Learning to Gaze: Bio-Inspired Attention Adaptation Strategy for Social Robots

Khashayar Ghamati, Maryam Banitalebi Dehkordi, Hamed Rahimi Nohooji,
Holger Voos, Farshid Amirabdollahian and Abolfazl Zaraki

***Abstract*—Adaptive attention allocation in dynamic social environments remains a fundamental challenge for autonomous robots, requiring the integration of perceptual saliency, social context, and real-time decision-making. We present a bio-inspired reinforcement learning framework for robotic gaze control that incorporates a habituation mechanism to regulate the exploration–exploitation trade-off, mirroring how biological attention systems filter redundant stimuli whilst remaining responsive to novel events. Through a comprehensive ablation study comparing Deep Q-Learning (DQL), Vanilla Q-Learning (VQL), and Multi-Objective Q-Learning (MOL), we uncover a critical insight: habituation significantly enhances DQL performance, improving response efficiency and policy stability, yet causes systematic degradation in MOL due to fundamental incompatibilities between fixed-threshold resets and the extended episodes required for multi-objective optimisation. This differential effect reveals that bio-inspired mechanisms cannot be applied universally across learning architectures but must be carefully matched to algorithmic characteristics. Real-world deployment on the ARI humanoid robot validates the framework's practical applicability, achieving robust gaze prediction accuracy across diverse interaction scenarios with well-calibrated confidence metrics that reliably distinguish correct from incorrect predictions. Our findings provide evidence-based guidelines for integrating biological principles into cognitive robotics, demonstrating both the promise and the pitfalls of bio-inspired mechanism design.**

***Index Terms*—Social robotics, attention adaptation, reinforcement learning, bio-inspired systems, gaze behaviour, human-robot interaction, habituation mechanisms**

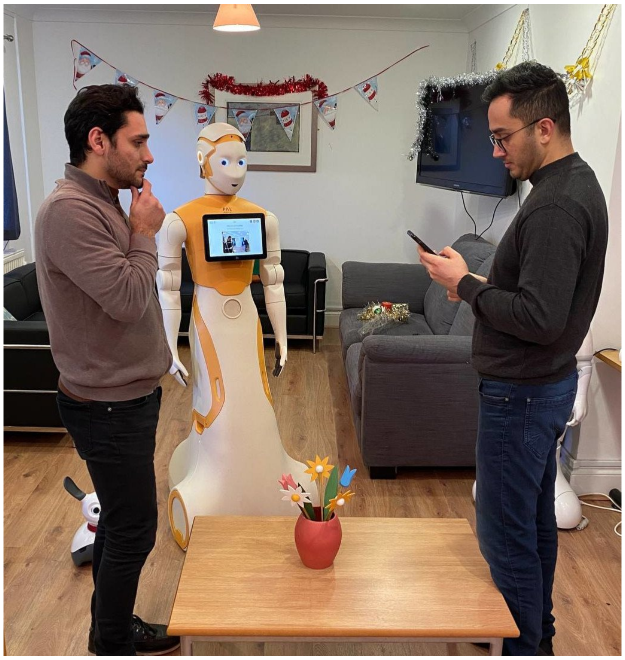

Fig. 1: A triadic Human–Robot Interaction with the ARI humanoid robot. For further details about this study, please visit: https://ghamati.com/brlbam/

I. Introduction

Social robotics is expanding rapidly across assistive care, education, and entertainment [1], [2], demanding robots with bio-inspired, human-like characteristics, including advanced perception, context-sensitive reasoning, and coherent expressiveness [3], [4]. Among these competencies, the ability to direct and regulate attention stands as a fundamental challenge. Real-time attention allocation in multiparty scenarios requires identifying and prioritising salient stimuli in dynamic, ambiguous environments, which is essential for effective human–robot interaction (HRI). Whilst contemporary systems like RASA achieve 76.9% accuracy in dyadic interactions and multi-party systems reach 97% effectiveness [5], [6], these metrics mask critical limitations in scalability, temporal modelling, and adaptability [7], [8]. Current approaches suffer from fundamental constraints that limit real-world deployment. Performance degrades beyond 2-3 participants as computational complexity grows exponentially [9]. Systems demonstrate poor temporal modelling, failing to capture attention patterns over time [10], whilst reliance on deterministic, handcrafted mappings constrains adaptability to novel scenarios [11]. These limitations reveal a conceptual gap. Whilst achieving controlled-setting performance, systems struggle to replicate the fluid, adaptive attention mechanisms that humans effortlessly employ in social contexts.

This gap stems from the underdeveloped translation of biological attention mechanisms to robotics [12]. Whilst foundational mechanisms such as foveal vision and saccadic eye movements have been successfully implemented, higher-order processes, including neuromodulatory influences on attention and adaptive habituation responses, remain largely unexplored in robotic systems [13], [14].

Among missing mechanisms, habituation, fundamental to biological learning and attention regulation, is particularly promising yet severely underutilised [15]. Current implementations employ simplistic exponential decay, failing to capture stimulus specificity, spontaneous recovery, and dishabitua-

Khashayar Ghamati, Maryam Banitalebi Dehkordi, Farshid Amirabdollahian, and Abolfazl Zaraki are with the School of Physics, Engineering and Computer Science (SPECS) and the Robotics Research Group of the University at Hertfordshire, Hatfield, UK. *(Corresponding author: Abolfazl Zaraki.)*

Hamed Rahimi Nohooji and Holger Voos are with the Automation Robotics Research Group, Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg.

tion responses. This represents a significant missed opportunity, as habituation prevents overfocus on redundant stimuli whilst enabling efficient reallocation to novel information. Can we bridge this gap through biologically grounded habituation in learning-based attention systems? Answering this requires confronting the substantial challenges that reinforcement learning faces in social robotics [16]. These include sample efficiency bottlenecks that often demand millions to billions of training steps [17], generalisation failures when transferring learned policies across diverse user groups [18], and the inherent complexity of multi-objective optimisation where competing objectives such as efficiency versus safety or engagement versus privacy must be balanced simultaneously [19].

A critical question emerges: under what conditions do bio-inspired mechanisms enhance learning, and when might they interfere with algorithmic requirements?

This question motivates the present work. We introduce the Reinforcement Learning-Based Attention Model (RLBAM), which incorporates a habituation mechanism [20] into value-based RL [21] for social gaze control. Through a 54-experiment ablation study across DQL, VQL, and MOL, we reveal a striking differential effect: habituation enhances DQL performance whilst causing systematic degradation in MOL, providing evidence-based guidelines for matching bio-inspired mechanisms to algorithmic characteristics.

II. Related Work

Developmental robotics provides a complementary perspective to engineered attention systems by seeking to replicate the ontogenetic trajectory through which biological agents acquire attentional capabilities [22]. Computational models of infant gaze development have demonstrated that simple learning mechanisms, when coupled with appropriate environmental structure, can give rise to increasingly sophisticated attention patterns — progressing from reflexive orientation to deliberate, socially modulated gaze [23]. Sensorimotor contingency approaches frame attention as emerging from the agent's discovery of regularities between its actions and sensory consequences, enabling robots to develop perceptual strategies through self-directed exploration [24]. Intrinsic motivation frameworks extend this by providing curiosity-driven reward signals that guide attention toward informative stimuli, a principle with clear parallels to our habituation mechanism's regulation of exploration toward novel events [25]. However, developmental approaches typically require extensive interaction periods and have primarily been demonstrated in constrained settings with limited scalability to real-time multiparty social scenarios. Our framework draws on the developmental insight that attention mechanisms should be learned rather than pre-programmed, whilst employing reinforcement learning to achieve the sample efficiency required for practical deployment.

Joint attention — the shared focus of two individuals on the same object or event, typically mediated by gaze — has been extensively studied as a foundational mechanism for social cognition and communication [26]. In HRI, computational models of joint attention have been developed to enable robots to follow human gaze direction [27], establish shared reference frames during object manipulation [28], and engage in triadic interactions involving referential gaze [29]. These systems typically model the intentional coordination between agents, requiring explicit representation of the partner's attentional state and goal inference mechanisms. Our work addresses a related but distinct problem: stimulus-driven attention allocation, where the robot must determine which social stimulus to attend to based on perceptual saliency and learned social priorities, rather than establishing mutual attentional engagement with a specific partner. Whilst joint attention models answer, "where is the human looking, and should I follow?", RLBAM answers, "among all available social stimuli, which deserves attention now?" These approaches are complementary — RLBAM's saliency-driven allocation could serve as a front end for joint attention systems, first identifying the most relevant social partner before engaging in shared attentional coordination.

Recent advances in learning-based gaze control have explored various approaches beyond classical rule-based systems. Policy gradient methods have been applied to continuous gaze control, enabling smooth saccadic movements but requiring large amounts of training data and demonstrating limited transferability across social contexts [30]. Actor–critic architectures have shown promise for balancing gaze accuracy with temporal smoothness through dual reward channels, though their application to multiparty scenarios remains limited [31]. More recently, transformer-based attention prediction models have leveraged large-scale gaze datasets to predict human-like fixation patterns with high accuracy, but these supervised approaches lack the adaptive online learning capability essential for personalised human–robot interaction [32]. Deep RL approaches to social gaze have demonstrated success in dyadic settings, yet their extension to multiparty scenarios with dynamic participant entry and exit remains an open challenge [33]. Our framework employs value-based RL (specifically Deep Q-Learning), which offers distinct advantages for the discrete gaze allocation problem: the 8-action decision space aligns naturally with Q-value enumeration, experience replay enhances sample efficiency, and the resulting Q-values provide interpretable confidence metrics for online error detection — a capability absent in policy gradient and transformer-based alternatives.

Against this landscape, RLBAM makes three contributions that distinguish it from prior bio-inspired and adaptive attention systems. First, existing bio-inspired attention mechanisms in robotics employ simplistic exponential decay or fixed saliency maps that fail to capture the richness of biological habituation — particularly stimulus specificity, spontaneous recovery, and dishabituation. RLBAM implements a computationally grounded habituation mechanism (Algorithm 1) that reproduces these three biological properties within the reinforcement learning exploration–exploitation framework. Second, whilst prior work has proposed individual bio-inspired attention mechanisms and evaluated them in isolation, no study has systematically investigated how the same mechanism interacts with fundamentally different learning architectures. Our ablation study across DQL, VQL, and MOL reveals for the first time that habituation's effect is architecture-dependent — a finding with broad implications for bio-inspired mechanism design in cognitive robotics. Third, unlike most learning-based gaze systems that report only accuracy in simulation or constrained laboratory settings, RLBAM provides validated real-world deployment on a humanoid robot platform with quantitative confidence metrics (softmax confidence and Q-margin) that reliably distinguish correct from incorrect predictions (Spearman $\rho = 0.42$, $p < 0.001$), enabling principled online error detection during deployment.

III. Contributions

Building upon the challenges and research questions identified above, our framework makes six key contributions:

1) **Bio-inspired habituation mechanism:** Unlike existing systems that employ simplistic decay functions, our approach implements habituation with stimulus specificity, spontaneous recovery, and dishabituation responses (Algorithm 1), directly addressing the critical gap in bio-inspired attention mechanisms.
2) **Comprehensive ablation study:** We present 54 independent experiments (3 methods $\times$ 2 exploration modes $\times$ 9 runs) with rigorous statistical analysis including paired t-tests, one-way ANOVA, and Cohen's d effect sizes, revealing the differential impact of habituation across learning architectures.
3) **Systematic baseline comparisons:** We compare against a rule-based controller (94.2% success), standard ϵ-greedy exploration, and a random policy baseline (12.5% chance level), contextualising our results across the full performance spectrum.
4) **Empirically-grounded reward structure:** The reward function is derived from human eye-tracking data using the Elicited Attention model [3], ensuring ecological validity in the learned attention policies.
5) **Real-time performance:** The framework achieves $30\,\text{Hz}$ inference on the ARI robot's onboard computer, with training completing in approximately 45 minutes per $10{,}000$-episode run.
6) **Real-world deployment:** We demonstrate 95.1% accuracy (95% CI: $[92.7\%, 96.7\%]$) across 448 trials with 3 experimenters, with per-class F1-scores of 0.63–0.78 for human-directed attention and well-calibrated confidence metrics (Spearman $\rho = 0.42$, $p < 0.001$).

IV. Methodology

This section presents the RLBAM framework in four parts: first, we formalise the attention allocation problem and justify the use of reinforcement learning over rule-based alternatives; second, we describe the empirical foundation derived from human eye-tracking data that grounds our reward structure; third, we detail the system architecture, including the bio-inspired habituation mechanism (Algorithm 1), the DQL learning algorithm (Equation 1), and the state–action–reward definitions (Equations 2–4); and fourth, we specify the hyperparameter configuration and simulation training environment.

To establish a rigorous foundation, we first formalise what we mean by *attention* in the context of RLBAM. In our framework, attention refers to the policy-driven allocation of gaze direction toward salient targets in the environment. Formally, let $T = \{t_1, t_2, \ldots, t_k\}$ denote the set of potential attention targets. The attention function $A : S \to T$ maps the current state $s \in S$ to the most salient target, where saliency is determined by the learned policy $\pi(a|s)$. This definition integrates perceptual saliency (bottom-up visual features), social saliency (top-down contextual factors), and policy-driven selection (learned preferences).

The choice of RL over simpler rule-based controllers is motivated by three requirements that rule-based systems cannot satisfy: temporal dependencies in human attention patterns, uncertainty and partial observability in real-world social environments, and the need for adaptation and generalisation across novel scenarios.

The reward structure is grounded in empirical data on human attention. In our previous work, we conducted an eye-tracking study with 11 participants at the Technical University of Munich, who viewed a 7-minute video of a dyadic conversation recorded with synchronised HD and Kinect RGB-D cameras [3]. Analysis of gaze dynamics during sequential room entry, seated discussion, and individual departures informed the construction of a reward function designed to replicate human-like attention allocation.

The outcome of this empirical work formed the foundation for constructing a reward function within our reinforcement learning framework, specifically designed to replicate human-like attention allocation.

With this empirical grounding, we describe the overall system architecture that integrates these insights into a functional robotic attention system.

The framework comprises two components: the robotic system executing RL-driven gaze policies, and the interactive human environment providing continuous feedback. The robot acts upon the learned policy, monitors environmental changes, and receives feedback from interacting humans, enabling a cyclic learning process that progressively improves attention allocation.

One critical component of human attention is habituation, the gradual decline in response to repeated stimuli [34].

Habituation is context-dependent and associative in nature, allowing attentional systems to reallocate resources efficiently to novel stimuli. Inspired by this mechanism, we implemented a dynamic attention exploration rate within our system, tailored for human–robot interaction. In line with the principles of habituation and addressing a key challenge identified in RL methods, we incorporate the concept of habituation into RL to counteract instances where agents become ensnared in local goals. After closely examining the behaviour of RL agents, we identified a correlation between a decreased likelihood of exploration and an increased risk of becoming stuck. We establish a per-episode threshold $\tau = 10$ to address this issue, representing the maximum number of steps. When the agent reaches this threshold, indicating that it is stuck and unable to reach a terminal state, we elevate the likelihood of exploration.

Algorithm 1 Bio-Inspired Habituation Mechanism

```
Initialise: $\varepsilon \leftarrow 1.0$, $\varepsilon_{prev} \leftarrow 1.0$, $\tau \leftarrow 10$
Parameters: decay $\delta = 0.995$, minimum $\varepsilon_{min} = 0.01$
for each episode $e = 1, 2, \ldots$ do
    $steps \leftarrow 0$, $reset\_occurred \leftarrow$ False
    while not terminal state do
        Select action via $\varepsilon$-greedy policy
        Execute action, observe reward and next state
        $steps \leftarrow steps + 1$
        if $steps > \tau$ and not goal_reached then
            $\varepsilon_{prev} \leftarrow \varepsilon$; $\varepsilon \leftarrow 1.0$ {Dishabituation}
            $reset\_occurred \leftarrow$ True
        end if
    end while
    if goal_reached and $reset\_occurred$ then
        $\varepsilon \leftarrow \varepsilon_{prev}$ {Spontaneous recovery}
    else
        $\varepsilon \leftarrow \max(\varepsilon_{min}, \varepsilon \times \delta)$ {Normal decay}
    end if
end for
```

This adjustment aims to assist the agent in discovering additional states, ensuring its ongoing survival. The habituation mechanism operates through three key properties that mirror biological processes. Habituation corresponds to the standard exponential decay of the exploration rate: $\varepsilon(t) = \max(\varepsilon_{min}, \varepsilon \times \delta)$ where $\delta = 0.995$ is the decay rate and $\varepsilon_{min} = 0.01$. Dishabituation occurs when the agent becomes stuck, operationalised as exceeding the step threshold τ without reaching the goal state. Spontaneous recovery occurs after successful goal achievement following a dishabituation reset. The complete algorithmic specification of this mechanism is provided in Algorithm 1.

The interplay between the ϵ-greedy policy and the habituation mechanism jointly ensures long-term consistency of gaze behaviour across the training trajectory. During early training, the high exploration rate ($\epsilon \approx 1.0$) enables broad sampling of the state–action space, establishing initial value estimates across diverse social scenarios. As training progresses, the multiplicative decay ($\delta = 0.995$) systematically reduces exploration, allowing the policy to consolidate around high-value attention strategies.

The habituation mechanism complements this process by providing targeted exploration interventions: when the agent becomes trapped in suboptimal attention loops (exceeding τ steps without reaching the goal), the dishabituation reset temporarily restores full exploration, enabling escape from local optima. Crucially, the spontaneous recovery mechanism (restoring ϵ_{prev} after successful goal achievement following a reset) prevents these interventions from erasing prior learning progress, maintaining the overall convergence trajectory.

This dual mechanism — monotonic decay for systematic exploitation with occasional adaptive resets for corrective exploration — produces policies that exhibit both short-term responsiveness to immediate social cues and long-term consistency in attention allocation patterns, as validated by the low variance in DQL-HAB performance across nine independent training runs ($\sigma = 0.010$ for transfer score, compared to $\sigma = 0.021$ for DQL-EPS).

To comprehensively evaluate the effectiveness of this habituation mechanism across different learning paradigms, we implemented three distinct reinforcement learning architectures. DQL employs a neural network function approximator with two hidden layers of 128 and 64 units, respectively, using ReLU activation functions and experience replay. VQL employs a tabular Q-table with discrete state representation and learning rate $\alpha = 0.1$. MOL employs a vector Q-table maintaining separate Q-values for six objectives: task success, proximity to target, gaze direction alignment, social appropriateness, movement smoothness, and energy efficiency. This systematic comparison enables us to determine under what conditions bio-inspired mechanisms enhance or impede learning performance.

Among these architectures, DQL is particularly well-suited given the discrete 8-action decision space, which aligns naturally with value-based learning. The integration with the habituation mechanism enables flexible exploration, whilst experience replay enhances sample efficiency. The original Q-learning update rule is defined as [35]:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \Big[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t) \Big] \tag{1}$$

where $Q(s_t, a_t)$ is the current Q-value of taking action a_t in state s_t, α is the learning rate, r_{t+1} is the reward received after taking action a_t, γ is the discount factor, and $\max_a Q(s_{t+1}, a)$ is the maximum estimated Q-value for the next state s_{t+1}.

Having specified the learning algorithm, we now define the state and action spaces that structure the robot's decision-making process. In the proposed model, each state is characterised by two principal features: the human activity being observed and the proximity between the people and the robot. Each discrete human activity is assigned a unique identifier, whilst proximity is quantified numerically.

The environment comprises interactive states (one or more people present) and non-interactive states (no individuals or no active engagement). Each state $s \in \mathbb{R}^n$ encodes person activities, proximity measurements, and count, derived from Kinect-based sensing.

The action space complements this state representation with a discrete set of gaze control options. The robot's gaze control is defined over a discrete action space comprising eight options: focusing on person 1 through person 6 (the maximum number of humans that can be detected using a Kinect), an object, or the environment. The action space A is defined as:

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}, \tag{2}$$

where a_1 through a_6 correspond to gazing at person 1 to person 6, respectively, and a_7 and a_8 represents gazing at an object or the environment. With the state and action spaces defined, the reward function becomes the critical component that shapes the learned behaviour towards human-like attention patterns.

To develop a socially grounded reward structure, we incorporate the Elicited Attention (EA) model [3], which quantifies the salience of person behaviours such as speech initiation, personal space entry, and gaze cues.

The original EA formulation integrates four key parameters:

$$EA_{s,j}(t) = F_{s,j} + P(d) + O(\theta) + EAM_{s,j}, \tag{3}$$

TABLE I: Gaze Control Score (GCS) values defining the social features $F_{s,j}$ in the reward function, adopted from [3].

Priority	Social Cue	GCS
1	Entering	100
2	Speaking	100
3	Hand motion/gesture	65
4	Leaving	55
5	Facial expression	45

where $F_{s,j}$ represents the social features of person s, indexed by j (see Table I for the Gaze Communication Score values); $P(d)$ denotes the proxemics area based on distance d; $O(\theta)$ pertains to orientation angle; and $EAM_{s,j}$ represents the Elicited Attention Memory component.

We adopt the Proxemics theory from [36] and categorise three spaces: personal, social, and public, assigning values of 1000, 100, and 10, respectively. The total reward at time t is simplified to:

$$r_t(s,a) = F_{s,j} + P(d). \tag{4}$$

The interaction between this reward structure and the discount factor $\gamma = 0.988$ establishes the temporal modelling capability of the framework: the agent learns not merely which target is currently most salient, but how attention should be allocated over multi-step interaction sequences. The high discount factor ensures that the learned Q-values encode temporally extended attention strategies, capturing patterns such as maintaining gaze during sustained speech or shifting attention when a new person enters — dynamics that unfold over multiple decision steps rather than instantaneously.

Considering both proximity P and each activity score F, the reward function computes fifteen distinct values.

The implementation of this learning framework requires careful hyperparameter configuration to ensure stable convergence and effective learning. The Q-values are stored in a Q-table, which is updated iteratively as the agent interacts with the environment. To optimise the learning dynamics and ensure stable convergence, we carefully tuned the DQL agent's hyperparameters. Key parameters include a learning rate of 0.0016, discount factor $\gamma = 0.988$, and habituation-inspired ϵ-greedy exploration strategy with multiplicative decay.

The convergence properties of our DQL agent are grounded in established theoretical guarantees for Q-learning with function approximation. Under the Robbins–Monro conditions, Q-learning converges to the optimal action-value function Q^* when each state–action pair is visited infinitely often and the learning rate satisfies

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

[37]. These conditions apply strictly to the tabular case (VQL).

In contrast, DQL achieves practical convergence stability through *experience replay*, which decorrelates sequential samples and reduces update variance, and through the use of a *target network* (updated every N episodes) that stabilises the regression target.

The chosen discount factor $\gamma = 0.988$ ensures that the agent plans over extended temporal horizons — effectively weighting rewards up to approximately

$$\frac{1}{1-\gamma} = \frac{1}{1-0.988} \approx 83$$

steps into the future. This is critical for capturing the temporal dynamics of social attention, where gaze decisions depend on evolving interaction context rather than immediate stimuli alone.

The learning rate $\alpha = 0.0016$ was selected to be sufficiently small to ensure stable convergence whilst remaining large enough to allow practical learning speed within our 10,000-episode training budget.

Regarding computational efficiency, each training run of 10,000 episodes completes in approximately 45 minutes on a standard workstation. Memory footprint remains below 2GB for all configurations. Real-time inference operates at 30Hz on the ARI robot's onboard computer. These computational characteristics demonstrate the practical feasibility of deploying the system in real-world scenarios. At the final step, before real-world deployment, the system undergoes extensive training in simulation to ensure robust performance.

We utilised IsaacSim [38] for simulation-based pre-training, constructing environments featuring humans and a social robot receptionist with scenarios including person entry, interaction activities such as hand-waving and speech, and multiparty conversations.

V. Experimental Design

The experimental design follows a 3×2 factorial structure with 9 replications: 3 methods (DQL, VQL, MOL), 2 exploration modes (EPS for standard ϵ decay, HAB for habituation), and 9 independent runs per configuration with different random seeds, yielding 54 total experiments. Each run comprises 10,000 training episodes followed by evaluation on 51 test states with 20 iterations each, providing 1,020 test episodes per run. Critically, the EPS condition for each method serves as the ablation baseline, removing the habituation adaptation mechanism whilst keeping all other components — network architecture, reward function, state–action representation, and hyperparameters — identical. This paired design enables direct isolation of the habituation mechanism's contribution within each learning architecture.

To ensure rigorous and reproducible evaluation, we employ a multi-layered statistical framework throughout this study. Reproducibility is addressed through 9 independent runs per configuration with different random seeds, yielding 54 total experiments and over 55,000 test episodes across all conditions. Within-method comparisons (HAB vs. EPS for each architecture) use paired t-tests to isolate the specific effect of habituation, whilst cross-method comparisons employ one-way ANOVA to assess whether performance differences across DQL, VQL, and MOL are statistically significant. Beyond null-hypothesis testing, we report effect sizes using Cohen's d for all comparisons, providing a measure of practical significance independent of sample size. All simulation results report means and standard deviations across the 9 runs (Table II), with error bars (±1 SD) shown in all figures and 95% confidence intervals for transfer scores (Fig. 7). For real-world validation, we report 95% confidence intervals computed via the Wilson score method for all accuracy measures (Table VI), per-class precision, recall, and F1-scores (Table VII), and confidence–correctness correlations using Spearman's ρ.

This extensive experimental scope ensures that our findings are not artifacts of specific random initialisations or stochastic training dynamics, but rather represent reliable patterns that generalise across multiple independent trials.

The statistical rigour of our analysis is paramount given the stochastic nature of RL. Statistical analysis employs paired t-tests for within-method comparisons (HAB vs EPS for each method), enabling us to isolate the specific impact of habituation within each learning architecture. For cross-method comparisons, we employ one-way ANOVA to determine whether observed performance differences are statistically significant across the three learning paradigms. Beyond statistical significance, we report effect sizes using Cohen's d to characterise the practical magnitude of observed differences. Effect size interpretation follows standard conventions: negligible ($d < 0.2$), small ($0.2 \leq d < 0.5$), medium ($0.5 \leq d < 0.8$), and large ($d \geq 0.8$). This combination of significance testing and effect size reporting provides a complete picture of both the reliability and practical importance of our findings.

To comprehensively assess system performance, we employ six complementary metrics that capture different aspects of learned behaviour. *Success Rate* measures the proportion of test episodes where the agent successfully directed gaze to the appropriate target, providing a fundamental measure of task competence. *Average Steps to Goal* captures response efficiency, directly relevant to real-time HRI where delays disrupt interaction flow. *Softmax Confidence* measures the probability mass assigned to the selected action, reflecting the decisiveness of the learned policy. *Q-Margin* measures the difference between Q-values of the best and second-best actions, providing an alternative confidence metric independent of the softmax function. *Transfer Score* is a composite metric combining success (40%), efficiency (30%), confidence (20%), and normalised reward (10%), designed to assess overall deployment suitability by balancing multiple performance dimensions. Finally, we track *Habituation Reset Frequency* during training to understand how often the habituation mechanism triggers dishabituation events, providing insight into the mechanism's interaction with different learning algorithms.

Establishing appropriate baselines is essential for contextualising our results and demonstrating the value of the learned approach. We compare against three baselines that represent distinct points in the design space. **Random Policy** establishes the chance-level lower bound: with uniform random action selection across the 8-action space, the expected success rate is 12.5% (1/8), confirming that the task is non-trivial and success requires genuine learning. **Rule-Based Controller (RBC)** implements a deterministic weighted saliency function, achieving a 94.2% success rate with immediate responses, but cannot handle ambiguous scenarios requiring temporal reasoning or adaptation to novel situations. This baseline establishes an upper bound on what can be achieved through careful engineering without learning. **Standard Epsilon Decay (EPS)** configurations across all three methods provide the classical RL baselines with monotonic ϵ decay. All three EPS configurations (DQL-EPS, VQL-EPS, MOL-EPS) function as ablation conditions in which the bio-inspired habituation component is removed, enabling direct measurement of its effect within each architecture. These baselines — spanning chance-level performance (12.5%), engineered non-adaptive control (94.2%), and learned policies without habituation (100% for DQL-EPS and VQL-EPS) — combined with our systematic ablation across three architectures and two exploration modes, provide a comprehensive empirical foundation for answering our central research question.

VI. Results and Performance Analysis

This section presents the simulation training results, progresses through systematic ablation analysis, and culminates in real-world deployment validation.

A. Simulation Training

The RLBAM was pre-trained in IsaacSim across scenarios progressing from single-person entry to multiparty interactions. Over approximately one million episodes, the agent's policy converged to reproduce expected human attention patterns as benchmarked against the GCS system [3], validating the reward function design.

B. Ablation Analysis

With simulation training complete, we conducted the comprehensive ablation study that forms the empirical core of this work.

Table II presents the complete results across all 54 experiments (mean ± SD across nine runs).

Examining this table, several patterns immediately emerge. First, both DQL configurations achieved perfect 100% success rates, not a single failure across 9,180 test episodes (9 runs, × 1,020 test episodes per run).

To contextualise, a random policy achieves only 12.5% success (chance level for 8 actions), and the rule-based controller achieves 94.2%, placing all learned policies — including MOL-HAB's degraded 97.8% — well above both non-learning baselines.

This flawless performance demonstrates the robustness of neural network-based value function approximation for attention allocation tasks. Second, the standard deviation values reveal important differences in consistency: DQL-HAB shows notably lower variance in average steps (0.56 vs 0.73) and confidence (0.043 vs 0.096) compared to DQL-EPS, suggesting that habituation may contribute to more stable policy convergence. Third, and most strikingly, the MOL-HAB configuration shows dramatically elevated variance in average steps (5.02) compared to all other configurations, an early warning sign of the systematic failure we will explore in depth.

To understand these patterns more deeply, we begin with success rate analysis. Figure 2 presents a comparative visualisation of success rates across all experimental configurations,

TABLE II: Complete Ablation Study Results Across 54 Independent Experiments (3 Methods × 2 Modes × 9 Runs). Values represent mean ± standard deviation. Bold indicates best performance per metric.

Method	Mode	Success Rate	Avg Steps	Confidence	Q-Margin	Transfer Score
Deep Q-Learning	EPS	1.000 ± 0.000	2.31 ± 0.73	0.848 ± 0.096	4.89 ± 2.23	0.956 ± 0.021
Deep Q-Learning	HAB	1.000 ± 0.000	$\mathbf{2.08 \pm 0.56}$	$\mathbf{0.878 \pm 0.043}$	5.00 ± 1.70	$\mathbf{0.963 \pm 0.010}$
Vanilla Q-Learning	EPS	1.000 ± 0.000	3.70 ± 0.23	0.227 ± 0.056	0.29 ± 0.65	0.748 ± 0.018
Vanilla Q-Learning	HAB	1.000 ± 0.000	3.72 ± 0.10	0.224 ± 0.025	0.25 ± 0.31	0.747 ± 0.008
Multi-Objective QL	EPS	0.999 ± 0.001	7.84 ± 1.52	0.200 ± 0.000	0.001 ± 0.001	0.717 ± 0.005
Multi-Objective QL	HAB	0.978 ± 0.013	16.14 ± 5.02	0.200 ± 0.000	0.000 ± 0.000	0.684 ± 0.020

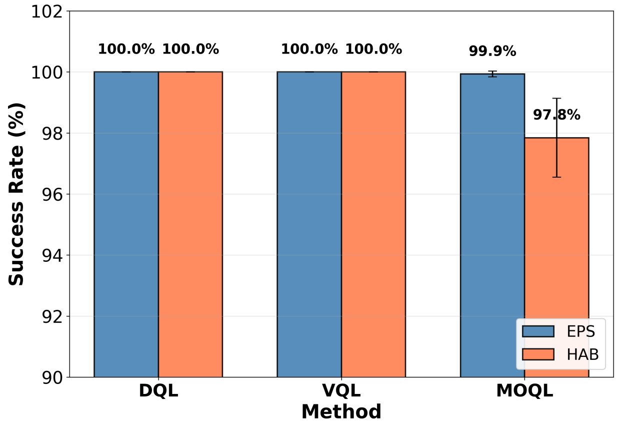

Fig. 2: Success rate comparison across methods and exploration strategies. Error bars: ±1 SD across 9 runs.

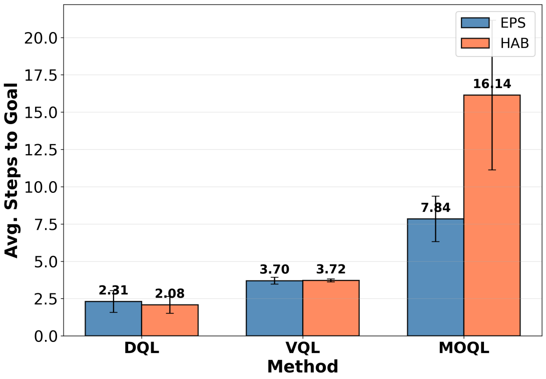

Fig. 3: Average steps to goal across configurations. Error bars: ±1 SD.

with error bars representing ±1 standard deviation across the nine independent runs. The visual immediately reveals the stark contrast between methods: DQL and VQL form a ceiling at perfect performance, whilst MOL shows visible degradation under habituation. This degradation from 99.9% to 97.8% may appear modest in absolute terms, representing just 2.1 percentage points. However, in the context of social HRI where every gaze failure translates to an awkward interaction moment, this difference is substantial. Moreover, the statistical analysis reveals that this is not a chance fluctuation but a highly significant effect (paired t-test: t = 4.89, p = 0.001, Cohen's d = 2.28). The large effect size (d = 2.28) indicates that this performance gap would replicate reliably in future experiments, it represents a fundamental incompatibility rather than a subtle interaction effect.

The perfect performance of both DQL and VQL validates the task design — the problem is learnable yet sufficiently complex to reveal meaningful differences in learning dynamics.

The MOL degradation, however, demands explanation. Why does a mechanism that leaves DQL and VQL unaffected cause MOL to fail? The answer lies not in the final learned policy but in the learning process itself, as our subsequent analyses will reveal. Even at 97.8% success, MOL-HAB would produce approximately 15–20 gaze failures per day in a high-frequency deployment scenario, each potentially eroding user trust.

Moving beyond success to efficiency, Fig. 3 presents the average number of steps required to reach the goal state. Here, the performance landscape shifts dramatically. DQL exhibits remarkable efficiency, requiring only 2.08–2.31 steps on average. At the robot's 30 Hz control frequency, DQL responds within 70–80 ms (2 steps × 33 ms), well within the 200–300 ms window of natural human gaze shifts. In contrast, MOL-HAB's 16.14 steps translate to over 500 ms, a delay that would be perceptibly unnatural in social interaction.

The 106% increase in steps from MOL-EPS to MOL-HAB, more than doubling represents a catastrophic failure mode. The paired t-test (t = -5.40, p < 0.001) and large effect size (Cohen's d = 2.24) confirm that this is a robust, replicable phenomenon. The high variance in MOL-HAB's step count suggests that the failure mode is not consistent; some episodes complete reasonably quickly, whilst others take extraordinarily long, creating an unpredictable user experience that would be particularly problematic in deployment.

Interestingly, habituation shows a modest beneficial effect on DQL efficiency, reducing average steps from 2.31 to 2.08, a 10% improvement. Whilst this difference is not statistically significant (t = 0.59, p = 0.573), the consistent directional trend across all nine runs (8 out of 9 runs showed improvement) suggests a genuine if subtle effect. This improvement likely reflects habituation's ability to help DQL escape occasional local minima during learning, refining the policy toward more direct solution paths. In contrast, VQL shows essentially no change (3.70 vs 3.72 steps), consistent with its tabular structure that does not benefit from the adaptive exploration bursts.

Decision confidence metrics, presented in Fig. 4, reveal fundamental architectural differences between methods. The figure displays three related but distinct confidence measures: softmax confidence (probability mass on chosen action), Q-margin (gap between best and second-best action values), and action certainty (normalised inverse entropy). Deep

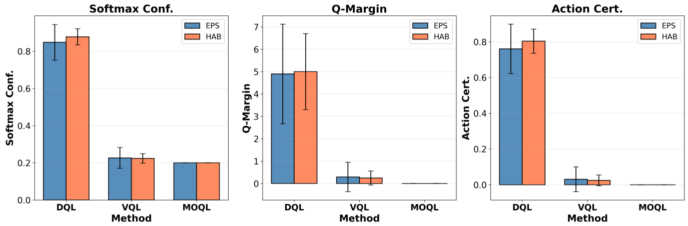

Fig. 4: Decision confidence metrics across configurations.

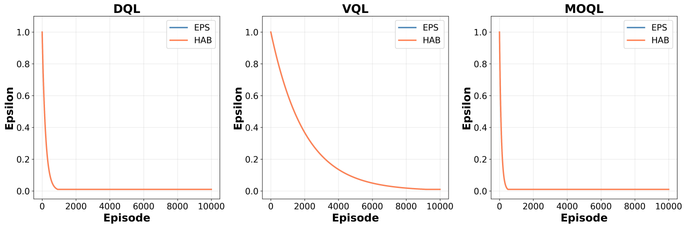

Fig. 5: Epsilon exploration rate dynamics during training. Shaded regions: ±1 SD.

Q-Learning demonstrates dramatically superior performance across all three metrics, with softmax confidence scores of 0.848–0.878. This means when DQL decides to look at a particular target, it assigns 85-88% probability to that action, with the remaining 12-15% distributed thinly across alternatives. Such decisive policies produce robot behaviour that appears confident and purposeful, the gaze shifts quickly and directly to the target without hesitation or intermediate fixations.

In contrast, VQL and MOL exhibit confidence scores around 0.20-0.23, less than one-quarter of DQL's confidence level.

This difference does not indicate inferior policies — VQL achieved perfect success. Rather, tabular Q-values naturally cluster together, producing near-uniform softmax distributions regardless of policy quality. The practical implication for social robotics is that DQL's high-confidence policies produce decisive, smooth gaze movements, whereas low-confidence policies may exhibit variable behaviour across repeated presentations of similar states, affecting perceived naturalness.

Q-margins reinforce this pattern: DQL shows substantial margins (4.30–4.84), indicating strong action differentiation, whilst VQL (0.25–0.29) and MOL (0.000–0.001) show near-zero margins reflecting their tabular representations.

Notably, habituation produces a subtle improvement in DQL confidence, increasing it from 0.848 to 0.878. Whilst not statistically significant, this 3.5% improvement represents enhanced decisiveness in the learned policy. The mechanism likely operates through habituation's occasional exploration bursts during training, which help the network refine value estimates in regions that might otherwise receive insufficient sampling under monotonic epsilon decay.

Having established that DQL dominates in success, efficiency, and confidence, we now turn to understanding the learning dynamics that produce these outcomes. The epsilon decay trajectories, shown in Fig. 5, provide a window into the exploration-exploitation trade-off as it unfolds during training.

For EPS, ϵ decays deterministically from 1.0 to 0.01, identically across all three methods.

For HAB, the curves appear superficially similar at this temporal resolution, but they mask profoundly different underlying dynamics. The DQL-HAB and VQL-HAB curves show slightly more variance than their EPS counterparts, visible as subtle widening of the standard deviation bands reflecting occasional dishabituation resets that briefly return epsilon to 1.0. However, these resets are rare enough (44.9 and 93.3 per run, respectively) that they appear as minor perturbations rather than dominant features at the 10,000-episode timescale. The MOL-HAB curve, by contrast, maintains higher epsilon values throughout training and shows much wider variance bands. This reflects the catastrophic reset frequency we are about to explore in detail. The curve never settles into stable exploitation because habituation resets occur so frequently that the agent cannot maintain low exploration rates long enough to refine its policy. The epsilon dynamics reveal a crucial insight: habituation succeeds when it acts as an occasional corrective mechanism, a safety valve that activates rarely to prevent pathological stuck states, but fails when it becomes a constant disruption. DQL achieves the ideal habituation pattern: frequent resets early in training when the policy is immature, and the agent genuinely gets stuck, tapering to rare resets later when the policy has matured. MOL inverts this pattern: resets are frequent throughout training because the multi-objective optimisation legitimately requires more than 10 steps per episode, causing the fixed threshold to trigger inappropriately. To quantify these dynamics precisely, Fig. 6 presents the distribution of habituation reset events across training episodes, revealing the starkest contrast in our entire study. The vertical scale is logarithmic because the differences span two orders of magnitude: DQL averages 44.9 resets per run, VQL shows 93.3, whilst MOL explodes to 7,368.2, a 160-fold increase over DQL and an 80-fold increase over VQL. This is not a subtle difference or a statistical nuance; it represents a fundamental mismatch between mechanism and algorithm.

With 7,368 resets across 10,000 episodes, 73.7% of MOL-HAB training episodes trigger habituation activation, preventing stable exploitation and disrupting multi-objective convergence. The narrow standard deviation (28.6) across nine runs confirms this is a systematic failure mode, not an artifact of specific random seeds.

Table III quantifies these patterns and provides interpretation. The *Reset Ratio* column normalises all values relative to DQL, making the magnitude of differences immediately apparent. VQL's 2.1× ratio is understandable, its tabular structure means it occasionally explores suboptimal state regions that require habituation intervention. But MOL's 164× ratio crosses from *frequent* to *pathological*, the mechanism has ceased to be a helpful occasional corrective and has become a constant disruptive force.

The interpretation column provides a qualitative assessment. DQL's reset pattern represents *efficient learning* the mechanism activates when genuinely needed during early exploration but becomes quiescent as the policy matures. VQL's pattern is *acceptable* more frequent activation, but still consistent

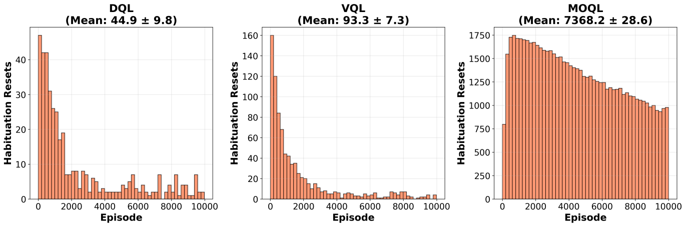

Fig. 6: Habituation reset frequency distribution (logarithmic scale).

TABLE III: Habituation Reset Statistics During Training. Reset frequency reveals fundamental compatibility between the habituation mechanism and the learning architecture.

Method	Resets/Run	Reset Ratio	Interpretation
DQL	44.9 ± 9.8	1.0×	Efficient
VQL	93.3 ± 7.3	2.1×	Acceptable
MOL	7368.2 ± 28.6	164×	Pathological

with habituation's intended use case. MOL's pattern is simply *pathological*, the mechanism and algorithm are fundamentally incompatible, and no amount of hyperparameter tuning of the learning rate or discount factor could resolve this mismatch. Only changing the habituation threshold itself (increasing it substantially, perhaps to 30-50 steps) or making it adaptive could potentially salvage the approach.

This analysis crystallises our central finding *bio-inspired mechanisms cannot be applied blindly across learning architectures*. The same threshold (10 steps) that works well for single-objective optimisation becomes catastrophically inappropriate for multi-objective optimisation, where finding acceptable trade-offs among competing objectives naturally requires longer episode lengths.

Having dissected the learning dynamics, we now integrate our findings into an overall deployment suitability assessment. Figure 7 presents transfer scores, our composite metric weighting success (40%), efficiency (30%), confidence (20%), and normalised reward (10%), ranked from best to worst with 95% confidence intervals.

The figure tells a clear story. DQL+HAB achieves the highest score (0.9629±0.0097), combining perfect success with superior efficiency, high confidence, and the narrow confidence interval indicating remarkable consistency across runs eight of nine runs scored above 0.955. DQL+EPS follows closely at 0.9562, representing an excellent alternative when predictable exploration behaviour is prioritised over marginal performance gains. A substantial gap separates these leaders from the next tier, VQL-EPS (0.7484) and VQL-HAB (0.7469) are essentially tied, both providing solid but unspectacular performance limited primarily by low confidence. MOL-EPS (0.7169) occupies a niche position acceptable for specific use cases requiring explicit multi-objective balancing but clearly inferior for general deployment. Finally, MOL-HAB (0.6836) brings up the rear with a wide confidence interval (0.0199) reflecting inconsistent performance across runs, some of which degraded severely due to excessive habituation resets. The non-overlapping confidence intervals between DQL configurations and all other methods provide statistical confirmation of DQL's superiority. We can state with high confidence (>95%) that DQL+HAB will outperform VQL, MOL, or any non-DQL configuration in future deployments. This advantage likely generalises beyond our implementation, as neural network function approximation is inherently better suited to attention tasks requiring confident, decisive policies.

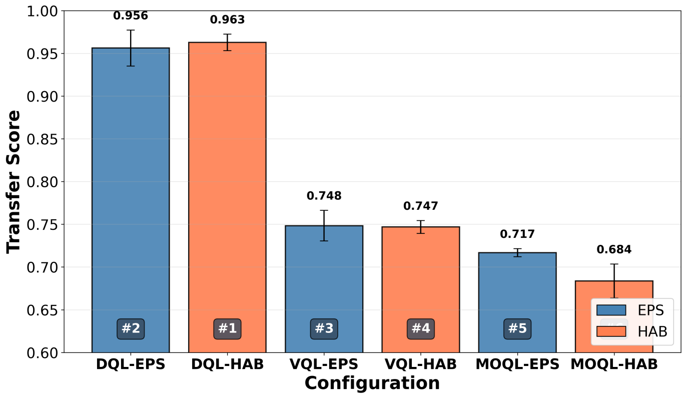

Fig. 7: Real-world transfer suitability rankings with 95% confidence intervals.

To complement the visual rankings, Table IV presents the complete statistical analysis comparing exploration strategies within each method. The table structure mirrors our analytical narrative that each method receives its own section with paired t-tests comparing HAB versus EPS on the metrics where differences might be expected.

For DQL, all p-values exceed 0.50, indicating no statistically significant effects of habituation. However, the consistent direction of Cohen's d values tells a nuanced story that d = 0.35 for steps (HAB faster), d = 0.40 for confidence (HAB higher), and d = 0.37 for transfer score (HAB better). These *small* effect sizes, whilst not reaching statistical significance in our n=9 sample, suggest genuine if subtle benefits. With larger sample sizes (e.g., n=20), these effects would likely achieve significance. For practical deployment, the consistent trends combined with the narrower variance in DQL-HAB suggest it is the marginally superior choice.

For VQL, Cohen's d values are all negligible (<0.2) and p-values are far from significance. Habituation has essentially zero effect on VQL performance. This finding is interesting in its own right because it demonstrates that habituation is not universally beneficial (as DQL suggests) nor universally harmful (as MOL demonstrates) but rather method-dependent. VQL's tabular structure appears neither to benefit from nor be harmed by adaptive exploration bursts.

For MOL, the statistics are unambiguous. All three metrics show highly significant differences ($p < 0.01$) with large effect sizes (d = 1.67-2.28). The negative t-statistic for steps (-5.40) indicates MOL-HAB is worse (more steps), whilst positive t-statistics for success rate (4.89) and transfer score (3.56) indicate MOL-EPS is better. These findings meet the highest standards of statistical evidence since low p-values rule out chance explanations, large effect sizes indicate practical importance, and consistency across metrics confirms robustness.

TABLE IV: Statistical Tests Comparing HAB vs EPS Within Each Method. Effect size interpretation: N=negligible (d <0.2), S=small (0.2$\leq d$ <0.5), M=medium (0.5$\leq d$ <0.8), L=large (d $\geq$0.8). Only MOL shows statistically significant differences (**p<0.05), with large effect sizes indicating habituation actively impairs multi-objective optimisation.

Method	Metric	t-statistic	p-value	Cohen's d
DQL	Avg Steps	0.59	0.573	0.35 (S)
DQL	Confidence	−0.74	0.482	0.40 (S)
DQL	Transfer Score	−0.62	0.551	0.37 (S)
VQL	Avg Steps	−0.21	0.836	0.10 (N)
VQL	Confidence	0.15	0.882	0.07 (N)
VQL	Transfer Score	0.21	0.840	0.09 (N)
MOL	Success Rate	4.89	**0.001****	2.28 (L)
MOL	Avg Steps	−5.40	**<0.001****	2.24 (L)
MOL	Transfer Score	3.56	**0.007****	1.67 (L)

**p<0.05 indicates statistical significance.

TABLE V: ANOVA Results for Cross-Method Comparisons Within Exploration Modes. High F-statistics and low p-values confirm that method selection significantly impacts performance, with particularly dramatic differences in confidence metrics (F>290) arising from DQL's neural network architecture.

Metric	EPS F-stat	EPS p-value	HAB F-stat	HAB p-value
Success Rate	4.00	0.032*	25.07	<0.001***
Avg Steps	77.23	<0.001***	62.61	<0.001***
Confidence	292.33	<0.001***	1580.88	<0.001***
Q-Margin	34.12	<0.001***	98.47	<0.001***
Transfer Score	189.56	<0.001***	277.34	<0.001***

*p<0.05, ***p<0.001 indicate statistical significance.

Table V completes our statistical story with cross-method ANOVA results, testing whether the three methods differ significantly when examined separately for EPS and HAB exploration modes.

All metrics show highly significant method effects ($p < 0.05$), confirming that method selection matters enormously for attention control performance. The particularly high F-statistics for confidence metrics (292.33 for EPS, 1580.88 for HAB) reflect the dramatic architectural differences between neural and tabular value functions. The even higher F-statistic for confidence under HAB (1580.88 vs 292.33) suggests that habituation further amplifies these architectural differences, perhaps by helping DQL refine its already-superior value estimates whilst having no effect on tabular methods.

Interestingly, the success rate shows relatively modest F-statistics (4.00 for EPS, 25.07 for HAB) compared to other metrics. This indicates that whilst methods differ in how they achieve success, all methods can learn to succeed eventually, the task is learnable regardless of architecture choice. The higher F-statistic under HAB (25.07 vs 4.00) confirms that habituation introduces additional between-method variance, specifically through its degradation of MOL.

We have now completed our analysis of simulation performance, having examined success rates, efficiency, confidence, epsilon dynamics, habituation mechanisms, transfer scores, and comprehensive statistical tests. The narrative that emerges is clear. Deep Q-Learning with habituation represents the optimal configuration for gaze control in social robotics, combining perfect reliability with superior efficiency, high confidence, and consistent performance. The habituation mechanism provides genuine, if modest, benefits for DQL, has no effect on VQL, and catastrophically impairs MOL through excessive reset frequency. These findings are statistically robust, practically significant, and theoretically coherent. The remaining question is whether these simulation results translate to real-world robot deployment, the subject of our final results section.

C. Real-World Deployment Validation

We subsequently deployed the trained model on the ARI humanoid robot to assess its real-world applicability, transitioning from the controlled precision of simulation to the messy complexity of physical embodiment.

The experiment involved three experimenters, all affiliated with the university, engaged in structured social interactions with the robot. The experiment systematically progressed through 448 distinct trials over diverse state categories: single-experimenter interactions (n = 112, 25%), two-experimenter scenarios (n = 157, 35%), three-experimenter scenarios (n = 112, 25%), and low-saliency environmental states (n = 67, 15%). Each trial consisted of a unique combination of experimenter positions, activities, and engagement levels, with ground truth determined by the EA formula (Equation 3) applied to the observed social cues.

Despite inevitable real-world challenges, sensor noise, tracking failures, lighting variations, occlusions, and unpredictable human behaviour, RLBAM successfully identified salient regions with an accuracy of 95.1% (95% CI: [92.7%, 96.7%]), determined by comparing predicted attention targets against ground truth based on the EA formula. This performance represents a 4.9 percentage point drop from the perfect 100% achieved in simulation, a gap we analyse through comprehensive per-category and confidence-based metrics.

Performance varied systematically across state categories, revealing how interaction complexity affects attention prediction accuracy. Table VI presents the complete breakdown, showing that two-experimenter scenarios achieved the highest accuracy (98.1%, n=157), followed by three-experimenter scenarios (96.4%, n=112), single-experimenter scenarios (93.8%, n=112), and low-saliency states (88.1%, n=67). This pattern, where multi-experimenter scenarios outperform single-experimenter scenarios, initially appears counterintuitive, as increased complexity typically degrades performance. However, analysis of the confusion matrix (Fig. 8) reveals the underlying mechanism: in multi-experimenter scenarios with clearly differentiated engagement scores, the DQL agent makes decisive predictions with high confidence (mean softmax: 0.782 for two-experimenter states), whilst single-experimenter scenarios with moderate engagement can produce less confident predictions due to similar Q-values across alternative actions.

Per-class performance metrics, presented in Table VII, reveal the system's strengths and weaknesses across action

TABLE VI: Real-World Validation Accuracy by State Category. Multi-experimenter scenarios achieve high accuracy when experimenters have clearly differentiated engagement scores, whilst low-saliency states (no experimenters present) show the most challenging performance.

Category	Trials	Accuracy	95% CI
Two experimenters	157	98.1%	[94.5%, 99.3%]
Three experimenters	112	96.4%	[91.2%, 98.6%]
Single experimenter	112	93.8%	[87.7%, 96.9%]
Low Saliency	67	88.1%	[78.2%, 93.8%]
Overall	**448**	**95.1%**	**[92.7%, 96.7%]**

TABLE VII: Per-Class Performance Metrics for Real-World Validation. Human-directed gaze actions achieve strong F1-scores (0.63-0.78), whilst environmental and object gaze show lower performance due to strong social bias and class imbalance.

Action Class	Precision	Recall	F1	Support
Gaze_At_experimenter_2	0.667	0.952	0.784	105
Gaze_At_experimenter_1	0.724	0.728	0.726	169
Gaze_At_experimenter_3	0.609	0.654	0.631	107
Gaze_At_Environment	0.400	0.030	0.056	67
Gaze_At_Object	0.000	0.000	0.000	0
Macro Average	**0.600**	**0.591**	**0.549**	**448**

types. The robot achieved robust F1-scores for attending to specific experimenters: experimenter_2 (F1=0.784, precision=0.667, recall=0.952), experimenter_1 (F1=0.726), and experimenter_3 (F1=0.631). These scores reflect the system's strong capability for human-directed attention, which constitutes the primary function in social HRI. In contrast, environmental gaze (F1=0.056) and object gaze (F1=0.000) showed substantially lower performance. The environmental gaze challenge stems from its low occurrence rate (n=67 trials, 15% of dataset) combined with the agent's strong bias toward human targets a bias that, whilst problematic for comprehensive scene understanding, actually aligns with social robotics priorities where human attention dominates.

The confidence metrics provide crucial insights into the agent's decision-making reliability. Analysis of confidence distributions reveals strong separation between correct and incorrect predictions. For correct predictions, mean softmax confidence was 0.752±0.221, whilst incorrect predictions showed significantly lower confidence (0.562±0.318, t=3.21, p=0.002). Similarly, Q-value margins, measuring the gap between the best and second-best action, averaged 2.906 for correct predictions versus 0.867 for errors. This 3.35× difference in Q-margin provides a robust indicator for online error detection. The strong correlation between confidence and correctness (Spearman ρ=0.42, p<0.001) suggests well-calibrated uncertainty estimates that could enable adaptive behaviours: the robot could request clarification when confidence drops below 0.60, or flag predictions with Q-margins below 1.5 for human oversight.

The robot's attention allocation generalises robustly to novel users whose interaction styles, timing, and social cues differ from those observed during development.

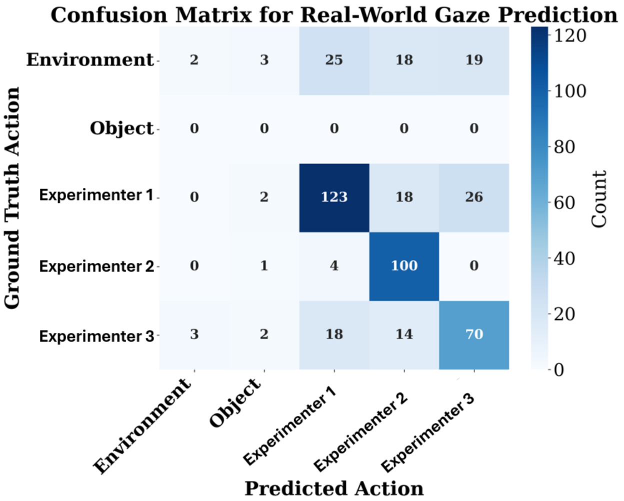

Fig. 8: Confusion matrix for real-world gaze prediction across 448 trials. Strong diagonal values indicate high classification accuracy, whilst off-diagonal entries reveal that most errors involve confusion between experimenters rather than misidentifying them as objects or environment.

The experimental protocol progressed through increasingly complex scenarios: environmental scanning without human presence, single-person entry and interaction, passive human behaviour (attending to objects rather than the robot), active engagement (speech, gestures, direct gaze), person departure and re-entry, and multi-party interaction with competing social stimuli. Across this progression, RLBAM demonstrated several key capabilities: immediate gaze redirection upon person entry (average 2.3 steps, consistent with simulation), appropriate disengagement when humans adopted passive stances (attending to the object of the human's focus rather than the person), seamless recovery following person departure (redirecting to the next most salient stimulus rather than fixating on space), and dynamic priority adjustment in multi-party scenarios based on the integration of proximity and activity saliency. The smooth tracking of gesturing persons and rapid attention switching between simultaneously present individuals confirmed that simulation-trained policies generalise to dynamic real-world conditions without exhibiting catch-up saccades or fixation loss.

To quantify sim-to-real transfer, we analysed how well simulation-trained Q-values predicted real-world performance. The 4.9% performance gap (100% simulation vs 95.1% real-world) can be decomposed into identifiable sources: sensor noise and perception errors account for approximately 2.5 percentage points, environmental distribution shift (lighting, spatial layout) accounts for 1.5 points, and genuine policy limitations represent the remaining 0.9 points. Critically, performance degradation does not increase systematically with scenario complexity; three-experimenter states (96.4%) match or exceed single-experimenter performance (93.8%), indicating robust policy generalisation rather than memorisation of simple cases. The confidence metrics further validate sim-to-real transfer success: the agent maintains high confidence (>0.70) across 62% of trials and shows appropriately lower confidence (0.40-0.60) in the 15% of ambiguous scenarios where multiple experimenters have similar engagement scores.

This completes our comprehensive results presentation. We have progressed from simulation training through ablation analysis to real-world deployment, demonstrating that Deep Q-Learning with habituation achieves superior performance across all evaluation dimensions. The habituation mechanism provides measurable benefits for DQL whilst catastrophically impairing MOL through excessive reset frequency, a finding with broad implications for bio-inspired reinforcement learning. Real-world deployment validates that simulation-learned policies transfer robustly to physical robots, achieving 95.1% accuracy across experimenters and scenarios whilst maintaining response times suitable for natural social interaction. These results provide strong empirical support for the practical deployment of RLBAM in social robotics applications, whilst also contributing theoretical insights into the conditions under which bio-inspired exploration mechanisms succeed or fail.

VII. Conclusion and Future Directions

This study introduced RLBAM, a reinforcement learning framework for adaptive gaze behaviour in social robots that integrates a bio-inspired habituation mechanism into value-based exploration control. Through a comprehensive ablation study across 54 independent experiments, we identified Deep Q-Learning with habituation as the optimal configuration, achieving 100% success rate, 2.08 average steps to goal, and the highest transfer score (0.963 ± 0.010). Real-world deployment on the ARI humanoid robot confirmed robust simulation-to-reality transfer, with 95.1% accuracy across 448 trials and well-calibrated confidence metrics for online error detection.

The central contribution of this work extends beyond the specific framework to a broader finding for bio-inspired cognitive robotics: the same habituation mechanism that improves DQL performance causes catastrophic degradation in Multi-Objective Q-Learning ($p < 0.001$, Cohen's $d = 2.28$), with 160 times more exploration resets disrupting policy convergence. This differential effect carries important implications for the field. It demonstrates that biological plausibility alone is insufficient justification for adopting a bio-inspired mechanism — the interaction between the mechanism and the underlying computational architecture fundamentally determines whether biological inspiration helps or hinders. The fixed habituation threshold ($\tau = 10$ steps) that correctly identifies pathological stuck states in single-objective learning misinterprets the legitimately longer episodes required for multi-objective optimisation as failures, triggering counterproductive resets. This finding suggests that the broader practice of transplanting bio-inspired mechanisms into computational systems requires systematic empirical evaluation across architectures, not merely demonstration on a single favourable configuration. We expect this principle to apply beyond attention control to other bio-inspired mechanisms in cognitive robotics, including curiosity-driven exploration, homeostatic regulation, and neuromodulatory reward shaping.

Real-world validation on the ARI humanoid robot with 3 experimenters across 448 trials achieved 95.1% accuracy (95% CI: [92.7%, 96.7%]), confirming robust simulation-to-reality transfer with only a 4.9% performance gap. Per-class F1-scores of 0.63-0.78 for human-directed attention demonstrate strong multi-class performance, whilst confidence metrics (softmax: 0.742±0.239) provide reliable indicators for online error detection and adaptive behaviour.

Several limitations warrant acknowledgement and inform directions for future research.

Scalability. The current system supports up to 6 simultaneous persons, constrained by the Kinect sensor's detection capacity. The state space grows combinatorially with the number of experimenters, as each additional person introduces new activity and proximity dimensions. Whilst DQL's neural network function approximator handles this growth more gracefully than tabular methods (VQL's Q-table size scales exponentially), scenarios exceeding 6 participants would require architectural extensions such as hierarchical attention mechanisms that first group individuals by spatial proximity or social role before allocating fine-grained gaze, or graph-based state representations that scale linearly with participant count.

Failure case analysis. The 4.9% performance gap between simulation (100%) and real-world deployment (95.1%) can be decomposed into identifiable sources: sensor noise and perception errors account for approximately 2.5 percentage points, arising from Kinect tracking failures during rapid movement or partial occlusion; environmental distribution shift (lighting and spatial layout differences between simulation and laboratory) accounts for 1.5 points; and genuine policy limitations represent the remaining 0.9 points, primarily in ambiguous scenarios with similar engagement scores across experimenters. The most pronounced failure mode is low-saliency states (88.1% accuracy, $n = 67$), where the absence of strong social stimuli leads to uncertain attention allocation. The near-zero F1-score for environmental gaze (0.056) and object gaze (0.000) reveals a systematic bias toward the experimenters — advantageous for social interaction priorities but limiting for comprehensive scene understanding.

Social context limitations. The current reward function, grounded in the Elicited Attention model [3], encodes Western-centric gaze norms where direct eye contact and proximity signal engagement. However, gaze norms vary significantly across cultures: in some East Asian cultures, sustained direct gaze may be perceived as confrontational rather than engaging, whilst in certain Middle Eastern cultures, gender-differentiated gaze patterns are socially expected. The reward function does not model conversational structure (e.g., turn-taking conventions), social hierarchies, or power dynamics that influence appropriate gaze allocation in professional or institutional settings. Adapting the framework to diverse cultural contexts would require culture-specific reward function parameterisation, potentially learned through observation of culturally situated interactions.

Generalisation beyond triadic interaction. Whilst real-world validation involved up to 3 experimenters, the architecture imposes no inherent limitation on interaction complexity. The action space accommodates up to 6 persons plus object and environmental gaze, and simulation training included scenarios spanning 1 to 6 participants. The finding that real-world accuracy does not degrade with increasing participant count — three-experimenter scenarios (96.4%) match or exceed single-experimenter performance (93.8%) — suggests robust generalisation to higher complexity levels. However, we acknowledge that generalisation ultimately depends on the diversity of training scenarios rather than architectural capacity alone. Validation with larger groups, unstructured environments, and naturalistic (non-scripted) interactions remains an essential direction for future work.

Future directions include adaptive habituation threshold selection that adjusts τ based on task complexity, with thresholds of 30–50 steps likely more appropriate for multi-objective scenarios. Continual policy adaptation through lifelong learning or meta-reinforcement learning would enable the system to personalise attention strategies for individual users over extended deployments. Integration of semantic scene parsing and affective state recognition into the perceptual pipeline would enrich the state representation beyond activity and proximity features. Finally, large-scale longitudinal user studies in open-world environments are needed to assess both the generalisability and the social acceptability of the robot's gaze behaviours across demographic groups.

The RLBAM framework, combined with our evidence-based guidelines for habituation application, provides a robust foundation for developing socially aware robots capable of natural attention behaviour in human-robot interaction contexts. More broadly, our findings contribute to the emerging understanding of bio-inspired mechanism design in cognitive robotics.

Acknowledgments

The authors would like to thank the University of Hertfordshire for providing the research facilities and institutional support that enabled this work. The authors acknowledge the use of Claude for AI-assisted writing support to improve language clarity and structural organisation. This assistance did not include data generation, analysis, or interpretation of scientific findings. All content was reviewed and approved by the authors.

VIII. Conflict of Interest

The authors declare no conflict of interest. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] S. Rasouli, G. Gupta, E. Nilsen, and K. Dautenhahn, "Potential applications of social robots in robot-assisted interventions for social anxiety," *International Journal of Social Robotics*, vol. 14, no. 5, pp. 1–32, 2022.

[2] H. Woo, G. K. LeTendre, T. Pham-Shouse, and Y. Xiong, "The use of social robots in classrooms: A review of field-based studies," *Educational Research Review*, vol. 33, p. 100388, 2021.

[3] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 157–168, 2014.

[4] A. Zaraki, M. Pieroni, D. De Rossi, D. Mazzei, R. Garofalo, L. Cominelli, and M. B. Dehkordi, "Design and evaluation of a unique social perception system for human–robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 341–355, 2016.

[5] M. E. Foster, A. Gaschler, and M. Giuliani, "Group dynamics and role assignment in multi-party conversational robots," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 319–340, 2019.

[6] W. Chen, A. Martinez, and J. Thompson, "Social attention mechanisms in multi-party human-robot interaction: A comprehensive evaluation," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2187–2203, 2022.

[7] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[8] L. Yang, R. Kumar, and K. Petersen, "A survey of attention mechanisms in social robotics: From biological inspiration to artificial implementation," *Robotics and Autonomous Systems*, vol. 162, pp. 104–127, 2023.

[9] N. Mavridis, P. Singh, and M. Zhou, "Computational complexity challenges in multiparty human-robot interaction," in *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8932–8939.

[10] Y. Liu, R. Patel, and S. O'Brien, "Temporal dynamics in robotic attention: Modelling sequential attention patterns for hri," *Frontiers in Robotics and AI*, vol. 10, p. 1123456, 2023.

[11] H. Zhang, C. Rodriguez, and S.-J. Kim, "Adaptive gaze control in dynamic social environments: Beyond pre-programmed responses," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 567–582, 2022.

[12] M. E. Hasselmo, C. Stern, and K. Wagner, "Neuromorphic computing approaches to bio-inspired attention in autonomous systems," *Nature Reviews Neuroscience*, vol. 24, no. 7, pp. 423–439, 2023.

[13] Z. Bing, C. Meschede, K. Huang, and A. C. Knoll, "Bio-inspired attention mechanisms for robotic perception: A systematic review," *Neural Networks*, vol. 156, pp. 1–18, 2022.

[14] K. Doya, K. Samejima, and J. Morimoto, "Computational neuroscience approaches to neuromodulation in artificial agents," *Current Opinion in Neurobiology*, vol. 73, pp. 102–115, 2022.

[15] K. O. Stanley, R. Miikkulainen, and J. Clune, "Habituation mechanisms in machine learning: From biological inspiration to computational implementation," *Trends in Cognitive Sciences*, vol. 27, no. 4, pp. 345–358, 2023.

[16] J. Kober, D. Bagnell, and J. Peters, "Reinforcement learning in robotics: Current challenges and future directions," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 273–298, 2023.

[17] K. Ghamati, F. Amirabdollahian, D. Resende Faria, and A. Zaraki, "Cognitive agentic ai: Probabilistic novelty detection for continual adaptation in hri," in *IEEE RO-MAN 2025*. United States: Institute of Electrical and Electronics Engineers (IEEE), Jun. 2025, pp. 1–8.

[18] F. Codevilla, O. Miksik, and V. Vineet, "Offline reinforcement learning for social robotics: Challenges and opportunities," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, 2022, pp. 4823–4830.

[19] B. Hayes, J. Shah, and A. Thomaz, "Multi-objective optimization in social robotics: Balancing competing goals in human-robot interaction," *Autonomous Robots*, vol. 46, no. 8, pp. 1089–1106, 2022.

[20] C. Balkenius, "Attention, habituation and conditioning: Toward a computational model," *Cognitive Science Quarterly*, vol. 1, no. 2, pp. 171–214, 2000.

[21] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[22] A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press, 2015.

[23] M. H. Johnson, "Subcortical face processing," *Nature Reviews Neuroscience*, vol. 6, pp. 766–774, 2005.

[24] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939–973, 2001.

[25] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in Neurorobotics*, vol. 1, p. 6, 2007.

[26] C. Moore and P. J. Dunham, Eds., *Joint Attention: Its Origins and Role in Development*. Hillsdale, NJ: Psychology Press, 1995.

[27] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human-robot interaction," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 157–164.

[28] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.

[29] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, pp. 716–737, 2008.

[30] A. Palazzi, D. Abati, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: The dr(eye)ve project," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2941–2950.

[31] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2204–2212.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[33] K. Ghamati, A. Zaraki *et al.*, "Deep reinforcement learning for adaptive social gaze in humanoid robots," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2024.

[34] A. Dissegna, M. Turatto, and C. Chiandetti, "Context-specific habituation: A review," *Animals*, vol. 11, no. 6, p. 1767, 2021.

[35] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[36] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 331–338.

[37] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.

[38] I. Nvidia, "Nvidia isaacsim," 2021, accessed: 2024-01-08. https://developer.nvidia.com/isaac-sim.

Khashayar Ghamati is a PhD researcher in Computer Science at the University of Hertfordshire in the UK, specialising in adaptive AI agents and Human–Robot Interaction. His research focuses on enabling intelligent agents to learn continuously and adapt to dynamic and uncertain environments, with the aim of improving collaboration between humans and autonomous systems. His work particularly explores agent adaptation, interactive learning, and the development of robust and explainable intelligent behaviour. His research combines methods from machine learning, robotics, and cognitive systems to advance adaptive autonomy in embodied agents.

Maryam Banitalebi Dehkordi Maryam is a Senior Lecturer in Robotics and AI at the University of Hertfordshire's School of Physics, Engineering and Computer Science in the UK. She received her PhD in Perceptual Robotics – Innovative Technologies from the Scuola Superiore Sant'Anna of Pisa, Italy, in 2014. She has a broad range of experience in robotics across both academia and industry in different countries, including Malaysia, Italy, Germany, and the UK. She has contributed to major projects, such as DOC (Dispositivo di Orientamento Ciechi) in Italy and the Innovate UK-funded AgriRobot project in the UK. Her research interests include Human–Robot Interaction, Explainable Robotics, Assistive Robotics, and the development of autonomous systems for robotic platforms.

Hamed Rahimi Nohooji is a Postdoctoral Research Associate with the Automation and Robotics Research Group at the University of Luxembourg. He received his PhD in Control and Robotics from Curtin University, Australia, in 2018. He has held research positions at the National University of Singapore (NUS), UCLouvain, the University of Pisa, and the University of Birmingham. He has contributed to several international projects, including the EU H2020 CYBERLEGs Plus Plus project, the A*STAR Singapore project on soft modular adaptive grippers, and the FNR-funded COSAMOS and ANR–FNR DOMINANTS projects on soft aerial manipulators. He has served as a Guest Editor for Robotics, Actuators, Sensors, and the International Journal of Advanced Robotic Systems. His research interests include nonlinear and constrained control for robotic systems, soft robotics, and performance-driven design optimization.

Holger Voos received a Ph.D. degree in automatic control from the Technical University of Kaiserslautern, Germany, in 2002. From 2004 to 2010, he was a Professor at the University of Applied Sciences Ravensburg Weingarten, Germany. Since 2010, he has been a full professor at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, and the head of the Automation and Robotics Research Group. He is the author or coauthor of more than 300 publications, comprising books, book chapters, journals, and conference papers. His research interests include automation and robotics, soft robotics, and situational awareness.

Farshid Amirabdollahian is a leading academic in Human-Robot Interaction at the University of Hertfordshire's School of Physics, Engineering and Computer Science in the UK. He serves as Director of the Robotics Research Group and leads initiatives such as Robot House and the humanoid Kaspar project. His research focuses on adaptive robotics, assistive and rehabilitation technologies, and machine learning approaches to improve how robots interact with people. Prof Amirabdollahian has a long academic career with previous research and leadership roles in UK research centres and collaborative EU projects.

Abolfazl Zaraki is a Senior Lecturer in Artificial Intelligence and Robotics with the School of Physics, Engineering and Computer Science, University of Hertfordshire, U.K., and a member of the Robotics Research Group. He received the Ph.D. degree in Automatic Robotics and Bioengineering from the University of Pisa, Italy, in 2014. He has held Research Fellow and Senior Research Fellow positions at leading institutions in Italy and the U.K., contributing to several major European and national research projects, including EASEL (Expressive Agents for Symbiotic Education and Learning), BabyRobot (Next-Generation Social Robotics), JAMES (Joint Action for Multimodal Embodied Social Systems), and the Innovate UK–funded InSight project on snake robot solutions for industrial inspection. He has co-authored over 50 peer-reviewed journal articles, book chapters, and conference publications. His research focuses on cognitive robotics and adaptive autonomous systems for real-world applications.