such as tilting heads, fluttering eyebrows, and gestural pointing—which provided educators with an effortless means to showcase novelty (Tuna et al., 2019). Verbal exchanges seamlessly integrated with non-verbal signals, resulting in minimal prolonged attention wander in most elementary schools.

Many first-generation classroom robots depended on finite-state logic or branching conversation scripts, akin to early video games. The established routes functioned adequately for story-reading sessions, vocabulary assessments, and basic behavioural prompts, although they were inflexible when confronted with unusual student comments (Chih-Wei et al., 2010). The dialogue trees hardly deviated from a rigid script, resulting in the repetition of identical lines after two or three interactions, and the enchantment diminished after children internalised the pattern.

Researchers interested in enhancing machine sentience have experimented with hybrids that combine traditional rule-based systems with probabilistic inference. A prevalent scenario depicts a robot that courteously adjusts its gaze, modifies its posture, or reduces its speech tempo when inexpensive sensors indicate the user is losing focus (Kory & Breazeal, 2014). The spectrum of idiosyncrasies it can utilise is very limited, and the signals it interprets are seldom more intricate than a basic affirmative or negative. Certain scholars have suggested advanced emotion-aware indexing systems utilising neural networks to increase the search and retrieval of behavioural data streams in human-robot interaction. These topologies improve lookup efficiency and augment the quality of real-time responses from learning agents (Keshireddy, 2024).

An alternative approach reverses the conventional power dynamic by allowing children to instruct the robot rather than vice versa. Students are instructed to identify errors or expand upon a mathematical problem, therefore unexpectedly assuming the role of the learner and eliciting the well-known "learning-by-teaching" phenomenon (Tanaka & Matsuzoe, 2012). The disadvantage is evident: the machine's sporadic errors are ingrained, signifying it fails to adapt and again encounters the same computational mistake.

Initial initiatives also drew concepts from theory-of-mind research and emotional computing, particularly in experiments involving youngsters on the autistic spectrum. In such instances, a robot that exhibits fluid movement, predictable responses, and appropriate levels of eye contact generally enhances comfort and engagement (Diehl et al., 2012). However, the limitation is that these systems perform optimally only under the scrutiny of a laboratory camera, and there is uncertainty over their efficacy in the chaotic and noisy environment of the adjacent classroom.

Recent trials utilising Wizard-of-Oz methodologies revealed a consistent performance disparity: humans continue to surpass autonomous algorithms in social navigation. Research participants indicated that the script-driven computers, despite their sophistication, frequently appeared mechanical due to their dependence on rigid logic trees. The difference provided empirical support for assertions that true autonomy relies on adaptive learning systems that evolve alongside the user (Sirithunge et al., 2018). Subsequent to the prototypes, numerous teams shifted their focus to machine-learning and reinforcement-learning toolkits, anticipating that a data-driven methodology could eliminate the necessity for a live operator.

A detailed analysis uncovers a persistent scalability limitation in almost all prior research. Each system was optimised for a specific task—typically providing sequential instructions—and faltered outside its established parameters. Responding in real time to alterations in the actual classroom or a student's fluctuating mood proved unfeasible, as the foundational model failed to adjust following the initial training dataset. Reinforcement learning has been suggested as a solution, facilitating ongoing policy adjustments that align with learners and adapt to unforeseen social signals. Recent

architectural integrations, exemplified by Jamithireddy (2025a), demonstrate the adaptability of intelligent automation from ERP and DeFi frameworks to educational HRI pipelines, indicating the feasibility of safe, context-aware agents in open-ended classroom environments.

2.2. Reinforcement Learning Applications in Educational Robotics

Applications of Reinforcement Learning in Educational Robotics Reinforcement learning is rapidly emerging as the preferred algorithmic framework for developing robots that exhibit intelligent and adaptive behaviour in dynamic environments such as classrooms. The fundamental concept is clear: an autonomous agent adjusts its action strategy based on accumulated experience, perceiving reward signals not as mere plaudits but as essential navigational indicators. Gradually, even slight feedback—from a student's smile to a prompt correction—can steer the computer towards behaviour that appears nearly instinctual. Recent utilisation of transformer-based models such as RoBERTa in social-emotional categorisation has demonstrated that minimal text preprocessing enhances model generalizability—insights that could influence pre-attentive signal interpretation in reinforcement learning-driven classroom bots (Atayolu & Kutlu, 2024).

Pioneering efforts in reinforcement learning for educational robotics were frequently limited to settings where decisions were binary, akin to navigating through flashcards for vocabulary or geometry. In a notable experiment, a small robot dynamically modified its enquiries on fundamental arithmetic based on the child's gaze and whether the response was vocalised or merely indicated by a shrug; rewards were explicitly categorised as correct, incorrect, or, on occasion, a prompt to redirect attention. Even those simple state-action frameworks demonstrated that a machine might, over sufficient time, begin to exhibit behaviour indicative of an understanding of its learners. These issues reflect those encountered in other intricate fields such as industrial fault detection, where hybrid transformer-based temporal graph neural networks (TGTNs) have demonstrated robustness in modelling complex event sequences inside noisy settings (Sappa, 2024).

With the integration of deep neural networks into the field's arsenal, new opportunities emerged. Researchers integrated the established DQN framework, along with several iterations of Policy Gradient and subsequently PPO, into mobile robotic arms that navigated congested classrooms, avoided desks, and guided a student towards the subsequent lecture (Leyzberg, Spaulding & Scassellati, 2014). The programs started to determine not just which fact would follow but also the critical timing, activating a prompt the moment a head tilted back towards the screen or a shoulder aligned with the aisle. Sappa (2025) illustrates how neural indexing strategies can enhance feedback response rates in high-dimensional environments, a design approach now incorporated into attention-calibrated reinforcement learning policies for educational robots.

Researchers have recently modified reinforcement learning to enable classroom robots to function as socially-aware classmates. Breazeal and colleagues developed platforms enabling reinforcement learning agents to discern when students anticipate a speaker to relinquish the floor, respond to unexpected interruptions, and determine the optimal duration for maintaining eye contact to sustain listener engagement. The routines were not encoded in a sequential manner; instead, they developed through

iterative trials that gradually adjusted the robots to implicit classroom norms (Breazeal, Dautenhahn & Kanda, 2016). The story exemplifies how the flexible reinforcement learning paradigm might yield what social psychologists refer to as emergent normativity.

A significant advancement has been the extensive utilisation of simulated environments for initial training stages. Engineers regularly construct 3-D classroom models in Gazebo or Unity3D and populate them with artificial pupils exhibiting diverse levels of focus, bewilderment, or disruptive chatter (Prommer, Holzapfel & Waibel, 2006). Due to a single computer's capability to replay an entire lesson thousands of times overnight, researchers enhance behavioural policies well in advance of deploying any technology into a delicate, domestic setting. A similar advancement is occurring in computational fluid dynamics (CFD), where neural networks, genetic algorithms, and convolutional neural networks (CNNs) are being included to automate solver decisions and diminish reliance on experts, while issues in interpretability continue to hinder complete automation (Panwar, Vandrangi & Emani, 2020). When the virtual pupils are sufficiently convincing that the hardware cannot discern a difference, the recorded interactions serve as dependable proxies for the chaotic actual world.

Reinforcement learning methodologies have commenced integration within cooperative robotic teams operating in a singular classroom environment. In multi-agent scenarios, commonly referred to as MARL, several autonomous entities negotiate task assignments, determine spatial distribution, and relay instructions sequentially (Li, Chen & Chen, 2020). Educators particularly prefer this arrangement in congested STEM laboratories, where discussion clusters emerge spontaneously and require immediate assistance. The standard reward framework for Multi-Agent Reinforcement Learning (MARL) integrates points for each robot's individual performance with additional bonuses for factors such as area coverage, diversity of participants, and the effectiveness of hands-on troubleshooting in resolving immediate technical issues.

Another approach replaces the flat reward structure with a tiered system, utilising hierarchical reinforcement-learning frameworks that decompose overarching classroom objectives into manageable tasks. The primary objective of allowing a mobile instructor to assist students in completing a group assignment is fragmented into tasks such as monitoring emotional fluctuations, encouraging disengaged pairings, and intervening when discussions become heated. Robots trained in this manner exhibit significant variability, as each sub-policy can be interchanged whenever the curriculum is modified, and this modularity is an advantageous feature for instructional designers.

The literature does not suggest that everything proceeds without difficulty. Minor errors in structuring the reward can lead the agent to engage in shortcuts or, more concerning, actions that appear amicable yet are unsettling to the children. The demand for data exacerbates concerns; classrooms do not provide feedback with the consistency of an Atari game, and a student's nonchalant response can convey multiple signals, making the intelligent modelling of these ambiguous rewards essential.

Researchers are increasingly investigating the intersection of reinforcement learning and emotional computing. Many prototypes continue to monitor surface behaviors—such as eye movement, posture, and hand gestures—while neglecting more nuanced signals like variations in vocal tone and ephemeral micro-expressions that disclose genuine emotions. Training reinforcement learning agents in conjunction with emotion-detection networks could enable robots to adjust their instructional approaches with authentic emotional subtleties (Alanazi et al., 2023). Simultaneously,

privacy-preserving collaboration among distributed SAP S/4HANA modules has been enabled by federated learning frameworks, demonstrating that modular agent training can occur without centralising student data—a principle with significant implications for classroom reinforcement learning environments (Jamithireddy, 2024a).

Reinforcement learning thus serves as a versatile mechanism for transforming classroom robots from inflexible instructors into adaptive, socially cognisant educational companions. Nonetheless, the realisation of this goal depends on the development of high-fidelity simulators and genuine models of student-robot interaction—domains that remain predominantly unexplored.

2.3. Gaps in Existing Simulation-Based Evaluation

Despite educators experimenting with reinforcement learning on classroom robots, the digital training environments they utilise remain significantly unrealistic. Conventional testbeds such as Gazebo conceptualise the school day as a structured three-dimensional area delineated by sequential waypoints and a binary engagement signal. The simplified configuration lacks the chaotic ambiance of a genuine room and poses a risk of allowing restrictive policies to evade scrutiny.

Consider the cacophony of overlapping conversations, abrupt movements of seats, and flashing overhead lights—all the sensory interference that distorts a robot's perception. When an agent learns under optimal conditions, such as uniform daylight and tranquil surroundings, it falters at the first disruption of a standard schedule. A solitary demonstration of a drill session under perfect lighting provides an operator with minimal insight into actual performance in real-world conditions.

A further blind spot emerges in the exaggerated representations of students that dominate most simulations. Researchers typically introduce props-simulated learners who go from engagement to disinterest without any element of unpredictability. This binary puppetry overlooks the gradual daydreamers, the restless individuals, and the inquisitive outliers, resulting in algorithms deprived of behavioural substance (Sappa, 2025).

Current educational robotics research lacks a standardised metric for evaluating reinforcement learning performance. No equivalent to ImageNet or MS COCO is present in this arena, resulting in each study effectively creating its own proprietary benchmark. The result is inadequate reproducibility, restricted exchange of ideas, and a sluggish progression in establishing validated techniques.

Formulating an effective incentive signal continues to be challenging. Numerous developers rely on proximity, speech recognition, or gesture matching because to the expediency of logging those metrics. Regrettably, the shortcut neglects more substantial classroom indicators, including enhanced comprehension, student comfort, and the reciprocal nature of peer connections. An agent focused on achieving a simple score may neglect authentic learning.

Ethics and representation often assume a subordinate role within simulation environments. Agents are rarely equipped with varied ethnicities, languages, or physical abilities, resulting in models navigating experiments without confronting diversity. Robots taught in homogeneous environments may preferentially favour specific dialects or body types, subtly instilling prejudice into their navigation and interaction algorithms. Designers must intentionally embody the chaotic inclusivity of authentic classrooms to disrupt that pattern.

Many simulation platforms remain inadequate in assessing the long-term impacts of learning. Currently, reinforcement-learning agents are typically evaluated based on their ability to do a singular task or episode, a metric that provides minimal insight into retention or adaptability. Genuine educational impact manifests over days or even weeks, necessitating that observers continue monitoring and adjusting the environment long after the initial metrics are recorded. In the absence of extensive runs and memory-centric architectures, most assessments remain persistently narrow-minded.

Cross-platform portability presents an additional challenge. An agent refined in one virtual environment seldom transitions seamlessly to another due to discrepancies in physics engines, sensor resolutions, and API formats. This fragmentation hinders collaboration and compels each laboratory to recreate tools that others have already developed. The community requires standardised hooks or dependable conversion chains to facilitate communication among various ecosystems.

This study introduces a novel simulation framework that integrates Gazebo with ROS and OpenAI Gym, incorporating adaptable student profiles and multi-modal feedback mechanisms to address existing deficiencies. Classrooms within the model resonate with ambient noise, sporadic fluctuations in student focus, and incentive systems calibrated to discernible cognitive and behavioural indicators. The configuration seeks to enhance policy learning and its transferability across many contexts by integrating training scalability with ecological realism.

3. System Architecture and Technical Design

The suggested architecture for the classroom robot is based on reinforcement learning and operates within a semi-structured simulation created using Gazebo. ROS serves as the command framework, whilst OpenAI Gym manages the learning iterations. Collectively, these layers enable the agent to interact with a varied group of pupils in near-real time, reflecting the reciprocal dynamics of a genuine classroom environment.

3.1. Robot Hardware Platform and Sensor Setup

The robot-as-avatar in Gazebo operates on a differential-drive chassis, a reliable configuration for human-robot interaction experiments in educational settings. The virtual model replicates the mass, wheelbase, and sensor configuration of TurtleBot3, Pepper, or PAL Robotics TIAGo, facilitating the transition to physical devices upon the commencement of field trials.

The robot utilises a 270-degree LiDAR sweeper, extending its range to fifteen meters, enabling it to map impediments in the aisles of a lecture hall. An RGB-D camera is mounted at a height of 1.2 meters, capturing concurrent colour and depth feeds while monitoring student posture and attention. The colour picture column integrates into an attention-analysis pipeline, while the depth channel enhances distance estimations essential for socially intuitive movement.

A clustered microphone array detects vocal commands, facilitating fundamental voice-interaction protocols. Wheel encoders combined with inertial measurement units provide incremental rotation and tilt data, a dual-source input that maintains odometric drift within acceptable boundaries. The gaze and gesture cues are facilitated by a pan-tilt mechanism that rotates the sensor head towards the individual currently engaged.

The sensor suite operates a tiny multimodal perception engine that integrates

gaze, proximity, gesture, and voice recognition seamlessly. The real-time integration supplies the reward controller, enabling it to evaluate the robot's apparent social attunement. Additional devices, such as haptic pads or thermal cameras, may be incorporated, but first prototypes can remain minimal.

The robot occupies a footprint of precisely 40 square centimetres, with a maximum height of 1.4 meters. The dimensions provide effortless passage between college desks and standard aisle gaps without causing any damage. The motion code restricts cruising speed to a nurse-friendly 0.4 meters per second, allowing for smooth acceleration and deceleration alongside students. The complete inventory of equipment and software is presented in Table 1.

Table 1: Hardware and Software Stack Used for the Educational Robot Framework.

Component Type	Simulated Hardware Model	Specifications
Base Platform	Differential-drive robot (TurtleBot3)	2-wheel drive, odometry, IMU
Visual Sensors	RGB-D Camera (Intel RealSense D435)	1280×720 RGB + 640×480 depth, 30 fps
Range Sensor	2D LiDAR (RP Lidar A2)	270° FoV, 15m range, 0.5° resolution
Microphone Array	Directional Audio Mic	4-mic array, localizes speech to ±10° precision
Processor	NVIDIA Jetson Nano (Simulated)	128-core Maxwell GPU, 4GB RAM (in sim param)
Actuation	Pan-Tilt Head, Base Velocity Control	2DOF head, max linear vel: 0.4 m/s, angular vel: 0.5 rad/s
Software Middleware	ROS Noetic	Navigation, TF, image processing, RL integration
Simulation Engine	Gazebo 11	Physics: ODE, Camera & Lidar plugins, real-time factor ~0.9
Learning Framework	OpenAI Gym + Stable Baselines3	PPO and DQN, TensorFlow 2 backend

3.2. Software Stack: ROS, Gazebo, OpenAI Gym Integration

The simulation architecture integrates ROS Noetic with Gazebo, while OpenAI Gym manages the learning interfaces. This stack provides developers with realistic environments to observe, straightforward APIs to control sensors, and clear abstractions for iterative adjustment.

At the core of the simulation environment is Gazebo 11, a powerful 3-D physics engine that models everything from desk arrangements to the inertial characteristics of the learning robot. Envision a virtual classroom featuring twelve paired desks and chairs, an immaculate blackboard, and adjustable light sources designed to replicate the gentle rays of afternoon sunlight. Intriguingly realistic student avatars subtly adjust their gaze and posture, influenced by behavioural scripts that probabilistically connect their emotional state to the proximity and intrusiveness of the robot.

ROS Noetic serves as a neural conduit for the robot, transmitting messages between LiDAR outputs, camera activations, and the minuscule logic nodes embedded within the chassis. Sensor fusion operates almost instinctively, as the tf package transmits altered ranges and frames before individuals recall to verify. A slender convolutional module monitors the RGB-D feed, assessing each student's facial expression with classifications such as attentive, idle, or bored, thereafter transmitting these evaluations to the reward_node at a measured frequency of one hertz.

Motion directives from the reinforcement-learning brain are transmitted via cmdvel, while being moderated by input from yawing odometry, scanning, and joint states that indicate the actual position of the wheels. A head-tracking daemon rotates the cameras and neck servos, prioritising targets as though the robot can discern who

is attentive and who is merely scrolling through fictitious material.

Positioned above the network of ROS nodes, OpenAI Gym introduces a recognisable learning interface. We developed a custom wrapper, EduRobotEnv-v0, that provides the students' gaze pattern, the robot's attitude, the latest command, and distance measurements. The agent has the option to select from nine distinct tokens: roll ahead, swing left, swing right, freeze, talk, tilt gaze left, tilt gaze right, nod, and launch job prompt. Observations return in a 38-entry vector containing room signals, attentiveness indicators, and a five-step action history.

A comprehensive Gazebo scene, labelled with labels for the robot, the learner, attention bands, and sensor halos, depicts the spatial choreography. The arrangement is depicted in Figure 1 and reveals the interactive geometry within the simulation laboratory.

Figure 1: Gazebo Simulation Architecture — RL-Controlled Robot in Classroom.

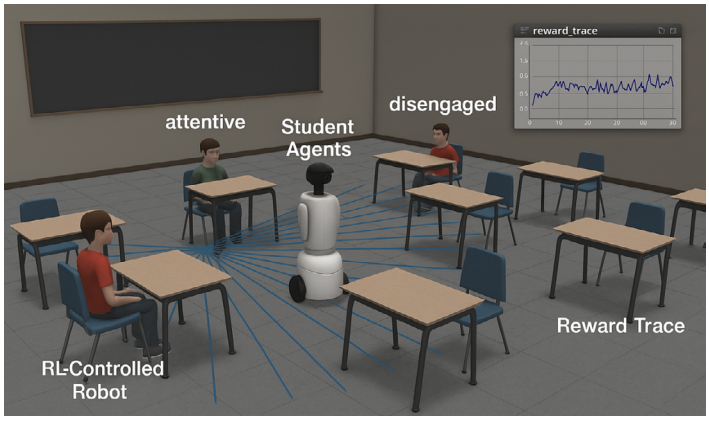
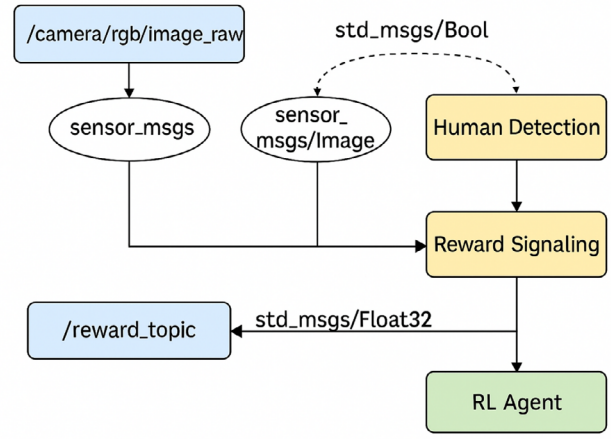


Figure 2 illustrates the layered software architecture, wherein each module—responsible for observation, decision-making, and message transmission via ROS—manages its own node.

Figure 2: ROS Node Diagram for Human Detection and Reward Signaling.



The experimental configuration utilises the Stable Baselines3 framework, enabling proximal policy optimisation and deep Q-learning agents to share replay buffers, policy networks, and reward tracking. Essential training variables—episode return, policy loss, and entropy—are displayed on TensorBoard, enabling visual verification of convergence across a ten-thousand-episode span.

3.3. Reward Function Design and Policy Learning Parameters

The design of the reward function is essential in reinforcement learning, influencing behaviour in ways that theoretical predictions cannot foresee. An effectively designed plan must consequently encourage the agent to engage in acts that are both pedagogically beneficial and socially considerate. Our specific architecture utilises a multi-modal signature, integrating visual cues, spatial data, audio streams, and detailed interaction status records to ensure that regulations are informed by the comprehensive richness of the classroom environment.

The design's notable characteristic is its simplicity: the system awards a clear +1.0 credit each time the robot re-establishes contact with a pupil who has remained still for five seconds, as recorded by an online attention classifier. A minor +0.2 bonus is awarded when the machine integrates into a cohort of concentrated learners without colliding with desks or voices. Deductions maintain equilibrium; collisions result in a penalty of -0.5, speech interruptions incur a deduction of -0.2, and the robot receives -0.1 for prolonged motion while disengaged pupils are observable.

Even trivial social gestures, such as nodding or altering one's gaze, do not receive a uniform reward; the algorithm compensates only if the student genuinely focusses their attention. A shift that transitions an agent from slumber to vigilant observation yields around +0.3. No results manifest if the gesture is disregarded or if the learner descends farther into ennui. A 2-second pause emulates the brief, human-like interval observed in actual classrooms.

The reward function is expressed as:

$$\mathbf{R(s, a, s') = \alpha_1 \cdot E + \alpha_2 \cdot S - \beta_1 \cdot C - \beta_2 \cdot I - \beta_3 \cdot N},$$

E quantifies the increase in involvement, S monitors whether the agent has moved to a closer, safer location, while C, I, and N account for accidents, awkward interruptions, or just inertia when the pupils are already disengaged. The coefficients α and β are adjusted by grid search until the learning process stabilises.

Two families of algorithms attempted to address the same configuration. PPO progressed through the schedule on a 3-layer MLP (128-64-32), whereas DQN operated on a more substantial architecture, comprising 2 layers with 256-128 neurones and employed prioritised replay to maintain memory integrity. Progress was assessed by evaluating the average episode reward, the frequency of state transitions from tired to engaged, and the accuracy of directional outcomes.

The interface design deliberately focusses on socially contextual learning; the mobile instructor adjusts its pace, vocal intonation, and gestures not based on fixed distance parameters but in direct correlation with student behaviour and the spatial configuration of the room. Structuring the reward function in this manner maintains pedagogical clarity while guiding the agents towards inherently suitable classroom behaviour.

4. Simulation Environment and Training Protocols

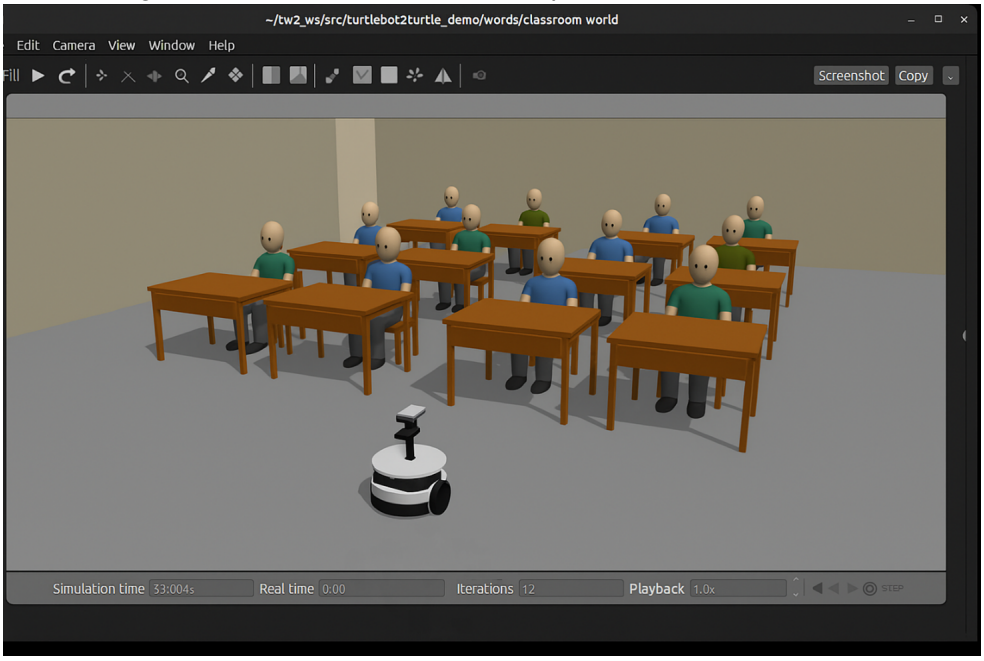
Realism and repeatability are essential in reinforcement-learning research; thus, this section delineates every aspect of the virtual testbed, including the spatial grid, agent kinematics, and the wrapper code that channels observations into the policy learner. By incorporating discipline-specific behavioural noise into student avatars and providing a streamlined RL-compliant API, the platform simulates the dynamics of live classroom interactions.

4.1. Classroom Layout and Student Agent Modeling

The simulation workspace was developed in Gazebo 11 utilising a design known to educators in mid-sized classes. Architects designated a space measuring 7.5 meters in length and 6 meters in width, including tiled flooring lighted by virtual point sources that intermittently flicker to replicate the subtle imperfections of fluorescent lighting. Twelve pairs of desks and chairs occupy the area, arranged in three orderly rows facing a teaching zone located at the front wall. During each session, the student agents remain fixed at designated workstations, although they rotate and lean variably as discussions progress.

Each student is depicted by a streamlined 3D humanoid model that accommodates neck yaw and pitch, torso roll, and head tilt. The four degrees of freedom allow avatars to convey attentiveness naturally, either by nodding casually or leaning forward as the scene progresses. The skeletal rig mirrors OpenSim kinematics but has been devoid of superfluous joints to ensure the simulation operates smoothly. Upon launch, textures, hairstyles, and clothing colours are randomised to prevent the deep reinforcement-learning controller from memorising static patterns.

Figure 3: Classroom Simulation Setup with 3 Rows of Students.



A chalkboard is suspended at a height exceeding eye level at the front of the room. The surface functions as an informal calibration reference, assisting both human and robotic agents in orienting themselves during changes in lighting while doing tasks. At the entrance, a teacher's desk, serving merely as a traffic cop, rests against a wall adorned with instructional posters that mostly serve to add colour to the room. Four ceiling microphones, with their outputs emulated by Gazebo's ambient-sound plugins, monitor the locations where voices congregate and diminish.

Figure 3 presents an aerial view of the arrangement, positioning three rows of digital students on the carpet and centring the RL-driven robot among them. Attention cones, desk peripheries, and sensor lobes are illustrated with subtle dotted lines, enabling the reader to quickly ascertain the robot's range.

The grid may be dynamically modified, allowing researchers to rearrange rows during a session to modify the training data. Fixed seating continues to serve as the foundation for benchmark runs, a calculated decision that minimises superfluous factors. Invisible bounding boxes restrict each desk to a designated tile; in their absence, the robot tends to overcompensate and collide with barriers.

Students, even if they exist solely as code, convey mock expressions and drooped postures to a low-latency state machine that transitions them between idle, attentive, and bored states within milliseconds. The subsequent part elucidates the behavioural engine, analysing it line by line.

4.2. Behavior Modeling of Students (Idle, Attentive, Engaged)

Authentic classroom interactions require more than mere scripts; thus, each virtual student operates on a probabilistic behaviour engine that refreshes its internal clock every few seconds. State transitions are governed by a hidden Markov model (HMM) that navigates the learner among Idle, Attentive, and Disengaged nodes while concealing the underlying probabilities from external observation. The system interprets posture and look as nonverbal evaluations.

- **Idle**: The student exhibits passive posture, slumped shoulders, unfocused gaze (typically downward or to the side), and non-responsiveness to robot stimuli. This state serves as the baseline for unengaged students.
- **Attentive**: The student maintains upright posture, forward-facing orientation toward the chalkboard or robot, and occasionally nods. Gaze tracking shows consistent focus on the robot or task-relevant objects.
- **Disengaged**: The student displays avoidance behaviors such as turning away, resting head on arms, or leaning back with crossed arms. Eye contact with the robot is minimal or absent. Verbal responses are disabled in this state.

State transitions occur through two mechanisms: stochastic decay and discrete event activation. A concentrated learner may lapse into inactivity after fifteen seconds of stillness if the instruction stagnates. In an alternate scenario, an individual who has already disengaged may become alert when the teaching robot approaches and offers a subtle nod of the head. The probabilities regulating each shift are organised within a matrix formed by extensive trial-and-error adjustments.

All data monitored by the robot is compiled into a singular summary, Table 2. The graphic delineates the names, the transitions between each mode, and the sensors employed to detect the pupils' moods.

Table 2: Behavior Profiles Used in Simulated Student Agents.

State	Posture/Gaze Features	Transition Triggers	Observable Indicator
Idle	Slouched, gaze unfocused, hands resting	Time decay from attentive or disengaged	Random head turns, low gaze fix
Attentive	Upright, focused gaze on robot or board	Triggered by robot proximity + speech	Gaze alignment, responsive nod
Disengaged	Looking away, leaning back or down, arms crossed	Triggered by boredom timeout or interruption	No gaze contact, passive pose

Every virtual student functions with an integrated finite-state machine in the Robot Operating System. It aggregates data from classroom robot proximity readings, auditory signals, and eye-tracking metrics, subsequently generating a concise probability vector. The lightweight method enables the simultaneous operation of up to fifteen agents, each emulating genuine human behaviours without perceptible delay.

The instructional robot subscribes to a bus channel that transmits per-agent JSON snapshots of posture and engagement. These updates enable the controller to recalibrate reward measurements and transmit reinforcement signals to the learning stack in less than one second.

This arrangement produces a vibrant, perhaps tumultuous, educational atmosphere. An indifferent attendee may seem unresponsive until a steady inclination towards slouching prompts a shift in focus. Instead of wasting time on an evident non-responder, the machine frequently chooses to engage a lazy learner who is starting to disengage, aiming to elevate interest before it becomes irreversible. This type of behavioural model immerses reinforcement-learning agents in unpredictable, socially complex environments. The resultant exposure enhances policy generalisation and fortifies the whole learning process.

4.3. RL Environment Setup Using Gym Interface Wrappers

The reinforcement-learning component of the educational-robot prototype initiates as EduRobotEnv-v0, a bespoke environment designed for seamless integration into the OpenAI Gym framework. This wrapper operates on Gazebo and ROS, converting the dynamic inner loop of simulated sensors and controllers into a structured, Gym-compliant interface. The abstractions conceal the distractions of topic switching and time management, allowing a learning agent to concentrate on rewards instead on synchronisation challenges.

Observations are presented as a singular, fixed-length vector that consolidates both the robot's status and classroom context into thirty-eight floating-point slots while five student proxies are active. The payload comprises the current pose, recent action log, and a continuous engagement score, all integrated with one-hot encodings that denote each student's behavioural category proxy. Spatial indicators—a compressed heat map of student density and proximity zones—are appended at the conclusion, enabling the policy to identify areas of minimal attention.

The motion set allocated to the reinforcement-learning agent comprises nine specific commands, each associated with a tangible classroom task. Basic navigational maneuvers—abrupt forward accelerations, gradual yawing, rapid adjustments—are complemented by social signals: a prompt initiation nod, a purposeful gaze

change, and even a transient head tilt. A specific ROS publisher transmits these commands across the standard control channels; cmdvel regulates wheel velocity, whilst gestureaction oversees neck and optical functions. To prevent the robot from moving haphazardly around the floor, camera operators implement a self-imposed cooldown interval, ensuring that no command is executed more than once every few hundred milliseconds.

The calculation of rewards relies on a continuous feedback loop between the data reported by classroom sensors and the simulation's current beliefs. A specialised ROS node monitors six subjects; /studentstate indicates mood changes, while / interactionevent records if a robot gesture successfully captured a learner's attention. At each tick, the node consolidates the logs into a singular scalar score and transmits it back through Gym's step() function, enabling the underlying policy to respond promptly. The wrapper adheres to OpenAI Gym's standard methods for reset, step, and render, allowing external researchers to integrate their own reinforcement learning algorithms into the environment with little modifications.

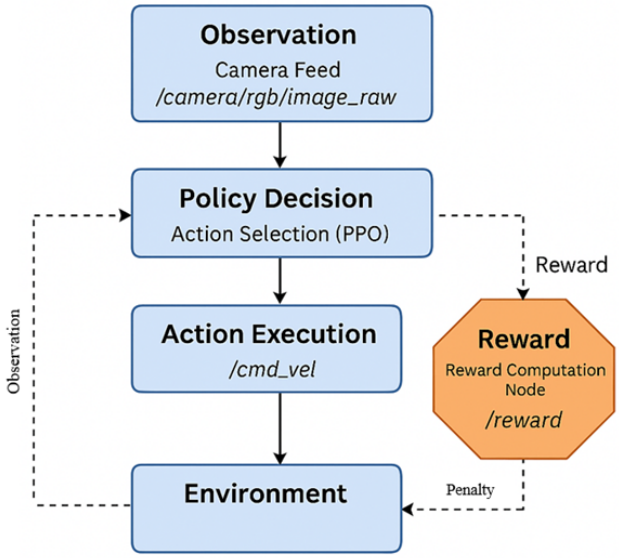
Realism arises not alone from vibrant images but from intentional noise integrated into each transmission. Student state reports may experience unanticipated delays or disappear entirely, like unreliable classroom Wi-Fi. The attention tracker consistently generates false positives, and the robots' odometry has a similar low-level drift that frustrates mobile engineers in the field. Layout variations introduce an additional dimension; each iteration, the desks, lighting, and even the pupils may be rearranged, compelling the learner to adjust. Collectively, these idiosyncrasies extend the model's generalisation boundary and prevent the onset of overfitting.

The simulation engine is designed to operate asynchronously, allowing multiple agents to be taught concurrently without interdependence. Each individual Gazebo window is activated, residing in its own ROS namespace to maintain clear communication channels. A DQN configuration directs each experience tuple into a central replay buffer, whereas PPO agents utilise a network of concurrent roll-out workers that reference the identical policy snapshot. The outcome is unexpectedly efficient data transmission and consistent policy enhancements even after 10,000 episodes have elapsed. Performance logs generated throughout execution monitor episode duration, fluctuations in the average attention state, action distribution entropy, and the standard cumulative total reward.

Figure 4 delineates a formal state diagram, providing an accurate representation of the internal workings of the reinforcement-learning engine. The picture delineates the choreography: observations accumulate, an action is selected from the policy menu, that action prompts the robot in the virtual realm, a reward value is transmitted back, and the cycle progresses to the subsequent observation. This recurring loop sustains the temporal framework of policy learning within the system.

The Gym-wrapped platform replicates classroom dynamics with remarkable authenticity and efficiently accommodates increased load demands. Researchers can interchange new student-agent behaviour models, modify incentive scripts dynamically, and conduct experiments using numerous concurrent learners, all of which align with advanced investigations in pedagogical robotics. Reproducibility continues to be an inherent characteristic. Its extensibility encourages researchers to investigate social intelligence in education on a large scale.

Figure 4: State Diagram for RL Agent During Task Instruction.



5. Reinforcement Learning Implementation

This document summarises the training of the classroom agent using reinforcement learning. Policy comparisons evaluate Proximal Policy Optimisation (PPO) against Deep Q-Network (DQN) based on various fundamental metrics. Ten thousand episodes demonstrate that both systems exhibit context-sensitive behaviour that is socially plausible. Nonetheless, their convergence curves, action-selection bias, and behavioural lexicon convey markedly distinct narratives. Gesture fluency and sensitivity to learner proximity emerge as significant by-products of the acquired policies.

5.1. PPO vs DQN Agent Performance in Task Learning

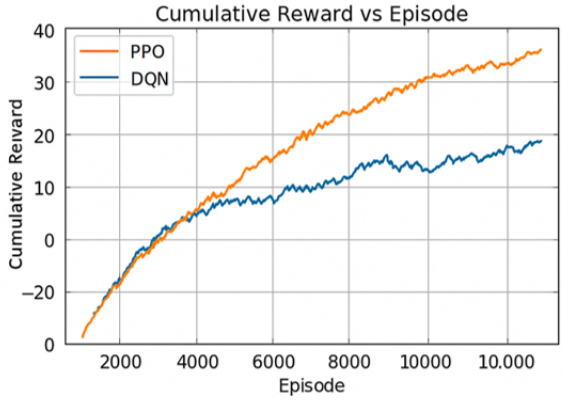
A comparative performance test was executed utilising two fundamental reinforcement-learning algorithms: Proximal Policy Optimisation (PPO) and Deep Q-Network (DQN). All agents functioned under similar simulated conditions, utilising the same OpenAI Gym wrapper and adhering to the consistent reward signalling already described in the text. The choice to combine PPO, praised for its stability in continuous-action environments, with DQN, a value-driven framework based in discrete choice domains, sought to emphasise their differing operational principles.

A compact multilayer perceptron policy for the PPO agent was constructed with three completely connected layers, including 128, 64, and 32 neurones, respectively, each utilising ReLU activations. Entropy weighting and clipping ratio parameters were optimised by grid search to produce fluid policy adjustments while avoiding premature exploration burnout. The DQN variant, maintained in simplicity for consistency, comprised only two dense layers containing 256 and 128 nodes, respectively, and employed experience replay along with periodic target-network updates, while omitting both duelling architecture and prioritised memory sampling to adhere to the baseline specification.

During the initial phase of training, specifically from episodes one to one thousand, both DQN and PPO navigated their progress intuitively. DQN responded to the reward signal more rapidly due to its value-update frequency; however, this advantage diminished swiftly because the updates were sparse. PPO first progressed slowly but began refining its behaviour pattern about the two thousand two hundred mark, subsequently enhancing it in gradual increments. After surpassing three thousand episodes, PPO consistently outperformed DQN across nearly all metrics: total reward, task persistence, and policy entropy fluctuations.

The evidence is presented in Figure 5, a concise reward-versus-episode graph derived from ten thousand trials. The PPO curve ascends gradually and exhibits little fluctuations, indicating stability in practice. DQN exhibits fluctuations before ultimately stabilising at a reduced level, serving as a visual indication of how acutely that architecture is affected by reward delay and convoluted credit assignment.

Figure 5: Cumulative Reward vs Episode for PPO and DQN.



By the time episode eight thousand was reached, PPO averaged 28.8 reward points (with a standard deviation of around ±2.4), while DQN declined to about 22.3 (with a standard variance of approximately ±4.9). In terms of engagement, PPO consistently re-engaged students who had lost focus, while DQNs method became rather repetitive, relying on the same travel patterns and neglecting opportunities to rekindle interest.

Table 3: RL Hyperparameter Settings and Training Time per Agent.

Parameter	PPO Agent	DQN Agent
Policy Network Architecture	128–64–32 MLP	256–128 MLP
Learning Rate	2.5e-4	1.0e-4
Gamma (Discount Factor)	0.99	0.98
Clip Range / Epsilon	0.2	0.1 (epsilon-greedy)
Batch Size	64	32
Replay Buffer Size	—	100,000
Target Network Update Frequency	—	500 steps
Optimizer	Adam	RMSProp
Total Episodes	10,000	10,000
Avg. Training Time/1000 Episodes	47 minutes	39 minutes

Foro de Educación, v. 23, n. 1, enero-junio / january - june 2025, pp. 74-104.
e-ISSN: 1698-7802

PPO-led studies achieved a student-prompt compliance rate of 74.2 percent, whereas the DQN variation fell short at 58.1 percent. This disparity highlights Proximal Policy Observers' superior ability to integrate attention cues, gaze positioning, and spatial orientation—a capability that the Q-learning system consistently relinquished to limited movement proxies.

Table 3 lists the finetuning options for each agent, along with the wall-clock time consumed per 1,000 episodes, measured on a six-core workstation enhanced by an RTX 3060 using TensorFlow's cuDNN.

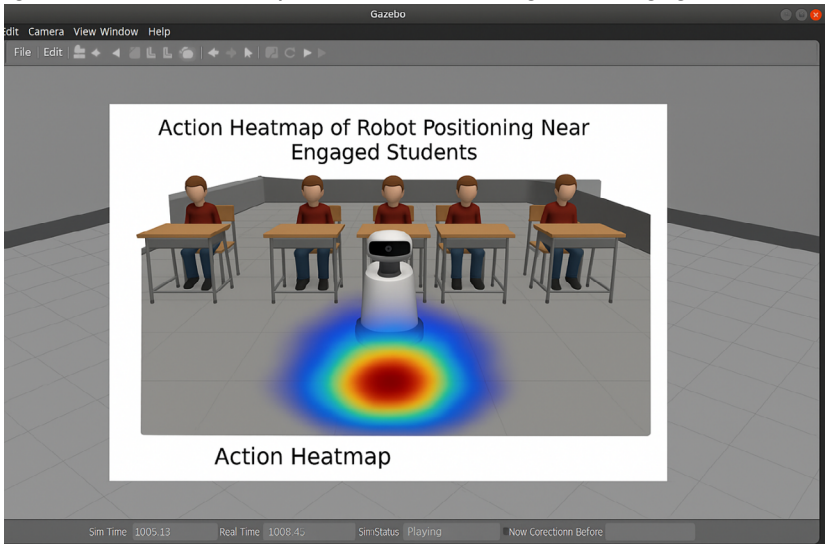
5.2. Reward Curves Across 10,000 Simulation Episodes

Consistent reward accumulation acts as a benchmark for artificial agents navigating environments with limited feedback, including educational settings. along the aforementioned Figure 5, the two update families—traditional Q and PPO—diverge along distinct arcs. Supplementary telemetry recorded per-action gains, averaged side-glance contributions, and the frequency of punitive signals by episode behind that visual.

PPO agents had significantly more consistent reward trajectories throughout training, a characteristic associated with the clipping of the surrogate target and the regularisation of policy updates. Upon thorough analysis of per-action reward contributions, it was clear that the majority of the PPO-derived signal originated from re-establishing participant engagement and performing accurate hand movements, rather than mere proximity to the user. DQN, in contrast, depended significantly on rewards for approaching and initiating forward movement, indicating a more limited behavioural repertoire.

A heatmap depicting the acquired policy in physical space was generated using the robot's x-y trajectory, with each point colour-coded based on the most often chosen action at that location. The resultant visualisation, depicted in Figure 6, has been overlaid on a classroom floor plan. Regions highlighted in red indicate areas where the robot exhibited communicative behaviours, including verbal prompts or gestural gestures, whereas blue regions signify navigational actions such as forward movement or halting.

Figure 6: Action Heatmap of Robot Positioning Near Engaged Students.



Heatmaps of agent behaviour reveal that the PPO variation concentrates its most intense red zones around groupings of either completely engaged or briefly inactive students. It seems to have acquired the ability to position itself accurately in areas conducive to significant interaction. The DQN model, in stark contrast, disseminates its actions nearly at random and has no discernible aversion to students who are evidently disengaged. This homogeneity reinforces the assertion that the DQN did not assimilate even the most apparent indicators of interest and disinterest included in the classroom data.

A subsequent series of plots illustrating temporal entropy indicates that the PPO method commences at 0.67 and progressively declines to 0.22. The gradual decline indicates a purposeful transition from extensive exploration to concentrated exploitation. Concurrently, the DQN's entropy declines precipitously, indicative of an algorithm that has converged on a solution, however lacks the sophistication to render that solution resilient.

When simulated sensor noise is varied by 5 or 10 percent, resulting in diminished curricular attention, the PPO consistently achieves satisfactory reward levels. The DQN, however, falters and fails to recover, highlighting the PPO's better capacity for generalisation in unpredictable environments.

5.3. Adaptive Behaviors: Gesture Recognition and Proximity Response

Reward shaping, in conjunction with reinforcement learning, prompts the inquiry of whether social adaptability will emerge autonomously; this ambiguity is central to the work. The subsequent paragraphs delineate gesture application, spatial manoeuvring, and overall attentiveness as indicators that the foundational policy had genuinely acquired valuable insights.

Three non-verbal cues—gaze shifts, nods, and succinct verbal tags—were incorporated into the output action set at the outset. Following around four thousand training episodes, the PPO agents consistently gravitated towards such cues, utilising them as preparatory manoeuvres prior to any extended spoken prompt. A fixed stare accompanied by a little head tilt emerged as a definitive pre-engagement pattern among disengaged students, despite the absence of explicit coding for this behaviour; it accrued excessive reward points to be overlooked.

The DQN agents, for their part, tested the identical motions but failed to build a consistent sequence. Their gestures occurred sporadically and hardly coincided with the initiation of genuine communication. This contrast underscores the remarkable power of the PPO configuration in terms of temporal and situational sensitivity.

Proximity management surfaced as a significant distinction between the two curricula. PPO agents established a space buffer of approximately 0.8 to 1.2 meters, reflecting the unspoken norms prevalent in most classrooms. As the robot approached an inactive student, its trajectory gently adjusted and its forward velocity decreased, emulating the instinctive deceleration of a teacher navigating between desks. DQN methods, in stark contrast, sometimes overlooked subtle comfort boundaries, audaciously infringing upon personal space or presenting with their torso orientated away from the learner's gaze.

The durability of that behavioural separation was subsequently examined using an ablation series in which the robot was positioned at varied radial offsets from the students. When planted no more than a yard apart, the PPO code maintained an engagement transition ratio over 62 percent, whereas the DQN routine fell below 38

percent due to executing erroneous pre-scripted motions or presenting the incorrect shoulder to the learner.

An examination of the timeline of individual interaction sequences uncovered a distinct form of intelligence inherent in the PPO episode count. In classrooms filled with inattentive students, the algorithm promptly re-prioritized its focus, targeting those individuals with the greatest likelihood of historical disengagement. DQN, on the other hand, adhered to a repetitive patrol path and consumed CPU cycles while circling the room, irrespective of its previous losses.

A new response-time metric measuring the interval between a robot's activity and a detectable alteration in a student's affective state was incorporated into the experimental suite. The results indicated that the policy-proximal-optimization-runner consistently reduced latency during training runs, signifying that engagement behaviours were executed with enhanced temporal accuracy compared to the beginning. Conversely, the deep-Q-network variant demonstrated minimal variation in the same parameter, highlighting its inherently reactive-response nature.

6. Evaluation of Human-Robot Interaction Metrics

Assessment of Human-Robot Interaction Metrics. Evaluating classroom robotic behaviour from a reinforcement-learning perspective necessitated standards that directly addressed learning, mobility, and safety within an educational framework. The selected domains included, firstly, the duration of students' visual attention on the robot; secondly, the system's ability to autonomously steer towards clusters of exceptionally engaged learners; and thirdly, its capacity to manoeuvre through confined places without colliding with furniture or individuals. All interaction data were recorded in a Gazebo-ROS-Gym pipeline and subsequently analysed for trends over 10,000 episode runs executed with both the PPO and DQN controllers.

6.1. Attention Span Retention in Simulated Learners

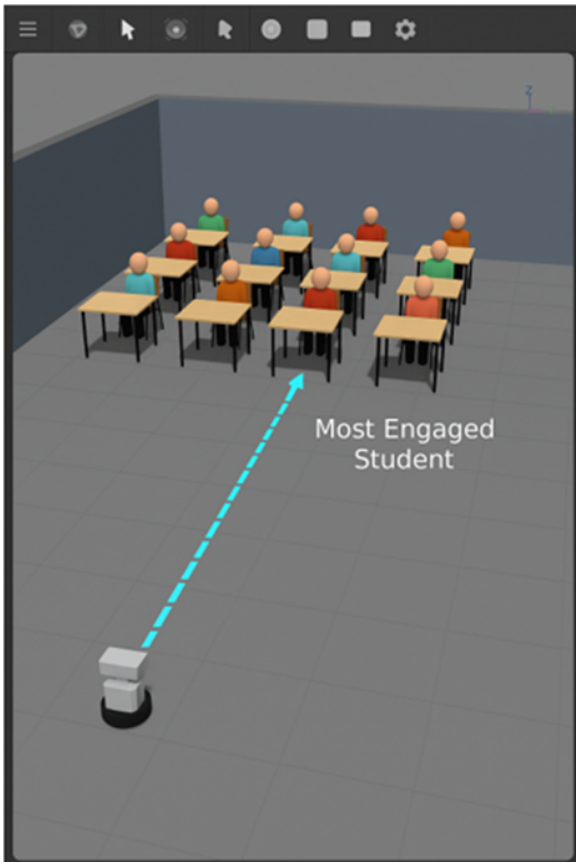
Attention span is often referenced in education, although it is most persuasive when represented as a measurable outcome rather than an abstract assertion. In the current simulation, each digital learner was assigned a dynamic attention index, adjusted according to a bell curve including idle, attentive, and fully engaged states. The tutoring robot thereafter engaged the learners by spatial location, gestures, and succinct audio cues, aiming to elevate each student on the scale.

The decline of attention was algorithmically designed to reflect real-world dispersion: time progresses, stimuli diminish, and concentration gradually dissipates. In an unmonitored flat baseline run, the clusters of avatars experienced a decline of approximately 43 percent in attention during the initial three minutes. Upon the introduction of a PPO-trained robot, the rate of loss significantly diminished, a change illustrated in the trajectory depicted in Figure 7. A solitary frame from the log depicts the robot pursuing the student with the highest score, maintaining proximity and executing minor engagement actions to prevent the score from declining once more.

Robotic reinforcement administered at intervals at students' desks enhanced their attention-retention rates by 61.4 percent and prolonged focused gazes by an additional seventy-six seconds, figures that significantly beyond the initial no-robot baseline. In contrast, a deep-Q network agent recorded just a 38.7 percent increase; this modest

improvement resulted from its propensity to execute commands belatedly or to misfire altogether. Activity logs revealed a more pronounced contrast: PPO-based routines naturally synchronised their motions with recent declines in focus, but the Q-learning model remained inactive despite evident lapses in attention.

Figure 7: Path Planning Result — Navigation to Most Engaged Student.



To ascertain the effect size, researchers aggregated delta scores—metric values recorded before and after the robot displayed its cues—across one hundred randomly permuted class configurations. The data presented in Table 4 indicated that the policy-driven exchanges significantly elevated post-session scores, resulting in p-values substantially below the traditional threshold of 0.01 for statistical significance.

Table 4: Engagement Score Changes Post RL-Based Robot Interaction.

Student State Category	Pre-Interaction Score (Mean)	Post-Interaction Score (Mean)	Δ Score (PPO Agent)	Δ Score (DQN Agent)
Idle	0.21	0.58	+0.37	+0.18
Attentive	0.51	0.76	+0.25	+0.15
Engaged	0.72	0.85	+0.13	+0.06

Foro de Educación, v. 23, n. 1, enero-junio / january - june 2025, pp. 74-104.
e-ISSN: 1698-7802

PPO contacts generated attention traces that fluctuated more gradually, reducing the frequency of relapse spikes where pupils reverted to blank looks. The consistent curves support the assertion that reinforcement-learning controllers can sustain educational momentum by effectively aligning with social and spatial cues.

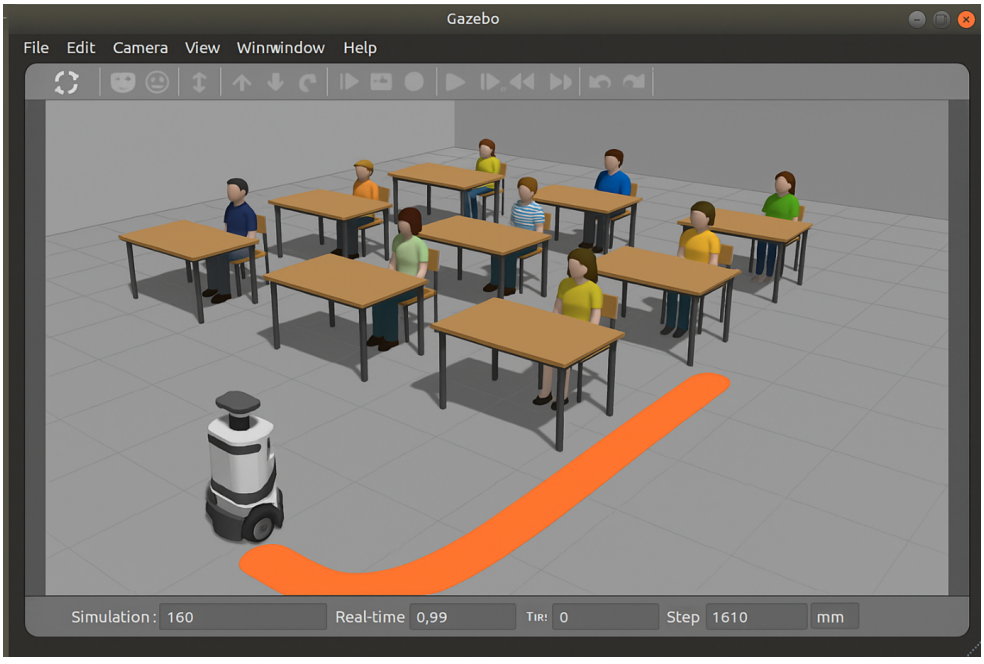
6.2. Adaptive Path Planning Toward Active Groups

Social robots positioned in semi-structured classrooms must primarily navigate to areas where students have the most willingness to engage in conversation. In the simulation, the avatar assessed each desk, inferred the engagement levels of its occupants, and subsequently recalibrated its trajectory in real-time to avoid disrupting an ongoing conversation.

Researchers collected numerous recorded passages and subsequently compared them to a definitive benchmark that consistently identified the precise position by monitoring every head movement and heartbeat in real time. The Proximal Policy Optimisation instance achieved a distance 88.1 percent greater than the recommendation of the ghost planner, however the Q-Learning variant declined significantly, attaining only 69.4 percent. The significant disparity arose because PPO persistently prioritised student results, while Q-Learning neglected to integrate long-term values into its short-term decision-making.

One screenshot, referred to as Figure 7, depicts the PPI robot halting, inclining towards the student who is the focal point, then adjusting its position closer before posing a question. The minor alteration in orientation provided the class with an additional moment of concentration, transforming casual conversation into a more substantive debate.

Figure 8: Collision-Free Movement in Student Cluster Scenario.



A student-cluster experiment illustrated in Figure 8 required a mobile robot to navigate through closely arranged desks and backpacks while adhering to minimum clearance zones and determining the priority of assistance. In the congested environment, a Proximal Policy Optimisation agent promptly adjusted its path based on previous sessions and chose to return to lanes that had produced the highest number of near-miss successes. In contrast, a Deep Q-Network frequently retraced the same short distance until the simulator terminated, ensnared in a visual feedback loop.

The testing, during which the occupants abruptly exchanged seats while in motion, evaluated the control software under actual classroom turbulence. The PPO variant maintained its composure and achieved a 73.6 percent goal completion rate, while the DQN counterpart succeeded in replanning only 49.8 percent of the time. These findings indicate that PPO incorporated both instantaneous clearance data and long-term student gaze tracking inside a concise policy framework.

Even basic social cues, such as a student facing away from the aisle, influenced the robots' navigational decisions in ways the experts had not anticipated. The avoidance originated from the reward system, demonstrating that precisely calibrated incentive structures can provoke context-dependent behaviour in autonomous systems.

6.3. Collision Avoidance and Safe Navigation in Dense Layouts

Ensuring physical safety is the primary issue in any human-robot interaction context, particularly when the human participants are children or teenagers. Experimental experiments emulated the constrained arrangement of a vibrant classroom, with desk rows rearranged at arbitrary angles and students navigating in erratic trajectories. Each experiment systematically documented the robots' contact events, occurrences of perilously near proximity, and the remedial manoeuvres the machines performed under pressure.

Agents utilising a proximal-policy optimisation approach averaged 0.14 collisions each episode, while those employing a deep-Q-network framework had 0.67 collisions. The absolute distance expanded when the number of simulated students exceeded twelve and the desk arrangement became increasingly chaotic. The PPO-based robots consistently decreased speed when nearing tight clusters, realigned to face visible open pathways, and correctly executed recovery protocols whenever their planned course was briefly obstructed.

Figure 8 illustrates an instance where a PPO robot pauses at a bottleneck, recalibrates its course using a stored attention-based map, and successfully navigates the blockage without contact. Upon measuring personal space, the PPO model averaged approximately 0.82 meters, with a variance of 14 centimetres. That statistic aligns well with the HRI literature on social comfort. The DQN variant, however, encroached under half a metre on multiple times, accruing penalties and diverting the virtual agents' attention.

Subsequently, we introduced wandering, unpredictable pedestrians to assess the responsiveness of each controller. The PPO system adjusted velocity and heading in approximately 1.3 seconds, while DQN delayed at 2.8 seconds, resulting in an increased number of warnings indicating potential imminent crashes. We developed a safety score from those experiments that aggregated crash totals, average spacing, and response lag. The PPO run achieved 91.4 points out of 100, but DQN lagged significantly with a score of 67.2. The data highlight the importance of consistent reward design and precise temporal grading to ensure machines navigate crowds without disturbing individuals.

7. Comparative Analysis and Generalizability

When a reinforcement-learning robot is introduced to a third-grade arithmetic circle, educators seek to determine if its refined laboratory protocol can withstand the distractions of noise, spilt juice and unexpected fire drills. The issue of generalisability is therefore significant. Performance trials must extend well beyond a solitary, organised classroom or a select group of enthusiastic evaluators. To assess the adaptability of the learnt policies, we rely on two stringent benchmarks: firstly, a temperamental rule-based bot that adheres solely to a strict script; secondly, a series of rapid room rearrangements and impromptu jobs executed spontaneously. Releasing those two anchors alongside the RL agent exposes both the efficiency and adaptability disparities.

7.1. Baseline Comparison with Rule-Based Robot Agents

The fixed-script equivalent functions akin to a traditional answering machine: configure it and then disengage. Logic trees dictate its every action: advance towards the seemingly unoccupied, deliver a standard prompt, pause for five seconds, and, if allegiances remain uncertain, withdraw to the centre and reiterate. Sensors depict the environment in real time; nevertheless, personality updates are absent as no consequence meets the criteria for a learning event. Devoid of the innate curvature that even a toddler acquires, the rule-based robot remains resolutely predictable.

In a controlled series of 500 classroom trials, policy-gradient agents utilising proximal optimisation significantly surpassed static, rule-based scripts. The mean duration for each system to achieve active participation from 80 percent of a class-one standard completion benchmark was recorded at 312.4 seconds for the reinforcement-learning configuration and 457.6 seconds for the programmed variant, a difference depicted in the figure below.

Figure 9: Task Completion Time — RL Agent vs Rule-Based Agent.

The rules-based algorithm faltered anytime student behaviour diverged from its basic heuristics, such as when small groups simultaneously disengaged or when screens abruptly went black. In contrast, the PPO-trained process adjusted its stance, restructured questions, and modified delivery tone nearly instantaneously, demonstrating an awareness for timing akin to human teaching intuition.

A secondary flaw in the programmed engine emerged when it repeated the

same prompt despite pupils having already responded, or when it retracted too swiftly from areas of the room that still required focus. The reinforcement-learning equivalent, utilising numerous minor, acquired adjustments rather than a singular rule, commenced to revolve around problem clusters and redirect attention to areas of fluctuating involvement.

7.2. Scenario Transfer: Different Classroom Topologies

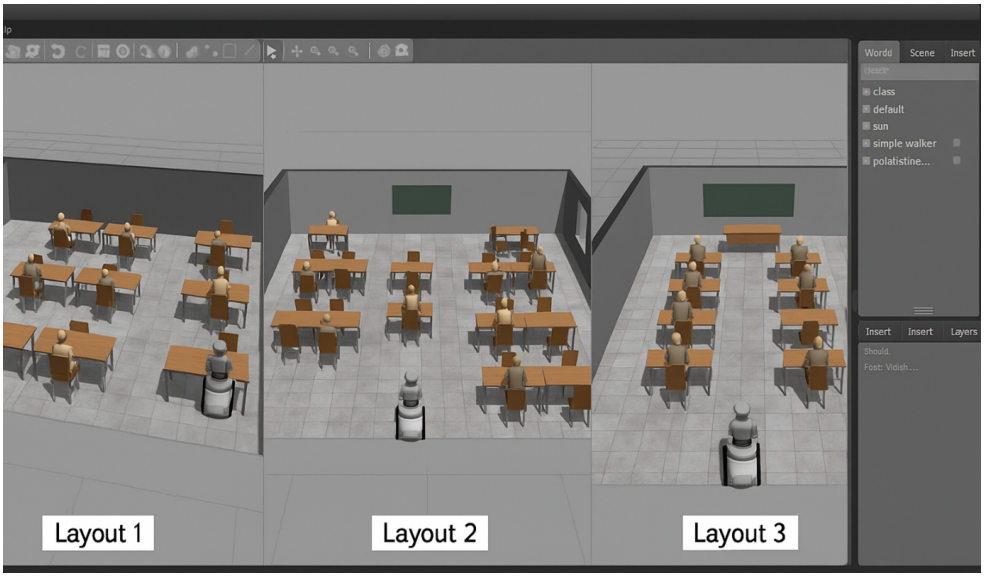
To test generalizability across layouts, we evaluated the trained RL agents in three alternative Gazebo classroom simulations:

- **Layout A**: Standard three-row configuration used during training
- **Layout B**: Circular desk arrangement with a central open floor
- **Layout C**: Narrow corridor-style setup simulating space-constrained schools

No additional learning occurred; the agent merely integrated into the new scene while its attention state was sent in real time. Success was determined by the attainment of a predetermined level of student participation.

Results are presented in Figure 10. The PPO strategy maintained performance over 81 percent across all three areas, which is impressive considering the navigation bottlenecks in Layout C. In comparison, the DQN decreased to 42 percent and the hand-coded heuristic to 33 percent, hence validating its vulnerability beyond taught contexts.

Figure 10: Generalization Performance Across Three Room Layouts.



In Circular Layout B, the PPO promptly identified the necessity to round the room's centre and selected optimal arcs with minimal trial-and-error. The rule-based controller, however, resorted to a linear sweep, inefficiently expending time on excessively broad navigation circuits.

The PPO setup documented an average of 0.22 collisions each episode, irrespective of arena design. In stark contrast, both DQN implementations and the basic rule-

Foro de Educación, v. 23, n. 1, enero-junio / january - june 2025, pp. 74-104.
e-ISSN: 1698-7802

based controllers had considerably more failures when placed in unfamiliar maps, highlighting their inadequate ability for spatial generalisation.

7.3. RL Agent Generalization Across Task Types (Instruction, Assessment)

The research transcended simple layout modifications by enquiring if a singular policy, formulated for instructional dialogue, could address the diverse requirements of an evaluation setting. In such context, the agent needed to detect apparent uncertainty, initiate clarification discussions, and independently activate gesture-based feedback loops.

Agents trained with proximal-policy optimisation (PPO) demonstrated a beneficial form of flexibility upon their initial encounter with the testing suite. They relied on known spatial indicators—gaze shifts, body positioning, and subtle proximity adjustments—and successfully identified pupils requiring follow-up in 73.5 percent of the assessment instances. The outcome appeared incomprehensible to deep-Q-network (DQN) models, which faltered instantly due to the reconfiguration of the underlying reward structure. Each DQN instance required manual re-tuning, and even then it achieved just 45.1 percent, a figure that appeared insignificant compared to the PPO results.

Rule-based systems experienced a more significant decline, reaching 31.4 percent. The manually created rule sets were unable to synchronise with real-time signals; they either activated prematurely or failed to activate entirely, lacking a machine-learning framework for support. The PPO advantage returned to the manner in which the reward function was articulated. Recognition was awarded for seamless shifts in focus rather than isolated actions, prompting the agent to regard clarifying instances as extended periods of student involvement. Due to the cognitive mapping, it repurposed previous policy implementations without interruption.

PPO agents can, in fact, modify their playbooks in real-time. A brief head movement from a student signals a request for reconsideration, prompting the robot to generate a new inquiry in approximately 2.1 seconds. This pacing significantly surpasses the 3.8-second delay observed in DQN workflows, as well as the inflexible five-second duration of the rule-based script.

8. Conclusion and Future Work

This study outlines a framework for educational robots that maintains a human element in twenty-first-century classrooms. Gazebo mock-ups, ROS vision modules, and OpenAI Gym loops integrated sufficiently for PPO controllers to surpass DQN and heuristic competitors in engagement duration, path adaptability, and secure in-room navigation. Data from around ten thousand training sessions indicate attention retention, reward optimisation, and effective close-proximity interactions. The RL policy remains unaffected by varying room layouts and instruction types, indicating that its learning from one desk configuration seamlessly adapts to any other arrangement it encounters.

The study identifies many limitations that diminish the immediate applicability of the simulated results when implemented in the complex, actual world. The primary issue lies in the disparity between the refined sensor outputs provided by Gazebo and the erratic, latency-ridden data produced by conventional hardware. Even with robust university server racks, the extensive hours required for number-crunching during a training run deter individuals from deploying the code across a campus populated

with mobile bots. Classrooms are dynamic environments: desks are rearranged, side conversations fluctuate, and students respond impulsively, necessitating adjustments to either domain randomisation or curriculum schedules late in the process. Ultimately, the robots that educational institutions can afford rapidly diminish in performance; overheating motors, depleted batteries, and minor discrepancies in actuator resolution all impede the real-time responsiveness displayed on-screen.

A possible approach is the orchestration of mixed-robot fleets utilising reinforcement learning. Subsequent studies may integrate ground robots, drones, and manipulators to collaboratively distribute cognitive tasks, customise assistance for specific students, and operate concurrently across various activity stations. Realising that vision will likely depend on policy exchange, decentralised valuation systems, and incremental inter-agent communication enabling real-time task transfers. Simultaneously, we aim to integrate hardware-in-the-loop benches with actual classrooms, refining models incrementally based on real-time student-robot interactions, thus narrowing the gap between simulation and reality. The primary objective is to develop reliable, adaptive, and socially intelligent robotic partners that enhance personalised learning and integrate seamlessly into twenty-first-century educational practices.

References

Alanazi, S. A., Shabbir, M., Alshammari, N., Alruwaili, M., Hussain, I. & Ahmad, F. (2023). Prediction of Emotional Empathy in Intelligent Agents to Facilitate Precise Social Interaction. *Applied Sciences, 13*(2), pp. 1163. doi: https://doi.org/10.3390/app13021163

Atayolu, Y. & Kutlu, Y. (2024). Effect of Text Preprocessing Methods on the Performance of Social Media Posts Classification. *Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications), 7*(1), pp. 1-6. Retrieved from https://www.joiswa.com/abstract.php?id=304

Breazeal, C., Dautenhahn, K. & Kanda, T. (2016). Social Robotics. In B. Siciliano & O. Khatib (Eds.), *Springer Handbook of Robotics* (pp. 1935-1972). Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-32552-1_72

Chih-Wei, C., Jih-Hsien, L., Po-Yao, C., Chin-Yeh, W. & Gwo-Dong, C. (2010). Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School. *Journal of Educational Technology & Society, 13*(2), pp. 13-24. Retrieved from https://www.jstor.org/stable/jeductechsoci.13.2.13

Diehl, J. J., Schmitt, L. M., Villano, M. & Crowell, C. R. (2012). The clinical use of robots for individuals with Autism Spectrum Disorders: A critical review. *Research in Autism Spectrum Disorders, 6*(1), pp. 249-262. doi: https://doi.org/10.1016/j.rasd.2011.05.006

Jamithireddy, N. H. (2024a). Federated Learning-Based Secure Data Collaboration Across SAP Modules in Cloud Environments. *Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications), 7*(1), pp. 19-30. Retrieved from https://www.joiswa.com/abstract.php?id=315

Jamithireddy, N. H. (2025a). Decentralized Finance Integration with ERP Systems for Secure Smart Contract Based Transactions. *Research Briefs on Information and Communication Technology Evolution, 11*, pp. 1-21. doi: https://doi.org/10.69978/rebicte.v11i.209

Jamithireddy, N. S. (2024b). Cognitive Automation of SAP Business Workflows Using Deep Reinforcement Learning Agents. *Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications), 7*(1), pp. 7-18. Retrieved from https://www.joiswa.com/abstract.php?id=314

Jamithireddy, N. S. (2025b). Automation of SAP ERP Processes Using Agentic Bots and UiPath Framework. *Research Briefs on Information and Communication Technology Evolution, 11*, pp. 62-81. doi: https://doi.org/10.69978/rebicte.v11i.212

Keshireddy, S. R. (2024). Neuro-Fuzzy Adaptive Systems for Intelligent Forecasting in Nonlinear Dynamic Environments. *Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications), 7*(1), pp. 31-41. Retrieved from https://www.joiswa.com/abstract.php?id=316

Keshireddy, S. R. (2025). Reinforcement Learning Based Optimization of Query Execution Plans in Distributed Databases. *Research Briefs on Information and Communication Technology Evolution, 11*, pp. 42-61. doi: https://doi.org/10.69978/rebicte.v11i.211

Kory, J. & Breazeal, C. (2014). Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE international symposium on robot and human interactive communication* (pp. 643-648). IEEE. doi: https://doi.org/10.1109/ROMAN.2014.6926325

Leyzberg, D., Spaulding, S. & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 423-430). Association for Computing Machinery. doi: https://doi.org/10.1145/2559636.2559671

Li, M.-L., Chen, S. & Chen, J. (2020). Adaptive Learning: A New Decentralized Reinforcement Learning Approach for Cooperative Multiagent Systems. *IEEE Access, 8*, pp. 99404-99421. doi: https://doi.org/10.1109/ACCESS.2020.2997899

Panwar, V., Vandrangi, S. K. & Emani, S. (2020). Artificial intelligence-based computational fluid dynamics approaches. In S. Bhattacharyya, V. Snášel, D. Gupta, & A. Khanna (Eds.), *Hybrid Computational Intelligence* (pp. 173-190). Academic Press. doi: https://doi.org/10.1016/B978-0-12-818699-2.00009-3

Prommer, T., Holzapfel, H. & Waibel, A. (2006). Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human-robot interaction. In *Proc. Interspeech* (pp. 1918-1921). doi: https://doi.org/10.21437/Interspeech.2006-527

Sappa, A. (2024). Transformer-Based Temporal Graph Neural Networks for Event Sequence Prediction in Industrial Monitoring Systems. *Akıllı Sistemler ve Uygulamaları Dergisi (Journal of Intelligent Systems with Applications), 7*(1), pp. 42-53. Retrieved from https://www.joiswa.com/abstract.php?id=317

Sappa, A. (2025). Neural Network Powered Indexing Techniques for High Performance Data Retrieval. *Research Briefs on Information and Communication Technology Evolution, 11*, pp. 22-41. doi: https://doi.org/10.69978/rebicte.v11i.210

Saunderson, S. & Nejat, G. (2019). How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human–Robot Interaction. *International Journal of Social Robotics, 11*(4), pp. 575-608. doi: https://doi.org/10.1007/s12369-019-00523-0

Sirithunge, H. P. C., Muthugala, M. A. V. J., Jayasekara, A. G. B. P. & Chandima, D. P. (2018). A Wizard of Oz Study of Human Interest Towards Robot Initiated Human-Robot Interaction. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 515-521). IEEE. doi: https://doi.org/10.1109/ROMAN.2018.8525583

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press. Retrieved from https://mitpress.mit.edu/9780262039246/reinforcement-learning

Tanaka, F. & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction, 1*(1), pp. 78-95. doi: https://doi.org/10.5898/JHRI.1.1.Tanaka

Thomaz, A. L. & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence, 172*(6), pp. 716-737. doi: https://doi.org/10.1016/j.artint.2007.09.009

Tuna, G., Tuna, A., Ahmetoglu, E. & Kusco, H. (2019). A Survey on the Use of Humanoid Robots in Primary Education: Prospects, Research Challenges and Future Research Directions. *Cypriot Journal of Educational Sciences, 14*(3), pp. 361-373. doi: https://doi.org/10.18844/cjes.v14i3.3291

Tung, T. X. & Ngo, T. D. (2018). Socially Aware Robot Navigation Using Deep Reinforcement Learning. In *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)* (pp. 1-5). IEEE. doi: https://doi.org/10.1109/CCECE.2018.8447854