IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 28 October 2024, accepted 20 November 2024, date of publication 2 December 2024, date of current version 10 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3508139

RESEARCH ARTICLE

Deep Reinforcement Learning for the Biologically Inspired Social Behaviour of Autonomous Robots Acting in Dynamic Environments

MARCOS MAROTO-GÓMEZ, MARÍA MALFAZ, ÁLVARO CASTRO-GONZÁLEZ, SOFÍA ÁLVAREZ ARIAS, AND MIGUEL ÁNGEL SALICHS, (Life Member, IEEE)

Systems Engineering and Automation Department, Carlos III University of Madrid, 28911 Leganés, Spain

Corresponding author: Marcos Maroto-Gómez (marmarot@ing.uc3m.es)

This work was supported in part by Robots sociales para mitigar la soledad y el aislamiento en mayores (SOROLI), funded by Spanishc State Research Agency [Agencia Estatal de Investigación (AEI)], Spanish Ministerio de Ciencia e Innovación, under Grant PID2021-123941OA-I00; in part by Robots sociales para reducir la brecha digital de las personas mayores (SoRoGap), funded by AEI, Spanish Ministerio de Ciencia e Innovación, under Grant TED2021-132079B-I00; in part by the Promoting research and development and innovation (R&D&I), funded by Spanish Ministry of Science and Innovation [Ministerio de Ciencia e Innovación (MCIN)]/AEI/10.13039/501100011033, under Project PDC2022-133518-I00; and in part by the European Union NextGeneration European Union (EU)/Recovery, Transformation and Resilience Plan [Plan de Recuperación, Transformación y Resiliencia (PRTR)].

ABSTRACT Robots are increasingly operating in highly complex and dynamic scenarios where they must continuously perceive their environment, learn from new experiences, and apply acquired knowledge to complete their tasks effectively. In these environments, the potential situations a robot encounters can become too vast to handle with predefined conditions. As a result, autonomous robots must incorporate learning methods that accurately represent the environment, make informed decisions, and optimize learning speed, task performance, and computational resources. Given the recent advancements of Deep Reinforcement Learning over classical Reinforcement Learning, this paper presents a Deep Reinforcement Learning system for biologically inspired, socially-driven decision-making in autonomous robots operating in such intricate environments with countless variations. This work formulates a learning framework as a Markov Decision Process, enabling robots to demonstrate adaptive social behaviour by integrating internal and external factors. The robot's state includes 11 variables derived from the robot's motivations, user perception, ambient light, and social norms, allowing the robot to select from ten possible actions autonomously. This study aims to develop fully autonomous robots that operate autonomously, learning and adapting to complex environments while maintaining an optimal balance between the robot's internal and social well-being. We compare eight state-of-the-art DRL algorithms to identify the best-performing approach and implement the learning system into our Mini social robot. The results highlight Rainbow as the most effective solution, enabling the Mini robot to exhibit highly adaptive, autonomous behaviour in challenging social environments. These results allow autonomous robots to increase their capabilities and reduce human supervision.

INDEX TERMS Autonomous decision-making, bio-inspired model, deep reinforcement learning, social robots.

I. INTRODUCTION

Autonomous robots with social skills have significantly improved in recent decades [1]. These robots find applications in various domains such as healthcare, education, entertainment, assistance, and companionship, where they assist humans and alleviate their workload [2]. This work arises from the growing demand for specialized workforce in specific sectors, prompting the interest in autonomous, socially adept robots capable of reducing human intervention [3]. Moreover, in scenarios involving Human-Robot Interaction (HRI), where robots interact with different individuals, it becomes crucial to equip the robot with adaptive behaviour to overcome user interaction challenges

The associate editor coordinating the review of this manuscript and approving it for publication was Poki Chen.

and potential limitations [4]. In HRI settings, robots often encounter hesitant individuals who may harbour mistrust or perceive them as a threat. There is a growing interest in modelling human biological functions within robots, enabling these systems to exhibit more natural and human-like behaviour that people can recognize and trust [5]. However, robots operating in the real-world face significant challenges in achieving optimal learning due to the many situations they must consider.

This paper presents an innovative and biologically inspired learning system designed for autonomous robots that operate as human assistants in dynamic scenarios. The main contribution of this research is combining a Deep Reinforcement Learning (DRL) model, stated as a Markov Decision Process (MDP) with a bio-system to enable robots to learn and exhibit autonomous behaviour in scenarios where the robot's state-action space is too large. To comprehensively represent the robot's state and train the agent, we introduce a biologically inspired environment built upon Open AI's Gymnasium. This environment captures the robot's state by incorporating artificial biological processes that simulate internal deficits like sleep or affection, the stimuli perceived from the surrounding environment, and social norms to ensure the robot adheres to human conventions.

The learning system is integrated into Mini [6], a social robot that assists individuals in diverse tasks such as cognitive stimulation therapies, companionship and entertainment. Through an extensive literature review in the subsequent section, we highlight the existing gap in biologically inspired methods that facilitate autonomous decision-making for social robots operating in large social environments over extended periods. Consequently, our DRL model aims to fill this void by endowing robots with biologically inspired learning capabilities, enabling them to consider numerous states and actions that enhance their skill set. By employing this approach, we intend to equip autonomous robots with the ability to learn from their environment, make independent decisions, and effectively adapt to a wide range of complex scenarios. This advancement paves the way for robots that autonomously can assist and interact with humans in unpredictable environments, pushing the boundaries of what social robots can achieve.

We previously explored the use of classical Reinforcement Learning (RL) to obtain autonomous decision-making in social robots. First, we used Q-learning [7] to maintain the physiological state of a social robot in good condition while interacting with a user. Later, we used Dyna-Q+ [8] as a model-based alternative to increase the environment complexity and extend our robot's state-action space. However, these approaches did not provide the expected outcomes in large state-action spaces.

Q-learning required more than 5 hours to learn how to behave using three discrete variables as input and three actions. Dyna-Q+ required almost the same amount of time in an environment with four discrete variables and eight actions. In a large environment like the one proposed in this manuscript, where seven continuous and four discrete variables are used with ten actions, these approaches would produce intractable training times. Some previous papers [9], [10], [11] compare classical RL methods with DRL methods in video games like Atari, with DRL surpassing classical performances. For this reason, this work overcomes our previous limitations and motivates the use of new variables to improve the robot's actuation skills.

The literature provides many DRL algorithms that produce positive learning results in different environments. This work compares those with better performance in the last years to find the best for our learning environment. We selected them due to their availability in Python libraries, their easy integration into our system, and their feasible computational requirements for our Mini robot. We compare the results provided by Deep Q-Network (DQN) [12], Double Deep Q-Network (DDQN) [13], Dueling Deep Q-Network (DuelDQN) [14], Double Dueling Deep Q-Network (DuelDDQN) [15], Soft Actor-Critic (SAC) [16], Proximal Policy Optimization (PPO) [17], Categorical DQN (C51) [18], and Rainbow [19]. We embed the best algorithm into the Mini social robot to analyze whether it correctly learns to survive in a challenging world by optimising its physiological and social well-being.

This paper continues in Section II with an analysis of similar papers that use DRL in social robotics and autonomous decision-making. Section III formulates the problem as an MDP and presents the algorithms we have compared to optimize our environment's learning. Then, Section IV presents the methodology of the paper and describes the Mini robot as the device used in this work. Section V shows the results we have obtained during learning and how Mini maintains its physiological and social well-being after learning. Next, Section VI discusses this work's primary results and Section VII enumerates its limitations. Finally, Section VIII closes this manuscript with the findings and future work that we will address to continue this research line.

II. DEEP REINFORCEMENT LEARNING IN AUTONOMOUS ROBOTS

Social robots have emerged as significant societal players, offering valuable assistance across various applications [20], [21], [22], [23]. With increasingly powerful sensors and actuators, these machines are becoming adept at perceiving and interacting with their environment. However, this advancement comes with a challenge - the growing number of variables that robots must consider when making decisions [24, p. 9–10]. DRL models have garnered considerable attention for tackling huge state-action spaces efficiently [24, p. 18].

While DRL has seen widespread application in various fields, its potential in autonomous social robots remains relatively unexplored. Most studies have focused on leveraging DRL for human-robot interaction (HRI), social navigation, and autonomous decision-making, aiming to create engaging, lifelike behaviours while accomplishing specific tasks. For instance, Qureshi et al. [25] employed a Multi-modal Deep Q-network (MDQN) to teach the Pepper social robot social skills using rich sensory information. After interacting with different users over 14 days, their results showcased the robot's ability to make social decisions in HRI autonomously. Building upon this work, the same authors [26] developed a DRL predictive model that enabled a social robot to learn human-like social skills. By considering its own and the user's states, the robot autonomously executed actions like handshaking, effectively guiding the interaction. Another notable study by Hong et al. [27] introduced an emotion-based architecture for producing autonomous affective behaviour, enhancing the robot's usability and facilitating the interaction.

In a more specific application, Lathuili'ere et al. [28] utilised DRL to teach a social robot gaze behaviour during social interactions. Integrated into the NAO social robot, the system leveraged visual information to detect user positions in multi-user domains and autonomously directed its gaze during interactions. Similarly, Gao et al. [29] employed DRL techniques to train the Pepper social robot in HRI for approaching behaviour towards small groups of people. The robot made appropriate decisions by utilising visual information while performing social-approaching behaviour. Expanding on this research, Cuay'ahuitl [30] developed a DRL-based dialogue system that supports multiple languages. Subsequently, this author [31] combined this system with other HRI skills to imbue a humanoid robot with multi-modal playing behaviour. By leveraging visual information, the robot trained a DRL framework to make optimal decisions.

Focusing on architectures employing DRL for social navigation, Chen et al. [32] proposed a social navigation system for crowded environments. Utilising meaningful sensor data, their model dynamically generated optimal paths that improved over time through experience. Additionally, the robot incorporated an attention system to extract features of pairwise interactions, enhancing decision-making. Similarly, Liu et al. [33] tackled social navigation in crowded scenarios by combining DRL with imitation learning. This novel approach mimicked human movements to improve social navigation in overcrowded environments. Following this research direction, Samsani and Muhammad [34] introduced a socially compliant robot navigation system for crowded spaces, leveraging DRL to predict people's movements and avoid collisions, producing smooth trajectories.

Researchers have made significant strides in designing learning systems that effectively regulate pedestrians' paths in crowded environments [35], [36], [37]. These innovative models leverage the power of DRL to tackle the intricate challenges posed by motion dynamics planning, which are derived from visual information. By harnessing the capabilities of DRL, these learning systems enable autonomous robots to navigate crowded spaces with finesse. They address the complexities of pedestrian movement, allowing robots to make informed decisions in real time. This approach leverages visual data to understand the dynamic environment and compute optimal paths, ensuring efficient and safe pedestrian interactions.

Introducing groundbreaking advancements in robotics, our comprehensive review sheds light on a significant gap in current research. The utilisation of DRL still needs to be improved in the context of autonomous and social robots for Human-Robot Interaction (HRI), social navigation, and autonomous decision-making. While some studies have focused on developing specific skills within controlled environments, the translation of these abilities into real, unpredictable settings still needs to be explored. Our research offers a compelling argument for integrating DRL into the autonomous decision-making processes of social robots. By imbuing these machines with biologically inspired behaviour, we unlock a new paradigm that champions their functional competence and ability to navigate complex social dynamics and foster positive human interactions.

III. BACKGROUND

DRL [38] is a Machine Learning method that combines Deep Learning [39] with Reinforcement Learning [40]. It addresses the problem of an agent learning to make decisions by using neural network predictions and observing the effects of its actions on the environment. Unlike classical RL approaches, DRL deals with significant large inputs, optimising a loss function $\mathcal{L}$ that works with a reward function representing the benefits of an action in a particular situation.

DRL problems are typically formulated as Markov Decision Processes (MDP) [41]. In an MDP problem, the learning task considers an agent interacting with a dynamic environment in discrete time steps t. The robot has to learn which action a from a predefined set $\mathcal{A} = \{a_1, \ldots, a_m\}$ fits better each state $s = \{s_1, s_2, \ldots, s_n\} \in \mathcal{S}$ (where s_i is each of the inputs that define the robot's state). The quality of executing a in s is represented by numerical reward r_t observed by the agent. Executing action a in state s leads the agent to a new state s' in the following time step $t + 1$. The transition from s to s' is not deterministic but depends on a transition probability $P(s, s') = Pr(s_{t+1} = s'|s_t = s, a_t = a)$ whose values are obtained from the weighted sum of past rewards the agent experienced in the same previous situation.

The goal of an MDP is to find a good behaviour policy π that enables maximising the accumulated discounted reward $R_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau$ during the life of the agent. To accomplish this, the agent has to explore the environment to build a behaviour policy based on experience. In this notation, the discount rate γ balances the importance of past and recent rewards. γ values close to 1 prioritize new experiences over past experiences, a situation that we consider in this work setting $\gamma = 0.99$ for all algorithms.

In DRL, the learning process relies on a neural network that aims to minimize a loss function $\mathcal{L}$ that translates numerical rewards into a gradient optimization problem. The loss function is generally defined by

$$\mathcal{L}(\theta) = \mathbb{E}_{s,a,r,s'}\left[(Q(s, a, \theta_i) - Y_i)^2\right], \tag{1}$$

where $\mathbb{E}[\cdot]$ represents the expectation, Y_i is the target value to update, and θ_i are a set of coefficients or weights that characterize the network layers.

Table 1 summarises the 8 DRL algorithms compared in this work and presents their main features and shortcomings according to the MDP problem we want to solve: maximising the robot's decision-making processes during extended periods in dynamic social environments. We selected these algorithms due to three reasons: their availability in Python libraries, their easy integration into our system, and their affordable computational resources.

A. HYPOTHESES

The formulation and background of the previous methods led us to evaluate a couple of hypotheses about the learning process in our learning scenarios. These hypotheses are:

1) Rainbow should be the algorithm with better performance and learning speed at the cost of increasing the computational complexity since it includes most of the advances presented in the other methods.
2) PPO and SAC should produce a good performance due to how they update the target function but can incur errors at the beginning of learning because of their exploration. Besides, PPO must be studied in detail because the sparse reward of our setting could be challenging for this algorithm.
3) Dueling approaches should perform well but occasionally fail when approximating the optimal solution. However, they require more computational resources than the other methods.
4) DQN, DDQN, and C51 should produce more stable learning, but the maximization of the reward should not be as good as in the other cases. Besides, C51 can find the environment challenging since sparse rewards lead to slow convergence.

IV. ENVIRONMENT

This section presents the Mini robot, where our DRL model has been integrated to produce autonomous and adaptive decision-making during long-lasting interactions. Besides, the section defines how the robot's and the environment processes evolve, the learning setup, and its evaluation.

A. MINI SOCIAL ROBOT

Mini [6] is a desktop robot used for entertainment and cognitive stimulation activities. Mini has a tablet device for displaying multimedia content and obtaining user selections using menus to perform its activities. The robot, shown in Figure 1, can verbally communicate with people using a microphone and a speaker. Mini has an Intel NUC 11 Pro computer with an Intel Core i7-1165G7 processor –8 GB of RAM and 1 TB of storage– which limits the execution of novel DRL algorithms on real-time. Moreover, it has three touch sensors in the belly and shoulders to sense tactile contact and a 3D camera to perceive the surroundings visually. Regarding its expressive functions, the robot has five degrees of freedom in the hip, arms, neck, and head, two animated screens that emulate its eyes, and four luminous devices in the heart, cheeks, and mouth.

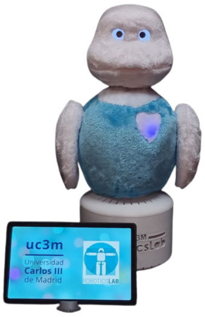

FIGURE 1. Mini social robot.

Since Mini was conceived to interact and assist people in the long run, reducing human intervention, it has to exhibit autonomous behaviour, identify the users' needs and produce a personalized and adapted interaction. Besides, to engage users in the interaction and increase their acceptability of the robot's actions, its behaviour has to be biologically inspired by living beings to seem natural and appropriate to the social norms generally accepted in social scenarios. Considering these ideas, we define the robot's state as a combination of artificial biological processes that evolve with time, external stimuli that the robot perceives, and social norms to satisfy the user's demands. As described in the following sections, Mini's goal is to select the action that produces the most positive effects on each situation, maximising its physiological and social well-being. Next, we define the robot's state and actions to learn an optimal behaviour policy using DRL to generate biologically inspired autonomous and adaptive behaviour in social contexts.

B. MINI'S STATE

The robot's state represents its internal, external, and social situation. It is defined by its artificial biological processes, the state and intensity of the stimuli the robot perceives, and the social events around it. We define Mini's state as the combination of seven internal motivations, the perception of the user's presence and light intensity, and social events that Mini has to communicate to the user. Next, we define how Mini's state changes due to the environment dynamics.

1) INTERNAL STATE: PHYSIOLOGY AND MOTIVATION

The physiological state in Mini consists of artificial biologically inspired processes that evolve with time. These processes simulate living beings' functions, intending to endow

TABLE 1. Comparison of the 8 DRL algorithms studied in this research.

Algorithm	Advantages	Shortcomings
Deep Q-Network (DQN) [12]	Utilizes Deep Learning to approximate Q-values, enabling effective learning in high-dimensional state spaces. Experience replay improves sample efficiency, and the target network stabilizes learning.	Prone to overestimate Q-values, leading to suboptimal policies. Struggles with sparse rewards may need more stability in complex or dynamic environments.
Double Deep Q-Network (DDQN) [13]	Mitigates the overestimation problem inherent in DQN by using separate networks for action selection and evaluation, leading to more reliable policy updates.	Still sensitive to hyperparameters and can be computationally demanding. May exhibit overestimation in highly sparse reward scenarios or environments with large rewards.
Dueling Deep Q-Network (DuelDQN) [14]	Separates the estimation of state values and action advantages, enhancing learning efficiency in environments where some actions are irrelevant. Improves convergence speed in many tasks.	Introduces greater computational overhead due to separate value and advantage networks. Requires careful tuning of additional hyperparameters. Performance may degrade if value and advantage components do not align well.
Dueling Double Q-Network (DuelDDQN) [15]	Combines the benefits of DDQN and DuelDQN, offering reduced overestimation and more efficient learning by focusing on relevant actions. Performs well in complex environments.	Computationally expensive due to the combination of dual Q-learning and duelling architecture. Increased complexity and slower convergence in certain environments.
Soft Actor-Critic (SAC) [16]	Achieves a balance between exploration and exploitation via entropy regularization. Efficient in continuous and discrete action spaces using the Gumbel-Softmax trick. Provides robust performance in many tasks.	High computational complexity due to the additional entropy term. They may struggle in environments with sparse or delayed rewards. Training can be sensitive to hyperparameter settings.
Proximal Policy Optimization (PPO) [17]	Ensures stable policy updates by limiting the update step size with a clipped objective, enhancing performance in complex environments. Simple to implement with good scalability.	Can fail in environments with sparse or delayed rewards. Requires careful tuning of the clipping parameter and other hyperparameters for optimal performance.
Categorical Deep Q-Network (C51) [18]	Extends DQN by estimating the full distribution of returns rather than a single expected value. Reduces overestimation bias and improves robustness to noise and stochasticity.	Assumes a stationary reward distribution, which can slow convergence in dynamic environments. It may be less effective in tasks with significant reward variation or non-stationarity.
Rainbow [19]	Combines multiple advancements (e.g., double Q-learning, duelling networks, distributional RL, and prioritized replay), resulting in superior performance across diverse environments.	High computational and memory requirements due to the integration of various techniques. Increased complexity makes it harder to tune and analyze the contribution of individual components.

the robot with natural and lively behaviour, motivating the robot to behave in a particular manner. In our environment, Mini has seven biological processes that evolve with time. These processes are:

- **Sleep:** Represents the evolution of the robot's sleep. It prevents the robot from being continuously active, especially at night, simulating its sleep.
- **Social entertainment:** This variable simulates Mini's desire to entertain the user. It promotes interaction with people to complete different activities.
- **Self-entertainment:** Related to the robot's need to entertain alone. It leads the robot to show lively and expressive behaviour, avoiding a continuous inactivity when the user is absent for long periods.
- **Cognitive interaction:** Defines the robot's need to receive positive cognitive social interaction. It is used to foster verbal communication with the user.
- **Physical interaction:** Variable that simulates the robot's need to receive positive physical contact. It is used to generate a bond with the robot.
- **Stress:** This process emulates the stress levels that appear in specific situations. It is used to avoid undesired situations with the user.
- **Energy:** Variable whose value decreases with time and the execution of activities. It regulates the robot's activity during the day and helps it stay calm after continuous interactions.

Biological processes evolve with time following a different rhythm (called variation –*vr*–) from the limits 0 to 100 units, as Table 2 shows. The natural variation of these variables makes their current value $cv_i(t)$ in time step t deviate from their ideal value (iv_i) at different rates. This difference causes a deficit $d_i(t)$ in Mini's internal state, which is computed using Equation 2. This equation takes inspiration from biologically inspired models for artificial agents [7], [42].

$$d_i(t) = |iv_i - cv_i(t)| \tag{2}$$

where

$$cv_i(t) = cv_i(t-1) + vr \tag{3}$$

The variation rates (vr) of physiological processes have been empirically set to obtain a specific robot behaviour. Depending on the urgency of the process, we define three different rates: high (±0.3), moderate (±0.2), and low (±0.1). In this approach, those variables related to social and self-entertainment have been defined as those with the highest urgency to motivate the robot to play faster. Processes

TABLE 2. Biological processes defining the internal state of the Mini robot. Biological processes evolve with time from 0 to 100, following specific variation rates. When they deviate from their ideal value, a deficit appears in the agent. Deficits and the stimuli Mini perceives influence motivation, urging behaviour.

Biological process	Range	Variation –vr–	Ideal value	Related motivation	Stimulus modulating the motivation	Modulation α_i
Sleep	0 to 100	$+0.2$	0	Sleep	Not light	2.5×10^{-3}
Social entertainment	0 to 100	-0.3	100	Play	User	2.5×10^{-3}
Self-entertainment	0 to 100	-0.3	100	Play alone	Not user	2.5×10^{-3}
Cognitive interaction	0 to 100	-0.2	100	Socialize	User	2.5×10^{-3}
Physical interaction	0 to 100	-0.2	100	Affect	User	2.5×10^{-3}
Stress	0 to 100	$+0.1$	0	Relax	None	None
Energy	0 to 100	-0.2	100	Rest	None	None

related to cognitive and physical interaction have a moderate variation rate to be less urgent than entertainment. Similarly, variables related to relaxation (sleep and energy) follow a moderate rhythm to avoid a continuous interaction. Finally, stress has the lowest variation rate since we consider this variable less urgent than the others.

The deficits in artificial biological processes can only be reduced by executing specific actions. Therefore, we define motivations as processes that urge behaviour to maintain an optimal internal state. Drawing on Lorenz's motivational theory [43], human motivation arises from internal deficits. Additionally, the stimuli we perceive from the environment are often important behavioural elicitors, increasing our motivation to execute action under particular situations (e.g., eating a delicious meal even if our hunger deficit is not high). As we describe in the following Section IV-B2, the user and the light intensity are the two stimuli Mini can perceive and boost motivational intensities. Using these ideas and the Equation 4 proposed by Ávila and Cañamero [44], the intensity of a motivation $m_i(t)$ in time step t is affected by internal biological deficits $d_i(t)$ that are modulated α_i by the intensity of the stimuli $s_i(t)$ we perceive in time step t.

$$m_i(t) = d_i(t) + \alpha_i \times d_i(t) \times s_i(t) \tag{4}$$

Mini has seven motivations, each dependent on an internal physiological deficit. Moreover, some of them are modulated by specific stimuli that it perceives from the environment. In our model, all motivations range from 0 to 100 units and evolve following the rhythms of their attached biological deficit and the intensity with which the robot perceives stimuli. The motivation with the highest intensity (dominant motivation) considers the stimuli the robot perceives from the environment.

Table 2 shows the relationship between motivations and biological processes. Moreover, it contains the stimuli that boost each motivation and the modulation factor α_i. Motivational intensities define the robot's internal state, being 7 of the 11 inputs of the learning system.

- **Motivation to Sleep:** The robot needs to sleep to regulate its biological sleep process. This motivation increases with low light levels.
- **Motivation to Play:** The robot needs to play with the user to reduce its deficits. This motivation increases when Mini perceives the user's presence.
- **Motivation to Play alone:** Mini has to play alone to reduce the self-entertainment deficit. This motivation increases when Mini does not detect the user's presence to promote interacting with the user.
- **Motivation to Socialize:** This motivation becomes active to reduce the cognitive interaction deficit. Perceiving the user amplifies the intensity of this motivation.
- **Motivation of Affect:** This motivation represents the robot's need to reduce its physical interaction deficit. Perceiving the user amplifies its intensity.
- **Motivation to Relax:** This motivation becomes active when the robot is stressed due to a threatening situation.
- **Motivation to Rest:** Becomes active if the robot's energy levels are low.

2) EXTERNAL STATE: STIMULI

The stimuli Mini perceives are essential for its decision-making. Consequently, they define the external component of its state. In this work, the influence of the stimuli on the robot depends on the perception intensity $s_i(t)$ from 0 to 100 units obtained in each time step t. This intensity affects motivation and behaviour selection. In our scenario, Mini can perceive two stimuli, whose features are summarized in Table 3.

- **Light:** light intensity perceived from the environment. The value is given by a photoelectric receptor placed in the robot's head. The light state can be dark if the intensity exceeds 5 units or lighted otherwise. Dark light allows the robot to sleep.
- **User:** person that interacts with the robot face-to-face. The information on whether the user is present or absent is provided by a 3D camera that performs face detection to faces in the scene. The intensity of this stimulus increases in 1 unit each second the robot detects the user and decreases at the same rate when not perceived (limited from 0 to 100). The possible states for this stimulus are user absent or present. The user presence enables Mini to execute some actions, such as playing.

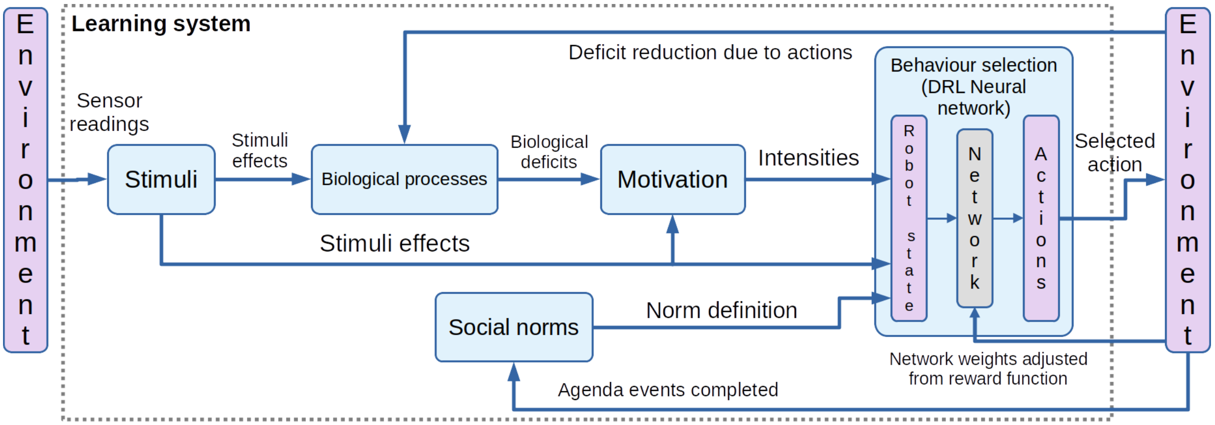

FIGURE 2. The environment designed for social robots learning to optimize its physiological and social well-being in HRI.

TABLE 3. Stimuli that Mini can perceive, their possible states, and their intensity range.

Stimulus	Possible states	Initial value	Intensity range
Light	Dark if intensity < 5 Lighted if intensity > 5	0	0 to 100
User	Absent if intensity $= 0$ Present if intensity > 0	0	0 to 100

3) SOCIAL STATE: SOCIAL NORMS

While making autonomous decisions, the robot has to fulfil the user demands and follow social norms to deploy a natural and socially interactive behaviour. Consequently, we have defined two discrete variables related to the users' agenda that define the users's social events, which we call social state. In this work, Mini considers the following variables in its social state:

- **Agenda ready:** This variable represents the need to execute specific activities or reminders to the user stored in the user agenda. The possible values for this variable are none, activity, or reminder.
- **Agenda priority:** This variable represents the priority with which the activity or reminder specified in the previous variable has to be executed. The possible values are none, low, moderate, or high.

As a result, the robot's state is a combination of its internal state (intensity of the seven motivations), external state (state of the two stimuli perceived from the environment), and social state (user social events). Mini's state consists of seven continuous variables defining the intensities of the motivations, two discrete variables for the two stimuli the robot can perceive, and two discrete variables representing the users' events stored in its agenda and their priority. Figure 2 shows the relationships between the processes involved in the biological model to maximize the robot's physiological and social well-being.

C. MINI'S ACTIONS

Mini can execute ten actions to show different skills, change the state of some stimuli, and reduce the deficits in biological processes. Next, we enumerate the robot's actions, their functionality, and whether the user has to be present to succeed in their execution.

- **Sleep:** The robot simulates it is sleeping by closing its eyes and performing expressions like snoring. This action reduces the sleep deficit.
- **Wait:** Mini rests waiting for upcoming events without executing a specific activity. This action restores the sleep deficit.
- **Play with the user:** The robot executes an entertainment activity that implies playing with the user. Inside this action are entertainment sub-actions like quiz games, showing photos, or playing music that is selected using a Preference Learning module, as described in [45]. This action reduces the social entertainment deficit but increases energy drop. The user must be present to play.
- **Play alone:** Mini executes playing activities that do not imply interacting with the user, like dancing. This action reduces the self-entertainment deficits but increases energy drop.
- **Talk:** Mini talks with the user about different topics and asks them to retrieve information that can be later used for improving the interaction. This action reduces the robot's cognitive social need but subtly increases energy drop. The user must be present to talk.
- **Meditate:** The robot closes its eyes and meditates to reduce stress levels. This action reduces the robot's stress.
- **Request affect:** The robot requests the user to provide affect by stroking its belly. If affection is provided, the physical social need is reduced. The user must be present to request affect. This action has a 70% chance of succeeding in the virtual environment.
- **Call the user:** This action attempts to bring a user to interact with the robot to enable the execution of some actions. It has a 50% chance of succeeding in the virtual environment. The robot must learn that this action enables playing or talking, which are essential for survival.

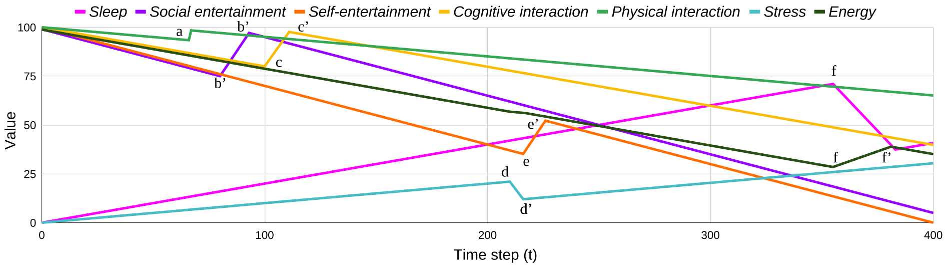

FIGURE 3. Evolution of the biological processes emulated in Mini and the effects of actions to reduce their deficits. In point a, the robot requests affect and receives a stroke from the user reducing its physiological and social needs. From b–b', the robot plays with the user, reducing its social entertainment deficit. From c–c', the robot talks with the user, reducing its social cognitive need. The robot meditates from d–d', reducing its stress levels. The robot dances from e–e', satisfying its self-entertainment biological process. Finally, Mini is sleeping from f–f', so its energy is restored and the sleep deficit simultaneously reduced.

- **Planned activity:** The robot starts an already planned activity stored in its agenda. After the activity, the agenda ready and agenda priority variables are set to none.
- **Tell reminder:** The robot tells the user a reminder already planned in the agenda. After the reminder, the agenda ready, and agenda priority variables are set to none.

The previous actions affect the environment and the robot's internal state differently. Additionally, some of them imply social effects whose compliance involves a reward for the robot. These effects are translated into the numeric reward for updating the DRL algorithm. At the beginning of training, all actions have the same weight and, therefore, the same probability of being executed. As learning progresses with new experiences gained by the robot, action weights change, increasing those of the actions that provide better benefits in each situation. Table 4 shows the action effects on the biological processes and the state of stimuli and social events. The effects of each action were empirically determined to reduce at least one deficit in an amount higher than the sum of the variation rates of the other processes. Using this method, it is possible to improve the robot's physiological well-being if the correct action is selected. As Table 4 shows, the action with effects that are not always produced (request affect) has a higher effect than the others to compensate for the probability of not applying the effect.

TABLE 4. Actions and their effects on biological processes, stimuli, and social events.

Action	Physiological effects	State changes	User	Light
Sleep	Sleep: -2 Energy: $+0.4$	None	–	Dark
Wait	Energy: $+2.3$	None	–	–
Play with user	Energy: -0.4 Social ent.: $+2.8$	None	Present	–
Dance	Energy: -0.2 Self-ent.: $+2.5$	None	–	–
Talk	Energy: -0.2 Cognitive int.: $+2.6$	None	Present	–
Meditate	Energy: $+0.1$ Stress: -2.2	None	–	–
Request affect	Phys. int.: $+3.2$ (70%)	None	Present	–
Call user	None	User present (50%)	–	–
Planned activity	None	Agenda ready and priority set to none	–	–
Tell reminder	None	Agenda ready and priority set to none	–	–

Figure 3 shows how the biological processes emulated in Mini evolve with time. Each time step t, if no action is executed, these variables deviate from their ideal values, producing a deficit. As we show in the figure with different examples, action execution reduces deficits, restoring the optimal state of the biological values. Each action has different effects on Mini, reducing specific deficits. Consequently, Mini's primary goal is to learn the behaviour policy that maintains them in the best condition to maximize its physiological and social well-being.

D. LEARNING METHODOLOGY

The robot's learning occurs in a virtual environment where the dynamics presented in the previous sections are generated. Learning starts with a robot choosing actions with equal probabilities (equal weights and the robot adjusting these weights with new experiences, increasing the probability of those actions that improve the robot's physiological and social well-being). Once the virtual environment learns an appropriate behaviour policy, we transfer the knowledge (trained model) to the real robot for its autonomous behaviour. The environment was designed as a contribution of this work using Open AI's Gymnasium methodology [46]. The algorithms were trained using TensorFlow, Torch, and Keras Python libraries.

The neural network (see Figure 4) used for learning consisted of an input flattened layer of size 1 × 11 (because 11 variables define the robot's state) connected to some dense hidden layers whose number varies depending on the algorithm. Some algorithms, such as Rainbow or DuelDQN, use a duelling network to estimate the value function (V-network) and a network to select the best action (A-network) in each situation are included between the hidden layers and the output layer, as Figure 4 highlights in orange.

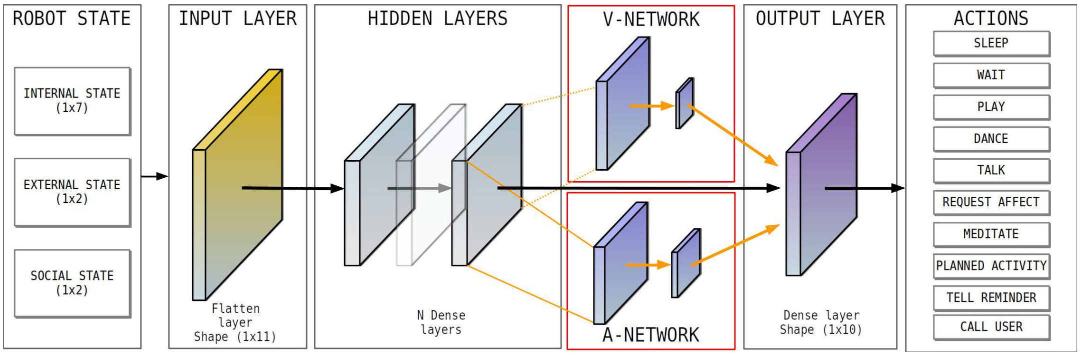

FIGURE 4. Neural network model used in our deep reinforcement learning approach. The straight connection in black represents those architectures not using a dueling type network. Highlighted in orange are the dueling layers used by methods such as DuelDQN or Rainbow.

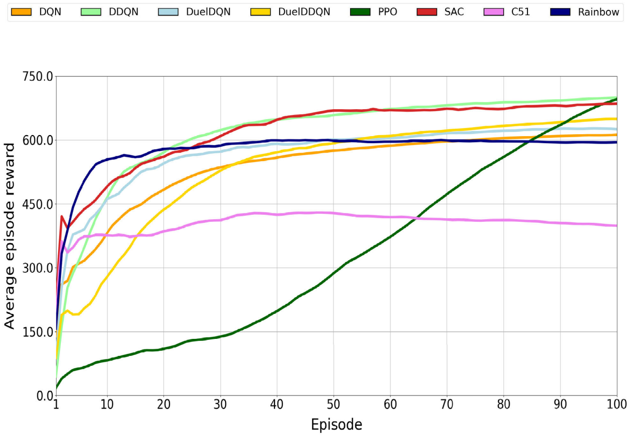

FIGURE 5. Average reward per episode.

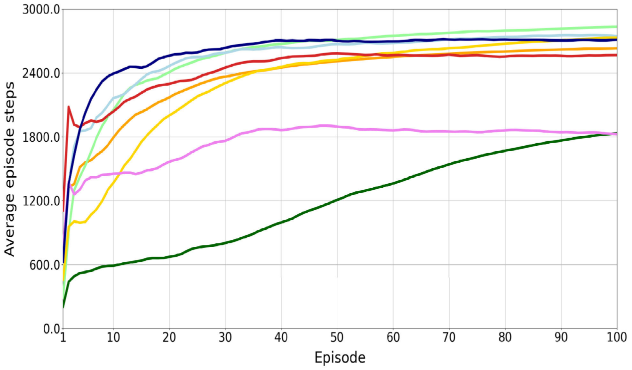

FIGURE 6. Average steps per episode.

We conducted a preliminary study to find the best performance of each algorithm. This preliminary test consisted of running a random search method to combine different hyperparameters in simulation 100 times randomly. The hyperparameter possibilities were obtained from the original algorithms provided in the Python libraries PyTorch, Keras, and TensorFlow. We saved each case's best hyperparameter combination to compare the results running each algorithm with the best hyperparameters ten times. The best hyperparameters are in Table 5.

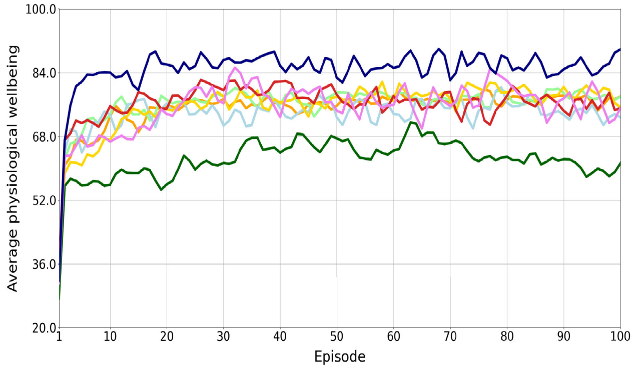

FIGURE 7. Average physiological well-being per episode.

1) REWARD FUNCTION

The reward function $\mathcal{R}$, which is a specific contribution of this work, aims to learn the behaviour policy that maintains the robot's physiological and social well-being (Wb) in the best possible condition. On the one hand, improving physiological well-being comes from executing actions that reduce internal deficits in biological variables. If the higher deficits are correctly reduced, the reward will be higher. On the other hand, social well-being improvement comes from fulfilling social norms related to communicating reminders and executing planned activities ordered by the user. Using this definition, the robot balances its action between a biologically inspired approach to show liveliness and natural behaviour and a social behaviour focused on the user. In this approach, the robot gives more importance to its physiological well-being since not reducing the deficits leads the agent to virtually pass away (a deficit reaches its maximum level of 100 units).

Equation 5 mathematically defines the reward function $\mathcal{R}$ used to update the algorithms integrated into our model. In each time step, the reward of executing an action is the

TABLE 5. The best principal hyperparameters for each DRL algorithm tested.

Parameter	DQN	DDQN	DuelDQN	DuelDDQN	SAC	PPO	C51	Rainbow
Learning rate	0.001	0.001	0.001	0.001	0.001	0.0003	0.01	0.2
Critic learning rate	–	–	–	–	0.01	0.001	–	–
Discount factor	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Hiden layers	3	3	3	3	2	1	2	4
Hidden neurons in layer	24	24	24	24	20	64	128	128
Memory limit	40000	40000	40000	40000	100000	100000	100000	100000
Target model update	0.01	0.01	0.01	0.01	1	1	100	100
Batch size	64	64	64	64	128	32	128	128
N step	–	–	–	–	–	–	–	3

sum of the weighted physiological well-being variance (Δ *Physiological Wb(t)*), a social reward (*Social Wb(t)*), and a *bonus* value that is +1 if the executed action reduces the deficit of the dominant motivation (motivation with the highest intensity) and −10 if the agent virtually passes away.

$$\mathcal{R}(t) = \Delta\ Physiological\ Wb(t) + Social\ Wb(t) + bonus \tag{5}$$

The weighted physiological well-being variance (Δ *Physiological Wb (t)*) is the difference between the physiological well-being in t and $t-1$. The weighted physiological well-being *Physiological Wb* (t) measures how good the robot's physiological state is in a certain time step by computing the average value of the robot's deficits. This value can be mathematically expressed using Equation 6,

$$Physiological\ Wb(t) = 100 - \frac{1}{N}\sum_{i=1}^{N} w_i(t) \times d_i(t), \tag{6}$$

where N is the number of biological processes in the robot, d_i is the deficit (see Equation 2) associated with a biological process, and $w_i(t)$ weights the current value of the process $cv_i(t)$ divided by its maximum value that for all the processes used in this work equals 100.

$$w_i(t) = \frac{cv_i(t)}{100} \tag{7}$$

We define social well-being *Social Wb* (t) as the reward obtained from fulfilling predefined social norms that aim at accomplishing user demands and social events. Table 6 shows the social norms and events defined for Mini's social behaviour, the conditions (actions and state of specific variables) to afford them, and the reward obtained in each case.

2) EVALUATION METRICS

The evaluation of the learning methodology presented in this contribution aims at maximising the reward obtained by the agent, maintaining its physiological and social state in the best possible condition, and surviving during extended periods without neglecting any physiological function. Consequently, the results presented in the following section show the evolution of the average episode reward, average steps per episode, average physiological well-being, and average social well-being. These metrics have the following role:

- **Average episode reward:** Represents the reward gained by the agent during the learning process. As the learning advances, the reward should increase with time.
- **Average steps per episode:** The learning methods must keep the agent alive by appropriately reducing internal deficits, so maximising the steps per episode is vital to correctly learning the behaviour policy.
- **Average physiological well-being:** Represents how well the learning system reduces the robot's deficits.
- **Average social well-being:** Represents how well the robot fulfils the social norms oriented to satisfy the user demands.

TABLE 6. Predefined norms are used to maintain the robot's social state in good condition and fulfill the users' demands. Each rule leads to a reward value if their associated condition is true.

Norm id	Action	Condition	Reward
0	Execute event	Event ready User present High priority	+4.0
1	Execute event	Event ready User present Moderate priority	+3.0
2	Execute event	Event ready User present Low priority	+2.0
3	Execute reminder	Reminder ready User present High priority	+4.0
4	Execute reminder	Reminder ready User present Moderate priority	+3.0
5	Execute reminder	Reminder ready User present Low priority	+2.0

3) EXPERIMENTAL SETUP

The experiment to obtain the best algorithm consisted of running each algorithm ten times and receiving the average results for the four metrics enumerated in the previous section. Each run had 100 episodes with a maximum of 3000 steps per episode. The episode ended if one of the

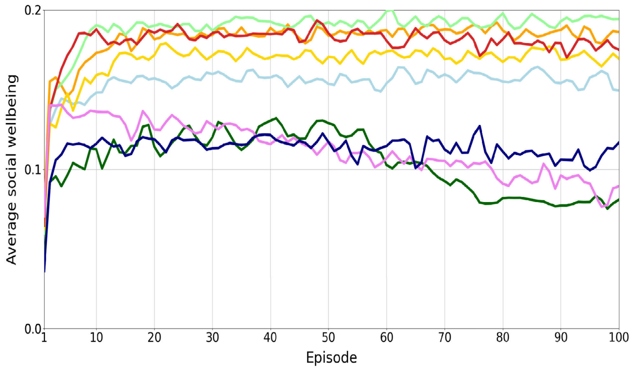

FIGURE 8. **Average social well-being per episode.**

biological deficits raised the worse value of 100 units, which means the agent virtually passed away. The initial state of the robot was generated randomly from all possibilities at the beginning of each episode, but preventing the biological processes of the robot from being very bad at the beginning to avoid early episode ends. We used the built-in TensorFlow and PyTorch methods to define random seeds for each run. The random seed was generated using Numpy's pseudo-random number generator from 0 to 1000. The plots have been smoothed using the *lfilter* method available in the Python library *Scipy*. This method applies a linear filter to a signal using a differential equation. It takes input arrays b and an as filter coefficients, representing the numerator and denominator of the filter's transfer function, respectively. In our approach, we have defined a = 1 and b as an array of 20 constant elements, all initialised to 1/20.

After obtaining the algorithm that better optimises the physiological and social well-being of the robot using the virtual environment, we transferred the trained model to the Mini social robot. We analysed the evolution of its physiological well-being during 3000 time steps for five runs. In the real test, each time step was set to 0.5 seconds (the total time was $3000 \times 0.5s = 1500s$), and the robot evaluated the best action to execute in each time step. If the new action the learning system selects differs from the current action, the robot changes its behaviour accordingly.

V. RESULTS

The first study's results were used to determine which algorithm produces the best outcomes in our scenario. Then, the results of the second study show the well-being optimization produced by the best algorithm in the Mini robot.

A. LEARNING COMPARISON

Figure 5 depicts the average reward per episode achieved by evaluating eight algorithms. DDQN and SAC stand out as the top performers, consistently generating the highest rewards. Alternative approaches like Rainbow and Dueling networks are closely followed. Rainbow demonstrates the quickest learning curve stabilization, accomplishing this feat in under ten episodes. Conversely, algorithms such as DQN or DDQN exhibit slower learning rates, necessitating around 40-50 episodes to stabilize. Regrettably, C51 and PPO struggle to optimize reward acquisition within the environment. C51 exhibits a significantly lower average reward than other algorithms, indicating sub-optimal learning. On the other hand, PPO achieves substantial reward gains in the final episodes but fails to stabilize its learning curve, which continually rises over the 100-episode duration.

The results depicted in Figure 6, showcasing the average steps per episode, align with the observations made regarding average rewards in Figure 5. The graph dynamics for each algorithm exhibit a similar trend. DDQN, DuelDQN, and Rainbow yield the most favourable outcomes, enabling the agent to survive for longer durations on average. Among them, Rainbow emerges as the most promising alternative due to its faster learning pace, mirroring the average episode reward graph findings. Figure 6 provides further insights into algorithms that struggle to solve the task, resulting in premature agent termination. For instance, as the average episode reward figure indicates, C51 and PPO barely achieve an average rate of 1800 steps per episode, indicating their failure to sustain the agent's survival in most trials.

As highlighted in previous sections, the primary objective of the learning system is to ensure the agent's survival while maximising physiological and social well-being. Hence, Figures 7 and 8 are critical in discerning which algorithm yields the most favourable outcomes within our environment. Rainbow demonstrates superior physiological maintenance, with the agent consistently exhibiting a physiological well-being score above 80 units on average during the learning process. Other algorithms, such as C51, SAC, or DQNs, effectively learn to maintain a good level of physiological well-being (around 70-80 out of 100), except for PPO, which yields lower ferior results with an average score around 60.

Analyzing the average social well-being, as depicted in Figure 8, offers valuable insights into learning performance. In this regard, DQN, DDQN, SAC, and DuelDDQN emerge as the top alternatives. While Rainbow excels at optimising physiological well-being, it slightly disregards social well-being. This pattern is also observed with PPO and C51, which exhibit moderately satisfactory regulation of social well-being. However, regarding physiological regulation, Rainbow proves to be excellent, in contrast to PPO.

B. WELL-BEING OPTIMIZATION

The insights derived from the preceding analyses strongly indicate that Rainbow is the optimal algorithm for optimising the agent's physiological well-being, while DDQN promotes social well-being within the virtual environment. Therefore, after careful consideration, we have integrated the Rainbow algorithm into Mini, as it guarantees the agent's survival in most episodes. While satisfying social events remains crucial, maximising physiological well-being takes precedence. The results obtained from the Rainbow algorithm demonstrate its ability to learn and regulate social well-being, albeit requiring more time to execute actions to fulfil the user's demands.

In Figure 9, the physiological and social well-being of the robot over 3000 time steps are showcased, utilising the behaviour policy learned by Rainbow from a randomly initialized state. Initially, the physiological well-being is moderately poor, hovering just below 60 units due to the random initialization. However, the robot's behaviour policy enables Mini to swiftly select the appropriate actions, restoring its physiological state and achieving a good condition within 300 steps. Similarly, the social well-being curve illustrates how social events are effectively handled, ensuring the robot successfully meets the user's demands.

VI. DISCUSSION

The comparison of the 8 DRL algorithms presented in Section V provides compelling evidence supporting the hypotheses outlined in the initial Section III of the paper. Our findings indicate that Rainbow is the most effective algorithm for optimising physiological well-being, aligning with our initial hypothesis. However, it is noteworthy that other algorithms yielded superior results when optimising social well-being. Consequently, these algorithms achieved higher average rewards (Figure 5) due to the greater value placed on social rewards than physiological rewards.

Figures 6 and 7 visually illustrate that the algorithms that excel in optimising physiological well-being (Rainbow, DDQN, and DuelDQN) also demonstrate superior efficiency in keeping the agent alive, resulting in a higher average number of steps per episode. Our hypothesis that duelling approaches would perform well yet occasionally falter due to the sporadic nature of high-value rewards in our environment was confirmed, particularly in the case of DuelDDQN. Dueling networks exhibit a stronger propensity for maximising physiological well-being over social well-being, yielding lower average rewards than other algorithms. However, these algorithms demonstrate better survivability. One plausible explanation for this observation pertains to our reward function's design. As the social rewards are significantly higher than physiological rewards to encourage the timely execution of social actions, the reward distribution deviates from the desired uniformity.

Finally, it is worth highlighting the performance of SAC, PPO, and C51. SAC emerges as a viable alternative as it strikes a balance between optimising physiological and social well-being. Conversely, PPO and C51 did not yield the expected results. On one hand, PPO failed to effectively learn how to maximise the robot's well-being, leading to continuous agent mortality. The average episode reward (Figure 5) suggests that performance may have improved with additional training episodes. However, the physiological and social well-being analysis contradicts this notion, as both indicators decrease as the number of episodes increases. Consequently, we decided to halt the learning process at 100 episodes, which was consistent with the other algorithms. On the other hand, C51 demonstrates competence in maintaining physiological well-being at a reasonably satisfactory level. However, social well-being deteriorates over time, indicating that the algorithm fails to fulfil social events.

The behaviour generation produced by the DRL model shows how the robot successfully maintains an optimal internal state while fulfilling the user's social events that dynamically appear. The system is intended to allow robots acting in social environments to select actions oriented to interact with their users. As stated earlier in this paper, previous works [47], [48] in this research line suggest that biologically inspired behaviour improves the robot's naturalness and user experience since people perceive the robot as more anthropomorphic and lively. Consequently, based on these results, this method defines the procedure to allow robots to learn the best behaviour policy to exhibit autonomous behaviour to regulate their physiological and social well-being. As we present in the conclusions, the user evaluation of the effects of this kind of system on robot users might be an interesting future line.

VII. LIMITATIONS

The shaping of the reward function is crucial for achieving optimal learning performance. It is essential to have a well-defined reward function that consistently yields desired outcomes to complete the learning process. However, there are instances where defining the reward function leads to inconsistent results, as the agent learns a behaviour policy that does not effectively solve the intended task. To address this challenge in our model, which focuses on maintaining the internal deficits of a robot while adhering to social norms and sustaining interaction, we have introduced a custom metric called "well-being." This metric measures whether the robot learns a behaviour policy that effectively solves the task.

Another significant limitation of Machine Learning models is the optimisation of hyperparameters to achieve the best results and the hardware where they run. In our approach, we optimise the hyperparameters for each specific algorithm, which becomes more complex due to the continuous nature of our scenario and the sparse rewards the agent receives. The sparsity of rewards implies that the agent may receive different rewards in similar situations, leading to slower and less stable learning. It is important to note that the Mini robot used in our study incorporates a biological model with predefined parameters, carefully chosen to represent specific dynamics such as the occurrence of social deficits and balancing between activity execution and rest. Currently, no standardised method for emulating and mathematically representing these biological processes in robots exists.

The hardware specifications of the Mini robot (Intel NUC with Intel I7 processor) where the DRL algorithms run are an important limitation in this work. Novel DRL algorithms consume many computation resources that are not feasible in our application. Therefore, we selected those algorithms with optimised training times and computational requirements.

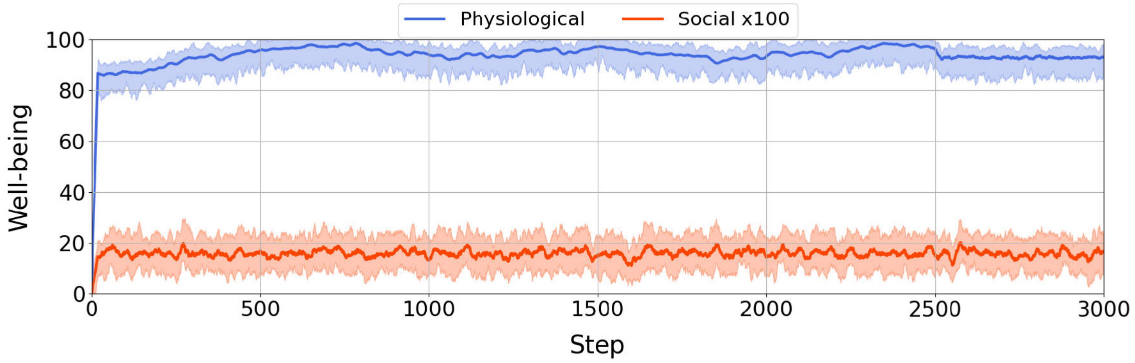

FIGURE 9. Average physiological and social (×100) robot well-being optimization using Rainbow after averaging the results of the learning process for 5 runs.

On the robot's side, there are limitations related to the information provided by its sensors. For example, measurements of the robot's position and light intensity can be noisy, resulting in an inaccurate representation of the robot's state. However, due to the adaptive nature of our model, the robot overcomes this issue by continuously selecting the most appropriate action in each situation. Choosing the correct action in a particular time step does not immediately harm the robot's physiological and social well-being, as they can be improved in subsequent steps. However, the continuous failure to restore these well-being factors ultimately leads to the agent's dire condition.

VIII. CONCLUSION

In conclusion, the advantages offered by DRL in terms of learning speed and stability for large state-action spaces present a vast array of possibilities for autonomous social robots. As robots advance in their sensory and actuation capabilities, it becomes crucial to equip them with various behavioural skills and adaptive mechanisms to navigate increasingly dynamic and complex environments. This research seizes upon these benefits as an opportunity to introduce a DRL model that empowers our social robot, Mini, to exhibit fully autonomous and adaptive behaviour.

To define a wide robot space, our model considers various input variables, including biologically inspired artificial processes, stimuli, and social norms. Through optimised action selection, it strives to maintain the robot's state in the best possible condition while satisfying user demands.

Building upon this research, we aim to expand the number of variables involved in the decision-making process of our autonomous robots. By doing so, we anticipate a substantial enhancement in their potential as they can manage a broader range of behaviours and situations. In future work, we will explore how to emulate biologically inspired processes, such as incorporating emotions into social robots and understanding their role in motivated decision-making. Additionally, to ensure user acceptance and usability during Human-Robot Interaction (HRI), we intend to incorporate user well-being into the loop and more social norms that align with those exhibited in formal environments. By continuously pushing the boundaries of research in this direction, we strive to create social robots that are not only highly capable but also seamlessly integrate into human-centric settings.

DECLARATION OF AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors used Grammarly to correct typos and spelling errors. After using this service, they reviewed and edited the content taking full responsibility of the publication.

REFERENCES

[1] M. Maroto-Gómez, F. Alonso-Martín, M. Malfaz, Á. Castro-González, J. C. Castillo, and M. Á. Salichs, "A systematic literature review of decision-making and control systems for autonomous and social robots," *Int. J. Social Robot.*, vol. 15, no. 5, pp. 745–789, May 2023.

[2] B. Lugrin, C. Pelachaud, and D. Traum, *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, Volume 2: Interactivity, Platforms, Application*. New York, NY, USA: ssociation for Computing Machinery, 2022.

[3] K. Baraka, P. Alves-Oliveira, and T. Ribeiro, "An extended framework for characterizing social robots," in *Human-Robot Interaction*. Cham, Switzerland: Springer, 2020, pp. 21–64.

[4] N. Gasteiger, M. Hellou, and H. S. Ahn, "Factors for personalization and localization to optimize human–robot interaction: A literature review," *Int. J. Social Robot.*, vol. 15, no. 4, pp. 689–701, 2021.

[5] A. Lambert, N. Norouzi, G. Bruder, and G. Welch, "A systematic review of ten years of research on human interaction with social robots," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 19, pp. 1804–1817, Nov. 2020.

[6] M. A. Salichs, Á. Castro-González, E. Salichs, E. Fernández-Rodicio, M. Maroto-Gómez, J. J. Gamboa-Montero, S. Marques-Villarroya, J. C. Castillo, F. Alonso-Martín, and M. Malfaz, "Mini: A new social robot for the elderly," *Int. J. Social Robot.*, vol. 12, no. 6, pp. 1231–1249, Dec. 2020.

[7] M. Maroto-Gómez, Á. Castro-González, J. C. Castillo, M. Malfaz, and M. A. Salichs, "A bio-inspired motivational decision making system for social robots based on the perception of the user," *Sensors*, vol. 18, no. 8, p. 2691, Aug. 2018.

[8] M. Maroto-Gómez, R. González, Á. Castro-González, M. Malfaz, and M. Á. Salichs, "Speeding-up action learning in a social robot with Dyna-Q+: A bioinspired probabilistic model approach," *IEEE Access*, vol. 9, pp. 98381–98397, 2021.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[10] P. A. Tsividis, T. Pouncy, J. L. Xu, J. B. Tenenbaum, and S. J. Gershman, ''Human learning in Atari,'' in *Proc. AAAI Spring Symp. Ser.*, 2017, pp. 1–4.

[11] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, ''A survey of deep reinforcement learning in video games,'' 2019, *arXiv:1912.10944*.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, ''Human-level control through deep reinforcement learning,'' *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[13] H. Van Hasselt, A. Guez, and D. Silver, ''Deep reinforcement learning with double Q-learning,'' in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–7.

[14] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, ''Dueling network architectures for deep reinforcement learning,'' in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1995–2003.

[15] M. Sewak, ''Deep Q network (DQN), double DQN, and dueling DQN,'' in *Deep Reinforcement Learning* (DQN). Cham, Switzerland: Springer, 2019, pp. 95–108.

[16] P. Christodoulou, ''Soft actor-critic for discrete action settings,'' 2019, *arXiv:1910.07207*.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, ''Proximal policy optimization algorithms,'' 2017, *arXiv:1707.06347*.

[18] M. G. Bellemare, W. Dabney, and R. Munos, ''A distributional perspective on reinforcement learning,'' in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 449–458.

[19] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, ''Rainbow: Combining improvements in deep reinforcement learning,'' in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[20] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, ''Social robots for education: A review,'' *Sci. Robot.*, vol. 3, no. 21, Aug. 2018, Art. no. eaat5954.

[21] R. van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. van der Ven, and P. Leseman, ''Social robots for language learning: A review,'' *Rev. Educ. Res.*, vol. 89, no. 2, pp. 259–295, Apr. 2019.

[22] A. A. Scoglio, E. D. Reilly, J. A. Gorman, and C. E. Drebing, ''Use of social robots in mental health and well-being research: Systematic review,'' *J. Med. Internet Res.*, vol. 21, no. 7, Jul. 2019, Art. no. e13322.

[23] C. Lytridis, C. Bazinas, V. G. Kaburlasos, V. Vassileva-Aleksandrova, M. Youssfi, M. Mestari, V. Ferelis, and A. Jaki, ''Social robots as cyber-physical actors in entertainment and education,'' in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2019, pp. 1–6.

[24] N. Akalin and A. Loutfi, ''Reinforcement learning approaches in social robotics,'' *Sensors*, vol. 21, no. 4, p. 1292, Feb. 2021.

[25] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, ''Robot gains social intelligence through multimodal deep reinforcement learning,'' in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2016, pp. 745–751.

[26] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, ''Intrinsically motivated reinforcement learning for human–robot interaction in the real-world,'' *Neural Netw.*, vol. 107, pp. 23–33, Nov. 2018.

[27] A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, X. Zhang, S. F. D. R. Alves, G. Nejat, and B. Benhabib, ''A multimodal emotional human–robot interaction architecture for social robots engaged in bidirectional communication,'' *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 5954–5968, Dec. 2021.

[28] S. Lathuilière, B. Massé, P. Mesejo, and R. Horaud, ''Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction,'' *Pattern Recognit. Lett.*, vol. 118, pp. 61–71, Feb. 2019.

[29] Y. Gao, F. Yang, M. Frisk, D. Hemandez, C. Peters, and G. Castellano, ''Learning socially appropriate robot approaching behavior toward groups using deep reinforcement learning,'' in *Proc. 28th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Oct. 2019, pp. 1–8.

[30] H. Cuayáhuitl, ''Simpleds: A simple deep reinforcement learning dialogue system,'' in *Dialogues With Social Robots*. Cham, Switzerland: Springer, 2017, pp. 109–118.

[31] H. Cuayáhuitl, ''A data-efficient deep learning approach for deployable multimodal social robots,'' *Neurocomputing*, vol. 396, pp. 587–598, Jul. 2020.

[32] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, ''Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning,'' in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6015–6022.

[33] L. Liu, D. Dugas, G. Cesari, R. Siegwart, and R. Dubé, ''Robot navigation in crowded environments using deep reinforcement learning,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 5671–5677.

[34] S. S. Samsani and M. S. Muhammad, ''Socially compliant robot navigation in crowded environment by human behavior resemblance using deep reinforcement learning,'' *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5223–5230, Jul. 2021.

[35] Z. Wan, C. Jiang, M. Fahad, Z. Ni, Y. Guo, and H. He, ''Robot-assisted pedestrian regulation based on deep reinforcement learning,'' *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1669–1682, Apr. 2020.

[36] Z. Zhou, P. Zhu, Z. Zeng, J. Xiao, H. Lu, and Z. Zhou, ''Robot navigation in a crowd by integrating deep reinforcement learning and online planning,'' *Appl. Intell.*, vol. 52, no. 13, pp. 15600–15616, Oct. 2022.

[37] X. Lu, H. Woo, A. Faragasso, A. Yamashita, and H. Asama, ''Socially aware robot navigation in crowds via deep reinforcement learning with resilient reward functions,'' *Adv. Robot.*, vol. 36, no. 8, pp. 388–403, Apr. 2022.

[38] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, ''An introduction to deep reinforcement learning,'' *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, 2018.

[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[41] M. Van Otterlo and M. Wiering, ''Reinforcement learning and Markov decision processes,'' in *Reinforcement Learning: State-of-the-Art*. Cham, Switzerland: Springer, 2012, pp. 3–42.

[42] M. Lewis, N. Fineberg, and L. Cañamero, ''A robot model of OC-spectrum disorders: Design framework, implementation, and first experiments,'' *Comput. Psychiatry*, vol. 2019, no. 3, pp. 40–75, Aug. 2019.

[43] K. Lorenz, *The Foundations of Ethology*. Cham, Switzerland: Springer, 1981.

[44] O. Avila-Garcia and L. Cañamero, ''Using hormonal feedback to modulate action selection in a competitive scenario,'' in *Proc. 7th Int. Conf. Simul. Adapt. Behav.*, vol. 8. Cambridge, MA, USA: MIT Press, 2004, p. 243.

[45] M. Maroto-Gómez, Á. Castro-González, J. C. Castillo, M. Malfaz, and M. Á. Salichs, ''An adaptive decision-making system supported on user preference predictions for human–robot interactive communication,'' *User Model. User-Adapted Interact.*, vol. 33, no. 2, pp. 359–403, 2022.

[46] S. Ravichandiran, *Hands-on Reinforcement Learning With Python: Master Reinforcement and Deep Reinforcement Learning Using OpenAI Gym and TensorFlow*. Birmingham, U.K.: Packt Publishing, 2018.

[47] M. Maroto-Gómez, Á. Castro-González, M. Malfaz, E. Fernández-Rodicio, and M. Á. Salichs, ''Modeling neuroendocrine autonomic responses in embodied autonomous robots,'' *Adv. Intell. Syst.*, vol. 5, no. 2, Feb. 2023, Art. no. 2200288.

[48] M. Maroto-Gómez, Á. Castro-González, M. Malfaz, and M. Á. Salichs, ''A biologically inspired decision-making system for the autonomous adaptive behavior of social robots,'' *Complex Intell. Syst.*, vol. 9, no. 6, pp. 6661–6679, Dec. 2023.

MARCOS MAROTO-GÓMEZ received the Ph.D. degree in robotics from the Carlos III University of Madrid, Madrid, Spain, in 2022. He is currently a Researcher and an Assistant Professor with the Carlos III University of Madrid. He belongs to the Robotics Laboratory Research Group and his actual research lines are related to human–robot interaction, decision-making, adaptation, autonomy, and machine learning applied to social robots.

MARÍA MALFAZ received the degree in physics science from La Laguna University, in 1999, the M.Sc. degree in control systems from the Imperial College of London, in October 2001, and the Ph.D. degree in industrial engineering, in 2007, and the topic was "Decision Making System for Autonomous Social Agents Based on Emotions and Self-learning." She is currently an Associate Professor with the Systems Engineering and Automation Department, Carlos III University of Madrid. Her research area follows the line carried out in her thesis and, more recently, she has been working on multimodal human–robot interaction systems. She belongs to international scientific associations: the IEEE Robotics and Automation Society (RAS), the International Association of Automatic Control (IFAC), and Comit Espaol de Automtica (CEA). She is a member of research networks, such as European Robotics Coordination Action (EURobotics) and Plataforma Tecnolgica Espaola de Robtica (HispaRob).

ÁLVARO CASTRO-GONZÁLEZ received the B.Sc. degree in computer engineering from the University of Leon, Leon, Spain, in 2005, and the M.Sc. and Ph.D. degrees in robotics and automation from the Carlos III University of Madrid, Madrid, Spain, in 2008 and 2012, respectively. He is a member of the Robotics Laboratory Research Group and an Associate Professor with the Department of Systems Engineering and Automation, Carlos III University of Madrid. He has been involved in several national, European, and corporate sponsored research projects. His current research interests include human–robot interaction, social robots, expressiveness in robots, decision-making, and artificial emotions.

SOFÍA ÁLVAREZ ARIAS received the degree in industrial technology engineering from the Polytechnic School of Gijón, University of Oviedo, Spain, in 2021, and the master's degree in industrial engineering from the Carlos III University of Madrid in 2023. Currently, she is a Researcher with the Robotics Laboratory, Carlos III University of Madrid, focusing on human-robot interaction, decision-making, adaptation, and learning in the context of social robotics.

MIGUEL ÁNGEL SALICHS (Life Member, IEEE) received the degree in electrical engineering and the Ph.D. degree from the Polytechnic University of Madrid. He is currently a Full Professor with the Systems Engineering and Automation Department, Carlos III University of Madrid. His research interests include autonomous social robots, multimodal human–robot interaction, mind models, and cognitive architectures. He was a member of the Policy Committee of the International Federation of Automatic Control (IFAC), the Chair of the Technical Committee on Intelligent Autonomous Vehicles of IFAC, a responsible of the Spanish National Research Program on Industrial Design, a Production Member of the Spanish Society on Automation and Control (CEA), and the Spanish Representative with the European Robotics Research Network (EURON). He is the Coordinator of the Secretariat of the Spanish Robotics Technology Platform (HispaRob).