Knowledge transfer in human-Artificial Intelligence collaboration.

by

Dimitrios Koutrintzes

Submitted
in partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence

at the

UNIVERSITY OF PIRAEUS

October 2023

Author Koutrintzes Dimitrios

II-MSc "Artificial Intelligence"

October 6, 2023

Certified by..................................

Maria Dagioglou
Research Associate
Thesis Supervisor

Certified by..................................

Georgos Vouros
Professor
Member of Examination Committee

Certified by..................................

Theodoros Giannakopoulos
Member of Examination Committee

Knowledge transfer in human-Artificial Intelligence collaboration.

By

Dimitrios Koutrintzes

Submitted to the II-MSc "Artificial Intelligence" on October 6, 2023, in partial fulfillment of the requirements for the MSc degree

Abstract

Socially aware AI agents should be able, among other things, to collaborate fluently with a human in tasks that require interdependent action in order to be solved. Towards enhancing mutual performance, collaborative AI agents should be equipped with adaptation and learning capabilities. However, co-learning requires long training intervals so that both partners learn and adapt to each other. To alleviate this, transfer learning methods could be explored to shorten training and improve performance. In the current thesis, we studied the experience and performance of human-agent teams in a task where a human and a Deep Reinforcement Learning (DRL) Soft-Actor-Critic (SAC) agent needs to learn in real-time how to collaborate in order to achieve a common goal. To test the benefits of transfer learning, a Learning from Demonstration method was used that utilized demonstration data from a human-agent expert team to facilitate the co-learning procedure. The proposed methods were evaluated through a study with 8 different human-agent teams, half of which played the game without transfer learning, while the rest with transfer learning. The results indicate that applying transfer learning in scenarios where the agent needs to collaborate with different humans has the potential to shorten training duration and improve the overall experience.

Thesis Supervisor: Maria Dagioglou
Title: Research Associate NCSR "Demokritos"

Acknowledgments

I would like to thank Dr. Maria Dagioglou for guiding me and helping me during the thesis. I would also like to thank Dr. Christos Spatharis for his useful comments and insight, as well as the RoboSKEL laboratory for providing me with the necessary hardware to complete my thesis.

Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the «funding body» or the view of University of Piraeus and Inst. of Informatics and Telecom. of NCSR “Demokritos”.

Contents

List of Figures

List of Tables

1 Introduction

Recent innovations in technology have enabled the development of agents capable of interacting with humans. Human-Agent Interaction (HAI) studies not only the development of relevant Artificial Intelligence (AI) methods but also the perceived experience of the human collaborators.

The development of Deep Reinforcement Learning (DRL) algorithms, specifically, has allowed Human-Agent Collaboration (HAC) in real-time scenarios, where human-agent teams collaborate towards achieving a common goal. Some examples include: games [2], [3], and robotic tasks in industrial environments [1], [4], and rehabilitation processes [5].

An important aspect of HAC teams is the ability to adapt to and learn from each other. With respect to learning, different categories of collaboration have been described in the literature, including: co-adaptation, co-learning, and co-evolution [6] depending on the time scales of learning, as well as the persistence and intention of the learning process. Co-adaptation focuses on how humans adapt to a situation in which they collaborate with an intelligent agent or robot (e.g. [7], [8]). Co-evolution refers to long-term real-world applications where both human and robot behaviour change in subtle ways over time (e.g. [11]). Co-learning refers to medium-term specific tasks that focus on improving performance or experience in human-agent collaboration (e.g. [9], [10]). Co-learning is the process where humans and agents learn how to learn with one another. Abich [12], in his work on the development of a human-agent co-learning tool for the United States Air Force (USAF), emphasises the importance of developing and maintaining trust, common ground, group awareness, communication and mutual adaptation for successful human-agent co-learning.

In this work, we focus on testing the benefits of using transfer learning in co-learning [30]. The purpose of the present study is to understand how transfer learning can impact the performance of the human-agent teams but also the experience of the human. The collaboration takes place in the form of a common task that requires the fluent cooperation of the human with an RL agent to achieve success. To accomplish the above, we evaluated the process to capture not only objective measures of collaboration but also subjective measures that capture the perceptions of the human participants.

In Chapter 2 of the thesis, an overview of the related work is presented including background knowledge in: Reinforcement Learning, Actor Critic agents, Co-learning Collaboration Environments, Transfer Learning, Subjective Measures of collaborative performance, as well as methods for capturing the personality of the human participants. In Chapter 3, the AI methodology, as well as all the related methods and material are described. Chapter 4 presents the results of our study. Lastly, Chapter 5 summarises the findings of the present work and discusses the limitations and potential future work.

2 Related Work

This chapter aims to review and analyse the current research on Human-Agent (HA) Co-learning. We will review the environments that have been used in HA co-learning, the challenges they have faced, and the modifications that have been proposed. Then we present a preview of different deep reinforcement learning methods that we can employ as our agent for collaboration, and finally transfer learning methods that we can use.

2.1 Reinforcement learning.

Reinforcement Learning (RL) [15] is a type of machine learning in which an agent learns by trial and error in an interactive environment, using feedback from its own actions and experiences. The feedback to the agent is in the form of rewards and punishments that signal positive and negative behaviour. The goal of the agent in RL is to learn by trial and error to maximise the total cumulative reward received from the environment. In the RL problem there are nine key terms that describe it:

- Environment: A digital or physical world in which the agent operates.
- Task: The objective which the agent needs to complete
- State (s_t): The situation of the agent in a given moment of time t . The state s belongs on a set of state S which includes every possible state in the environment.
- Action (a_t): The action the agent does to move to the next state based on the policy in a given point of time t . Every action a is from an action space A of all the possible actions.
- Path: A series of actions taken to reach a target point.
- Policy (π): Method of mapping the state of the agent to actions.

- Transition probability (T): probability that an action a_t at a state $s_t$will yield the state s_{t+1} .
- Reward (R): Feedback from the environment.
- Action Value function: An expected future reward that an agent could receive by taking an action in a specific spate, and then following its policy.

Figure 2.1 presents a RL task that is modelled as a Markov Decision Process (MDP). MDP is a discrete-time stochastic control process and is applied to situations where outcomes are partly random and is defined as a tuple $(S, A, T, R)^2$. In this model, for a given moment t , the agent receives the current state s_t and uses its policy to select the action a_t that will lead to a new state s_{t+1} and returns a reward signal. To create the values the policy has a state-action value function. We will refer to it as a simple value function, and it creates values for each possible action in a given state. These values specify how good is for an agent to perform in a particular state. Then, after executing the selected action transits to a new state s_{t+1}, and the environment returns a reward R_{t+1}. The agent can then use the reward to compare with the predicted value that it had previously and create an error that it can use to update its policy.

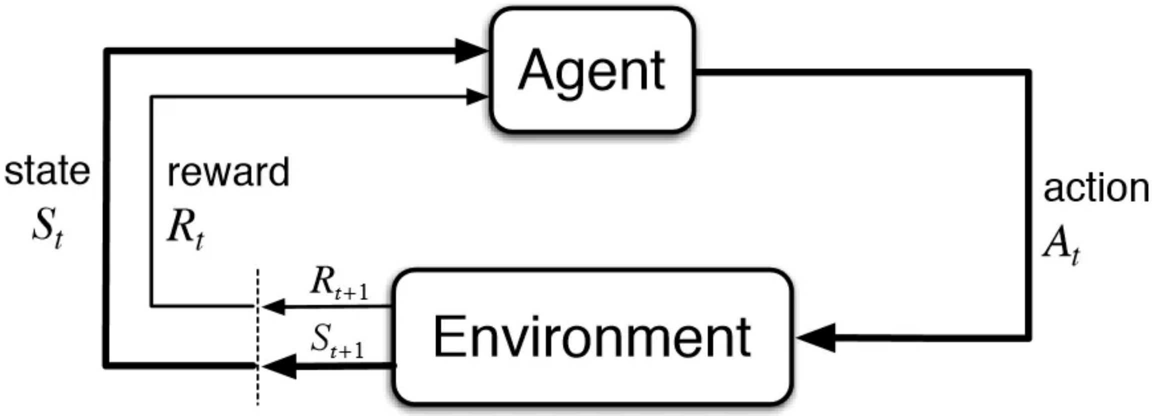


Figure 2.1: The cycle of interaction between the agent and the environment [15].

As we said earlier, policy dictates the action to be taken in a given state. In general machine learning modelling, there are two major approaches, deterministic and stochastic [14]. In the context of RL, the difference between the two approaches lies in the way the policy makes a decision.

- Deterministic Policy: A deterministic policy is a policy that maps each state to a single action with certainty. In other words, the agent will always take the same action in a given state. This policy is represented by a function $\pi : S \rightarrow A$, where S is the state space and A is the action space. The deterministic policy function maps each state $s \in S$ to a single action $a \in A$.
- Stochastic Policy: A stochastic policy chooses from a probability distribution over actions for each state. This means that the agent may choose a different action for the same state. The policy is represented by a function $\pi : S \times A \rightarrow [0,1]$, where S is the state space and A is the action space. This function returns a probability for each possible action $a \in A$ for a given state $s \in S$.

An example of a deterministic policy is using the argmax function, which selects the action with the best-expected value. An agent for a given state s uses a value function to predict the values for each possible action. Using these values, the argmax function chooses the action with the biggest value. This means, if there is no change in the value function, the action selected in a state will always be the same. In contrast, a stochastic policy uses a probabilistic method, like softmax, that uses the values for each possible action, to create a categorical distribution to use in order to choose the final action in random. This means that the policy selects an action with a probability that is connected with the values that the actions return.

Each approach has benefits and drawbacks that are based on the environment and the requirements of the task. A deterministic approach is better in tasks that require precision and any deviation from the optimal action can significantly impact the outcome. A stochastic approach is better when there is uncertainty and requires exploration of the environment to accomplish the task.

In general RL, when deciding on the approach, the main concern is how to achieve the task in the environment based on the factors we described above. In co-learning we also include the human cooperator as a factor. In co-learning, both the agent and the human need to learn how to complete the task and how to cooperate with each other. This means that the agent needs to learn the changeable behaviour of the human. This makes stochastic approaches more appropriate for use in co-learning.

We refer to the value function as the tool that in any approach dictates the decisions the agent takes. This means that it tries to predict the cumulative reward received if it takes that specific action and follows its policy thereafter. There are two stages in how to use the values, exploration and exploitation.

- Exploration: The agent chooses to move around the environment to states that the policy might not be selected otherwise, in order to “test” these states and update the value function to return more accurate values.
- Exploitation: The agent chooses to move to states that it believes to be the most profitable based on its current policy, aiming to maximize immediate rewards.

Figure 2.2 depicts the dilemma between exploration and exploitation. Let's say there are two restaurants where the agent wants to choose the best place to eat. The first is a place the agent has tested and knows is good and predicts a good value and the second is a place it has never tested, so it predicts a zero value. If it chooses the usual place, it knows that it will have a good experience, and a positive reward, but if it tries the new place, the expected reward can vary. If the new place is bad and the agent has a bad experience (negative reward), then it can update the value closer to the reward it took. But if the experience is good (positive reward) then it could update the value to better known for future choices that it is a good option or even the better option. After the agent has a good knowledge of all possible states (the two restaurants in our example) then it can move to exploit what it knows to maximize the overall reward it receives.

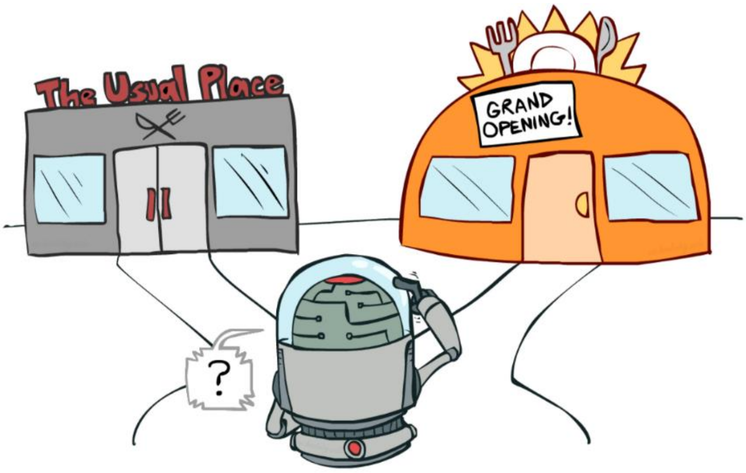


Figure 2.2: Choice dilemma [63]. The choice between an already known place or something new.

Based on this logic, there must be a balance between exploration and exploitation. Minimal exploration will cause the value function to produce inaccurate values, while too much exploration forces the agent to sacrifice rewards while having nothing more to learn. We will focus on the three most common approaches, random exploration, epsilon-greedy exploration and Boltzmann exploration. For a deeper analysis, see [16] for a complete survey about exploration methods.

- Random: The agent selects the actions randomly, with an equal probability for all possible actions, regardless of the expected values. This means for the entirety of the exploration the policy works based on the equation below.

$$\pi(s) = random\ action\ from\ A(s)$$

- Epsilon-Greedy: Greedy refers to the agent that focuses on exploiting. In an ε-greedy policy, the agent has a probability 0≤ ε ≤ 1 to select a random action. The decision-making is shown in the equation below,

$$\pi(s) = \begin{cases} random\ action\ from\ A(s) & if\ \xi < \varepsilon \\ \pi(s) & otherwise \end{cases}$$

with the ξ as a random number between [0,1] drawn in each time step. Based on the e the agent selects to either use the current policy or to select an action at random. A bigger e means that the agent mainly explores while a smaller one means mainly exploitation.

- Boltzmann exploration: It is a softmax exploration method that utilizes action-selection probabilities. In softmax instead of the policy producing an action base on the state for the agent to act, it produces a probability for each possible actions in the given state. The probability for each action is determined by ranking the value function estimates using a Boltzmann distribution as shown below,

$$\pi(a|s) = \Pr\{a_t = a|s_t = s\} = \frac{e^{\frac{Q(s,a)}{\tau}}}{\sum_b e^{\frac{Q(s,b)}{\tau}}}$$

where Q is a table that contains the values for each state-action combo. The temperature score τ regulates exploration if it is high, or exploitation if it is low.

Each approach has its weaknesses and benefits. Random exploration, while widely used, it can be inefficient and wasteful. An example of this is that if you are close to the target, and only one possible action can achieve the task, random exploration will take a long time to find it. A benefit of random exploration is that it guarantees that it will explore the environment more than any other method.

Epsilon-greedy can solve some of the problems of random exploration. The main difficulty is to choose an optimal epsilon value, in order to avoid just making a random exploration. But with an optimal value, epsilon-greedy can use both the policy, and be random at the same time. Compared to the random, the epsilon-greedy uses the policy, to increase the probability of making the right choice when needed. At the same time, it explores the environment in a way that is not affected by the policy and the value function.

Boltzmann exploration has similar difficulties with epsilon-greedy. Basically, a big temperature score makes the exploration random, so it needs to be optimized. It uses the softmax function and thus can be used in stochastic policies. It can be used in deterministic enviroments, by using first a stochastic policy for exploration and switching to a deterministic policy for exploitation. The main disadvantage of the other two methods is that the probability of action to be selected is based on its value. This means that unless the temperature score is too high, all selected actions are affected by their values. This method is beneficial when the purpose of exploration is to explore action/paths that lead to similar rewards. As in the exploration, the agent will commonly select these paths it will train the value function faster than an epsilon-greedy approach.

Another factor in RL algorithm is how the value function and policy are structured. Previously we saw some simple approaches in the form of simple argmax/softmax policies and value functions that are directly connected with the reward received. In general, there are many ways to approach the modelling of these functions. Goodman [17] presented three categories of RL algorithms in three families of algorithms. The three families are the following:

- Actor-Only (Policy-based): Methods that typically work with a parameterised family of policies over which the optimisation procedure can be applied directly. A parameterised policy is like a set of instructions that can generate a wide range of continuous actions. This can benefit the Actor-Only as it can generate a spectrum of continuous actions, but the optimisations typically used, such as policy gradient methods, suffer from high variance in the estimates of the gradient, leading to slow learning.

- Critic-Only (Value-based): Methods that use temporal difference (TD)[15] learning and have lower variance in the estimate of expected returns. In TD learning, the value function is estimated based on the current state and the observed immediate reward, as well as the estimated value of the next state. A simple policy derived from critic-only methods is greedy exploitation, where the agent chooses the action that yields the best value. However, this requires an exploration run to find the action that leads to the optimal value. This can be computationally expensive.

- Actor-critic: Methods that aim to combine the advantages of actor-only and critic-only methods. While a parameterised actor has the advantage of computing continuous actions without the need for optimisation procedures on a value function, the advantage of the critic is that it provides the actor with low-variance knowledge of performance. More specifically, the critic's estimate of expected returns allows the actor to update with gradients that have a lower variance, thereby speeding up the learning process. The lower variance is traded for a larger bias at the beginning of learning, when the critics' estimates are far from accurate. Actor-critic methods tend to have good convergence properties, in contrast to pure critic methods.

Actor-critics have shown excellent results compared to Actor-only and Critic-only results and have increasingly been used in HAC studies [18,19].

2.1.1 Actor-Critics

In table 2.1 we present some Actor-Critics we focus on our selection process. The table includes the Actor-Critic2 or AC2 [20] the Trust Region Policy Optimization or TRPO [52], the Proximal Policy Optimization Actor-Critic or PPO AC [21] and the Soft Actor-Critic or SAC [22]. More RL algorithms can be found in the Goodman survey including Actor-only and Critic-only methods.

Table 2.1: Actor-Critic's

Algorithm	**Architecture**	**Main Benefits**
AC2 [20]	Policy Gradient Actor with Primary and Secondary Critics	Manages both bias and variance in policy gradients. Secondary critic focuses on problematic states (upper 95 percentile) for stable performance. Concentrates training on problematic states for variance reduction with tolerable bias.
TRPO (Trust Region Policy Optimization) [52]	Uses a trust region to limit policy updates and ensure stability.	Provides more stable policy updates by constraining the changes in each update. Can achieve better sample efficiency and convergence than vanilla policy gradient methods.
PPO AC (Proximal Policy Optimization Actor Critic) [21]	Alternates between sampling data and optimizing a 'surrogate' objective function with stochastic gradient ascent.	Combines benefits of policy gradient methods with multiple epochs of minibatch updates. Stable and reliable like trust region methods, but simpler to implement (only a few code changes required). Competitive results compared to other actor-critic algorithms.

SAC (Soft Actor Critic) [22]	Off-policy Maximum Entropy Deep RL with Stochastic Actor	Utilizes maximum entropy reinforcement learning to maximize expected reward while maximizing entropy. Off-policy updates with a stable stochastic actor-critic formulation. Achieves state-of-the-art performance on continuous control tasks. Outperforms both on-policy and off-policy methods (PPO, TD3, etc.).

2.1.2 Soft Actor-Critic and SAC discrete.

Soft Actor-Critic (SAC) was introduced by Haarnoja [22] and [61] as an off-policy maximum entropy soft actor-critic deep RL method, with a stochastic actor. The purpose was to combine both pattern-efficient learning and stability. In his paper [22], Haarnoja proves that SAC accomplice a convergence for policy iteration in the maximum entropy framework. He also provides empirical results, in comparison with previous work, that include both off-policy and on-policy methods and show a significant improvement in both performance and sample efficiency.

Entropy is a term that was originally used in physics to denote the lack of order within a system. In RL [23], the definition of entropy is repurposed to describe the unpredictability of the action that an agent takes given a policy. This means that the more random the agent the higher the entropy and vice versa. In SAC the purpose of the entropy is to help converge to the optimal policy and to capture multiple modes of near-optimal behaviour.

The general formula of calculating the entropy is given in EQ 1. The negative summary of the probability of x and the logarithm of the probability of x, also known as the surprise of x. In RL this equation is taking place in a state level. In Eq 2, for any given state, the entropy is calculated based on the probability of the policy for each possible action, and the surprise of the action.

$$H(X) = -\sum_{x \in X} P(x) log P(x) \qquad (1)$$

$$H\big(\pi(\cdot\,|s_t)\big) = -\sum_{\alpha \in A} \pi(\alpha|s_t) \log \pi(\alpha|s_t) \qquad (2)$$

In SAC the entropy is used as a soft value, similar to what we explained in the Boltzmann exploration in section 2.1. This makes the soft value to vary based on the randomness of the agent. Based on this, when the agent is more random and mostly explores the environment, produces a smaller soft value, allowing the value function to train based on the environment rewards. This will make the agent move to a more exploiting stage, where it will move in paths that produce the best rewards but increase the entropy. This switch will mean that the agent will start converging to actions that produce similar rewards, and the entropy will ensure that all actions have similar probabilities.

The SAC algorithm is divided into two steps, policy evaluation and policy improvement. In the policy evaluation step of the soft policy iteration, the value of the policy π is calculated based on the maximum entropy objective in EQ 3.

$$J(\pi) = \sum_{\tau=0}^{T} E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + aH(\pi(\cdot\,|s_t))] \qquad (3)$$

Basically, the value is equal to the expected reward given in a state action pair and the entropy of the policy in the same state, regulated by the temperature alpha.

The temperature alpha (α) parameter determines the relative importance of the entropy term versus the reward, and controls how much the entropy will affect the value of actions. In his original work, Haarnoja [22] intended for the temperature value to be a hyperparameter tuned by the user, but since finding the optimal value is non-trivial. In his findings, the temperature needed to be tuned for each task. To solve this, he introduced a gradient update using Eq 4 to update the α parameter in training, where H is the minimum expected entropy.

$$J(\alpha) = E_{a_t\sim\pi_t}[-\alpha \log \pi_t(a_t|s_t) - \alpha\overline{H}] \qquad (4)$$

For a fixed policy, the soft Q-value is computed starting from an arbitrary function $Q : S \times A \rightarrow R$ and repeatedly applying a modified Bellman backup operator T^{π} given by Eq 5.

$$T^{\pi}Q(s_t,a_t) \triangleq r(s_t,a_t) + \gamma E_{s_{t+1}\sim\rho}[V(s_{t+1})] \qquad (5)$$

Here, the Bellman of a Q-value of an action a_t in the state s_t is defined by the acquired and the discounted expected soft state value of the next state s_{t+1}. The soft state value function is Eq 6. This function calculates the value of a given state

s_t, based on the expected Q values of that state and the entropy of the actions in that state.

$$V(s_t) = E_{a_t \sim \pi}[\mathrm{Q}(s_t, a_t) - \alpha \log \pi(a_t|s_t)\,] \qquad (6)$$

We can obtain the soft value function for any policy π by repeatedly applying T^{π}as formalized in Lemma 1 .

Lemma 1 (Soft policy evaluation). Consider the soft Bellman policy operator T^{π} in Equation 2 and a mapping $Q_0\colon S \times A \ \rightarrow\ R$ with $|A| \ < \ \infty$, and define $Q_{k+1} \ = \ T^{\pi}\ Q_k$. Then the sequence Q_k converges to the soft Q-value of $\pi\ as\ k \ \rightarrow\ \infty$.

In the policy improvement step, SAC updates the policy towards the exponential of the new Q-function. This choice of updates results in an improved policy in terms of its soft value. To make the policies more tractable, an additional restriction on the set of policies Π is used. For example, this can correspond to a parameterized family of distributions such as Gaussians. To account for the constraint that $\pi \ \in\ \Pi$, the set of policies is projected using the Kullback-Leibler divergence (D_{KL}). In other words, in the policy improvement step, for each state, the policy is updated according to Eq 6, which is defined by the D_{KL} of the policies in a given state and the Q-value of the state for all possible actions, and $Z(s_t)$, a partition function normalizing the distribution, which is generally intractable but does not contribute to the gradient with respect to the new policy and can therefore be ignored. The formalized result is given in Lemma 2.

$$\pi_{new} = \arg\min_{\pi' \in \Pi} D_{KL}(\pi'(\cdot\,|s_t)||\,\frac{\exp(\frac{1}{\alpha}Q^{\pi_{old}}(s_t,\cdot))}{Z^{\pi_{old}}(s_t)}) \qquad (7)$$

The full soft policy interaction algorithm alternates between soft policy evaluation and soft policy improvement steps, and will probably converge to the optimal maximum entropy policy among the policies in Π (Theorem 1).

Lemma 2 (Soft policy improvement). Let $\pi_{old} \in \Pi$ and let π_{new}be the optimiser of the minimisation problem defined in Equation 4. Then$Q^{\pi_{old}}(s_t, a_t) \geq Q^{\pi_{new}}(s_t, a_t)$ for all $(s_t, a_t) \in S \times A$ with $|A| < \infty$.

Theorem 1 (Soft Policy Iteration). Repeated application of soft policy evaluation and soft policy improvement from any $\pi \in \Pi$ converges to a policy π^* such that $Q^{\pi^*}(s_t, a_t) \geq Q^{\pi}(s_t, a_t)$ for all $\pi \in \Pi$ and $(s_t, a_t) \in S \times A$, assuming $|A| < \infty$.

In a sizeable continuous domain, it is necessary to derive a practical approximation to soft policy iteration. For this reason, SAC uses a function approximator for both the Q-function and the policy, and instead of running evaluation and improvement to convergence, it alternates between optimizing both networks with stochastic gradient descent. Based on this, SAC uses a parameterised state value $V_\psi(s_t)$, a soft Q-function $Q_\theta(s_t, a_t)$ and a trackable policy $\pi_\varphi(s_t|a_t)$. The parameters of these networks are ψ, θ and φ.

The state value function approximates the soft value because it is related to the Q-function and the policy according to Eq 6. There is no need for a separate function approximator for the state value, but in practice, including a separate function approximator for the soft value can stabilize the training and is convenient to train simultaneously with other networks. The soft-value function is trained to minimize the residual squared error as shown in Eq 8.

$$J_{V(\psi)} = E_{s_t \sim D}[\frac{1}{2}(V_\psi(s_t) - E_{a_t \sim \pi_\varphi}[Q_\theta(s_t, a_t) - a\log\pi_\varphi(a_t|s_t)])^2] \qquad (8)$$

This means that the error is defined by the difference between the soft value of the given state s_t and the expected value of the Q-functions for s_t with the action taken a_t,and the entropy in the policy in s_t. D represents the distribution of previously sampled states and actions, or a replay buffer. Its gradient can be estimated with an unbiased estimator as in Eq 9, where the actions are sampled according to the current policy instead of the replay buffer.

$$\widehat{\nabla}_{\psi} J_V(\psi) = \nabla_{\psi} V_{\psi}(s_t)(V_{\psi}(s_t) - \; Q_{\theta}(s_t, a_t) + \alpha \log \pi_{\varphi}(a_t|s_t)) \quad (9)$$

The soft Q-function parameters can be trained to minimize the soft Bellman residual as shown in Eq 10, where the error is determined by the expected difference between the value of the Q-function in state s_t, for action a_t, and the soft Bellman residual $\hat{Q}$ as shown in Eq 11.

$$J_Q(\theta) = E_{(s_t, a_t) \sim D}[\frac{1}{2}(Q_{\theta}(s_t, a_t) - \hat{Q}(s_t, a_t))^2] \qquad (10)$$

$$\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma E_{s_{t+1} \sim \rho}[V_{\bar{\psi}}(s_{t+1})] \qquad (11)$$

$$\widehat{\nabla}_{\theta} J_Q(\theta) = \nabla_{\theta} Q_{\theta}(s_t, a_t)(Q_{\theta}(s_t, a_t) - \; r(s_t, a_t) + \gamma V_{\bar{\psi}}(s_{t+1})) \qquad (12)$$

The Q is calculated based on the reward given in the state s_t for action a_t and the discounted expected soft value of the next state s_{t+1}. The gamma discount factor γ is a hyperparameter set by the designer, with a value between 0 and 1. The soft value function can be optimised with stochastic gradients as shown in Eq 12. This update uses the target value network V_{ψ}, where ψ can be an exponential moving average of the value network weights, which has been shown to stabilise training. As an alternative, the value function weights can be periodically updated to match the current value function weights. Finally, the policy parameters can be learned by directly minimising the expected KL-divergence, as shown earlier in Eq 7. The error is defined as shown in Eq 13 by the D_{KL} of the policy for all possible actions in state s_t, and again by the Q-function value in state s_t for all possible actions in this state and the partition function Z_{θ}. Since in this case the Q-function is represented by a neural network, a typical solution for policy gradient methods such as the ratio gradient estimator [24] would not work, so a re-parameterisation trick is used that results in a lower variance estimator. To do this, the policy is re-parameterized using a neural network transformation as shown in eq 14.

$$J_\pi(\varphi) = E_{s_t \sim D}[D_{KL}(\pi'(\cdot\,|s_t)| \left| \frac{\exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)} \right)] \qquad (13)$$

$$a_t = f_\varphi(\epsilon_t; s_t) \qquad (14)$$

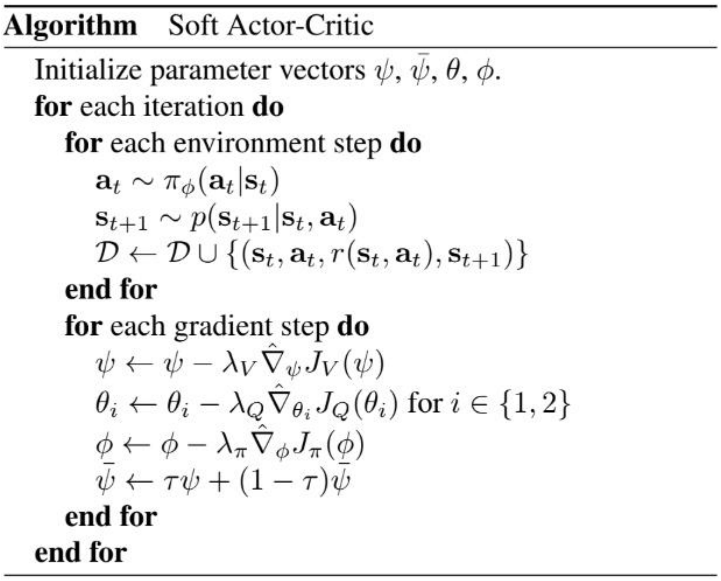
The ϵ_t is an input noise vector sampled from a fixed distribution, such as a spherical Gaussian. The is implicitly defined in terms of f_φ, and it is noted that the partition function is independent of φ and can be omitted. The new equation is eq 15.

$$J_\pi(\varphi) = E_{s_t \sim D, \epsilon_t \sim N}[\alpha \log \pi_\varphi\left(f_\varphi(\epsilon_t; s_t)\middle|s_t\right) - Q_\theta(s_t, f_\varphi(\epsilon_t; s_t))] \qquad (15)$$

The approximation of the gradient of this is in eq 16. Here the a_t is evaluated at $f_\varphi(e_t : s_t)$. This unbiased gradient estimator extends the DDPG-style policy gradient [25] to any tractable stochastic policy.

$$\hat{\nabla}_\varphi J_\pi(\varphi) = \nabla_\varphi \log \pi_\varphi(a_t|s_t)$$
$$+(\nabla_{a_t} \log \pi_\varphi(a_t|s_t) - \nabla_{a_t} Q_\theta(s_t, a_t))\nabla_\varphi f_\varphi(\epsilon_t; s_t) \qquad (16)$$

SAC also uses two Q-functions to mitigate the positive bias in the policy improvement trick, which is known to degrade the performance of value-based methods. The two Q-functions are parameterized with θ_i and trained independently to optimize the $J_Q(\theta_i)$. During the value gradient Eq 9 and the policy gradient Eq 16, the minimum value of the two networks is used. This method was proposed by Fujimoto et al [26]. The final complete algorithm is described in Figure 2.3.



```
Algorithm  Soft Actor-Critic
  Initialize parameter vectors $\psi$, $\bar{\psi}$, $\theta$, $\phi$.
  for each iteration do
    for each environment step do
      $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
      $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
    end for
    for each gradient step do
      $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
      $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
      $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
      $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$
    end for
  end for
```

Figure 2.3 Representation of the Soft Actor-Critic algorithm

In his work, Christodoulou [27] introduces some changes to make SAC work in a discrete action setting. In a discrete action space, the $\pi_\varphi(a_\tau|s_\tau)$ now outputs a probability instead of a density. Therefore, the tree objective functions $J_Q(\theta)$, $J_\pi(\varphi)$, $J(a)$ still apply, but there are five important changes to the process of optimizing these objective functions.

Instead of giving the Q-function the action as input and output its Q-value, it is simpler to output the Q-values for all possible actions at once, which means that the Q-functions change from $Q\colon S \times A \rightarrow R$ to $Q\colon S \rightarrow R^{|A|}$. In a continuous setting, this is not possible, as there are infinitely many possible actions we could take.

For the same reason, the policy no longer needs to output the mean and covariance of our action distribution, instead it can output the action distribution directly. The policy therefore changes from $\pi\colon S \rightarrow R_2\,|A|$ to $\pi\colon S \rightarrow [0,1]\,|A|$, where the network has a softmax function in the last layer of the policy to ensure that it outputs a valid probability distribution.

The soft value function in the continuous setting involved taking an expectation over the action distribution, while in the discrete setting this is not necessary as we can recover the full action distribution. This means that the soft value functions, as shown in Eq 17, include the output vector of the policy.

$$V(s_t) := \ \pi(s_t)^T[Q(s_t) - \alpha \log\left(\pi_\varphi(s_t)\right)] \qquad (17)$$

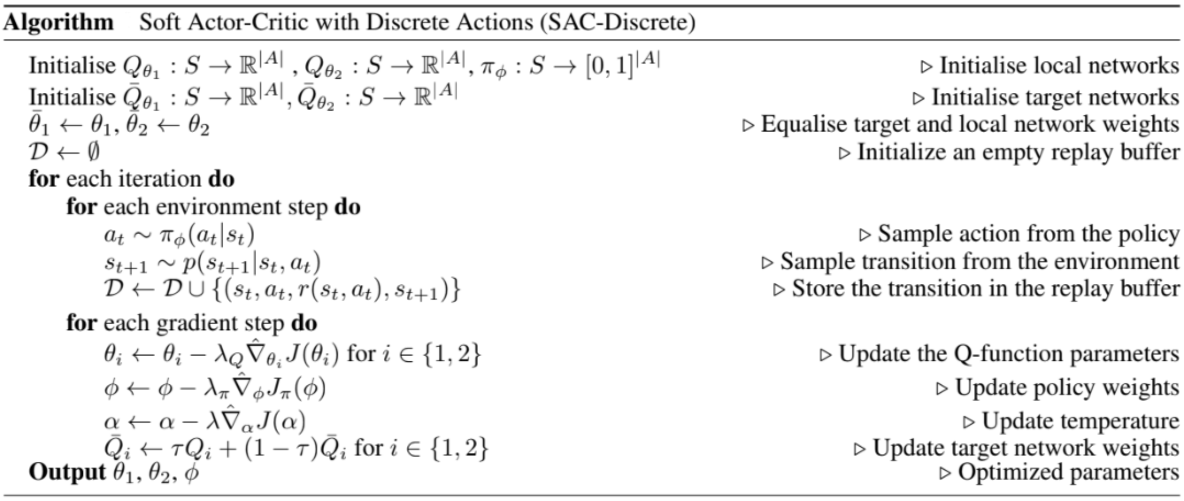
Similarly, the temperature loss equation has been modified to also reduce the variance. The new temperature objective is now as shown in Eq 18, where we again include the output vector of the policy.

$$J(\alpha) = \ \pi(s_t)^T[-\alpha(\log\left(\pi_\varphi(s_t)\right) + \ \bar{H})] \qquad (18)$$

The final change is to the objective of the policy, as shown earlier it required a re-parameterization trick to allow gradients to pass through the expectation operator. Now that the policy outputs the exact action distribution, it is possible to compute the expectation directly. The new objective for the policy is shown in Eq 19, where the expected value of the policy output vector is the difference between the regulated entropy in state st and the Q-functions value fors_t.

$$J_\pi(\varphi) = E_{s_t \sim D}[\pi(s_t)^T[\alpha \log\left(\pi_\varphi(s_t)\right) - \ Q_\theta(s_t)]] \qquad (19)$$

After the changes applied by Christodoulou [27] the new discrete Soft Actor Critic algorithm is described in Figure 2.4



```
Algorithm  Soft Actor-Critic with Discrete Actions (SAC-Discrete)
Initialise $Q_{\theta_1} : S \rightarrow \mathbb{R}^{|A|}$ , $Q_{\theta_2} : S \rightarrow \mathbb{R}^{|A|}$, $\pi_\phi : S \rightarrow [0,1]^{|A|}$    ▷ Initialise local networks
Initialise $\bar{Q}_{\theta_1} : S \rightarrow \mathbb{R}^{|A|}$, $\bar{Q}_{\theta_2} : S \rightarrow \mathbb{R}^{|A|}$    ▷ Initialise target networks
$\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2$    ▷ Equalise target and local network weights
$\mathcal{D} \leftarrow \emptyset$    ▷ Initialize an empty replay buffer
for each iteration do
    for each environment step do
        $a_t \sim \pi_\phi(a_t|s_t)$    ▷ Sample action from the policy
        $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$    ▷ Sample transition from the environment
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$    ▷ Store the transition in the replay buffer
    for each gradient step do
        $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J(\theta_i)$ for $i \in \{1,2\}$    ▷ Update the Q-function parameters
        $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$    ▷ Update policy weights
        $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$    ▷ Update temperature
        $\bar{Q}_i \leftarrow \tau Q_i + (1-\tau)\bar{Q}_i$ for $i \in \{1,2\}$    ▷ Update target network weights
Output $\theta_1, \theta_2, \phi$    ▷ Optimized parameters
```

Figure 2.4 Representation of the discrete Soft Actor-Critic algorithm

2.2 HAC and co-learning

In a HAC environment, a human and a reinforcement learning (RL) agent work together to complete a common task. Each one acts differently in many ways, from the difficulty to be completed, the learning curve required of the human and agent, and the way the human and agent interact. Semeraro summarised some categories of HAC [28], in the context of an industrial setting and the use of cobots. The three categories are separated based on the role of the human and the agent in the task they have to perform. The three categories are object transfer, collaborative assembly and collaborative manufacturing.

- Object handover: the robot's role in the interaction is to provide the human with objects to complete the task, the focus is more on a specific part of the interaction without considering the overall complex task. In this case, it is important for the robot to understand the user's intentions and expectations when receiving the object.
- Collaborative Assembly: Humans and robots are tasked with assembling a complex object through sequential sub-processes. The human and the robot interact on the same task, in the same workspace and at the same time.
- Collaborative manufacturing: the robot makes a permanent physical change to an object in collaboration with the human as part of a manufacturing process.

These categories summarize some ways where humans can collaborate with RL agents in their daily lives. In co-learning, humans and agents need to learn to work as a team in order to maximise their performance. As shown in Figure 2.5, in a co-learning process, both sides need to self-learn and reflect based on the advice and feedback they receive from the collaboration. In the case of the agent, these steps have a form of reward from the environment and training processes, illustrated more in section 2.1. In contrast, a human adapts based on his previous experiences and his personality.

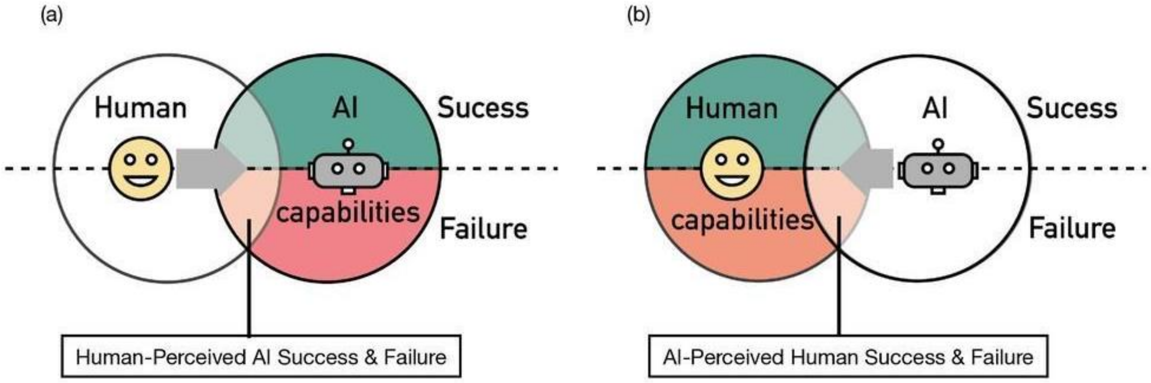

(a)
Human
AI
Sucess
capabilities
Failure
Human-Perceived AI Success & Failure
(b)
Human
AI
Sucess
capabilities
Failure
AI-Perceived Human Success & Failure

Figure 2.5: The two perceptions of humans and AI in HAC [64].

In the quest to create an environment that can be used to monitor and evaluate human-agent collaborations, non-trivial and only solvable collaboration platforms have been proposed and used in research [29], [30] and [31]. These platforms, both robotic and virtual, provide an excellent way to evaluate co-learning methods, but can also provide insight into human behaviour during collaboration.

In Shafti et al. environment [29], the goal is to move a ball from a corner to a target by tilting a platform, as visualised in Figure 2.6. The human controls one axis and the robot controls the other, and the platform contains a series of obstacles that the human and robot must work together to overcome. Similarly, Tsitos et al. [30] uses a robotic arm to create an environment in which a human and an agent must cross certain paths to reach the goal [30], as shown in Figure 2.7. In his work, he used a number of subjective measures in addition to the objective measure. Using subjective measures, he was able to show how the experience of participants was improved in many aspects of the collaboration using transfer learning.

Lygerakis [31] created a virtual simulation of [29] that was used to evaluate different training approaches. Differences between Shaftis and Lygerakis environments are shown in Table 2.2

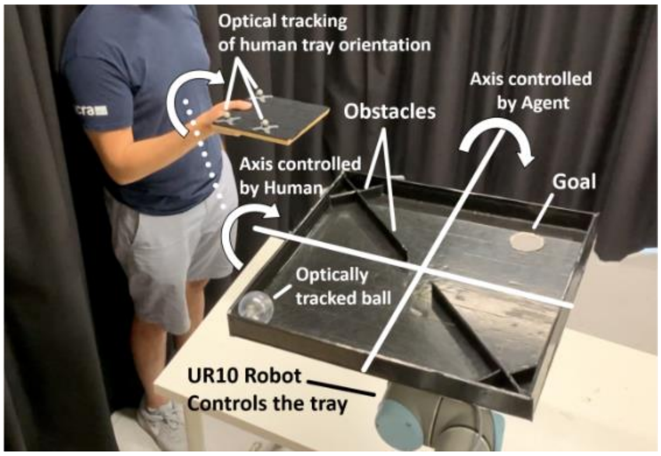


Figure 2.6: Human-Robot co-learning setup [29]: A ball and maze game is designed to require two players for success; one player per rotation axis of the tray. One axis is teleoperated by a human player, and the other axis by a deep RL agent. The game can only be solved through collaboration.

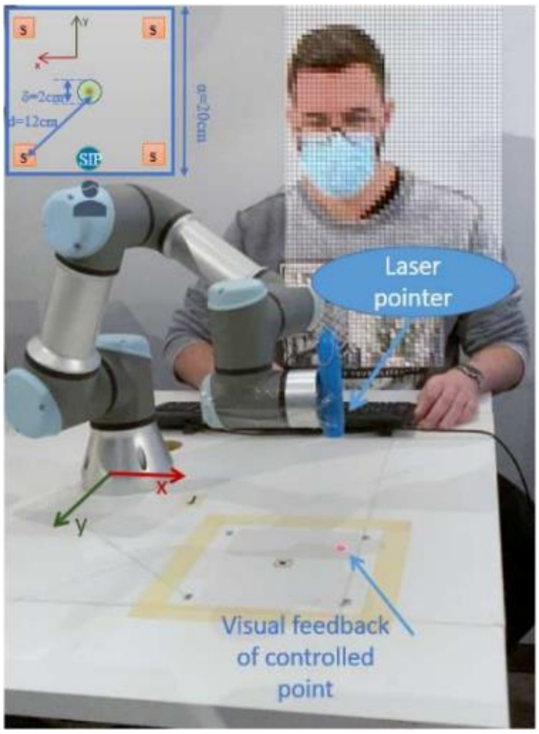


Figure 2.7: Human-Robot Collaboration setup [30]. The robot's movements are onstrained within a 20cm × 20cm area - a schematic representation of the area is presented in the upper left corner of the figure. The EE is placed in one of the four starting ('S') positions and the HR team has to bring the EE in the centre (green area) of the square. A laser pointer attached to the EE of the robot provides to the human visual feedback about the position of the EE that is controlled.

Table 2.2 Shaftis and Lygerakis environment settings and game settings.

Enviroment	**Shafti et al [29]**	**Lygeraki et al [31]**
Platform	Real-world, with a Universal Robots UR10 as the robot manipulator	3D Virtual world, using Unity version 2020.3.13f1
Tray	50 cm x 50 cm	10 x 10 Unity units
Ball size	6 cm	1 Unity unit
Opening between obstacles	9 cm	1.4 Unity unit
Target Hole	5 cm	1 Unity unit
Method of control	Smaller tray with optical markers using a motion capture system consisting of Optitrack flex 13 cameras	Keyboard
Action space	Continuous	Discrete
Balls starting point	3 corners above the obstacles (rotating in each trial)	3 corners above the obstacles (random in each trial)
Control frames	200	200
Size of control frame	200ms	200ms
Size of a trial	40s	40s
Replay buffer	1000	1000000
Reward	10 on goal, -1 every other state	10 on goal,-1 every other state
Score	200 minus 1 point for each control frame, if goal not reach the score reaches zero	200 minus 1 point for each control frame, if goal not reach the score reaches zero

2.3 Transfer learning

Transfer learning (TL) [15] occurs when an existing model/policy is used to solve a new challenge or problem. In RL, transfer learning involves capturing knowledge gained from interacting with the environment to complete a task and using that knowledge to improve learning and performance in another related task. In HAC, and specifically in co-learning, the agent has to both learn the task and learn to work efficiently with the human collaborator. TL can be used to increase the efficiency and performance of the team's learning process. However, the use of TL must happen without creating a negative transfer.

Negative transfer occurs when transfer learning can make later problems more difficult to solve. A TL method could help the agent to learn its part of the task. However, this step should not discourage the agent from learning to cooperate with a human more efficiently.

TL can take many forms and each can have different effects in the training and co-learning process. Zhungadi et al. [33] gave an overview of different approaches to TL research . He categorised approaches to TL based on the information that each method provides in transfer:

- Reward Shaping (RS): This technique uses external knowledge, obtained from a domain expert or other sources, and is used to influence the reward provided by the environment to encourage desirable behaviour and discourage undesirable actions, with the aim of guiding the agent's policy learning (e.g. [53, 54]).

- Learning from Demonstration (LfD): This technique involves the transfer of knowledge in the form of demonstrations provided by a human expert, a previously learned expert policy, or even a suboptimal policy. The agent can observe this demonstration and imitate the behaviour demonstrated (e.g. [55])

- Policy Transfer (PT): In this approach, the external knowledge takes the form of pre-trained policies from one or more source domains. These policies represent a mapping from state to action and embody the agent's decision-making process. By transferring policies learned in related tasks, the agent can benefit from the knowledge gained and apply it to the target task .

- Inter-Task Mapping (ITM): This approach uses mapping functions between the source and target environments to support knowledge transfer. These mapping functions provide a way to align the representations or properties of the data between the two environments, thereby facilitating knowledge transfer (e.g. [56]).

- Representation Transfer (RT): Knowledge is transferred in the form of feature representations learned during the training process. Feature representations play a crucial role in capturing relevant patterns and information from the input data. By transferring representations learned in a source domain to a target domain, the agent can benefit from shared knowledge that may be useful for the target task.

Zhungadi et al [33] also provided 6 questions that someone needs to consider before choosing a transfer learning method for their study. The questions are

1. What knowledge is being transferred?
2. Which RL frameworks are compatible with the transfer learning approach?
3. What is the difference between the source and target domains?
4. What information is available in the target domain?
5. How pattern-efficient is the transfer learning approach?
6. What are the goals of transfer learning?

2.3.1 Deep Q from Demonstration

Deep Q from Demonstration (DQfD) [34] was introduced with the aim of using the data already collected in existing systems. More specifically, Hester targeted systems that do not have an accurate simulation, but have already been used by human operators. In these cases, data is collected from the human operator to provide demonstrations for RL training. The aim was to create a methodology that learns as much as possible from the demonstration before running in the real system. This method is divided into two parts, the pre-training phase and the interaction with the environment.

In the pre-training phase, the agent samples mini-batches from the demonstration data and updates the network by applying four losses, a one-step double Q-learning loss, an n-step double Q-learning loss, a supervised large margin classification loss, and an L2 regularisation loss. The supervised loss is used to aid pre-training, as the demonstration data covers a narrow part of the state space and does not represent all possible actions, many state actions have no data to give them realistic values. If the model were pre-trained using only Q-learning, the network would update towards the maximum value of the next state and the network would propagate the highest of these unfounded values throughout the Q-function.

The authors add a large margin classification loss based on Eq 20, where a_E is the action taken by the expert demonstrator in state s and$l(a_E, a)$ is a margin function that is 0 if $a = a_E$ and positive otherwise. The aim was to force the values of the other actions to be at least one margin lower than the values of the demonstrator's action. Adding this loss grounds the values of the unseen actions to reasonable values and makes the greedy policy induced by the value function imitate the demonstrator. By adding only this supervised loss, the algorithm in pre-training will have nothing to prevent the values between successive states and the Q-network will not satisfy the Bellman equation. This is essential to improve the policy on-line with temporal difference (TD) learning.

$$J_E(Q) = \max_{a \in A}[\, Q(s, a) + l(a_E, a)] - Q(s, a_E) \qquad (20)$$

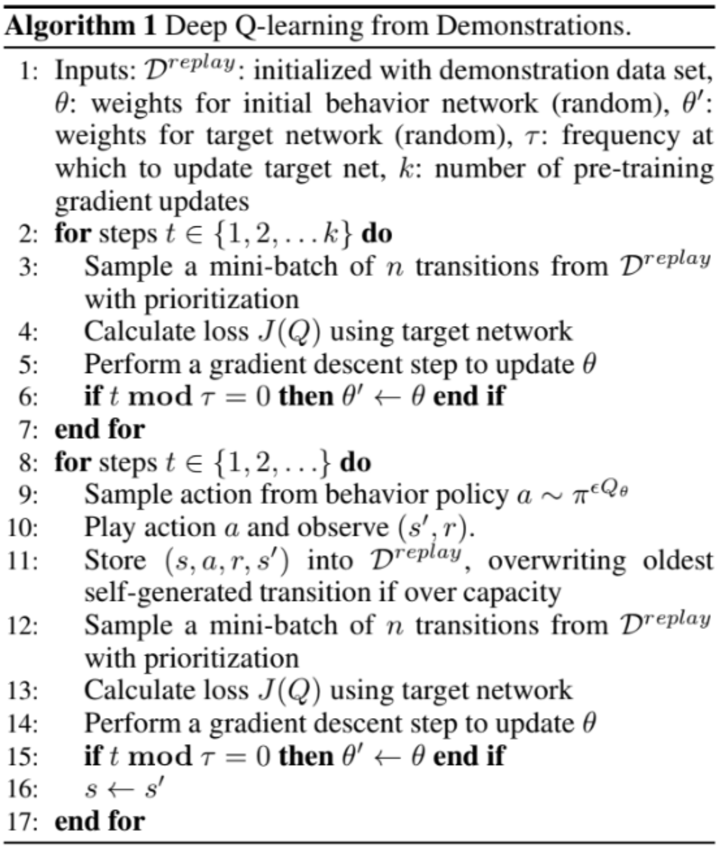
In order to propagate the values of the expert trajectory to all earlier states, an n-step return (with n=10) has been added, which leads to better training. The n-step return is in Eq 21.

$$r_t + \gamma r_{t+1} + \cdots + \gamma^{n-1} r_{t+n-1} + max_a \gamma^n Q(s_{t+n}, a) \qquad (21)$$

The authors also added a L2 regression loss applied to the weights and biases of the network to prevent it from overfitting on the relatively small demonstration dataset. The total loss used to update the network is a combination of all four losses Eq 22. The λ parameters control the weighting between the losses.

$$J(Q) = J_{DQ}(Q) + \lambda_1 J_n(Q) + \lambda_2 J_E(Q) + \lambda_3 J_{L2}(Q) \qquad (22)$$

After the pre-training phase, the agent begins to interact with the environment and collect self-generated data. All data is collected in a replay buffer (D^{replay}), which contains both self-generated and demonstrated data. When the replay buffer is full, the agent overwrites older self-generated data. The agent never overwrites demonstration data. For proportionally prioritized sampling, various small positive constants, e_a and e_d, are added to the priorities of the agent and demonstration transitions to control the relative sampling of demonstration data versus agent data. All losses are applied to the demonstration data in both phases, while the supervised loss is not applied to self-generated data (2 = 0). The final algorithm is Figure 2.8.



Algorithm 1 Deep Q-learning from Demonstrations.

```
Inputs: $\mathcal{D}^{replay}$: initialized with demonstration data set, $\theta$: weights for initial behavior network (random), $\theta'$: weights for target network (random), $\tau$: frequency at which to update target net, $k$: number of pre-training gradient updates
for steps $t \in \{1, 2, \ldots k\}$ do
    Sample a mini-batch of $n$ transitions from $\mathcal{D}^{replay}$ with prioritization
    Calculate loss $J(Q)$ using target network
    Perform a gradient descent step to update $\theta$
    if $t \bmod \tau = 0$ then $\theta' \leftarrow \theta$ end if
end for
for steps $t \in \{1, 2, \ldots\}$ do
    Sample action from behavior policy $a \sim \pi^{\epsilon Q_\theta}$
    Play action $a$ and observe $(s', r)$.
    Store $(s, a, r, s')$ into $\mathcal{D}^{replay}$, overwriting oldest self-generated transition if over capacity
    Sample a mini-batch of $n$ transitions from $\mathcal{D}^{replay}$ with prioritization
    Calculate loss $J(Q)$ using target network
    Perform a gradient descent step to update $\theta$
    if $t \bmod \tau = 0$ then $\theta' \leftarrow \theta$ end if
    $s \leftarrow s'$
end for
```

Figure 2.8 Representation of the Q-learning from Demostration algorithm.

2.4 Subjective measures

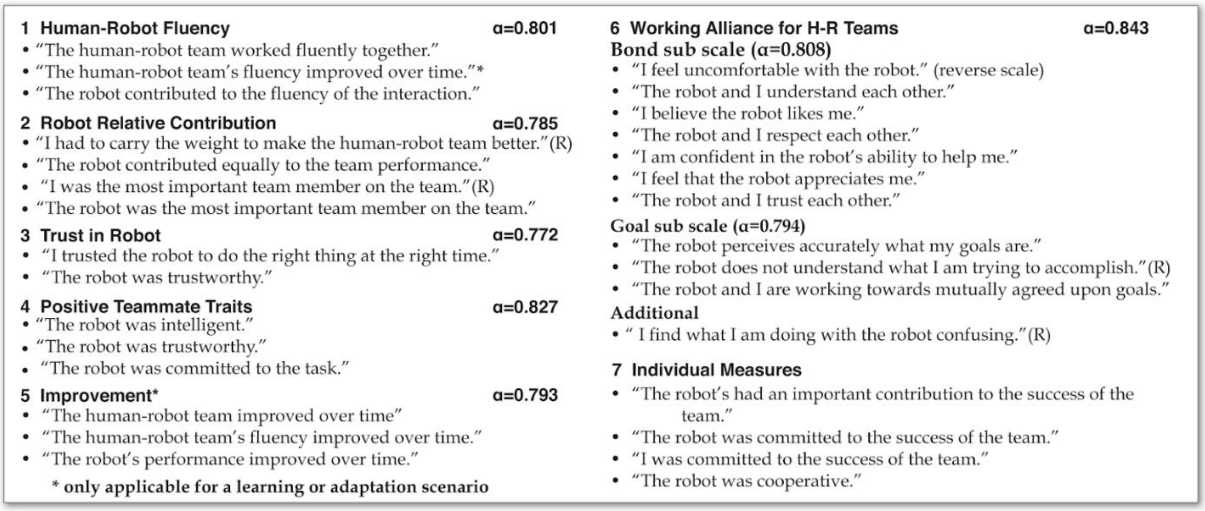
In the context of a co-learning task, the experience and the behaviour of human participants is affected not only by the "objective performance" of the agent, but also by their personal beliefs and perceptions towards the AI agent. The latter is captured through the use of subjective measures.

Subjective measures refer to evaluations based on personal opinions, perceptions or individual experiences. They aim to capture aspects that are difficult to measure objectively, such as emotions, attitudes, preferences, satisfaction, or quality of experience. Many studies have presented subjective measures to evaluate better the experience of the human. Works in conversation agents [35], dialogue systems [36] and explainable Ai [37] are some examples of focus for using subjective measures. In each study, the subjective measures are altered to better reflect the objective in which humans and agents collaborate. Riedelbauch et al [38], in their work presented many of the questionnaires that have derived from studies in subjective measures. They also categorized these questionnaires based on the focused subject. Our focus is on the validation of human-agent teaming fluency, so we focused on the proposed questionnaire by Hoffman. Riedelbauch also presents what human factors and teamwork aspects each questionnaire covers.

Hoffman [39], in his work about evaluating fluency in HRC, presented a complete questionnaire to evaluate seven different aspects of collaboration. These measures were gathered from existing works and his proposed additions to the fluency scale. The measures are:

- Human-Robot Fluency
- Robot Relative Contribution
- Trust in Robot
- Positive Teammate Traits
- Improvement
- Working Alliance for human-robot teams
- Individual measures

The questions for each category are shown in Figure 2.9



1 **Human-Robot Fluency** α=0.801

- "The human-robot team worked fluently together."
- "The human-robot team's fluency improved over time."*
- "The robot contributed to the fluency of the interaction."

2 **Robot Relative Contribution** α=0.785

- "I had to carry the weight to make the human-robot team better."(R)
- "The robot contributed equally to the team performance."
- "I was the most important team member on the team."(R)
- "The robot was the most important team member on the team."

3 **Trust in Robot** α=0.772

- "I trusted the robot to do the right thing at the right time."
- "The robot was trustworthy."

4 **Positive Teammate Traits** α=0.827

- "The robot was intelligent."
- "The robot was trustworthy."
- "The robot was committed to the task."

5 **Improvement*** α=0.793

- "The human-robot team improved over time"
- "The human-robot team's fluency improved over time."
- "The robot's performance improved over time."

*** only applicable for a learning or adaptation scenario**

6 **Working Alliance for H-R Teams** α=0.843

Bond sub scale (α=0.808)

- "I feel uncomfortable with the robot." (reverse scale)
- "The robot and I understand each other."
- "I believe the robot likes me."
- "The robot and I respect each other."
- "I am confident in the robot's ability to help me."
- "I feel that the robot appreciates me."
- "The robot and I trust each other."

Goal sub scale (α=0.794)

- "The robot perceives accurately what my goals are."
- "The robot does not understand what I am trying to accomplish."(R)
- "The robot and I are working towards mutually agreed upon goals."

Additional

- " I find what I am doing with the robot confusing."(R)

7 **Individual Measures**

- "The robot's had an important contribution to the success of the team."
- "The robot was committed to the success of the team."
- "I was committed to the success of the team."
- "The robot was cooperative."

Figure 2.9 Hoffman's [39] proposed subjective measures.

In each study the questionnaire changes to fit the objective. Paliga [32] used some of these measures in her work about evaluating the relationship between human-robot interaction fluency with job performance and job satisfaction. Similarly, Yang et al [57] used some measures to evaluate a classification in a handover task. Tsitos used six of these measures, excluding the individual measure, in order to evaluate the difference in experience by using a transfer learning methodology to enhance the performance of a human-robot team.

2.5 Personality

In HAC the personality and personal views of the humans play a big role to their approach to collaboration. This also affects how each human answers the subjective measure.

In [40, 42], the Big Five personality trait questionnaire is presented to better understand the experience of the human in HAC. Furthermore, in [41] the arthors used a custom scale to capture personality traits for the same purpose. The use of a custom questionnaire provides a better optimization to the objective of the study. While this is true, using a well-established scale like the big five that has decades of studies guarantees a better validity to the results. In addition to this, Schepman [43] presented in his work a scale about attitudes toward AI. He also presented results about the correlation with the Big Five personality traits. This scale can provide a view of the attitude in which the human approaches a collaboration with an agent. In addition to these, we will review the use of the personal view questionnaire about personal human values.

2.5.1 Big Five

In order to capture some aspects of the participant's personality, we focus on the Big Five personality trait questionnaire [58]. Also known as the Five Factor Model (FFM) or the OCEAN model, the Big Five is a widely used psychological framework for assessing personality traits. It is based on the idea that personality can be described and categorised into five basic dimensions. Each dimension represents a broad human trait and is measured on a continuum from low to high levels of the trait. These five dimensions are

- Openness to Experience (sometimes referred to as Intellect/Imagination): This dimension reflects a person's inclination towards new experiences, imagination, and intellectual curiosity. Individuals with increased openness tend to be creative, adventurous, and open-minded, while those decresed openness are often more conventional and prefer routine.

- Conscientiousness: This dimension refers to a person's level of organization, responsibility, and self-discipline. Highly conscientious individuals are typically diligent, dependable, and detail-oriented, whereas those low levels of conscientiousness may be more spontaneous and less focused on following rules.

- Extraversion: This dimension captures the extent to which a person seeks social interaction and stimulation. Extraverts are typically outgoing, energetic, and sociable, whereas introverts tend to be more reserved and prefer solitude or smaller social settings.

- Agreeableness: This dimension reflects an individual's tendency to be cooperative, compassionate, and considerate towards others. Highly agreeable individuals are often empathetic, kind, and willing to compromise, while those low levels of agreeableness may be more competitive and sceptical of others' motives.

- Neuroticism (sometimes referred to as Emotional Stability): This dimension measures a person's emotional stability and resilience to stress. High levels of neuroticism are associated with anxiety, mood swings, and a higher susceptibility to negative emotions. Conversely, individuals low levels of neuroticism tend to be emotionally stable, calm, and less reactive to stressful situations

The Big Five personality questionnaire is the result of decades of research by several psychologists. There are many versions of the questionnaire, varying greatly in the number of items and the wording of these items. We focus on the version of Goldberg et al [44] which was translated by Tsaousi et al [45]. This version contains 50 items, 10 for each dimension. All the items, categorised in each dimension, are shown in Table 2.3.

Table 2.3 Big Five Questions Categorized on the personality traits

Greek	English	Pos/Neg
Openness to Experience		
Έχω ένα πλούσιο λεξιλόγιο.	Have a rich vocabulary	Pos
Δυσκολεύομαι να κατανοήσω αφηρημένες ιδέες.	Have difficulty understanding abstract ideas	Neg
Έχω ζωηρή (ζωντανή) φαντασία.	Have a vivid imagination.	Pos
Δεν ενδιαφέρομαι για αφηρημένες ιδέες.	Am not interested in abstract ideas	Neg
Έχω εξαιρετικές ιδέες.	Have excellent ideas	Pos
Δεν έχω καλή φαντασία.	Do not have a good imagination	Neg
Είμαι γρήγορος/η στο να καταλαβαίνω πράγματα.	Am quick to understand things	Pos
Χρησιμοποιώ δύσκολες λέξεις.	Use difficult words	Pos
Αφιερώνω χρόνο για να αξιολογώ τα πράγματα (που κάνω).	Spend time reflecting on things	Pos
Είμαι γεμάτος/η ιδέες.	Am full of ideas	Pos

Conscientiousness		
Είμαι πάντοτε προετοιμασμένος	Am always prepared	Pos
Αφήνω τα πράγματά μου ολόγυρα.	Leave my belongings around	Neg
Δίνω προσοχή στις λεπτομέρειες	Pay attention to details	Pos
Τα κάνω άνω κάτω	Make a mess of things	Neg
Κάνω τις «αγγαρείες» αμέσως.	Get chores done right away	Pos
Συχνά ξεχνώ να βάζω τα πράγματα πίσω στη σωστή τους θέση.	Often forget to put things back in their proper place	Neg
Μου αρέσει η τάξη.	Like order	Pos
Αποφεύγω αυτά που πρέπει να κάνω (τα καθήκοντά μου).	Shirk my duties	Neg
Ακολουθώ ένα πρόγραμμα.	Follow a schedule	Pos
Είμαι ακριβής στη δουλειά μου.	Am exacting in my work	Pos
Extraversion		
Είμαι η ζωή σε ένα πάρτι.	Am the life of the party.	Pos
Δεν μιλώ πολύ.	Don't talk a lot	Neg
Αισθάνομαι άνετα όταν βρίσκομαι ανάμεσα σε ανθρώπους.	Feel comfortable around people	Pos
Προτιμώ να μένω στο παρασκήνιο	Keep in the background	Neg
Αρχίζω συζητήσεις.	Start conversations	Pos
Έχω ελάχιστα πράγματα να πω.	Have little to say	Neg
Μιλώ με πολλούς διαφορετικούς ανθρώπους στα πάρτι.	Talk to a lot of different people at parties.	Pos

Δεν μου αρέσει να προσελκύω την προσοχή πάνω μου.	Don't like to draw attention to myself	Neg
Δεν με ενοχλεί να είμαι το επίκεντρο της προσοχής.	Don't mind being the centre of attention	Pos
Είμαι ήσυχος/η όταν βρίσκομαι ανάμεσα σε ξένους.	Am quiet around strangers	Neg
Agreeableness		
Αισθάνομαι μικρό ενδιαφέρον για τους άλλους.	Feel little concern for others	Neg
Ενδιαφέρομαι για τους ανθρώπους.	Am interested in people	Pos
Προσβάλλω τους άλλους.	Insult people	Neg
Συμπάσχω με τα συναισθήματα των άλλων.	Sympathize with others' feelings	Pos
Δεν ενδιαφέρομαι για τα προβλήματα των άλλων.	Am not interested in other people's problems	Neg
Έχω μαλακή καρδιά.	Have a soft heart	Pos
Δεν ενδιαφέρομαι πραγματικά για τους άλλους ανθρώπους.	Am not really interested in others	Neg
Βρίσκω χρόνο για τους άλλους.	Take time out for others	Pos
Αισθάνομαι τα συναισθήματα των άλλων.	Feel others' emotions	Pos
Κάνω τους ανθρώπους να αισθάνονται άνετα.	Make people feel at ease	Pos

Neuroticism		
Αγχώνομαι εύκολα.	Get stressed out easily	Neg
Είμαι χαλαρός/ή τις περισσότερες φορές.	Am relaxed most of the time	Pos
Ανησυχώ για διάφορα πράγματα.	Worry about things	Neg
Σπάνια νοιώθω μελαγχολία.	Seldom feel blue	Pos
Ενοχλούμαι εύκολα.	Am easily disturbed	Neg
Αναστατώνομαι εύκολα.	Get upset easily	Neg
Η διάθεσή μου αλλάζει διαρκώς.	Change my mood a lot	Neg
Έχω συχνές εναλλαγές στη διάθεσή μου.	Have frequent mood swings	Neg
Εκνευρίζομαι εύκολα.	Get irritated easily	Neg
Συχνά αισθάνομαι μελαγχολικά.	Often feel blue	Neg

2.5.2 Perception of AI

In his study, Schepman [43] built a tool to capture the general attitude towards AI, and created 3 questionnaires.

- The first was divided into two parts with general questions about the characteristics of AI, one with positive questions like "*There are many beneficial applications in artificial intelligence*" and the second with negative questions like "*The rise of artificial intelligence poses a threat to people's job security*". This questionnaire has a total of 32 questions.
- The second asked 42 questions about participants' comfort with AI applications, such as "*Translating speech into different languages in real time*" and "*Helping a police force predict the risk of reoffending when making bail decisions*".
- The third contained the same questions as the second part, but on a different scale about the capabilities of specific AI applications compared to humans.

From the first questionnaire, 20 items remained after removing 7 because of high correlation with other questions and 5 because of exploratory factor analysis on Jamovi[65]. This questionnaire has 12 positive and 8 negative questions. We refer to this questionnaire as *AI Attitude Scale* and all questions are in Table 2.4.

In this paper, Schepman validates the effectiveness of using this questionnaire to capture general attitudes towards AI by cross-validating it with the second and third questionnaires, which are more application-specific, using a sample size of 100. To demonstrate the use of the final 20-item questionnaire, Schepman published a second paper in which they showed whether psychological factors could correlate with general attitudes toward AI [46]. He had a sample size of 300 and used the Big Five personality traits to extract personal characteristics.

Table 2.4 AI Attitude scale questions.

Greek	**English**	Pos/Neg
Θα προτιμούσα να αλληλεπιδρώ με ένα σύστημα ΤΝ παρά με έναν άνθρωπο για τις συναλλαγές της καθημερινής ζωής.	For routine transactions, I would rather interact with an artificially intelligent system than with a human.	Pos
Η ΤΝ μπορεί να προσφέρει νέες οικονομικές ευκαιρίες για τη χώρα μου.	Artificial Intelligence can provide new economic opportunities for this country.	Pos
Οργανισμοί χρησιμοποιούν την ΤΝ με ανήθικο τρόπο.	Organisations use Artificial Intelligence unethically.	Neg
Τα συστήματα ΤΝ μπορούν να βοηθήσουν τους ανθρώπους να αισθάνονται πιο ευτυχισμένοι.	Artificially intelligent systems can help people feel happier.	Pos
Είμαι εντυπωσιασμένος από το τι μπορεί να κάνει η ΤΝ.	I am impressed by what Artificial Intelligence can do.	Pos
Νομίζω ότι τα συστήματα ΤΝ κάνουν πολλά λάθη.	I think artificially intelligent systems make many errors.	Neg
Ενδιαφέρομαι να χρησιμοποιώ συστήματα ΤΝ στην καθημερινή μου ζωή.	I am interested in using artificially intelligent systems in my daily life.	Pos
Θεωρώ ότι η ΤΝ είναι κακόβουλη.	I find Artificial Intelligence sinister.	Neg
Η ΤΝ μπορεί να πάρει τον έλεγχο από τους ανθρώπους.	Artificial Intelligence might take control of people.	Neg
Νομίζω ότι η ΤΝ είναι επικίνδυνη.	I think Artificial Intelligence is dangerous.	Neg

Η ΤΝ μπορεί να έχει θετικές επενέργειες στην ευημερία των ανθρώπων.	Artificial Intelligence can have positive impacts on people’s wellbeing.	Pos
Η ΤΝ είναι συναρπαστική.	Artificial Intelligence is exciting.	Pos
Θα σας ήμουν ευγνώμων αν μπορούσατε να επιλέξετε Συμφωνώ απόλυτα.	An artificially intelligent agent would be better than an employee in many routine jobs.	Pos
Σε πολλές εργασίες ρουτίνας ένα σύστημα ΤΝ θα ήταν καλύτερο από έναν άνθρωπο.	There are many beneficial applications of Artificial Intelligence.	Pos
Ανατριχιάζω από δυσφορία όταν σκέφτομαι τις μελλοντικές χρήσεις της ΤΝ.	I shiver with discomfort when I think about future uses of Artificial Intelligence.	Neg
Τα συστήματα ΤΝ μπορούν να αποδώσουν καλύτερα από τους ανθρώπους.	Artificially intelligent systems can perform better than humans.	Pos
Μεγάλο μέρος της κοινωνίας θα επωφεληθεί από ένα μέλλον γεμάτο ΤΝ.	Much of society will benefit from a future full of Artificial Intelligence	Pos
Θα ήθελα να χρησιμοποιήσω ΤΝ στη δική μου δουλειά.	I would like to use Artificial Intelligence in my own job.	Pos
Άνθρωποι σαν και μένα θα υποφέρουν αν η ΤΝ χρησιμοποιείται όλο και περισσότερο.	People like me will suffer if Artificial Intelligence is used more and more.	Neg
Η ΤΝ χρησιμοποιείται για την κατασκοπεία των ανθρώπων.	Artificial Intelligence is used to spy on people	Neg

2.5.3 PVQ 21

Shalom H. Schwartz is a social psychologist, cross-cultural researcher, and creator of the theory of basic human values. In his work, Schwartz identified ten basic human values, each distinguished by its underlying motivation or goal. To measure these values, he constructed the Schwartz Value Survey (SVS) [47], which has been used in studies in over 65 countries and contains 56 items, and later introduced the Portrait Values Questionnaire PVQ [48] as a simpler version of the SVS to make it more understandable. There are several versions of the PVQ that have been developed over time to meet different research needs. One of the most commonly used versions is the PVQ-21, which contains 21 items, while other versions include the PVQ-40 and PVQ-57.

All versions measure the ten basic human values introduced by Schwartz and the main difference is the number of questions used to measure each value. For our purposes, we focus on the PVQ-21 [49] because of its smaller size, so that our participants would not get tired of the questionnaire and lose focus. While the PVQ-40 and PVQ-57 can provide a more comprehensive assessment of the values, the PVQ-21 has been shown to provide reliable results and is sufficient for our work.

The ten basic human values identified by Schwartz represent different motivational goals and aspirations. Here are the ten values with a brief explanation of each one

- Self-Direction: This value emphasizes independent thought, creativity, and autonomy. Individuals who prioritize self-direction value their freedom of choice, enjoy exploring new ideas, and strive for personal growth.
- Stimulation: This value reflects a desire for excitement, novelty, and variety. People who prioritize stimulation seek adventure, enjoy taking risks, and actively seek out new experiences.
- Hedonism: Hedonism represents the pursuit of pleasure and enjoyment in life. Individuals who prioritize hedonism seek fun, seek gratification, and prioritize their own happiness and pleasure.
- Achievement: This value focuses on personal success through demonstrating competence, gaining recognition, and striving for

excellence. People who prioritize achievement value ambition, set high goals, and are driven to succeed.

- Power: Power values involve the desire for control, influence, and social status. Individuals who prioritize power seek leadership positions, enjoy being in control, and strive for dominance and authority.
- Security: This value emphasizes safety, stability, and order. People who prioritize security value a sense of stability, seek predictability, and prioritize the avoidance of risks and uncertainties.
- Conformity: Conformity values centre around adhering to social norms, traditions, and expectations. Individuals who prioritize conformity value obedience, respect for authority, and strive to fit in with societal expectations.
- Tradition: Tradition values reflect respect for customs, cultural heritage, and traditional values. People who prioritize tradition value maintaining customs, preserving societal norms, and showing respect for cultural heritage.
- Benevolence: Benevolence values revolve around caring for others, empathy, and concern for the welfare of others. Individuals who prioritize benevolence value kindness, compassion, and strive to promote the well-being of others.
- Universalism: Universalism values focus on social justice, equality, and concern for the welfare of all people. People who prioritize universalism value justice, equality, environmental sustainability, and strive to make the world a better place.

All the questions, both in English and in Greek, are presented in Table 2.5 according to the value that each item contributes to its measurements.

Table 2.5 Personal Values Questionnaire (PVQ) questions

English Question (Male Version)	Greek Question (Combine Male and Female Version)
BENEVOLENCE	
It's very important to him to help the people around him. He wants to care for other people.	Είναι πολύ σημαντικό για αυτήν/όν να βοηθά τους ανθρώπους που την/τον περιβάλλουν. Ενδιαφέρεται για το καλό των άλλων.
It is important to him to be loyal to his friends. He wants to devote himself to people close to him.	Είναι σημαντικό για αυτήν/όν να είναι πιστή/ος στους φίλους της/του. Θέλει να αφοσιώνεται στους ανθρώπους που βρίσκονται κοντά της/του
UNIVERSALISM	
He thinks it is important that every person in the world be treated equally. He wants justice for everybody, even for people he doesn't know.	Πιστεύει πως είναι σημαντικό όλοι οι άνθρωποι στον κόσμο να αντιμετωπίζονται ισότιμα. Πιστεύει ότι όλοι πρέπει να έχουν ίδιες ευκαιρίες στη ζωή
It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.	Της/Του είναι σημαντικό, να ακούει ανθρώπους με διαφορετικές απόψεις από τις δικές της/του. Ακόμα και όταν διαφωνεί θέλει να μπορεί να τους κατανοεί.
He strongly believes that people should care for nature. Looking after the environment is important to him.	Πιστεύει ακράδαντα ότι οι άνθρωποι πρέπει να προστατεύουν τη φύση. Η προστασία του περιβάλλοντος είναι πολύ σημαντική για αυτήν/όν.

SELF-DIRECTION	
Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.	Είναι πολύ σημαντικό για αυτήν/όν να έχει καινούργιες ιδέες και να είναι δημιουργική/ος. Τον/Την αρέσει να κάνει πράγματα με τον δικό της/του πρωτότυπο τρόπο.
It is important to him to make his own decisions about what he does. He likes to be free to plan and to choose his activities for himself.	Είναι σημαντικό για αυτήν/ον να λαμβάνει τις δικές της/του αποφάσεις για ότι πρόκειται να κάνει. Θέλει να είναι ελεύθερη/ος και να μην εξαρτάται από άλλους.
STIMULATION	
He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.	Της/Του αρέσουν οι εκπλήξεις και θέλει να κάνει πάντα καινούρια πράγματα. Πιστεύει ότι στη ζωή είναι σημαντικό να κάνεις πολλά διαφορετικά πράγματα
He looks for adventures and likes to take risks. He wants to have an exciting life.	Αναζητεί την περιπέτεια και είναι ριψοκίνδυνος. Θέλει η ζωή της/του να είναι συναρπαστική
HEDONISM	
Having a good time is important to him. He likes to “spoil” himself.	Η καλοπέραση είναι σημαντική για αυτήν/όν. Της/Του αρέσει να καλομαθαίνει τον εαυτό της/του.
He seeks every chance he can to have fun. It is important to him to do things that give him pleasure	Πάντα ψάχνει ευκαιρία για γλέντι. Είναι σημαντικό για αυτήν/όν να κάνει πράγματα που την/τον ευχαριστούν.

ACHIEVEMENT	
It is very important to him to show his abilities. He wants people to admire what he does.	Είναι πολύ σημαντικό γι’ αυτήν/όν να δείχνει τις ικανότητές της/του. Θέλει ο κόσμος να θαυμάζει αυτό που κάνει.
Being very successful is important to him. He likes to impress other people.	Η επιτυχία της/του, είναι πολύ σημαντική για την/τον ίδια/ιο. Ελπίζει ότι ο κόσμος θα αναγνωρίσει τα επιτεύγματά της/του.
POWER	
It is important to him to be rich. He wants to have a lot of money and expensive things.	Είναι σημαντικό γι’ αυτήν/όν να είναι πλούσια/ιος. Θέλει να έχει πολλά λεφτά και ακριβά πράγματα.
It is important to him to be in charge and tell others what to do. He wants people to do what he says.	Είναι σημαντικό για αυτήν/όν να την/τον σέβονται οι άλλοι. Θέλει οι άλλοι να κάνουν αυτό που τους λέει.
SECURITY	
It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.	Είναι πολύ σημαντικό για αυτήν/όν να ζει σε ένα ασφαλές περιβάλλον. Αποφεύγει οτιδήποτε θα μπορούσε να θέσει σε κίνδυνο την ασφάλειά της/του.
It is very important to him that his country be safe from threats from within and without. He is concerned that social order be protected.	Είναι πολύ σημαντικό για αυτήν/ον η κυβέρνηση να μπορεί να εγγυηθεί για την ασφάλειά της/του. Θέλει ένα κράτος ισχυρό, ικανό να προστατεύσει τους πολίτες του.

CONFORMITY	
He believes that people should do what they're told. He thinks people should follow rules at all times, even when no-one is watching.	Πιστεύει ότι οι άνθρωποι πρέπει να κάνουν αυτό που τους λένε. Πιστεύει ότι οι άνθρωποι πρέπει πάντα να τηρούν τους κανόνες, ακόμα και όταν κανείς δεν τους βλέπει.
It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.	Είναι σημαντικό γι’ αυτήν/όν να συμπεριφέρεται πάντα σωστά. Προσπαθεί να αποφύγει οτιδήποτε θα έλεγε κανείς ότι είναι λάθος.
TRADITION	
He thinks it's important not to ask for more than what you have. He believes that people should be satisfied with what they have.	Είναι σημαντικό γι’ αυτήν/όν να είναι ταπεινή/ός και μετριόφρων. Προσπαθεί να μην τραβά την προσοχή.
Religious belief is important to him. He tries hard to do what his religion requires.	Η παράδοση είναι κάτι πολύ σημαντικό για αυτήν/όν. Προσπαθεί να τηρεί τα ήθη και τα έθιμα.

3 Methodology

In this chapter, we present the overall methodology of the present thesis. Specifically, in Section 3.1, the Human-Agent collaborative task is presented. Section 3.2 specifies the setting of the discrete SAC agent. In Section 3.3 the two groups of the collaborative study are discussed. Then Section 3.4 presents the methodology of initializing the models of the discrete SAC agent. In Section 3.5 the collaboration measures are introduced, followed by the personality questionnaires in Section 3.6 and the description of the participation process in Section 3.7

3.1 Human-Agent collaborative task:

The human-agent collaborative (HAC) task was based on previous work [31]. During the HAC task, a human participant collaborated with an RL agent to control a virtual tray. The goal was to move a ball from a starting position to a target state. The virtual environment, shown in Figure 3.1, contained a square 10x10 unit tray, enclosed in a 1-unit high barrier around all four sides. Additionally, there were two diagonally placed obstacle walls with a 1.4-unit 'gate' in the centre. There was a 1-unit diameter hole in the bottom right corner, which served as a target for the rolling ball to fall into. The ball was 1-unit in diameter. The three possible starting positions are in Figure 3.1 and were:

- bottom right
- top right
- top left.

All starting positions are above the two obstacles. In order for the target state to be achieved the ball must fall into the hole, not roll over it.

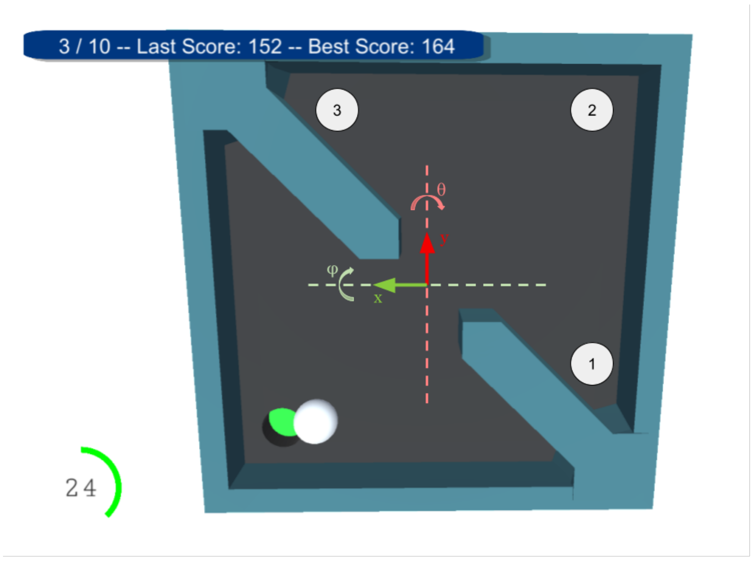


Figure 3.1. HAC virtual environment. It includes the main platform that the human and the agent control, the white ball they need to move to the target, and the green hole that the ball needs to fall into to reach the target. At the top left there is information for the player about the current game, last score, and best score. At the bottom left is the time left of the current game. The figure also depicts the three possible starting locations. The starting positions are; one bottom right, two top right and three top left. The tray rotates around the two axes, with the player controlling the y-axis and the agent controlling the x-axis. The angels (θ,φ) depict the angels each member applies in their axis.

The game consists of 200 control frames, where each control frame represents a period during which an action is continuously applied and lasts for 200ms. Therefore, the total game duration was 40 seconds. The game ended when the ball reached the goal or when the 40 seconds elapsed. The agent receives a reward of -1 for each control frame, except for the target state, where it receives a reward of 10.

$$r(s,a) = \begin{cases} +10, \textit{goal reached} \\ -1, \qquad \textit{otherwise} \end{cases}$$

In the task, the human participant was responsible for controlling the rotation of the tray around the axis y, while the RL agent was responsible for controlling the axis x. This is depicted in Figure 3.1. The human participant was responsible for applying an angle θ on the y-axis and the agent was responsible for applying an angle φ on the x-axis. By rotating the two axes the Human-Agent team aimed to move the ball from its starting position to the target. The human participant controlled the rotations of the tray via keyboard. Both team members could provide three discrete actions.

1. Rotating the tray clockwise
2. Keeping the current rotation angle
3. Rotating the tray counter-clockwise

For the human participants this meant that they had the following options:

- By pressing 'Right Arrow' (>), they rotate the tray clockwise.
- By pressing 'Left Arrow' (<), they rotate the tray counter-clockwise.
- By pressing nothing the tray keeps the current rotation angle.

The RL agent had a 1-dimensional action space of a = {-1,0,1}. Similar to the human actions, the agent actions resulted in counter-clockwise, or clockwise rotation (-1 and 1 accordingly), or maintenance of the tray rotation angle (0).

The agent observed the environment using the following 8-dimensional state space:

- The ball *position* on the tray (x, y)
- The ball *speed* on the x-axis and y-axis (sx, sy)
- The *angle* of the tray around the x-axis and y-axis (φ, θ)
- The *acceleration* of rotation around the x-axis and y-axis (sφ, sθ)

$$\text{state} = (x, y, sx, sy, \varphi, \theta, s\varphi, s\theta)$$

The tray rotates 30 degrees toward both sides, and each discrete action from both members applies a change of around 5 degrees.

Figure 3.2 depicts the directions of the ball based on the angle of the tray. If one of the members idles around a 0-degree angle, the other member can move the ball in a vertical or horizontal way. To move the ball diagonally, both members need to rotate the platform. Based on the starting positions of Figure 3.1, to reach the target there must be some diagonal movements. Because of this, the human-agent team can reach the target only if they collaborate.

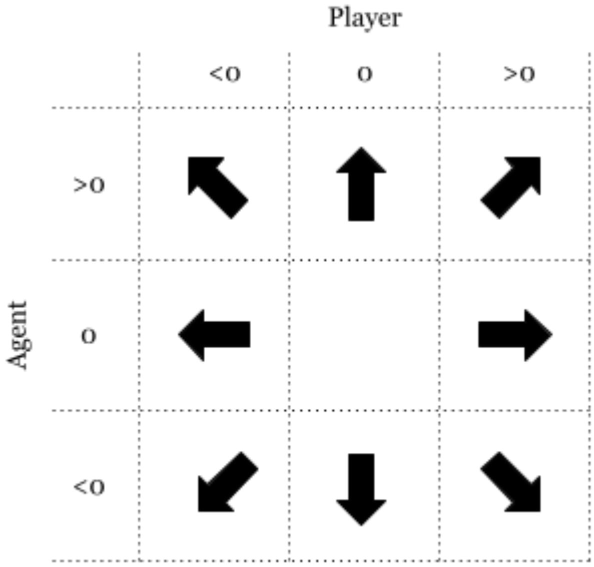


Figure 3.2 Movement of the ball base of angles of the tray.

To facilitate the needs of our study we made some changes to the original setup [31]. Our changes were focused on the information the participants had and keyboard controls. First, we changed the information provided in the top left of the screen (see Figure 3.1). We removed the live score that was displayed before and added the current game in the block, the last score, and the best score achieved in the entire process. Secondly, during the training session, a progress bar appeared on the screen. During the training sessions, the participants need to answer some questions and a progress bar could be distracting. As seen in Fig 3.3, during the training session, the participants now only see a *“Please wait” message*. The third change can be found in the restart process after the training session. Initially, upon completing the training session, the game would restart immediately.

To allow as much time as needed to answer any questions, the participants were given the control of restarting the game. After completing the training session, the message "Press Space to Continue" appears on the screev as seen in Figure 3.3. Then the player can press ‘space’ on the keyboard to continue playing.

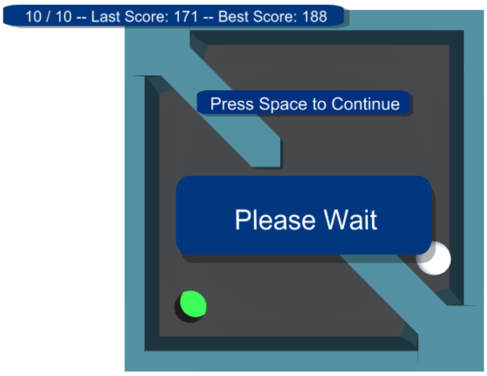


Figure 3.3. The screen when the training has ended and the player can continue the game.

3.2 RL agent

3.2.1 Discrete Soft Actor-Critic

The agent is a discrete SAC [27]. More information about the agent can be found in section 2.1.2. The implementation settings are shown in Table 3.1. All models used the same architecture, consisting of 2 hidden layers of 32 kernels and an output layer of size 3 (the number of available actions). The agent was trained using off-line sessions with 4000 gradient updates, and each update had a batch size of 256. The replay buffer had a size of 1M, to hold as much information as possible.

Discrete SAC has a stochastic policy for decision-making. It uses a softmax function to create the distribution of probabilities of the actions. The action then is selected randomly using the distribution. For exploration, discrete SAC has a soft policy. This means that all actions have a probability to be selected. This probability is calculated using both entropy and the alpha temperature. Entropy is influenced by the probabilities of the actions, while the alpha temperature is a parameter that undergoes training during offline sessions based on entropy and target entropy. The starting value of the alpha temperature is 1. The target entropy is calculated using the Entropy target equation in Table 3.1, where |A| is the number of possible actions. In our case with 3 possible actions the target entropy is 0.679, and it dictates the entropy if all action have the same probability of being chosen.

Table 3.1 SAC discrete study settings.

Hyperparameter	**Value**
Layers	2 fully connected layers, 1 output layer
Fully connected layer units	32, 32, number of moves available: 3
Batch size	256
Replay buffer size	1000000
Discount rate	0.99
Learning rate Actor	0.0003
Learning rate Critic	0.0003
Learning rate Alpha temperature	0.001
Optimizer	Adam
Weight initializer	Xavier initialization
Fixed network updates per off-line training	4000
Loss	Mean squared error
Entropy target	0.98 * (-log (1 / |A|))

3.2.2 Transfer Learning Methodology

The participants were randomly assigned into one of two groups. In the first group, the participants collaborate with an agent that has no transfer learning (TL); this is the No_TL group. The agent is a discrete SAC agent as described in section 3.2.1. In the other group, the collaborative agent includes transfer learning (TL group). This group uses the same discrete SAC agent, but with the addition to a TL algorithm. Our proposed TL algorithm is based on the DQfD algorithm presented in section 2.3.1. Based on this, there were two phases. Initially, there was a pre-training phase, which involved offline training using the demonstration replay buffer. In the second phase, the agent began to collaborate with human players and interact with the environment. During the off-line training the agent was trained with both the expert demonstration data and the participant-generated data.

The expert demonstrations were collected while the expert player collaborated with the expert agent. It is important to clarify that the term "expert agent" here refers to the agent that collaborated with the expert player and does not imply expertise on the agent's part. The expert player had 30-35 hours of hands-on experience in the collaboration task. The expert collaborated with a discrete SAC agent, as described in the previous section. The demonstrations collected were the action-state transitions of this expert agent during the collaboration with the expert player. All actions that were collected came from winning games.

In the first phase, the new agent was trained in an off-line session using the demonstration replay buffer. Contrary to the original methodology [34], w*e do not use additional loss functions to increase the efficiency of the pre-training session*. The large margin classification loss, which is used to boost the demonstration actions against other possible actions, could hinder the chances of personalisation to the new user. The n-step double Q-learning loss and L2 regularisation loss were intended to boost generalisation and sample efficiency. SAC has shown good sample efficiency and generalization in unseen environments. Both Haarnoja [22] and Christodoulou [27] showed examples of these capabilities. Based on this, we made the decision not to utilize any of these loss functions. Consequently, the new agent underwent a standard offline training session using the demonstration replay buffer.

In the second phase, the agent interacts with the environment and the new player. We use a second replay buffer to collect the self-generated interactions. The agent trains only through off-line sessions. In total, there were four off-line sessions in our setup. In each session, the agent used a percentage of data from the demonstration replay buffer and the rest from the self-generated replay buffer. This works by splitting the data in each batch that the agent uses to update the weights. The percentage was fixed for all 4000 updates in each session and decreased after each session. It starts from 80%, then 60%, 30%, and 10%. For example, in the first session, in each batch of size 256, there were 80% (203) data from the demonstration replay buffer and 20% (51) from the self-generated replay buffer.

3.3 Human-Agent Collaborative Study

3.3.1 Selection of participants and statistics

The study was advertised via electronic communication towards academic PhD and MSc programs. The announcement included general information about what they would be doing in the study, how much time it would take, and some Q&A questions. The participation was on a voluntary basis and did not include any reward for completing the study. The only criteria that would exclude a participant was if they had participated in similar studies before. The group they were assigned to, was selected based on an alternate order.

The final testing sample consisted of 8 participants (male: 7, female: 1, ages: 24-39, avg. age: 31, all right-handed). Additionally, two more participants were recruited during the pilot of our design. Based on the feedback of these two participants we made changes to our initial design. The data from the two participants will not be used in the results, as some of the conditions were different.

3.3.2 Groups

As mentioned in section 3.2.2, the participants were split into two groups. The groups are referred to as No_TL and TL. Both groups played in total of 60 games as follows: 10 games with a random agent and 50 games with the agents of each condition. The games were divided into blocks of 10. Between the blocks, there was an off-line training session.

The procedure for the No_TL group is shown in Figure 3.4. The expert player also followed the same procedure for collecting demonstrations for the transfer learning algorithm. The extra step of the expert's playthrough is shown with the dotted arrows.

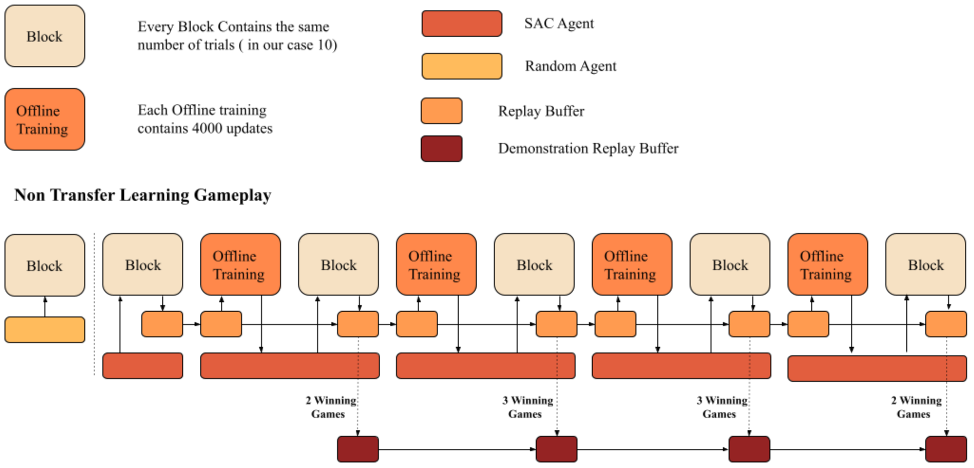


Fig 3.4 Diagram of experiment with no transfer learning.

At the start of the process, participants played a block with a random agent. This block is used as a baseline between the two groups and we refer to this as the baseline block. In the second block, the participants started to collaborate with the discrete SAC agent, we will refer to this block as the first block, and the rest as the second, third, etc. Here all actions of the agent were collected to the replay buffer in order to be used in training. When the participants completed a block, an off-line training session started using the data stored in the replay buffer. The data in the replay buffer were not deleted and were kept in the replay buffer upon completing the off-line training session and the starting of the next block. In total the participants collaborated with the discrete SAC agent in 5 blocks (50 games) and the agent was trained in 4 off-line training sessions.

For the demonstration data, we opted to use games across the collaboration in order to include games in which the expert's agent followed a suboptimal policy. This provides the demonstration buffer with a variety of strategies and in a sense an exploration of some different ways it is possible to win. This, combined with the excellence in generalisation of the discrete SAC agent, should provide a policy that is open to following any new behaviour that the new players may have. In total 10 games were captured from all the blocks except the first. Specifically, as shown in image 3.4, 2 games were captured from the 2nd block, 3 games were captured from the 3rd block, 3 games were captured from the 4th block, and the last 2 games were captured from the 5th block.

The TL group follows a similar structure to the procedure. As shown in Figure 3.5, the participants in the first block collaborated with a random agent. This block, as mentioned before, is used to capture the baseline performance of the two groups. After that, the discrete SAC agent has the first phase of the transfer learning algorithm. In this step, the agent had an off-line training session using only the demonstration data.

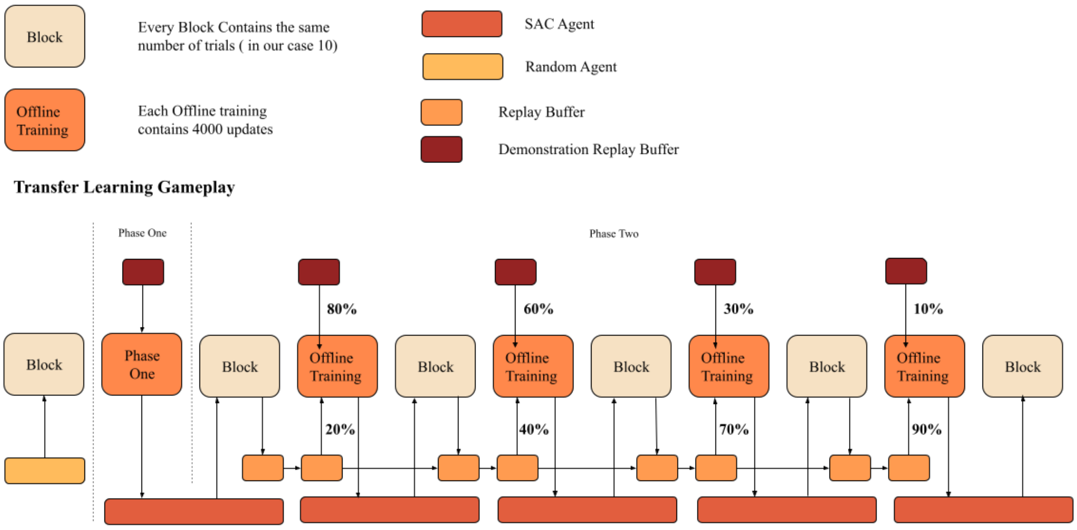


Fig 3.5 Diagram of experiment with transfer learning.

Upon completion of the off-line training session, the participants started to collaborate with this agent in the second block. Here also starts the second phase of the transfer learning algorithm. Again, all interactions during the block were saved in the replay buffer and were not deleted after any training sessions. When the block was completed, the off-line training session started. In each off-line training session, the agent used both data from the demonstration buffer and the participant's self-generated replay buffer. At the first off-line training session, the agent mostly sampled data from the demonstrations’ replay buffer. This decreased in each off-line training session and at the end the agent mostly sampled from the participant's self-generated replay buffer. More information about how this step works in section 3.2.2. Upon completing the off-line training session the participants started the next block. Again as the NO_TL group, in total the participants collaborated with the discrete SAC agent in 5 blocks(50 games) and the agent was trained in 4 off-line training sessions.

3.3.3 Familiarisation and baseline

Each participant in both groups was given a period of familiarisation with the keyboard controls and the environment. This included five periods of 3-minutes with control of both axes. These blocks can also be exploited to assess if the performance of the users in between the two groups are comparable. If any participant were to fail to score 3 out of the 5 games, he/she would be disqualified, as a potential outlier in his/her group. In our sample, none of the participants failed this step. At the same time, based on how fast they scored in this step, we can evaluate the skills of the participants and compare them with their performance in then collaboration.

After finishing the familiarisation games, participants start the main collaboration process. As mentioned in section 3.3.2, the collaboration starts with a baseline block where they cooperate with a random agent. All actions have the same probability. The participants were not informed that this step is with a random agent, or that there is any difference with the rest of the collaboration until the whole process is completed.

3.4 Initialisation of the models

3.4.1 Initialization procedure

Between development and the pilots that we conducted, we observed a variance in the performance of the initialised agent. To better observe these behaviours, we conducted a test where the expert played with 15 initialised agents to see how different the performance each time was. The expert played a block with each agent and the resulting average scores are shown in Figure 3.6. Based on these results, we observe that the teams of expert-initialised agents can perform from almost like having a trained policy (Runs 3, 5) to incapable of scoring (Runs 2, 7, 13, 14).

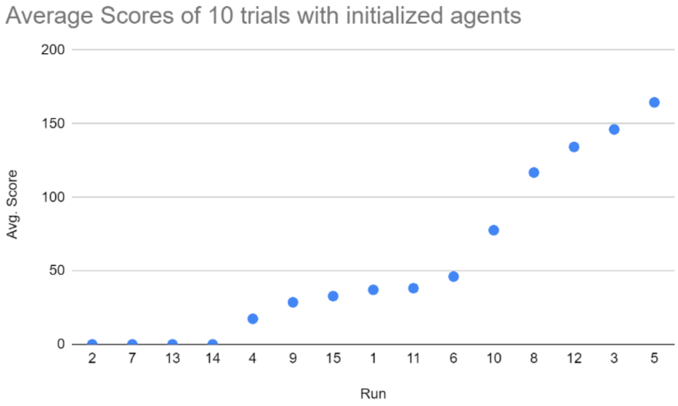


Figure 3.6 Average score of each run with an initialized agent.

In our study, we aim to observe the difference in performance and experience between the two groups. We wanted the main differences between all the participants to be their personalities, skills and approach to collaboration. A solution to the variance in the initialization would be to increase our sample enough in order for this variance to be absorbed. This is the ideal solution, but almost impossible within the reach of our study, due to the duration and thus commitment required from the participants. Typical sample sizes in studies like ours are usually around 10-20 people per group.

Another solution would be to test different initialization processes that may produce a smaller variance. While there are examples that we could follow [50], we opted not to. The main problem with this is that any fundamental change in even the initialization will change the entire co-learning experience. There are already works [29, 30] that used the same or similar settings. So any change in this level would not allow us to make any comparison with those works. Based on the above, our solution was to use the same initialization to all participants.

We selected to use the median performant agent from the 15 we tested in fig 3.6. The final agent we used was *Run 1* which performed with an average score of 37.1. This initialization is used in the No_TL group, the expert, and also the TL group before the pretraining with the demonstration data. The benefit of this is that now all participants will have the same starting experience, meaning that the rest of the co-learning process is affected to a greater degree by their approach to the collaboration.

Based on this logic, we follow a similar approach to the TL group. During the offline training in the first phase with the demonstration data, the same initialization as the one used in the No_TL group is employed. The demonstration data are the same for all participants in the TL group. Additionally, we conducted the first training, with the demonstration data once and used the resulting agent for all the participants.

3.4.2 The game experience

As mentioned in the above section, all participants of the No_TL group and the agent collaborated with the same initialised agent. Through the expert's collaboration with the agent, we can observe its behaviour and the strategy that the expert followed.

For information about how the environment work and the state space of the agent see section 3.1. We separate the environment into two areas; above the obstacles(upper area) and below the obstacles (goal area). The purpose of the game is to move the ball from the upper area through the middle “gate” into the goal area and finally drop the ball into the target hole. When the ball is in the upper area, the agent can have two different behaviours. If the ball is close to the obstacles or close to the wall on the right side of the tray, it will rotate the platform counterclockwise to send the ball to the bottom. While roaming towards the centre of the area or the upper wall the agent rotates clockwise and sends the ball to the top. When the ball passes through the “gate” toward the goal area, the agent rotates the platform only anticlockwise and sends the ball to the bottom.

We divided the paths into 3 parts based on each of the three starting positions that the ball had during the first block. In this block, the expert started four times in the bottom right corner, four times in the top right corner and two times in the top left corner. In Figure 3.7, the paths that had as a starting point the bottom right corner are presented. In general, the initial strategy here involves sliding the ball along the obstacle to reach the "gate" positioned between the two obstacles, allowing the ball to pass into the goal area. At this point, the agent needs to act in a way that tilts the platform towards the target. Nevertheless, this strategy alone does not guarantee successfully reaching the target. As we mentioned above, after passing the “gate” the agent follows a policy that only rotates the platform counterclockwise.

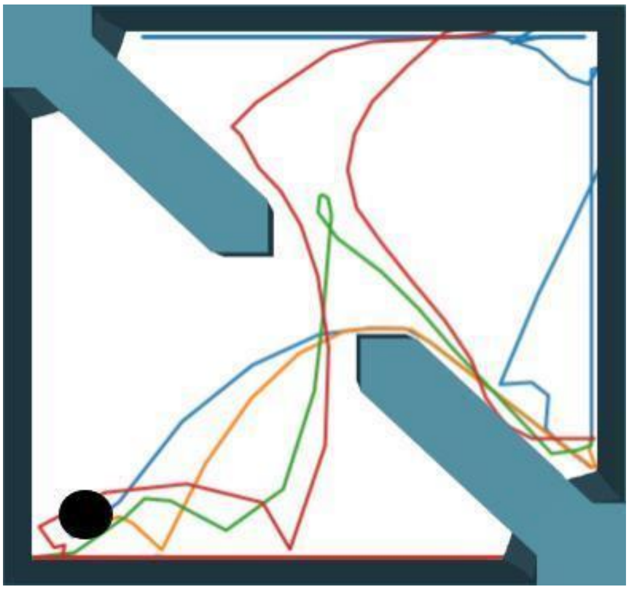


Figure 3.7 Paths during four games starting from the bottom right position

Based on our experience, the human can contribute to exploring the states of the environment, in order to assist the agent to find and exploit certain favourable states (of position, velocity of the ball and the tilt of the tray). By favourable states, we mean states where the agent seems to be changing its actions. In practice this means that the human player keeps affecting the states of the environment by changing frequently his actions, until the agent changes its action.

In the four games presented in Figure 3.7, the team managed to win two times (blue and orange traces) while failed in the other two attempts (green and red traces). In the unsuccessful games, the expert (human) player attempted to assist the agent (as described above) by tilting the tray both clockwise and counterclockwise in the dimension he controlled (traces at the bottom of the tray - the green trace is overlapped by the red one). This could provide the opportunity to the agent to take the appropriate action. During the successful games, the actions and the environment were such that they allowed the ball to move either directly or via a bounce to the target. During one of the failed attempts (red trace), we can see that the agent bounced the ball and passed though the target. This is because while reaching the target, in order for the ball to fall inside, it has to have a lower speed, or else it passes above the hole.

In Figure 3.8, we can see the paths starting from the top right. In this scenario, the expert did not score. Additionally, there are numerous paths that lead to the target but do not guide the ball into the hole. We also see (blue trace) that the expert also tried to bounce the ball to the left wall with no success. There is a difference in possible strategy from starting on the top right side. In general, you can pass the “gate” without using the obstacles that cut the speed of the ball. In one of the games (green trace), the expert passed the ball without using the obstacles. As mentioned in the beginning when the agent is in the center of the upper area, it rotates anticlockwise. In order to pass this and not send the ball to the top, it needs to acquire speed to leave fast from that area and through the gate. This means that it will reach the target with too much speed and as in this case fail to win.

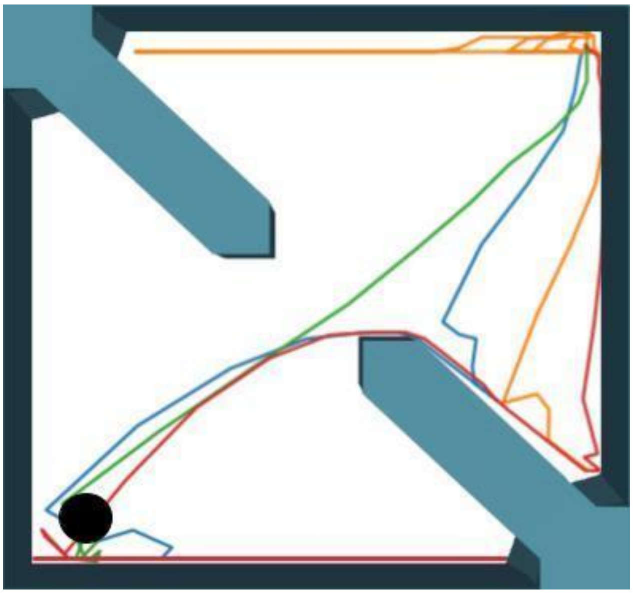


Figure 3.8 Paths during four games starting from the top right position

In Figure 3.9, for the paths starting in the top left we see similar paths to the other starting points. Keeping the ball close to the obstacles and trying to reach the target neither directly or via bounce. For a different view of these three cases, we provide the heatmaps in Figures 3.10, 3.11 and 3.12.

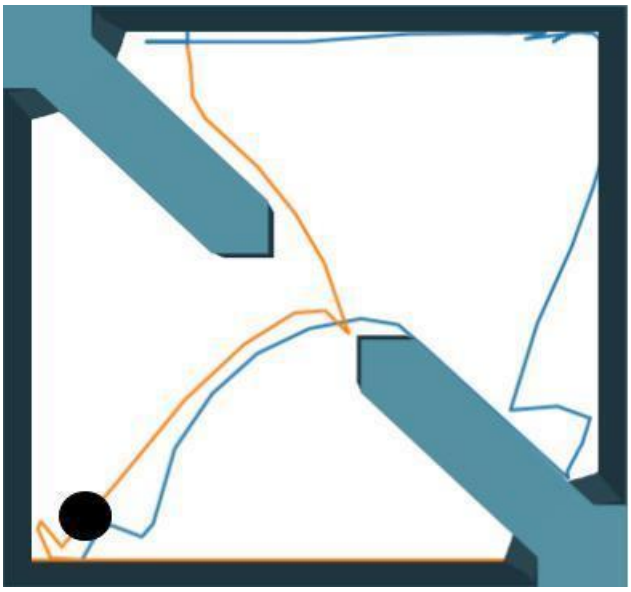


Figure 3.9 Paths during four games starting from the top left position

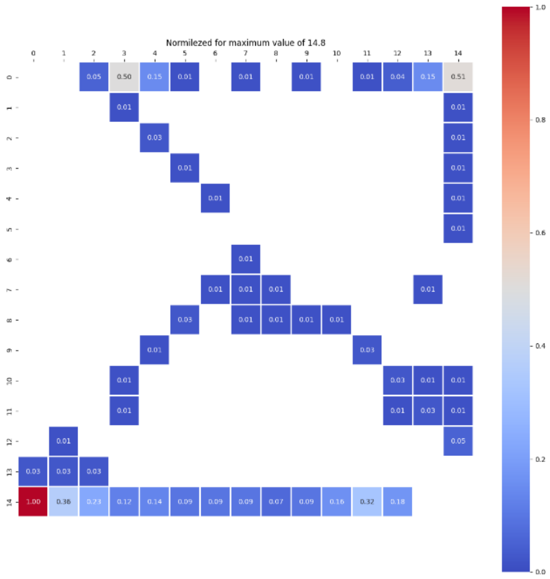


Figure 3.10 Heatmap during four games starting from the bottom right position

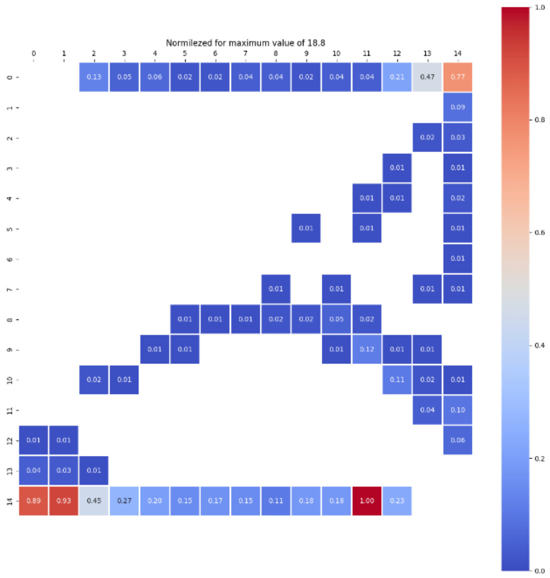


Figure 3.11 Heatmap during four games starting from the top right position

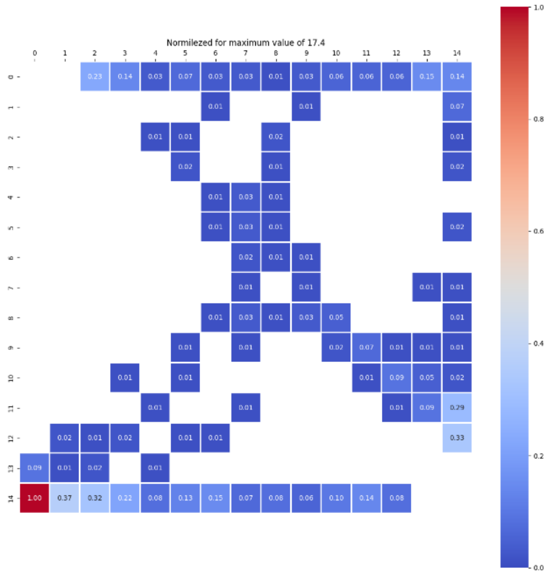


Figure 3.12 Heatmap during four games starting from the top left position

3.5 Collaboration measures

In a collaboration between human and agent the "objective performance" of the team is not the only important aspect. Each participant can have a different perception of the collaboration even if the objective measures are similar. For this reason, we also used subjective measures to capture information about different aspects of the collaboration from the side of the participant.

3.5.1 Objective measures

Objective measures are observable and measurable criteria used to assess or evaluate something in a standardised and unbiased way. We mainly focus on four measures:

- Scores: Score of each game. The score starts at 200 at the start of a game and then it is subtracted by one for each control frame played.
- Wins: The number of wins achieved in a block of games.
- Normalized Travel Distances: The travelled distance is the distance that the ball travelled during a game. The travel distance is multiplied by the percentage of the total control frames played in a game. This normalized travelled distance was used to account for the games that the ball was driven to a side of the tray and the team never managed to bring it back into the game, providing an erroneously short distance.
- Travel Speeds: The average speed the ball had during a game.

3.5.2 Subjective measures

We are using a questionnaire to capture subjective measures of the participant's experiences during the collaboration process. We focused on six aspects of the collaboration, *Human-AI Fluency*, *AI Contribution*, *Team Improvement*, *Trust*, *Teammate Traits*, and *Alliance*. We follow the steps of Tsitos et al. [30]. The questionnaire initially presented by Hoffman [39] and Tsitos provided feedback on how these measures could improve to provide better results.

In his findings, Tsitos noted a lack of internal consistency in the responses regarding the robot's contribution. The main problem seems to be the lack of framing of the questions in terms of low or high performance. As an example, in the case of the No_TL group, participants could either agree that the robot was the most important team member 'in terms of not achieving high performance' or disagree that the robot was the most important team member 'in terms of successful games'. It is also mentioned that there was an imbalance between questions about the robot's contribution and questions about the human's contribution.

Based on this, we decided to create two branches in the *Human-Agent Contribution* category. First, we ask 4 questions in the context of the performance in the last block and then two questions in the context of the whole process. This gives all participants the same mindset when answering these questions, while at the same time capturing both the human-agent contribution in the training process and their contribution to the final result.

The English version of the first 4 questions asked in the context of the performance in the final block are as follows

1. How do you judge the team's performance in the last ten tests
2. I was primarily responsible for this performance.
3. This performance was a joint effort of the team
4. The AI system was primarily responsible for that performance.

The first of these questions was added in order to produce a subjective measure of the performance of the team. By asking participants to judge the team's performance, we can observe how they view the objective performance, thus making a link between the objective outcomes and the participants' perceptions of those outcomes. The remaining three questions are similar to those used by Tsitos et al [30] and introduced by Hoffman[39], with minor changes to the wording to better suit our study.

The main change we made in questions 2-4 is the clarification to participants that they are answering in the context of their performance in the final block. Another change is in the way we group these questions in our results. While in both prior works [30,39] questions 2-4 are grouped together, we chose to group only questions 2 and 4 with question 2 having a reverse scale. This group provides a view of the perceived *contribution of AI in the final block*.

While we will provide results from the *AI Contribution* based on the grouping of questions mentioned above, we believe that this is not the best way to present these results. Instead, we can use these questions to create a narrative based on the participant's responses. For example, if in the first question, the participant judges the performance negatively, we can deduce that the context for the team's effort and responsibility is "who is to blame", however, in a positive performance the context is "who helped more". Following this, the third question can show whether the participant believed that both members had an equal share in the final performance or not, and finally the second and fourth questions can see how the participant credits or blames the performance.

After asking these four questions in the context of the performance in the last block, the participants were asked the remaining questions for all other categories included in the context of the whole collaboration process. In this context, the participant answered two more questions about the *contribution of the human agent in the collaboration*. The two questions are as follows

1. Throughout the interaction, I was the most important member of the team.
2. Throughout the interaction, the AI system was the most important member of the team.

While these questions should give a more explicit answer as to whom the participant credits as the more important member, and using the first question in a reverse scale we can group them together and use them as a group to represent the changes in *human-agent contribution* between the two groups, we still believe that the internal consistency is still lacking and that they are best used individually to tell a story rather than as a combined metric.

The other change was in the *Improvement* category where we added two new questions asking participants about the importance of each member in improving the team. These questions are subject to evaluation in terms of their correlation with this category and whether any questions should be reversed, but we feel that these questions are important to capture participants' perceptions of the importance of each member in improving the co-learning experience. All questions are translated into the native language of our participants, Greek, and are shown in Table 3.2.

In addition to these questions, the participant is asked to *judgment of control* question between each block. This question originates from the work of Dewey et al [12], and we modified it to both make sense in our setup and also, as it translated into the native language of the participants, Greek, we wanted to make sure that everyone interprets the objective of the question the same. The question is the last item in Table 3.2.

Table 3.2 All questions for the subjective mesures separeted into each measure.

English	**Greek**	**Neg/Pos**
FLUENCY		
The human-TN team worked together seamlessly (flowing/harmoniously, EN-fluent).	Η ομάδα ανθρώπου -ΤΝ συνεργάστηκε απρόσκοπτα (με ροή/ αρμονικά, EN-fluent).	Pos
The team's cooperation has become more fluid over time.	Η συνεργασία της ομάδας έγινε πιο εύρυθμη με τη πάροδο του χρόνου.	Pos
The AI system contributed to the fluid collaboration of the team.	Το σύστημα ΤΝ συνεισέφερε στην εύρυθμη συνεργασία της ομάδας.	Pos
CONTRIBUTION		
AI CONTRIBUTION ALL GAMES		
I had the main responsibility for this performance.	Εγώ είχα την κύρια ευθύνη γι’ αυτήν την επίδοση.	Neg

The AI system was primarily responsible for this performance.	Το σύστημα ΤΝ είχε την κύρια ευθύνη γι’ αυτή την επίδοση.	Pos
AI CONTRIBUTION LAST 10 GAMES		
Throughout the interaction, I was the most important member of the team	Καθ’ όλη τη διάρκεια της αλληλεπίδρασης, ήμουν το πιο σημαντικό μέλος της ομάδας.	Neg
Throughout the interaction, the AI system was the most important member of the team.	Καθ’ όλη τη διάρκεια της αλληλεπίδρασης, το σύστημα ΤΝ ήταν το πιο σημαντικό μέλος της ομάδας.	Pos
TRUST		
I had confidence in the AI system that it would do the right thing at the right time.	Είχα εμπιστοσύνη στο σύστημα ΤΝ ότι θα έκανε το σωστό πράγμα τη σωστή στιγμή.	Pos
There was mutual trust between me and the AI system.	Υπήρχε αμοιβαία εμπιστοσύνη ανάμεσα σε μένα και το σύστημα ΤΝ.	Pos
TEAMMATE TRAITS		
The AI system was intelligent.	Το σύστημα ΤΝ ήταν ευφυές.	Pos
The AI system was trustworthy.	Το σύστημα ΤΝ ήταν αξιόπιστο.	Pos
The AI system was dedicated to achieving the goal.	Το σύστημα ΤΝ ήταν αφοσιωμένο στην επίτευξη του στόχου.	Pos
The AI system was cooperative.	Το σύστημα ΤΝ ήταν συνεργάσιμο.	Pos
IMPROVEMENT		
The human-AD team improved over time.	Η ομάδα ανθρώπου - ΤΝ βελτιώθηκε με την πάροδο του χρόνου.	Pos
My performance improved during the experiment.	Η επίδοσή μου βελτιώθηκε κατά τη διάρκεια του πειράματος.	Pos
The performance of the AI system improved during the experiment.	Η επίδοση του συστήματος ΤΝ βελτιώθηκε κατά τη διάρκεια του πειράματος.	Pos

I had the main responsibility for the improvement of the team.	Εγώ είχα την κύρια ευθύνη για την βελτίωση της ομάδας.	Pos
The AI system had the main responsibility for the improvement of the team	Το σύστημα ΤΝ είχε την κύρια ευθύνη για την βελτίωση της ομάδας	Pos
ALLIANCE		
I believed that the AI system could help me.	Πίστευα ότι το σύστημα ΤΝ μπορούσε να με βοηθήσει.	Pos
The AI system could perceive my intentions.	Το σύστημα ΤΝ μπορούσε να αντιληφθεί τις προθέσεις μου.	Pos
The AI system didn't understand what I was trying to achieve.	Το σύστημα ΤΝ δεν καταλάβαινε τι προσπαθούσα να πετύχω.	Neg
I think working with the AI system was confusing	Θεωρώ ότι η συνεργασία με το σύστημα ΤΝ ήταν μπερδευτική	Neg
Extra Items		
This performance was a joint team effort.	Αυτή η επίδοση ήταν από κοινού αποτέλεσμα της ομάδας.	
How do you judge the team's performance in the last ten tests?	Πως κρίνετε την επίδοση της ομάδας στις τελευταίες δέκα δοκιμές;	

What did you think of working with the AI system?
Do you have any comments on the experiment?
	Πως σου φάνηκε η συνεργασία με το σύστημα ΤΝ;
	Έχεις κάποια σχόλια για το πείραμα;	

3.6 Personality

To reinforce our results, we used a series of questionnaires to gather more information about our users. In general, the focus is on how personal characteristics affect the participant's perception of their interaction with AI agents. Before the start of the collaboration process, the participants are asked to complete a questionnaire that included the following questionnaires:

- Big five personality traits (50 questions, Table 2.3)
- Schwartz Portrait Values Questionnaire (21 questions, Table 2.5)
- AI Attitude Scale (20 questions, Table 2.4)

Additionally, we asked some questions about themselves. These additional questions are in Table 3.3 and cover the following subjects:

- Personal information (age, gender, dominant hand, eye/neurological problems) (5 questions)
- Experience in gaming (2 questions)
- Knowledge about AI (3 questions)

3.7 Process

To ensure a fluent and accurate process, all participation was on-site. We provided different locations close to the study/work location of each participant. Each location was an open office or similar environment, where participants had limited distractions and would feel more comfortable. The equipment used was the same for all participants. The entire process is presented in the Table 3.4

Table 3.3 Question about personal information, experience in gaming and knowledge in AI.

English	**Greek**
Personal information	
Gender	Φύλο
Age	Ηλικία
Dominant hand	Επικρατές χέρι
Diagnosed neurological condition	Διαγνωσμένη νευρολογική πάθηση
Use of myopia glasses/lenses	Χρήση γυαλιών/φακών μυωπίας
Experience in gaming	
What experience do you have with gaming?	Τι εμπειρία έχεις με παιχνίδια (gaming);
What is your preferred platform (Tap none if you have no experience with games)?	Ποιά είναι η προτιμώμενη πλατφόρμα σας (Πατήστε καμία αν δεν έχετε καθόλου εμπειρία με παιχνίδια);
Knowledge about AI	
How would you describe your relationship with AI?	Πως θα χαρακτηρίζατε τη σχέση σας με την ΤΝ;
Do you come into contact with AI applications in your daily life?	Έρχεστε σε επαφή με εφαρμογές ΤΝ στην καθημερινότητά σας;
What is the main source of information on developments around AI issues?	Ποια είναι η κύρια πηγή ενημέρωσης των εξελίξεων γύρω από θέματα ΤΝ;

Table 3.4 Procees of Participation during the study.

	Step	Description
Pre-Study	**Arrival**	• Welcomed participants. • Ensured they read the announcement info. • Reiterated important information.
	Consent Form	Provided the participants with the consent form regarding their involvement in the study.
	Initial Questionnaire	• Participants filled out a questionnaire with questions from section 3.6. • Duration: 15-20 minutes.
	Information Video	Showed participants a video detailing the rest of the process to ensure consistent information.
	Q&A after Video	Asked participants if they needed additional information before proceeding.
Study	**Familiarization Game**	For details, refer to section 3.3.3.
	Collaboration Process	For details, refer to section 3.3.2.
Debriefing	**Final Questionnaire**	Participants filled out a questionnaire containing subjective measures from section 3.5.2.
	Debriefing	Engaged in a conversation-type debriefing. Answered questions or addressed observations from participants.

4 Results

In this chapter, we present the results of the HAC study. The chapter is divided into four sections: the first one, describes the participants, the second focuses on the objective measures, the third demonstrates some individual and group behaviours during the HAC and the fourth focuses on the subjective measures.

4.1 Participants

Our final sample consisted of 8 participants (male: 7, female: 1, ages: 24-39, avg. age: 31, all right-handed). Regarding the rest of the descriptive variables of the group:

-Gaming experience: Seven participants had more than 5 years of gaming experience, while one 1 had less than 1 year

-Preferred platform: Five participants preferred a PC with a keyboard, while three preferred a smartphone

-Knowledge of AI: Two participants had limited knowledge, one had knowledge of the latest developments and 5 were students or professionals in the AI sector.

-Source of information about AI: Seven participants had their postgraduate studies, while one had social media.

Tables 4.1 and 4.2 show the results of the AI Attitude Scale for No_TL and TL respectively. The scale produces 2 scores: a score for the positive questions and a score for the negative questions. Schepman et al. [59] presented multiple ways the scale can be used. In this work, we focus simply on the difference between the positive value and the negative value.

Table 4.1 No_TL participants attitude towards AI

Participant	Positive	Negative	Difference (P-N)
P_NTL_1	4.08	3.57	0.51
P_NTL_2	3.92	4.71	-0.80
P_NTL_3	3.00	3.29	-0.29
P_NTL_4	3.83	3.86	-0.02
Average	3.71	3.86	-0.15

Table 4.2 TL participants attitude towards AI

Participant	Positive	Negative	Difference (P-N)
P_TL_3	4.25	3.14	1.11
P_TL_4	4.33	3.57	0.76
P_TL_2	3.75	3.43	0.32
P_TL_1	4.33	4.14	0.19
Average	4.17	3.57	0.60

Our two groups are dissimilar in their attitude towards AI. The No_TL group showed an attitude towards AI of -0.15, and the TL group showed an attitude of 0.6. This difference in attitude could affect the confidence with which they approach the collaboration with the agent.

Similar to the attitude towards AI, the Big Five and personal values scales are used for comparisons between the two groups. In Figure 4.1 shows the five factors from the Big Five questionnaire, for the No_TL and TL groups. The two groups showed some differences in some factors. More specifically the TL group had mostly higher values in the Agreeableness factor, a mostly higher intellect imagination and some participants had a lower emotional stability. In extraversion and conscientiousness, the two groups are relatively similar.

In Figure 4.2 are the ten factors from the personal values questionnaire, for the No_TL and TL groups. We do not focus on each factor, but again we see some differences between the two groups.

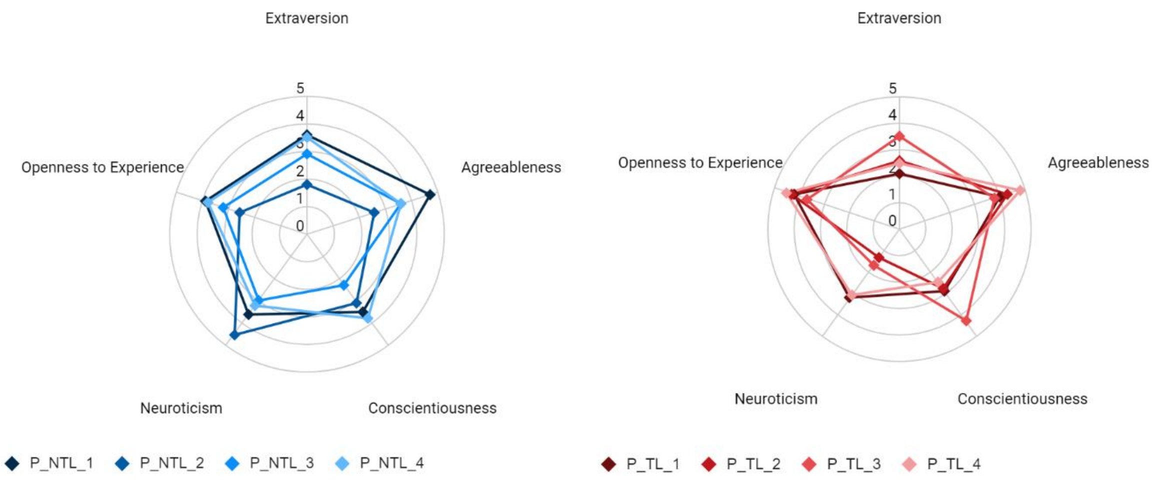


Figure 4.1 Big five personality traits of the No_TL group (left) and the TL group (right)

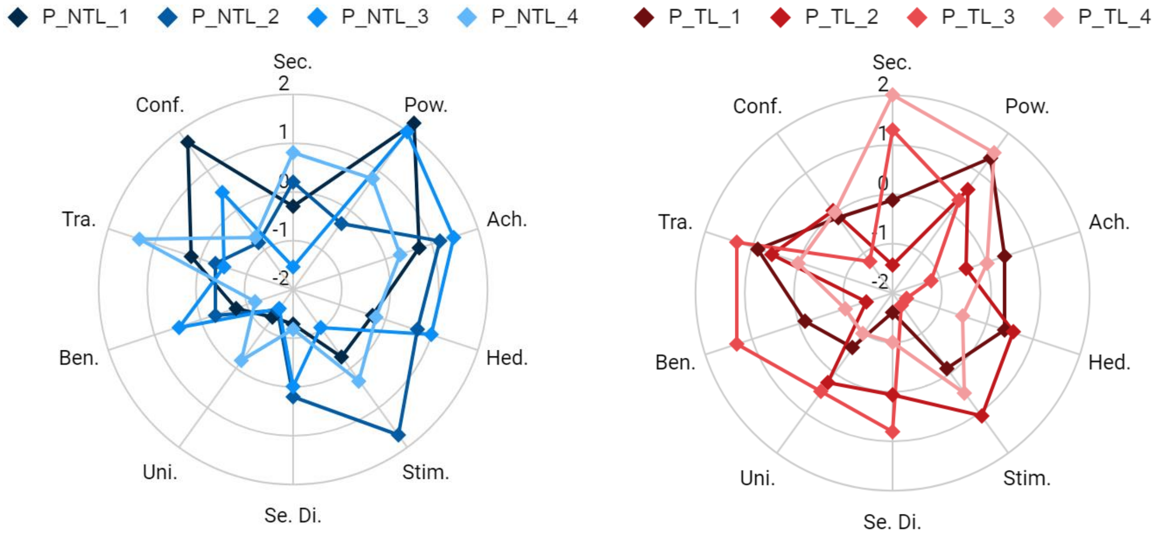


Figure 4.2 Schwartz Personal Values of the No_TL group (left) and TL group (right)

At the beginning of the study, before the actual collaboration experience, the participants played on their own (without collaborating with an agent) 5 games to familiarize themselves with the keyboard controls. This process has been described in details in section 3.3.3. Data collected during this part of the study, can be used to evaluate the skills of the participants. In Table 4.3 are the result times (in seconds) for the No_TL and TL groups respectively. In both groups, there is a varying performance among participants. The best-performing participants were P_NTL_2 and P_TL_2 for the No_TL and TL groups respectively. The worst-performing participants were P_NTL_3 and P_TL_1 for the No_TL and TL groups respectively

Table 4.3 Mean, Mix and Max values of the Times in seconds for each participant in the Familiarization process. In the mean values, green represents the best value and red the worst value in each group. (Pt means Participant)

Group	**No TL**				**TL**			
Pt	P_NTL_1	P_NTL_2	P_NTL_3	P_NTL_4	P_TL_1	P_TL_2	P_TL_ 3	P_TL_4
Mean	53	**48**	**72**	61	**158**	**59**	82	85
Max	165	82	**254**	127	**325**	280	186	152
Min	**20**	26	45	33	56	53	34	**25**

4.2 Objective measure results

4.2.1 Game scores, Wins, Distance travelled and Ball Speed

Figures 4.3 and 4.4 show the game scores distribution of the human-agent teams across all blocks for the No_TL and TL groups respectively. In all figures in this section, we also include the expert's team results (light green). The results of the baseline block are not presented due to the loss of the data. The TL teams managed to achieve a performance comparable to the expert towards the end of the study. Also, the TL teams managed to reach a competitive level from the first block, something that the expert achieved in the third block. And while the expert achieved better and more consistent game scores at the end, the TL teams are not far in performance. On the other hand, the No_TL teams showed an inconsistent performance; some teams achieved game scores near the expert’s team but they still presented a big variance in their performance. Also in the No_TL group, a team (P_NTL_3) failed to increase the performance across the blocks.

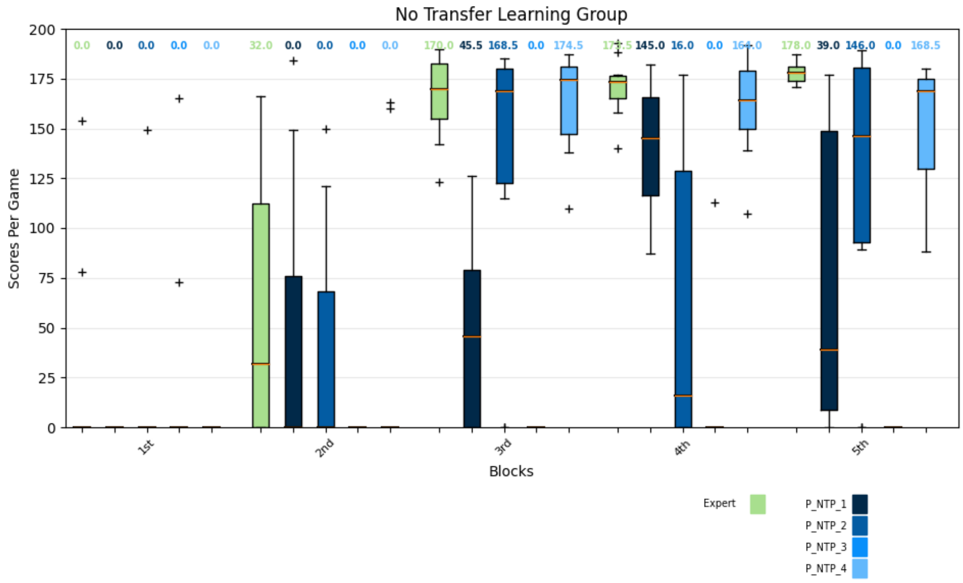


Figure 4.3 Distribution of scores, in each block, for participants of the No_TL group and the expert

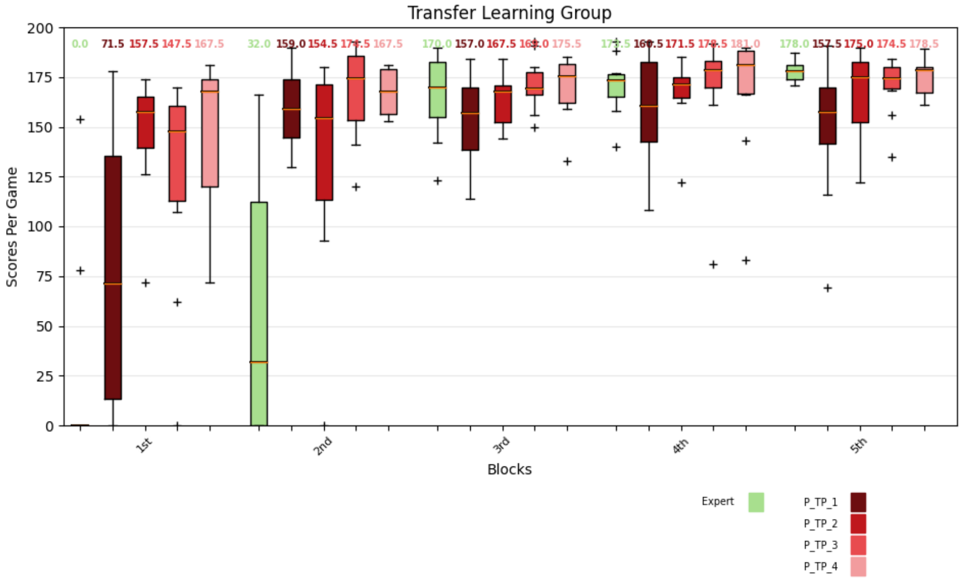


Figure 4.4 Distribution of scores, in each block, for participants of the No_TL group and the expert

Figures 4.5 and 4.6 show the wins in each block of the human-agent teams for the No_TL and TL groups respectively. Similar to the previous variable, the TL group achieved better results than the No_TL group. The biggest difference here is that in the TL group, the "worst" performing team, failed in only 3 games across all the blocks, while in the No_TL group, the worst-performing team achieved only 3 wins.

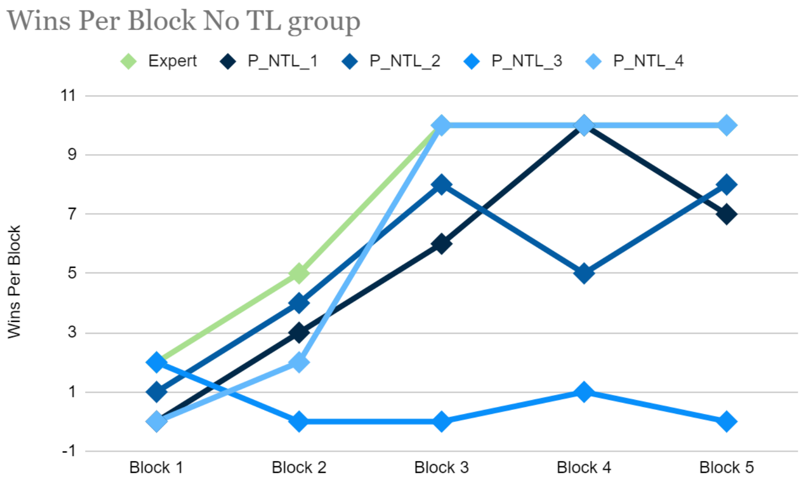


Figure 4.5 Wins per block for the No_TL group

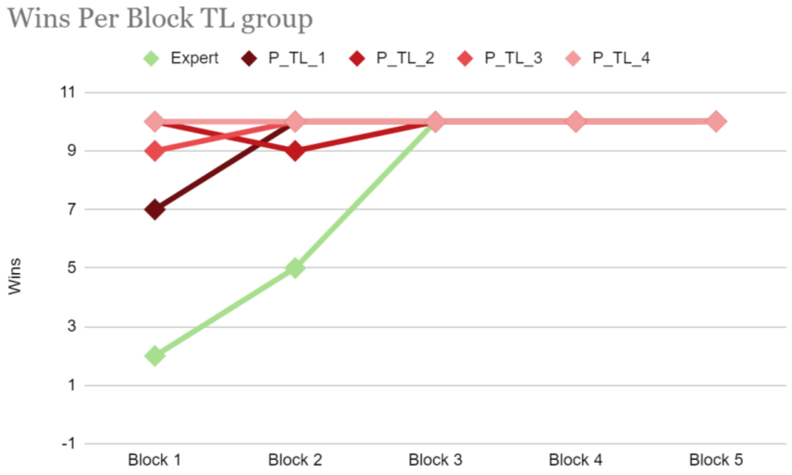


Figure 4.6 Wins per block for the TL group

Figures 4.7 and 4.8 have the distributions of the normalized travel distance for each block, for the No_TL and TL blocks respectively. For more details about the normalization see section 3.5.1. In the No_TL group, the observations are similar to the game scores, with an inconsistent performance between the teams. In the TL group, the distances are smaller compared to the No_TL group, with more consistency and much closer to the expert. An interesting observation, for both groups, is the variance in distances metric during the first block. As explained in section 3.4.1, all participants and experts collaborated with the same initialised agent. Nevertheless, this does not prevent individual performance from being exposed.

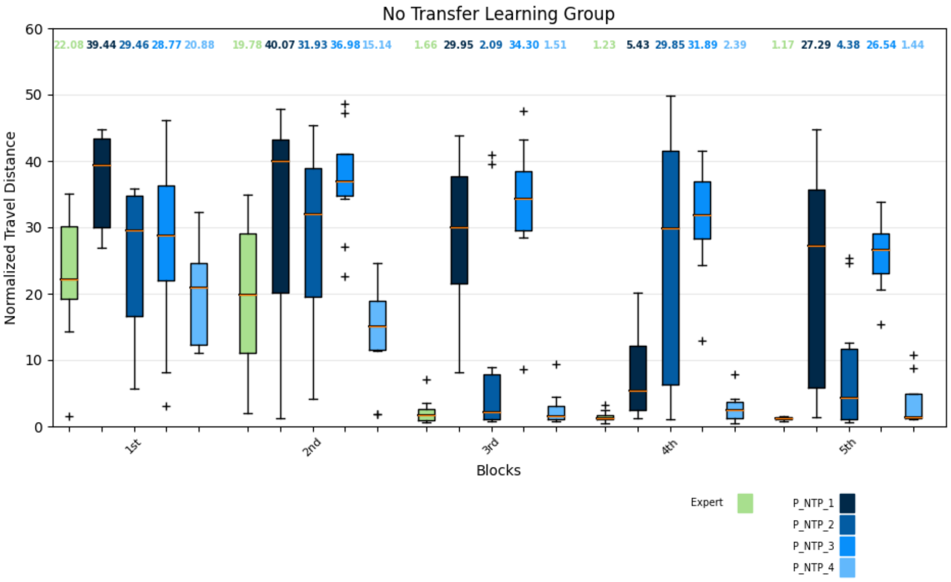


Figure 4.7 Normalized Travel distances of NO_TL in each block

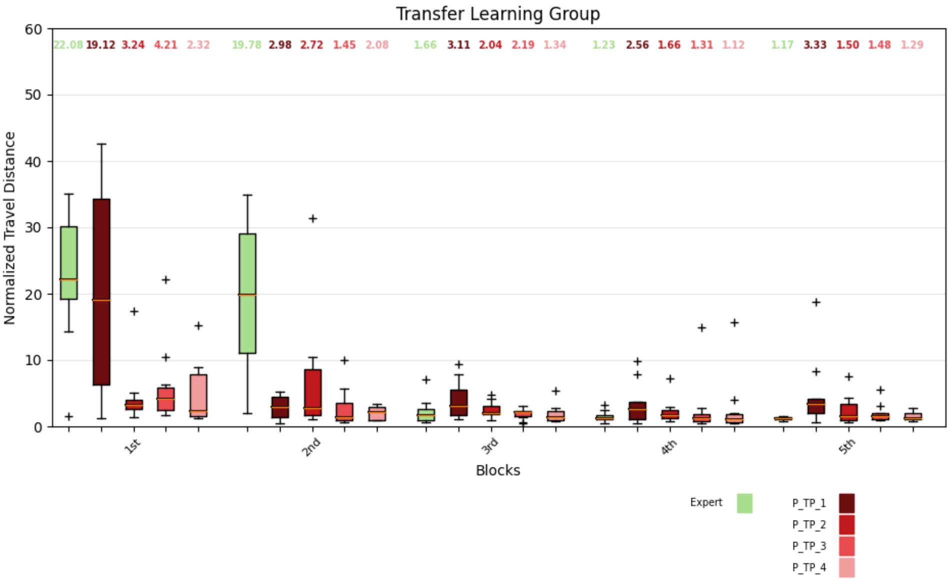


Figure 4.8 Normalized Travel distances of TL in each block

Figures 4.9 and 4.10 have the distribution of the average game speed of the ball, for each block, for the No_TL and TL blocks respectively. As in the above objective variables, here again, in the No_TL group, the speeds have a much bigger variance compared to the TL group. The TL group, while having a small variance in speeds at the final block, overall they are close to the results of the expert.

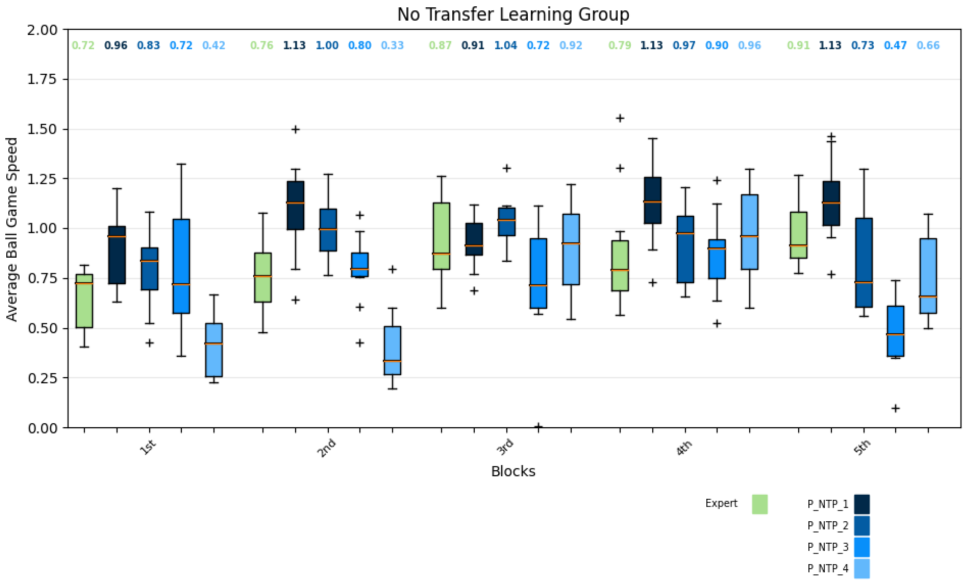


Figure 4.9 Average Travel Speeds of NO_TL in each block

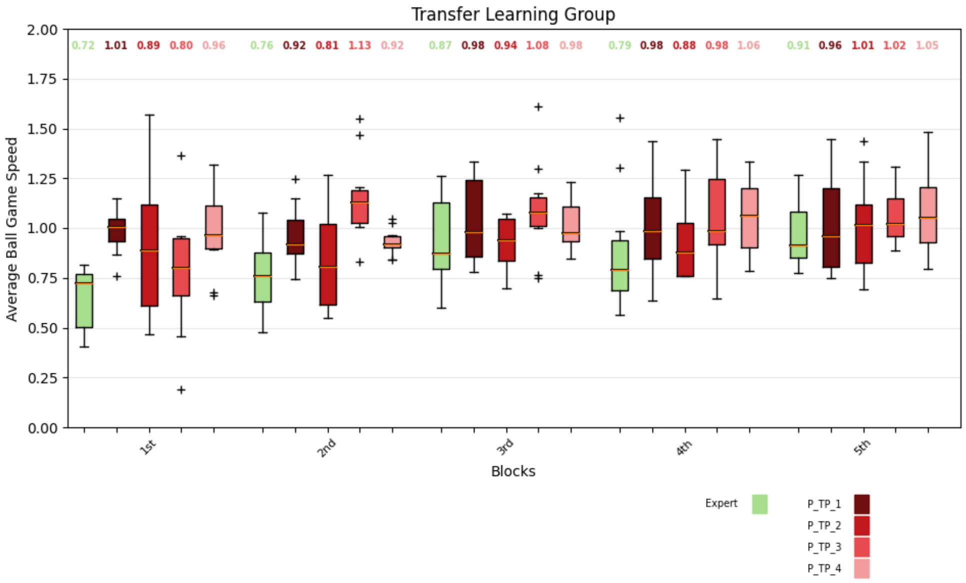


Figure 4.10 Average Travel Speeds of TL in each block

The total training time each participant spent in the study are shown in Table 4.4. Overall, the No_TL group spent an average of 34.9 minutes in collaboration with the agent and the TL group spent an average of 17.35. These times exclude the time each participant spent in the baseline games.

Table 4.4 Times in minutes for the entire collaboration period with the SAC agent of each participant

No_TL Group					
Participant	P_NTL_1	P_NTL_2	P_NTL_3	P_NTL_4	Average
Time in minutes	34.6	32.4	45.1	27.5	34.9
TL Group					
Participant	P_TL_1	P_TL_2	P_TL_3	P_TL_4	Average
Time in minutes	20	17.3	16.5	15.6	17.35

For the No_TL group, an interesting observation in all the above measures is a drop in performance between some blocks. More specifically, in the game scores in Figure 4.3, of the team P_NTL_2, in the 4th block, the performance dropped before rising again in the next block. This behaviour is apparent to most teams in the No_TL group, either on the 4th or the 5th block. In the expert's team, during the 4th block, while the overall performance didn't drop, there was a reduction of the variance in the scores. During that block, the expert experienced that the agent started to perform much more fluently and control the tray more precisely. This meant that the agent was performing smaller movements and fewer errors. This resulted in a reduction in travel distance but also a drop in the speed of the ball. Both are obvious in the normalized travel distances in Figure 4.7 and speed in Figure 4.9. Generally in the rest of the participants, this change in the agent seemed more aggressive and didn't benefit them as much.

4.3 Behavior

4.3.1 Gameplay

In this section, we present some behaviours of the participants, using the paths created by the movement of the ball during the collaboration. Each included figure contains all paths for the 1st, 3rd and 5th blocks. Figure 4.11, shows the paths during the collaboration of the expert and the expert agent. The main observation here is that as the collaboration proceeded, the team passed the ball through the "gate" sooner and spent less time around the target.

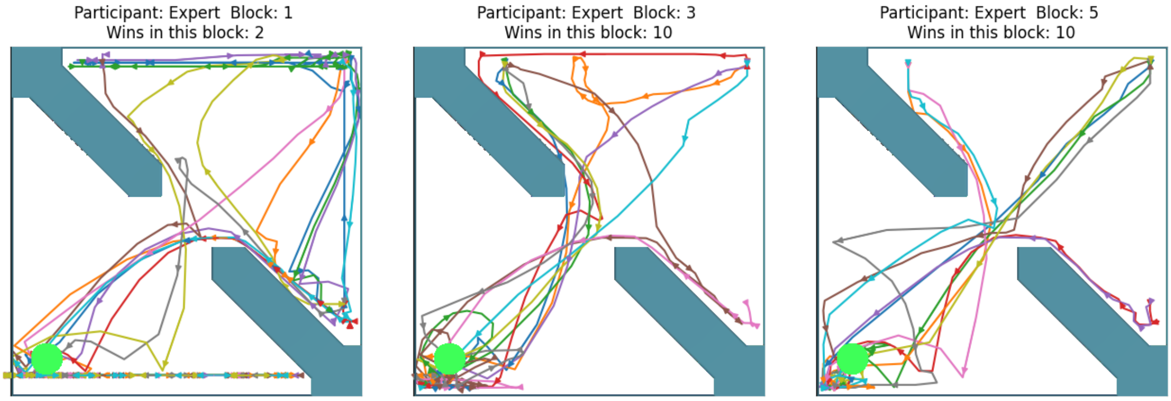


Figure 4.11 The paths of the ball from the 1st, 3rd, and 5th block during the Expert's collaboration with the agent.

In Figure 4.12 there are the paths of the team of participant P_NTL_1. Similar to the expert's team, in the 3rd block the paths got better, passing the "gate" with ease and not spending too much time around the target. But in the final block the paths are all over the tray. The participant controls the y-axis of the tray, see section 3.1 for more information. The main thing to understand here is that rotating the y-axis sends the ball towards the left and right sides of the tray. In the paths of the 5th block, the ball was aggressively moving between the left and right sides. This shows the aggressive control the participant adopted during the process.

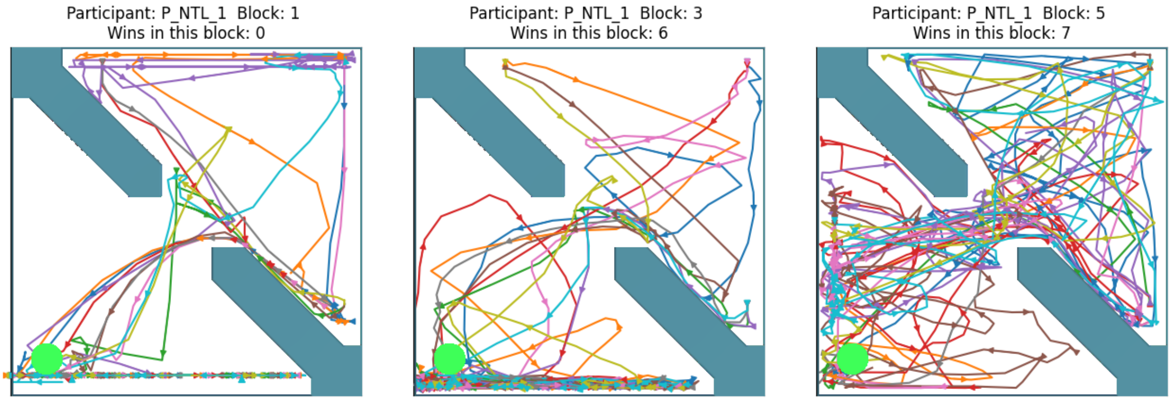


Figure 4.12 The paths of the ball from the 1st, 3rd, and 5th block during the P_NTL_1's collaboration with the agent.

In Figure 4.13 are the paths of the team of participant P_NTL_3. This team failed to increase its performance and achieved only 3 wins. Two of those wins were on the first block. Interestingly, in the first block, the paths do not show any significant difference to the expert. But after that, in the 3rd and 5th block, the team struggled to pass the ball through the 'gate' and most of the time the ball was stacked on the top of the tray.

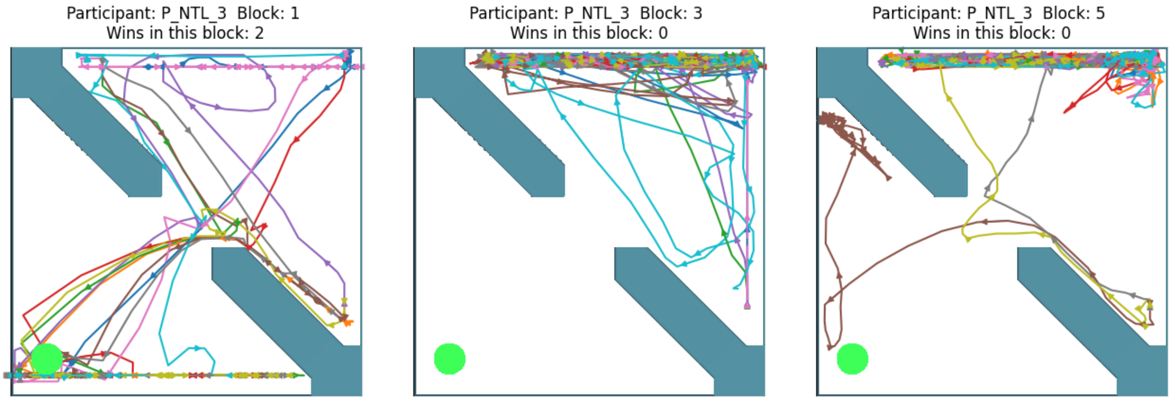


Figure 4.13 The paths of the ball from the 1st, 3rd, and 5th block during the P_NTL_3's collaboration with the agent.

In contrast to the above two teams of the No_TL group, Figure 4.14, shows the paths of the team of participant P_TL_1 of the TL group. In the 1st block, the team passed the ball through the "gate" consistently but with a little difficulty and then the team mostly spent time around the target. As the process proceeded, the teams got better at those two things. In the 5th block, the team spent the most time in the middle, before passing the 'gate' and at the bottom, around the target.

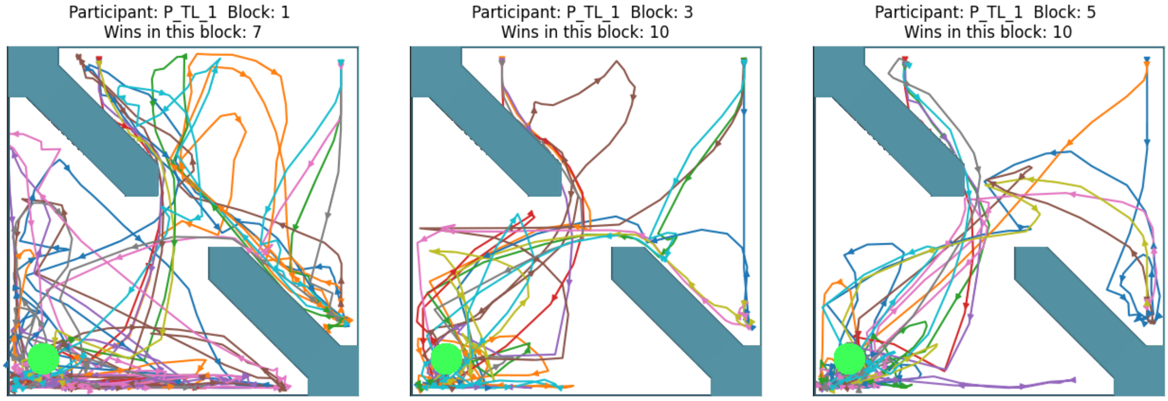


Figure 4.14 The paths of the ball from the 1st, 3rd, and 5th block during the P_TL_1's collaboration with the agent.

Paths for all the participants and heatmaps of all the blocks for each participant can be found in Appendix 7.1

4.3.2 Exploration with variable temperature

One aspect that affects the exploration of the agent is the value of temperature α. The temperature α is a parameter that defines the degree of entropy affecting the soft value in the Bellman equation. This affects the exploration-exploitation trade-off of the agent. For more details see section 2.1.2. The temperature α is defined based on the entropy of the potential actions and the target entropy. In our case, the target entropy is 0.67. This means that the AI agent, based on the entropy, has control over when to switch from exploration to exploitation, and vice versa. In this section, we review how the temperature α and the entropy changed during the off-line training (OLT) sessions.

Figure 4.15 shows the resulting temperature and entropy of the expert agent. In the first two OLT sessions, the temperature rises if the entropy is lower than the target entropy. In both of these OLT sessions, when the entropy reached around the target entropy, the temperature, at the end of the off-line training, fell to ~0.7 and ~0.4 in the first and second OLT sessions, respectively. During the third OLT session, in order to keep the entropy stable the temperature rises from a value of ~0.4 to a value of ~0.6 and it remains stable at that value during the last OLT session. The low spikes in entropy on the first and second off-line training sessions are because of the new states the agent explored in the previous block of actual games. In the 3rd and 4th off-line training sessions, the agent had already explored the map, therefore the entropy of the states was already close to the target value.

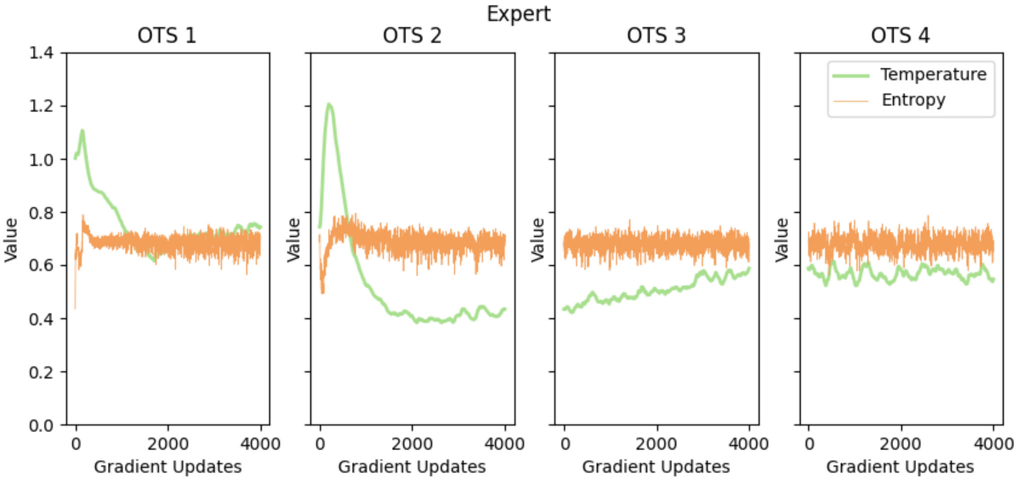


Figure 4.15 Exper’s Temperature and entropy trajectories during the off-line training sessions (OTS).

Figures 4.16 and 4.17 are the temperatures of the agents for the No_TL and TL groups respectively. The entropy is not further presented, as the changes in the entropy look similar to those of the expert (Figure 4.15) and do not provide any additional information. In the No_TL group, the temperature has some peaks in the first and second OLT sessions before reaching a stable value in the last two OLT sessions. Between the participants, there is a variance in the value where the temperature is stabilised. In the team of participant P_NTL_4, who had the closest game scores to the expert, the agent’s temperature stabilized to a value of

~0.5, close to that of the expert agent. In contrast, in the team of participant P_NTL_3, who performed the worst, the agent's temperature stabilized to a value of ~0.35, the lowest or similar to the lowest.

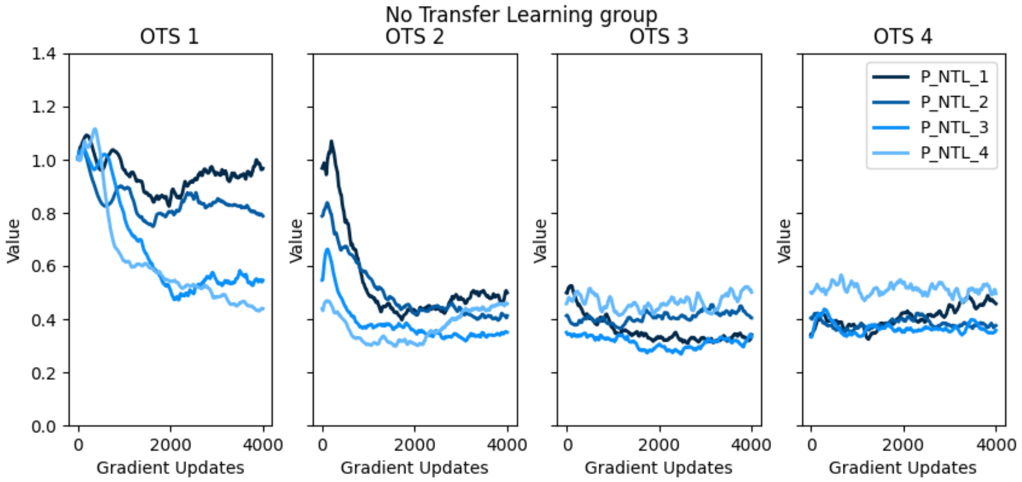


Figure 4.16 No_TL groups Temperature trajectories during the four off-line training sessions (OTS).

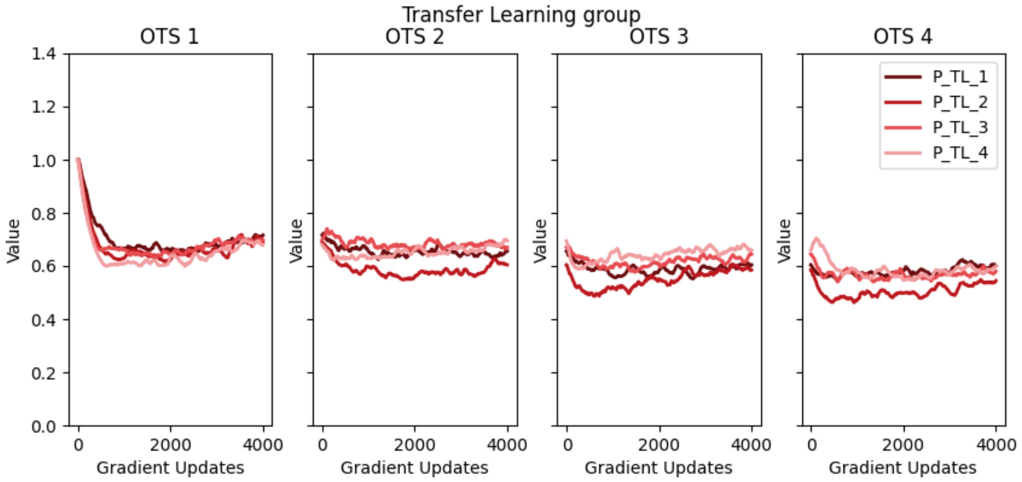


Figure 4.17 TL groups Temperature trajectories during the four off-line training sessions (OTS).

For the TL group, we do not include the off-line training session in the first phase of the TL method. In the four OLT sessions, the temperature remains stable across all the participants with a small spike on the first off-line training session. As there is no need for the same level of exploration as the No_TL group, this was expected. The most important observation here is that the temperature in all agents is stabilized at a value of ~0.6 with the exception of one that falls close to ~0.5.

Based on the above results, there is a possible correlation between the temperature and the performance that the teams had. In the No_TL group, the agent of the best-performing team had a similar temperature to the expert, and the same to all the agents of the TL group. The lower performance could be the cause of the lower temperature, meaning that those were the optimal temperatures for those cases. But it could be the opposite. Could a fixed temperature based on the data from the expert, be better than a trained one?

In the No_TL group we can observe the following:

a. the agent of the best-performing team (P_NTL_4) reached a similar temperature to that of the expert, and the agents of the TL group.
b. the rest of the three agents reached a lower temperature (compared to: P_NTL_4, the expert, the agents of the TL group)

Based on the above observations, there is a possible correlation between the temperature and the performance that the teams reached at the end of the study. This is further discussed in section 5.2.

4.3.3 Keyboard Controls

In the collaboration task, each member of the agent-human team controls the rotation of an axis of the tray. The rotation of the tray causes the ball to move, and the two members are responsible for moving the ball towards the target. The collaboration task is presented in section 3.1 for more information. In this co-learning task, both members need to learn the task while collaborating with the other. In this section, we want to present how the human side collaborates through the use of keyboard controls.

Figures 4.18 and 4.19 are the distributions of normalized changes that the participant made during a game for the No_TL and TL groups respectively. The normalised changes are the number of control frames that the player applied input to make a change to the tray, divided by the total number of control frames played in the game. This measure does not present an objective view of the performance of the participant, but it shows the involvement that a participant has during the game.

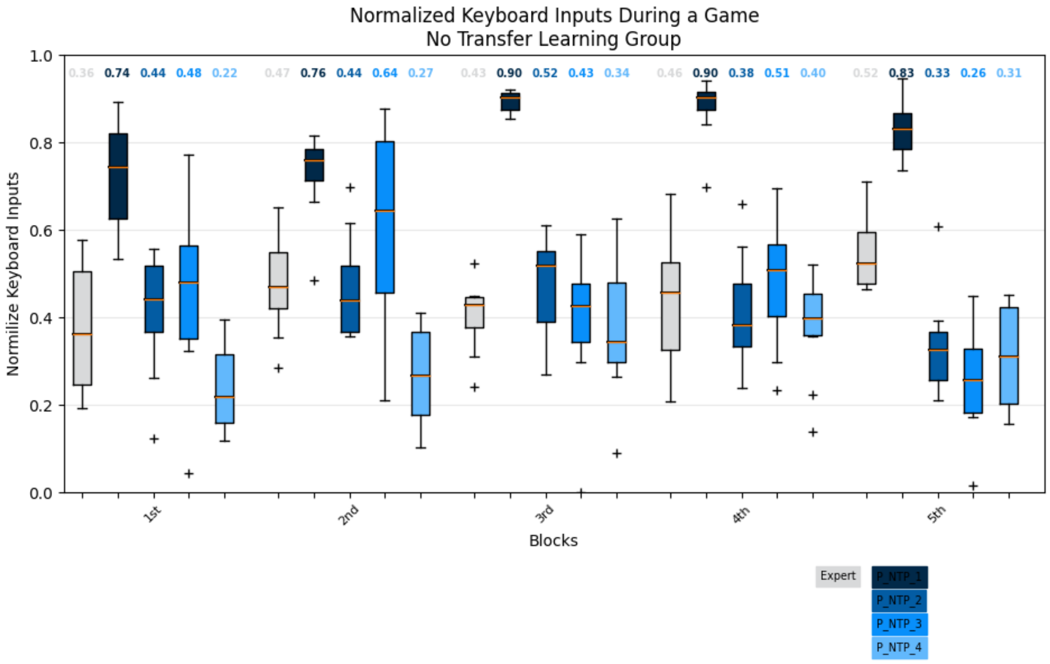


Figure 4.18 No_TL groups Normalized Keyboard Control Inputs during games for each block

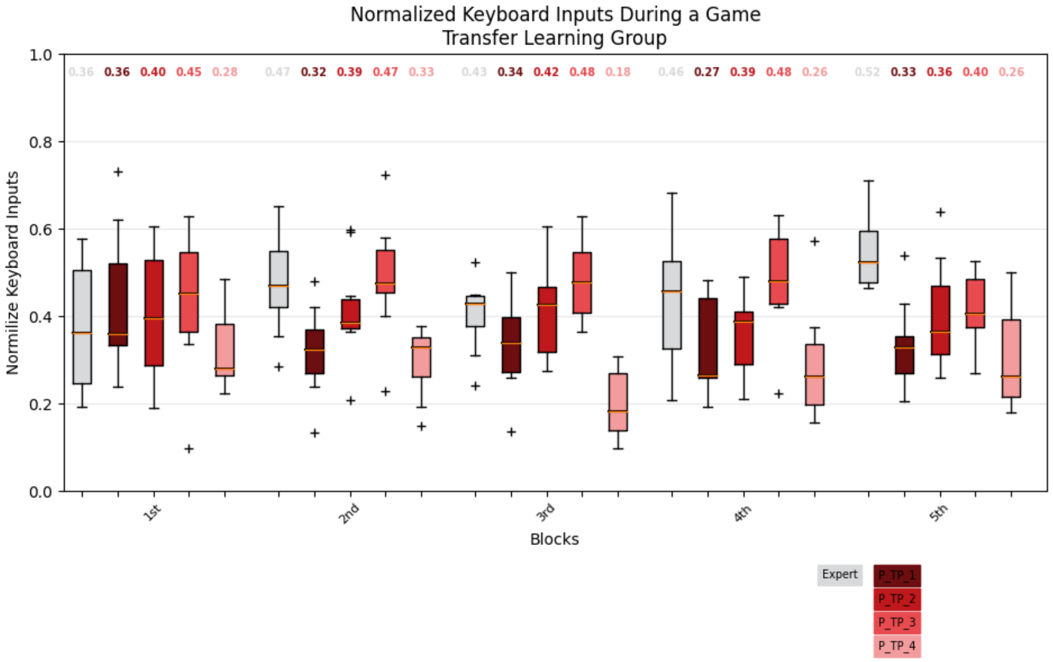


Figure 4.19 TL groups Normalized Keyboard Control Inputs during games for each block

It must be noted that these inputs do not show the way a person applied the inputs but only in how many control frames the person applied an input. For example, if an input was applied in 4 consecutive control frames, we count that as 4 interactions. In this scenario, the participant could simply hold the keyboard input down for the duration of those 4 control frames, or the participant could re-press the keyboard input as little as 2 times or more. Also, the figures (4.18, 4.19) include changes that were applied by the participant, but the tray was already to maximum angle and therefor there was no change in the angle of the tray.

Throughout the collaboration, the expert shows a stable total interaction with values around 0.4 and 0.5. While during the first block, the interactions are a little lower than that, we see it more as an outlier. Overall in both groups, while there is a variance between each participant, toward the end their total interaction are similar or lower to the experts. In the last block, all participants except one had a lower total interaction than the expert.

The most interesting behaviour in this graph is that of participant P_NTL_1. In the section 4.3.1 we presented the paths of the ball during the 1st, 3rd and 5th blocks. In that section, we showed that this participant in the 5th block had aggressive behaviour. We also mention that in the third block, the paths were similar to the experts. Here its obvious that the behaviour of the participants was overall different to the expert's behaviour. Even in that 3rd block where the ball moved in similar ways. This shows that the strategy that this participant followed was from the beginning different to the others.

Participant P_NTL_1 was the only one adopting this strategy, meaning we can't say if it is a bad strategy or not. Using this graph we can see how difference between participants in their approach to the game. This measure is not perfect, as we said earlier it does not show how the inputs are applied and what changes, if any, they made. With that in mind, through these graphs (figures 4.18 and 4.19), we were able to present another view of the different behaviours of our participants. In a bigger sample using measures like this, it is possible to decide if behaviours like this are outliers in the study or common behaviours that a group of people follow.

4.3.4 Rotation of the Tray

In the above section we focus on the total interaction of the participant using the keyboard controls. In this section, we present a view of how both the agent and the participant rotate the tray and how this affects the final result. In section 3.1, we described how the rotation of the tray affects the ball. In this section, we present the rotation of the tray in one randomly selected game from the 1st, 3rd and 5th blocks. For our purpose, in this section, we show only the games the participants P_NTL_1 and P_NTL_3. In the graphs, we also include the inputs of the participants. For an insight into all other participants, we provide more figures in the Appendix 7.2.

In Figure 4.20 are the angles of participants P_NTL_3 team. In the 1st block we see that the agent rotated towards negative 30 degrees and then stayed there. This behaviour of the agent is explained better in section 3.4.2. In the 3rd and 5th blocks, the agent collaborated more. In the previous section, we show how this participant had the most total interactions, in this graph we can see how those interactions were applied.

For more information about the changes both members can make and how they affect the ball see section 3.1. In the first block, the participant made big constant changes clock/anticlockwise. In the 3rd and 4th blocks the participant continues to apply big constant changes but this time more often than in the first block. It seems that the agent does try to follow the sudden continuous changes meaning the agent did learn to follow the behaviour of the participant.

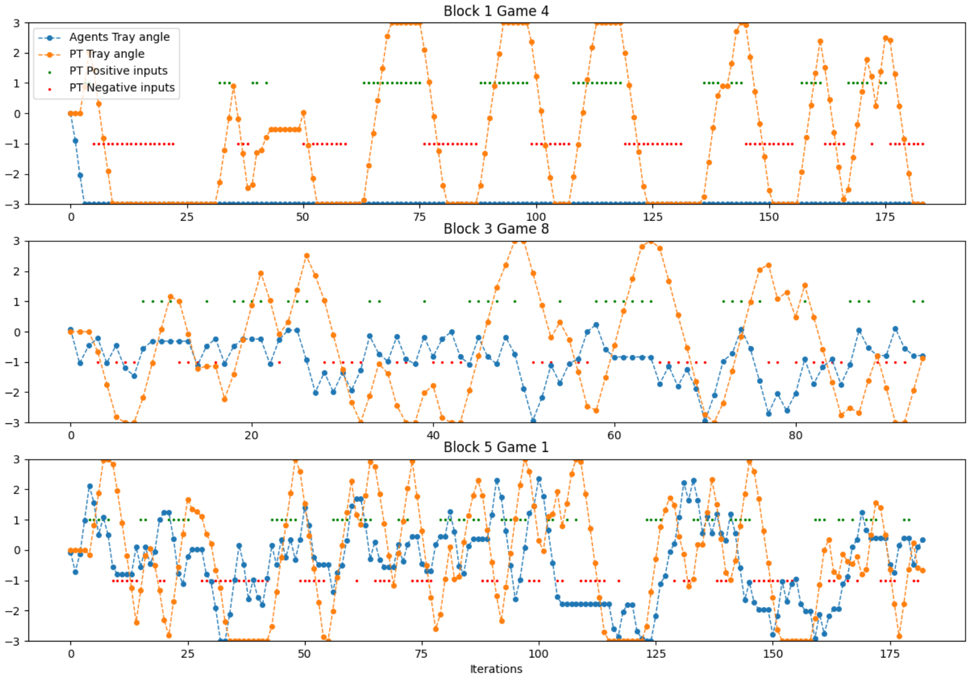


Figure 4.20 Angles of the tray during 3 games of the P_NTL_1 collaboration process. Blue traces for the angle around the x-axis (agent controls), and Orange traces for the angles around the y-axis (participant controls)

In Figure 4.21 are the angles of the participant's P_NTL_3 team. This team failed to increase its performance during the study and achieved victory only 3 times. In the 1st block, the results are similar to others with the agent being the main problem, with it choosing to remain at an angle of negative 30 degrees in the game. In the 3rd and 5th blocks, the agent seems to rarely rotate past 0 degrees into negatives. But also, it was not constantly applying a single decision like the 1st block. While this graph does not provide an explanation on this behaviour, using it we can visualize the problem the participant faced.

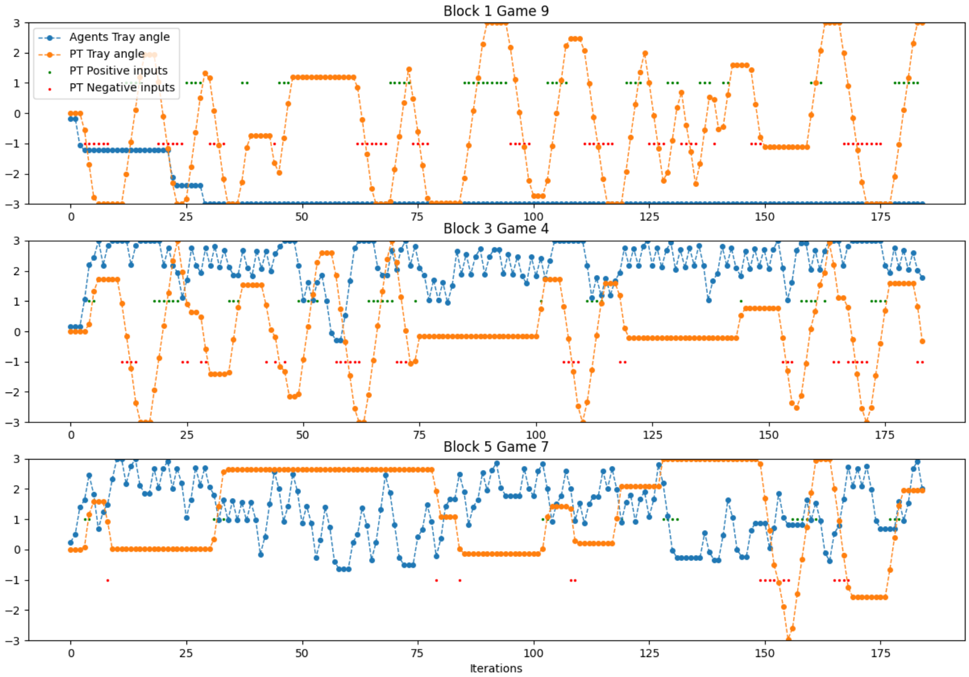


Figure 4.21 Angles of the tray during 3 games of the P_NTL_3 collaboration process. Blue traces for the angle around the x-axis (agent controls), and Orange traces for the angles around the y-axis (participant controls)

In Figure 4.22 are the angles of the participant's P_TL_1 team. In the 1st block, here we see the massive difference in the action of the agent. In both cases above for the No_TL group, the agent fixed the tray into a position and did not collaborate more after that. Here the agent is far more active during the game. In the 3rd and 5th blocks, both members show unison in their action and the games are much smaller in duration.

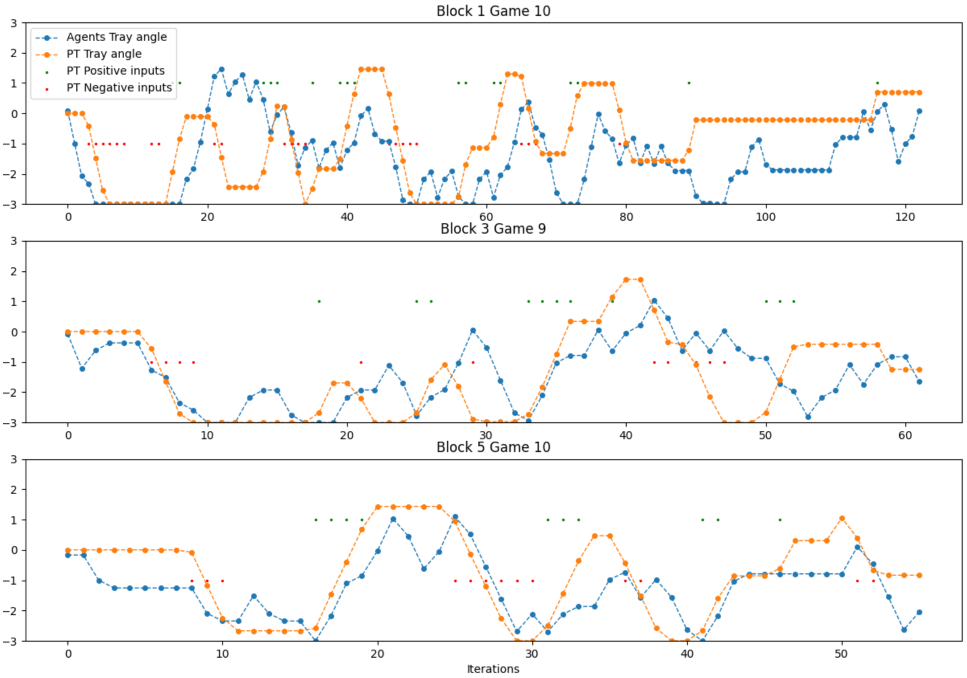


Figure 4.22 Angles of the tray during 3 games of the P_TL_1 collaboration process. Blue traces for the angle around the x-axis (agent controls), and Orange traces for the angles around the y-axis (participant controls)

4.4 Subjective measures:

4.4.1 Judgement of control:

Figures 4.23 and 4.24 show the *Judge of Control* (JoC) responses for each block, for the No_TL and TL groups respectively. Although we have the baseline results here, we do not focus on them because we do not have the objective results to compare with. At each block, the participants responded on a scale from '1' to '9'; '1' meaning having absolutely no control and '9' being total control.

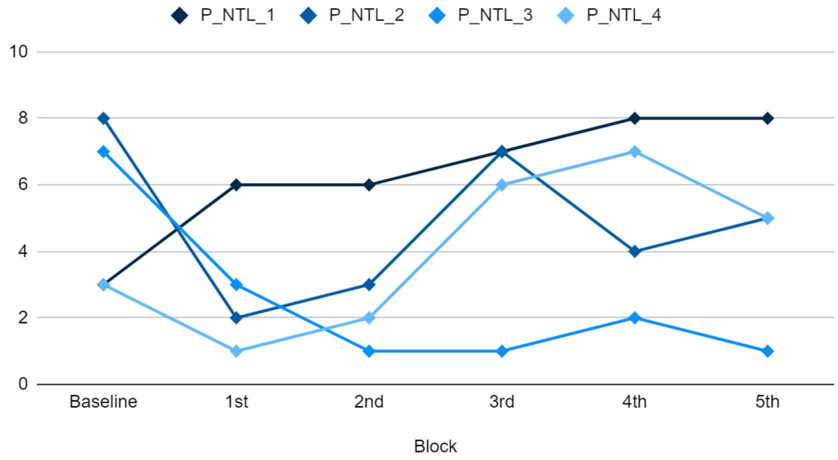


Figure 4.23 Judgement of control for the No_TL group.

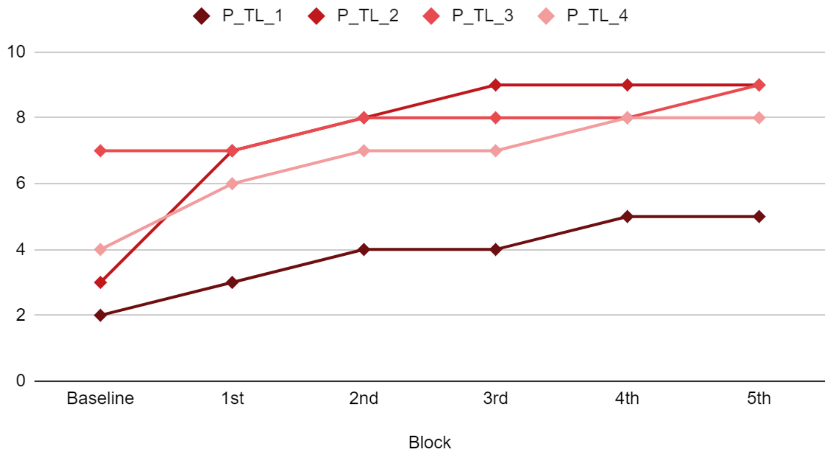


Figure 4.24 Judgement of control for the No_TL group.

In the No_TL group, most participants in the first block reported a low *JoC*. After the first block almost all started to report an increasing *JoC*, with the exception of P_NTL_3. Similar to the drop of performance observed in the objective measures (section 4.2) some participants reported a drop in the JoC during the study.

In comparison the TL group had overall a better JoC. Most participants reported a good JoC during the 1st block and after that the JoC increased for the rest of the study with no drops. Participant P_TL_1 had the lower reported JoC in this group and by far. One possible explanation could be that the experience with the TL algorithm did not have the expected positive effect. But there are also other possible explanations. It could be that the skill of the participant is overall lower and therefore did not have the same JoC with the others. In section 4.1 showing the familiarization results, the participant P_TL_1 was the lower performer by far in the TL group.

In most participants the JoC seems related with their objective teams performance. The exceptions are participants P_NTL_1 and P_TL_1. We described above reasons that could explain the JoC reported by P_TL_1. Similar reasons could had affected the answers of P_NTL_1. Another possible explanation is that these participants could be overall more pessimistic or optimistic. While the Big five scale, does not produce an optimism/pessimism trait, in his work Sharpe et al [60], indicated a strong positive relationship between optimism and the traits of Emotional Stability, Extraversion, Agreeableness, and Conscientiousness. In this work, we do not focus on further analyses of this aspect.

4.4.2 Collaboration metrics

For the subjective collaboration measures, in Figure 4.25 are the measures of *Fluency*, *Trust*, *Teammate Traits*, *Improvement* and *Alliance*. For more information about the included questions on each measure see section 3.5.2.

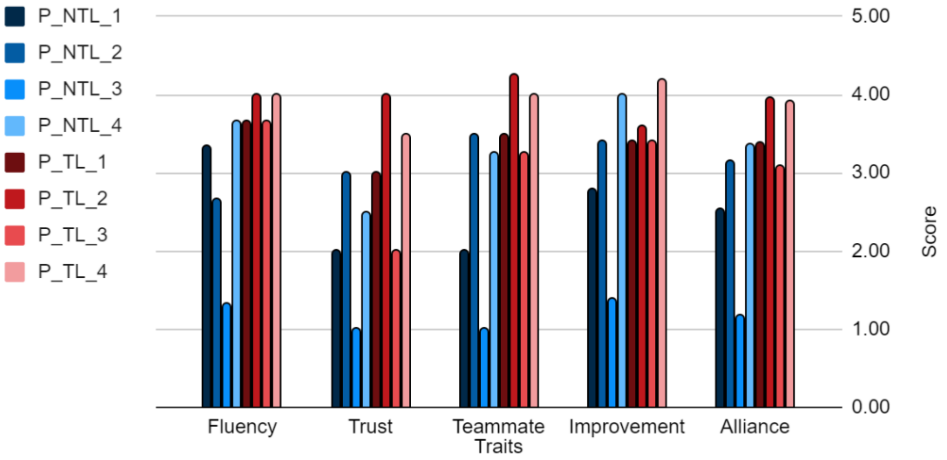


Figure 4.25 Subjective measures of fluency, trust, teammate characteristics, improvement and alliance for all participants.

For all these measures a higher value indicates a more positive attitude. In *Fluency*, *Teammate Traits*, *Improvement* and *Alliance* there seems to be an overall better experience within then TL group. In the No_TL group, there are participants who reported similar experiences to the TL group. Notable P_NTL_4, which had the best performance in the No_TL group, reports a similar experience to the TL group. This indicates that the overall subjective experience is related to the objective team performance.

Regarding the *Trust* that the participants had in the AI, we observe a variance in both groups. In the TL group, on average, participants trusted the AI more than in the No_TL group, but in general, each participant had a very different experience. This might indicate that this measure is not solely related to the experience during the game and it needs to be interpreted in combination with the overall attitude of the participants towards AI.

Figure 4.26 shows the participants' perception regarding the two measures of the *AI Contribution*. In section 3.5.2, we explained the reasons for using two measures instead of one and also the questions each measure has. The two measures are AI contribution in all games (*AIC-all games*) and AI contribution in the last 10 games (*AIC-10 games*).

In these measures, a value higher than 3 (4 or 5) means higher perceived contribution of the AI. A value of 3 means an equal perceived contribution of both members. And a value lower than 3 (1 or 2) means higher perceived contribution of the human participant.

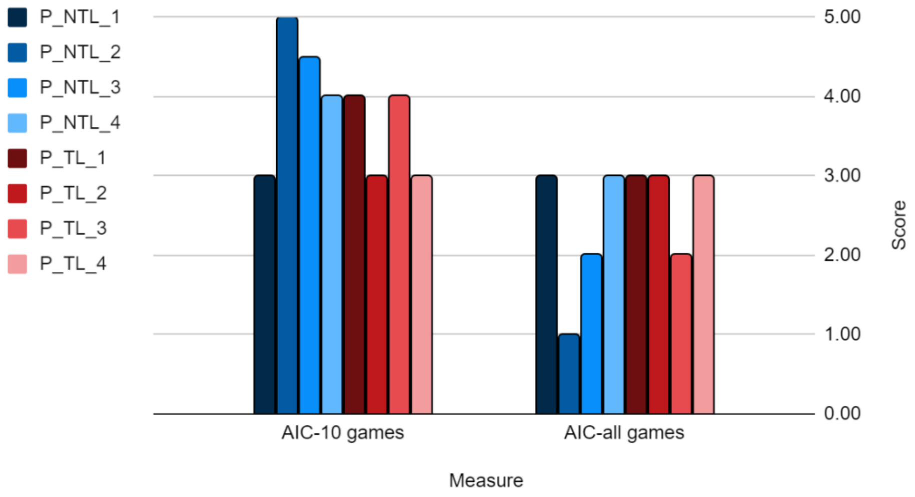


Figure 4.26 Two Subjective measures of AI contribution for all participants. The fiist measure is the AI contribution on the last (10) games (AI C.L. games) and the second is the AI contribution on all games (AI C.A. games)

In *AIC-all*, most participants reported an equal contribution between the AI and themselves. Participant P_NTL_2 had the lower value (1) of all, meaning that this participant perceived that contributed more than the AI. Generally *AIC-all* does not show the same relation with the objective teams performance as other measures shown at the begging of the section.

In *AIC-10,* both groups have overall an increase to the perceived contribution of the AI. Even participant P_NTL_2 where in the *AIC-all* had the lower value, here has the higher value (5).

To better understand the perceived AI contribution during the last 10 games, we used the four questions that were asked in the context of the last 10 games. The four questions are:

1. How do you judge the performance of the team in the last ten games?
2. This performance was a joint result of the team?
3. I had the main responsibility for this performance.
4. The AI system had the main responsibility for this performance.

The second and fourth questions are used for the AIC-10. More information in section 3.5.2.

Each question is a five item Likert scale. The scales are in Figure 4.27. In the trees of Figures 4.28 and 4.29 these answers are clustered in 3 groups: less than 3 (1 or 2), three and more than three (4 or 5). Three refer to the middle answers, which a neutral answer to the question.

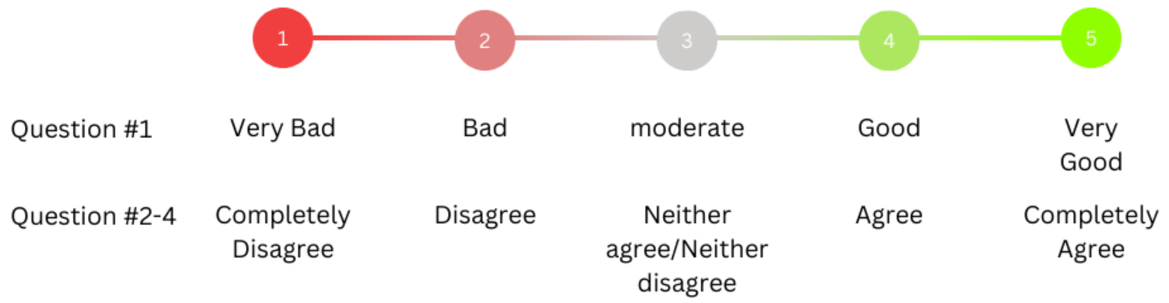


Figure 4.27 Likert scales for questions

In the No_TL group in Figure 4.28, all participants rated the performance as neutral or negative (3 or below 3). The two participants who rated the performance negatively, P_NTL_2 and P_NTL_3, answered that they did not agree that the performance was the result of the team. Both disagreed that they were responsible for the performance and agreed that the AI was responsible for the performance. In the open question at the end, the participant P_NTL_3 said that “the agent is not a smart AI system”. This shows how the participant blames the agent for the final performance. P_NTL_2 did not answer the open question.

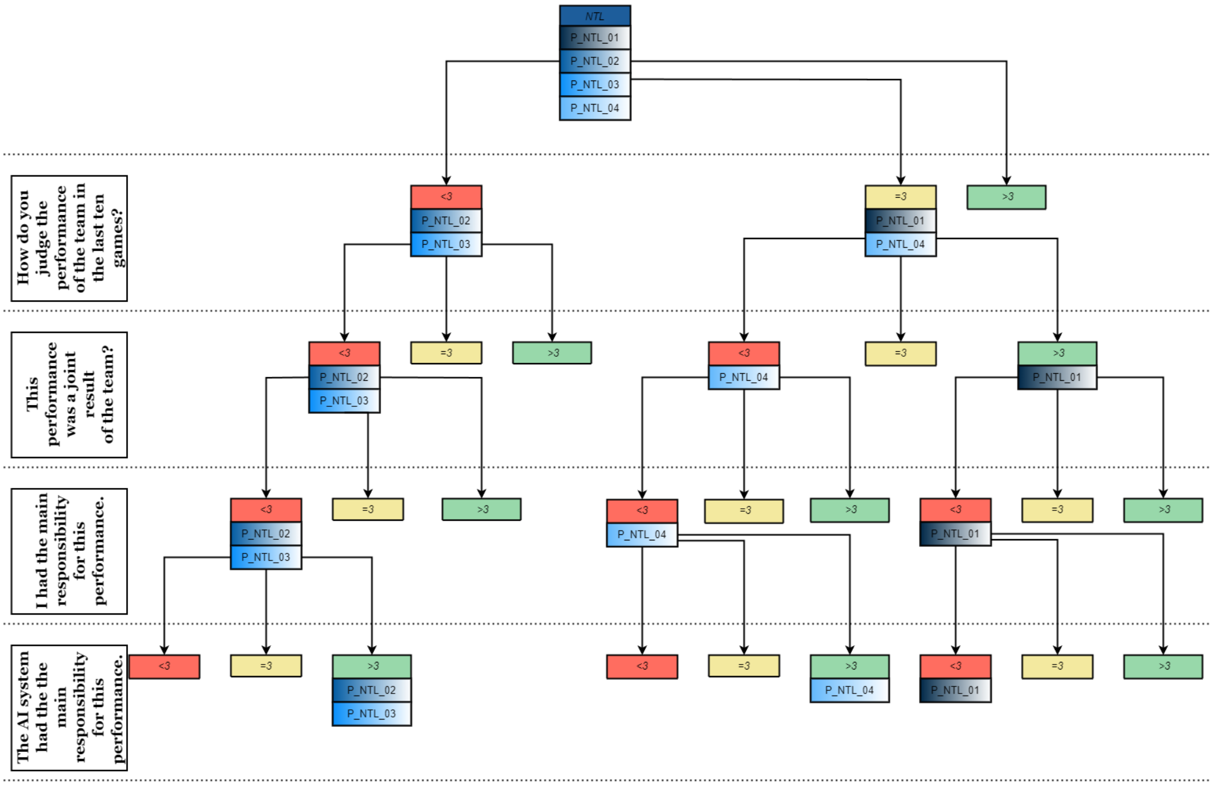

NTL
P_NTL_01
P_NTL_02
P_NTL_03
P_NTL_04
How do you judge the performance of the team in the last ten games?
<3
P_NTL_02
P_NTL_03
=3
P_NTL_01
P_NTL_04
>3
This performance was a joint result of the team?
<3
P_NTL_02
P_NTL_03
=3
>3
<3
P_NTL_04
=3
>3
P_NTL_01
I had the main responsibility for this performance.
<3
P_NTL_02
P_NTL_03
=3
>3
<3
P_NTL_04
=3
>3
<3
P_NTL_01
=3
>3
The AI system had the the main responsibility for this performance.
<3
=3
>3
P_NTL_02
P_NTL_03
<3
=3
>3
P_NTL_04
<3
P_NTL_01
=3
>3

Figure 4.28 No_TL groups answer tree for the question of the Final 10 games.

For the other two participants who rated the performance as 3, P_NTL_1 and P_NTL_4 had a split perception of their experience. P_NTL_4 answered that the performance was not a joint result and that of the two members, the AI was mainly responsible for the performance. In contrast, P_NTL_1 replied that the performance was a joint result and that neither member was predominantly responsible. In the open question, participant P_NTL_1, said that “the agent didn't learn over time, and it could be more cooperative and intelligent in order to achieve our goal”. P_NTL_4 said “ In the end, the improvement slowed down, maybe because I expected the agent to continue learning, that is, to get used to waiting for the agent to make idle movements.”.

In the TL group in Figure 4.29, all participants rated the performance as neutral or higher (3 or above 3). This directly shows the better experience that all participants had. Participant P_TL_1 was the only one who rated the performance with a 3. He also answered that the performance was not a joint result and that the AI was most responsible. This participant did not leave a comment in the open question therefore we do not have any extra information about the experience during the study.

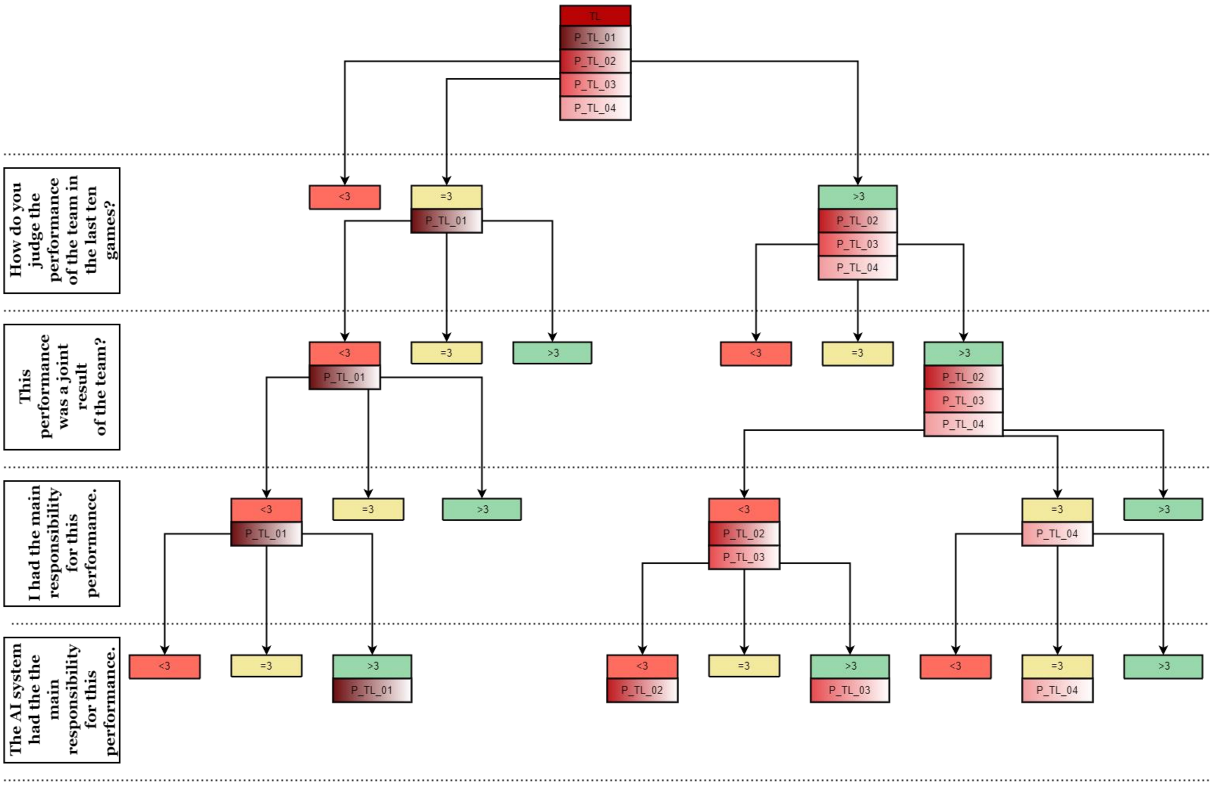

TL
P_TL_01
P_TL_02
P_TL_03
P_TL_04
How do you judge the performance of the team in the last ten games?
<3
=3
P_TL_01
>3
P_TL_02
P_TL_03
P_TL_04
This performance was a joint result of the team?
<3
P_TL_01
=3
>3
<3
=3
>3
P_TL_02
P_TL_03
P_TL_04
I had the main responsibility for this performance.
<3
P_TL_01
=3
>3
<3
P_TL_02
P_TL_03
=3
P_TL_04
>3
The AI system had the the main responsibility for this performance.
<3
=3
>3
P_TL_01
<3
P_TL_02
=3
>3
P_TL_03
<3
=3
P_TL_04
>3

Figure 4.29 TL groups answer tree for the question of the Final 10 games.

Of the other 3 participants who rated the performance higher than 3, all of them answered that the performance was a joint result. P_TL_4 answered with 3 to both questions about which member was more important to the performance, while P_TL_2 answered with lower than 3 to both questions. P_TL_3 answered that the AI was the most important member for the performance. This was an unexpected answer as it contradicts the answer about the performance being a joint result. In the open question, the participant said “It was overall a good cooperation, it was evident that there was a learning process which gradually improved the system's actions”.

5 Conclusions & Discussion

In this chapter, we present a conclusion to our study. Additionally, we also provide subjects for discussion and suggestions for future expansion.

5.1 Conclusion

The aim of this study was to present the benefits of using transfer learning in a co-learning collaboration task between a human participant and an AI agent. In the objective measures, the TL group showed an advantage compared to the No_TL group. The TL group managed to complete the collaboration task in almost all games and in the end, the results were much closer to the expert team. Also on average, the TL group needed less than half the time to complete all games compared to the No_TL group.

On the subjective measures, the TL group also showed a better experience than the No_TL group. In the measures of F*luency*, *Teammate Characteristics*, *Improvement* and *Alliance* the TL group had overall a better experience, but here the difference between the two groups was smaller than that of the objective measures. Some participants of the No_TL group reported similar results to the TL group in some measures. While not having a clear advantage, overall, the TL group showed a more consistent experience than the No_TL group. In the *Trust* measure, again the TL group had better results, but both groups showed a variance among the participants. This could mean that other aspects affect the *trust* that the participants show to the agent. With a bigger sample, we could use the Big Five, personal values, AI attitude questionnaires to better understand what could affect *trust*.

On the *Contribution of AI*, we opted to create two categories that measure the *Contribution of AI in all games and in the final 10 games*. In the *Contribution across all games*, the TL group had consistent results with almost all participants answering that there was an equal contribution from themselves and the AI. The No_TL group overall rated that they contributed more to the performance. For the *AI contribution in the last 10 games*, in section 4.4.2 we showed a different way to present the results. There we see that the No_TL group, while rating that the AI contributed more, this was in a negative narrative, with all participants rating the results in the final 10 games mostly negatively. In contrast, the TL group rated the results of the final 10 games mostly positive and the contribution was equal to both members.

Tsitos et al [30], in a similar work, presented results of using a transfer learning method called probabilistic policy reuse (PPR). The environment and collaborative task were different therefore direct comparison is not possible. PPR helped participants to achieve results similar to the expert after many blocks of games. In comparison, the No_TL group in that work did not manage to achieve results any similar to the expert. Our TL methodology achieved similar results to the expert much faster, but at the same time, some participants from the No_TL group were not far in performance. In the subjective measures, there was a similar increase in the results of the PPR group with our TL group. We did not have a similar sample of participants, and the collaborative task is much different in making any conclusion about the strengths of each method. With that said, the use of demonstration data and a pre-train step in our study allows a better performance in the initial blocks but it could come with the cost of personalization of the agent to the participant. In contrast, PPR could allow for better personalization but require more time to achieve that result. It would be interesting in future work to compare the two methodologies in the same collaboration task.

Finally, in our study, we used questionnaires to capture the personal traits and values of the participants and their attitudes towards AI. While we did not use these extensively in our work, mainly due to the small sample of participants, such material could allow further analysis of our results in future work. We have mentioned in the chapter 4, how we could use the personalities of the participants to explain some results we saw. Another use can be to compare the sample of participants, among different works, based on these personality measures.

5.2 Discussion

In section 4.3.2, we present results that showed a possible correlation between the temperature α and the performance of the teams. Based on these results two questions emerge:

- Does a low performance cause the lower temperature to be (and stay) low, or does the stabilization to a low temperature prevent further improvement of the performance?

- Could a fixed temperature based on the data from the expert, be better than a trained one?

In the first work introducing the SAC agent, Haarnoja et al. [22] used a fixed temperature α as a hyperparameter. This hyperparameter was meant to be optimised by the user based on the needs of the task at hand. In a follow-up work, Haarnoja et al. [61], introduced a temperature α that is updated based on a target entropy. The purpose was to make it easier to find the optimal value. The temperature α and the error that was used to update it, are discussed in section 2.1.2. Based on our results, the use of the variable temperature could be the reason for the agent's negatively change of behaviour during the games of the No_TL group.

A crucial point in any environment, in the context of RL, is how the rewards are applied. In our specific case, all states are rewarded with a -1 except the target state which is rewarded with 10. More information about the environment is in section 3.1. The reward is combined with the Bellman equation, described in section 2.1.2, meaning that when the agent finds the target, during the gradient updates it will start to spread the reward in the rest of the states in the Q-function. As a result, in order to maximize the reward, the agent learned to minimize the path towards the target by passing through as few as possible states, minimizing in this way the '-1' rewards (penalties). The longer paths that are less rewarding are visited less often and therefore do not affect as much the training procedure. This makes the agent focus on the paths that produce the best cumulative reward.

Yet, in the context of SAC, the goal of the agent was to reach the target entropy. The entropy was affected by the probability of the actions in a given state. When the system reached the target entropy, all the actions had the same probability of being selected. This meant that the soft value $(aH(\pi(\cdot\,|s_t)))$ combined with the stochastic policy, allowed paths that are longer, but still reached the target, to have a possibility of being selected. In an extreme scenario, where the target entropy is reached without ever reaching the goal (the ball falls into the target and the agent is rewarded with 10) the policy is only trained on -1 rewards and most probably without exploring the entire state space. This behaviour can explain the bad performance of the No_TL group, while having reached the target entropy.

Considering that in the situation just described above, the preferred policy for an agent would be to start selecting actions that go to states that are not yet visited, we ask the following questions:

- Would a fixed temperature solve the problem?
- Would another description of the target entropy, that allows actions to be selected based on a range of probabilities, instead of equal ones, facilitate further exploration?
- Is there another parameter that could prevent the system from reaching the target entropy without limiting exploration?

In the context of using the SAC in a discrete setting, some works [62,13] already showed that using a different description of the target entropy can provide better results. In our results, it seems that the use of TL can provide a solution to this problem, but could the use of the changes referenced above allow better performance in both groups?

Another aspect that can be changed in the RL agent is the exploration during the 1st block. Generally, discrete SAC uses a soft policy to explore using a soft state value function. This is described in section 2.1.2. Specifically, the soft value $(aH(\pi(\cdot\,|s_t)))$ affected the agent during the training process when we calculate the residual squared error. This meant that during the first block, where the agent had no prior off-line training session, the agent followed the randomly initialized policy. This initialized policy had no function to boost exploration and therefore it simply exploited the random policy. Based on this, during the No_TL group the agent in the 1st did not follow any exploration and was disadvantaged compared if there was an extra function for exploration. The TL group was not affected from this as there was a pre-trained step in the first phase of the TL method.

In his work, Tsitos [30] used a random agent in the first block of No_TL. This would help exploration in the first block and benefit the first training session. We can not compare the No_TL group in our work with the results of Tsitos, due to the difference to the collaboration task, but the use on any extra function for exploration would not have help positive our participants in any way.

In section 3.4.1 we provided results of the first 10 games with 15 different initializations. The use of the random agent in the first block meant that the difference in the policy of the initialized agent would play less role in the overall performance. At the same time, it also prevented the good policies that come in the initialization to provide better results. In the section 3.4.2, we show how the agent interacted in the first block during the games of the expert. In those paths the ball could reach the target relatively simple and then in only required a little more help from the agent to win. Based on this, the use of an e-greedy exploration method in the first block seems a better solution. E-greedy would allow the use of the policy, and therefore if it's a good policy would help the performance of the team, while at the same time provide a needed help if the initialization was worse than the one we used.

5.3 Limitations

Our results and conclusions are limited by the sample of people who attended our study. While we tried to provide enough data to support our findings, what we presented cannot be overstated and a follow up study with a bigger sample is required for a better conclusion. Overall, the main focus on this study should be the use of the different measures that we used and the information that the measures provided to the explanation of the different observation we made on our results.

Another limitation was on the information we could collect from the interaction of the participant with the keyboard. In section 4.3.3, we present graphs that show the changes on the tray from the user. While those results provide a view on the difference in approach of some participants, it is overall incomplete. In our data we cannot present how the participant interacted with the keyboard and also, we cannot show the latency between the moment the participants pressed a button and when their actions changed the tray. In future works more information from the keyboard controls would provide interesting information about the approach of the participants.

6 Bibliography

[1] Matheson, Eloise, et al. "Human–robot collaboration in manufacturing applications: A review." Robotics 8.4 (2019): 100.

[2] Sfikas, Konstantinos, and Antonios Liapis. "Collaborative agent gameplay in the pandemic board game." Proceedings of the 15th International Conference on the Foundations of Digital Games. 2020.

[3] Daronnat, Sylvain, Leif Azzopardi, and Martin Halvey. "Impact of agents' errors on performance, reliance and trust in human-agent collaboration." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 64. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2020.

[4] Kragic, Danica, et al. "Interactive, Collaborative Robots: Challenges and Opportunities." IJCAI. 2018.

[5] Chiriatti, Giorgia, Giacomo Palmieri, and Matteo Claudio Palpacelli. "A framework for the study of human-robot collaboration in rehabilitation practices." Advances in Service and Industrial Robotics: Results of RAAD. Springer International Publishing, 2020.

[6] Van Zoelen, Emma M., Karel Van Den Bosch, and Mark Neerincx. "Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning." Frontiers in Robotics and AI 8 (2021): 692811.

[7] Ehrlich, Stefan K., and Gordon Cheng. "Human-agent co-adaptation using error-related potentials." Journal of neural engineering 15.6 (2018): 066014.

[8] Nikolaidis, Stefanos, David Hsu, and Siddhartha Srinivasa. "Human-robot mutual adaptation in collaborative tasks: Models and experiments." The International Journal of Robotics Research 36.5-7 (2017): 618-634.

[9] Lee, Chang-Shing, et al. "Intelligent agent for real-world applications on robotic edutainment and humanized co-learning." Journal of Ambient Intelligence and Humanized Computing 11 (2020): 3121-3139.

[10] Lee, Chang-Shing, et al. "Ontology-based fuzzy markup language agent for student and robot co-learning." 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2018.

[11] Döppner, Daniel A., Patrick Derckx, and Detlef Schoder. "Symbiotic co-evolution in collaborative human-machine decision making: Exploration of a multi-year design science research project in the Air Cargo Industry." (2019).

[12] Abich IV, Julian, and Eric Sikorski. "Taking a Constructivist Approach to Human-AI Co-learning Design."

[13] Xu, Yaosheng, et al. "Target entropy annealing for discrete soft actor-critic." arXiv preprint arXiv:2112.02852 (2021).

[14] Deterministic vs. Stochastic Policies in Reinforcement Learning https://www.baeldung.com/cs/rl-deterministic-vs-stochastic-policies#:~:text=The%20primary%20difference%20between%20a,over%20actions%20for%20each%20state.

[15] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[16] Amin, Susan, et al. "A survey of exploration methods in reinforcement learning." arXiv preprint arXiv:2109.00157 (2021).

[17] Grondman, Ivo, et al. "A survey of actor-critic reinforcement learning: Standard and natural policy gradients." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012): 1291-1307.

[18] Mathewson, Kory W., and Patrick M. Pilarski. "Actor-critic reinforcement learning with simultaneous human control and feedback." arXiv preprint arXiv:1703.01274 (2017).

[19] Gabrielli, Guglielmo, and Cristian Secchi. "An actor-critic strategy for a safe and efficient human robot collaboration." 2021 20th International Conference on Advanced Robotics (ICAR). IEEE, 2021.

[20] Labao, Alfonso B., and Prospero C. Naval. "AC2: A policy gradient actor with primary and secondary critics." 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018.

[21] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[22] Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." International conference on machine learning. PMLR, 2018.

[23] Maximum Entropy Policies in Reinforcement Learning & Everyday Life https://awjuliani.medium.com/maximum-entropy-policies-in-reinforcement-learning-everyday-life-f5a1cc18d32d#:~:text=Because%20RL%20is%20all%20about,the%20actions%20an%20agent%20takes.

[24] Fu, Michael C. "Chapter 19 Gradient Estimation." Simulation 13 (2006): 575-616.

[25] Deep Deterministic Policy Gradient https://spinningup.openai.com/en/latest/algorithms/ddpg.html#deep-deterministic-policy-gradient

[26] Fujimoto, Scott, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." International conference on machine learning. PMLR, 2018.

[27] Christodoulou, Petros. "Soft actor-critic for discrete action settings." arXiv preprint arXiv:1910.07207 (2019).

[28] Semeraro, Francesco, Alexander Griffiths, and Angelo Cangelosi. "Human–robot collaboration and machine learning: A systematic review of recent research." Robotics and Computer-Integrated Manufacturing 79 (2023): 102432.

[29] Shafti, Ali, et al. "Real-world human-robot collaborative reinforcement learning." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020.

[30] Tsitos, Athanasios C., and Maria Dagioglou. "Enhancing team performance with transfer-learning during real-world human-robot collaboration." arXiv preprint arXiv:2211.13070 (2022).

[31] Lygerakis, Fotios, Maria Dagioglou, and Vangelis Karkaletsis. "Accelerating human-agent collaborative reinforcement learning." The 14th PErvasive Technologies Related to Assistive Environments Conference. 2021.

[32] Paliga, Mateusz. "The Relationships of Human-Cobot Interaction Fluency with Job Performance and Job Satisfaction among Cobot Operators—The Moderating Role of Workload." International Journal of Environmental Research and Public Health 20.6 (2023): 5111.

[33] Zhu, Zhuangdi, et al. "Transfer learning in deep reinforcement learning: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
[34] Hester, Todd, et al. "Deep q-learning from demonstrations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[35] Silvervarg, Annika, and Arne Jönsson. "Subjective and objective evaluation of conversational agents in learning environments for young teenagers." Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Vol. 84. 2011.

[36] Foster, Mary Ellen, Manuel Giuliani, and Alois Knoll. "Comparing objective and subjective measures of usability in a human-robot dialogue system." Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009.

[37] Silva, Andrew, et al. "Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction." International Journal of Human–Computer Interaction 39.7 (2023): 1390-1404.

[38] Riedelbauch, Dominik, Nico Höllerich, and Dominik Henrich. "Benchmarking Teamwork of Humans and Cobots–An Overview of Metrics, Strategies, and Tasks." IEEE Access (2023).

[39] Hoffman, Guy. "Evaluating fluency in human–robot collaboration." IEEE Transactions on Human-Machine Systems 49.3 (2019): 209-218.

[40] Cerekovic, Aleksandra, Oya Aran, and Daniel Gatica-Perez. "How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits." Human Behavior Understanding: 5th International Workshop, HBU 2014, Zurich, Switzerland, September 12, 2014. Proceedings 5. Springer International Publishing, 2014.

[41] Matthews, Gerald, et al. "Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems." Personality and individual differences 169 (2021): 109969.

[42] Hussain, Sadaf, et al. "Trait based trustworthiness assessment in human-agent collaboration using multi-layer fuzzy inference approach." IEEE Access 9 (2021): 73561-73574.

[43] Schepman, Astrid, and Paul Rodway. "Initial validation of the general attitudes towards Artificial Intelligence Scale." Computers in human behavior reports 1 (2020): 100014.

[44] Goldberg, Lewis R. "An alternative" description of personality": the big-five factor structure." Journal of personality and social psychology 59.6 (1990): 1216.

[45] Ι. Τσαoύσης, Μ. Βακόλα, Σ. Γεωργιάδης. ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ ΑΥΤΟ-ΑΞΙΟΛΟΓΗΣΗΣ ΤΗΣ ΠΡΟΣΩΠΙΚΟΤΗΤΑΣ

[46] Schepman, Astrid, and Paul Rodway. "The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust." International Journal of Human–Computer Interaction 39.13 (2023): 2724-2741.

[47] Schwartz, Shalom H. "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries." Advances in experimental social psychology. Vol. 25. Academic Press, 1992. 1-65.

[48] Schwartz, Shalom H. "An overview of the Schwartz theory of basic values." Online readings in Psychology and Culture 2.1 (2012): 11.

[49] Vecchione, Michele, et al. "Personal values and political activism: A cross-national study." British journal of psychology 106.1 (2015): 84-106.

[50] Narkhede, Meenal V., Prashant P. Bartakke, and Mukul S. Sutaone. "A review on weight initialization strategies for neural networks." Artificial intelligence review 55.1 (2022): 291-322.

[51] Sharpe, J. Patrick, Nicholas R. Martin, and Kelly A. Roth. "Optimism and the Big Five factors of personality: Beyond neuroticism and extraversion." Personality and Individual Differences 51.8 (2011): 946-951.

[52] Schulman, John, et al. "Trust region policy optimization." International conference on machine learning. PMLR, 2015.

[53] Guo, Yaohui, X. Jessie Yang, and Cong Shi. "Reward Shaping for Building Trustworthy Robots in Sequential Human-Robot Interaction." arXiv preprint arXiv:2308.00945 (2023).

[54] Huang, Bingling, and Yan Jin. "Reward shaping in multiagent reinforcement learning for self-organizing systems in assembly tasks." Advanced Engineering Informatics 54 (2022): 101800.

[55] Argall, Brenna D., et al. "A survey of robot learning from demonstration." Robotics and autonomous systems 57.5 (2009): 469-483.

[56] Fachantidis, Anestis, et al. "Transfer learning via multiple inter-task mappings." Recent Advances in Reinforcement Learning: 9th European Workshop, EWRL 2011, Athens, Greece, September 9-11, 2011, Revised Selected Papers 9. Springer Berlin Heidelberg, 2012.

[57] Yang, Wei, et al. "Human grasp classification for reactive human-to-robot handovers." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020.

[58] Barrick, Murray R., and Michael K. Mount. "The big five personality dimensions and job performance: a meta-analysis." Personnel psychology 44.1 (1991): 1-26.

[59] Schepman, Astrid, and Paul Rodway. "The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust." International Journal of Human–Computer Interaction (2022): 1-18.

[60] Sharpe, J. Patrick, Nicholas R. Martin, and Kelly A. Roth. "Optimism and the Big Five factors of personality: Beyond neuroticism and extraversion." Personality and Individual Differences 51.8 (2011): 946-951.

[61] Haarnoja, Tuomas, et al. "Soft actor-critic algorithms and applications." arXiv preprint arXiv:1812.05905 (2018).

[62] Zhou, Haibin, et al. "Revisiting discrete soft actor-critic." arXiv preprint arXiv:2209.10081 (2022).

[63] Avery Parkinson, The Epsilon-Greedy Algorithm for Reinforcement Learning https://medium.com/analytics-vidhya/the-epsilon-greedy-algorithm-for-reinforcement-learning-5fe6f96dc870

[64] Huang, Yi-Ching, et al. "Human-AI Co-learning for data-driven AI." arXiv preprint arXiv:1910.12544 (2019).

[65] Şahin, Murat, and Eren Aybek. "Jamovi: an easy to use statistical software for the social scientists." International Journal of Assessment Tools in Education 6.4 (2019): 670-692.

7 Appendix

7.1 Participant paths

In this appendix we provide the paths of the ball of all participants. Section 4.3.1 has a further analysis about the experts along with the participants P_NTL_1, P_NTL_3, and P_TL_1 . The paths of those participants and the experts are not included in this appendix.

In Figures 7.1 and 7.2 are the remaining participants of the No_TL group. And In Figures 7.3, 7.4 and 7.5 are the remaining participants for the TL group.

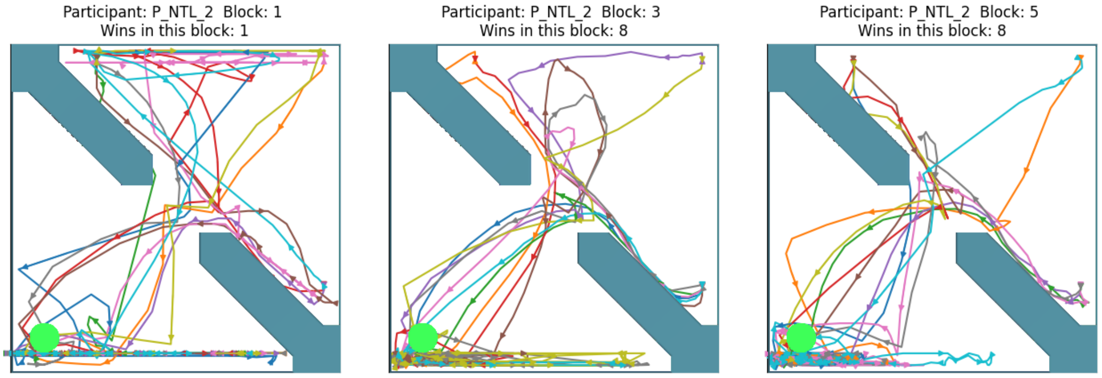


Figure 7.1 The paths of the ball from the 1st, 3rd, and 5th block during the P_NTL_2's collaboration with the agent.

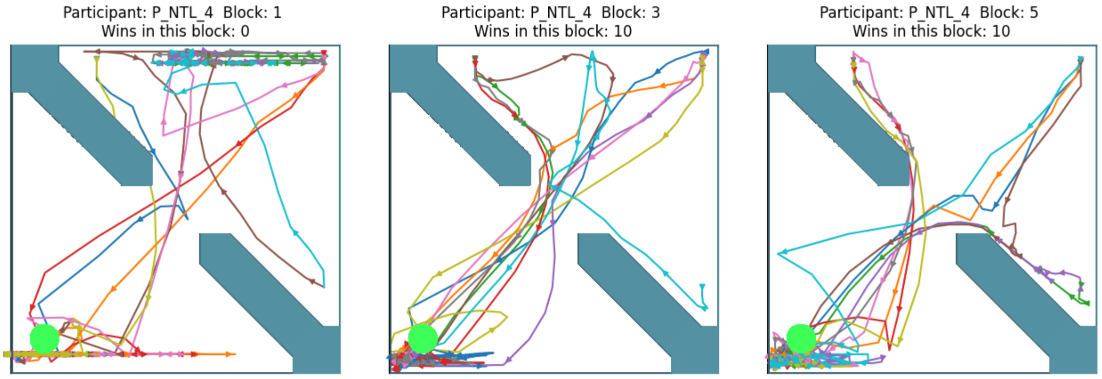


Figure 7.2 The paths of the ball from the 1st, 3rd, and 5th block during the P_NTL_4's collaboration with the agent.

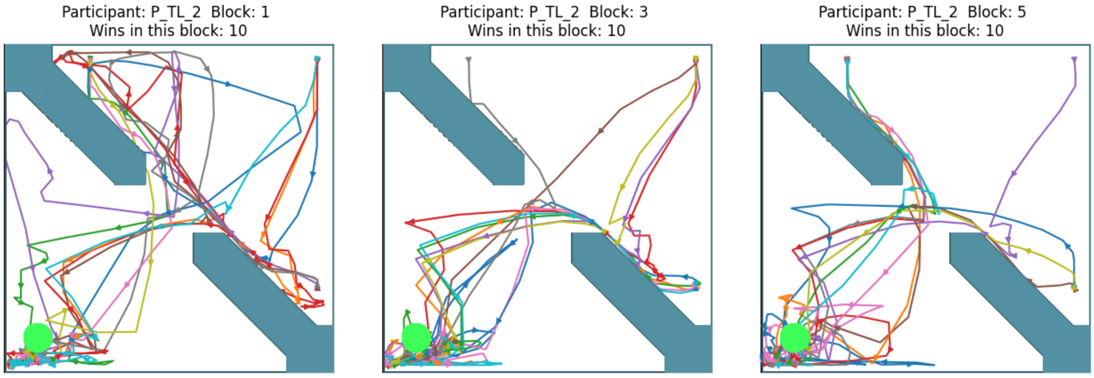


Figure 7.3 The paths of the ball from the 1st, 3rd, and 5th block during the P_TL_2's collaboration with the agent.

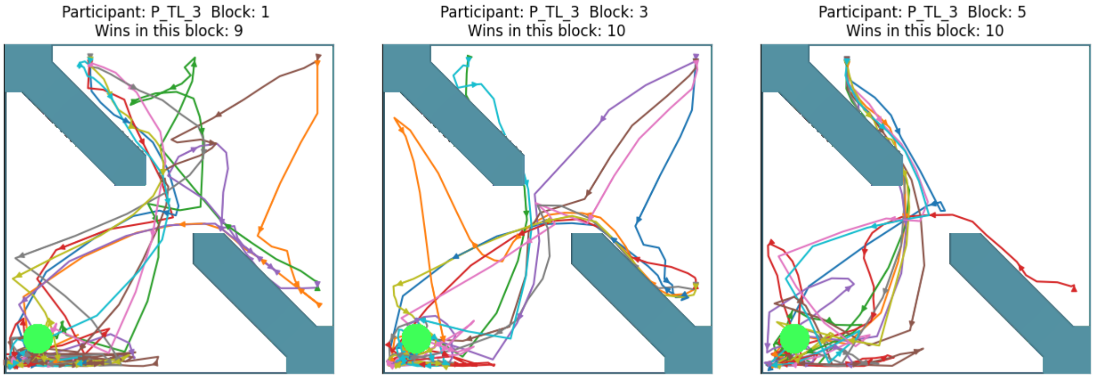


Figure 7.4 The paths of the ball from the 1st, 3rd, and 5th block during the P_TL_3's collaboration with the agent.

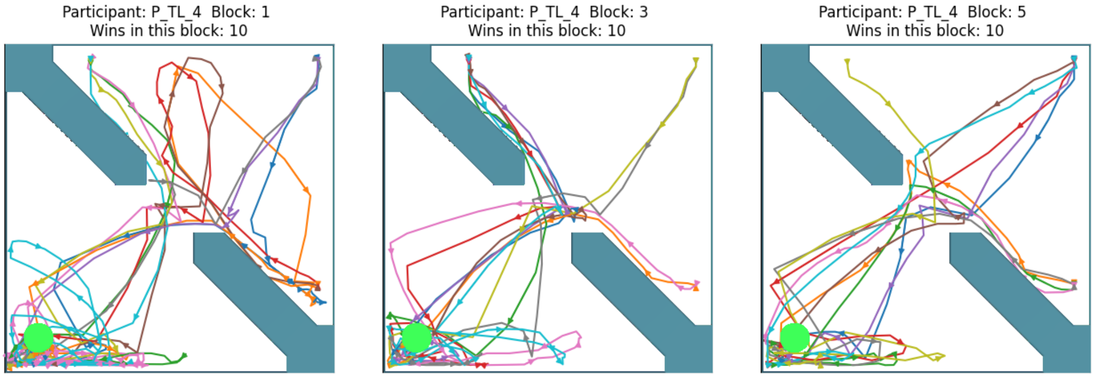


Figure 7.5 The paths of the ball from the 1st, 3rd, and 5th block during the P_TL_4's collaboration with the agent.

7.2 Control Inputs and Tray Changes

In this appendix we provide the changes in angle of the tray graphs from 3 games across the first, third and fourth blocks. Section 4.3.4 has a further analysis about the participants P_NTL_1, P_NTL_3 and P_TL_1.

In Figure 7.6 has 3 games from the experts playthrough. Figures 7.7 and 7.8 have the remaining participants of the No_TL group . Figures 7.9, 7.10 and 7.11 have the remaining participants of the TL group.

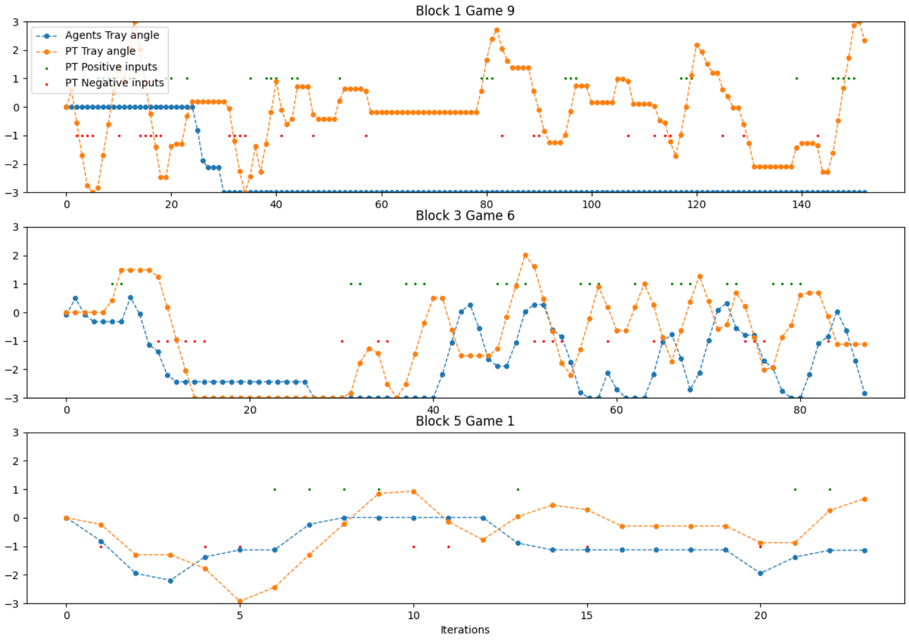


Figure 7.6 Tray angles and players controls during the experts collaboration.

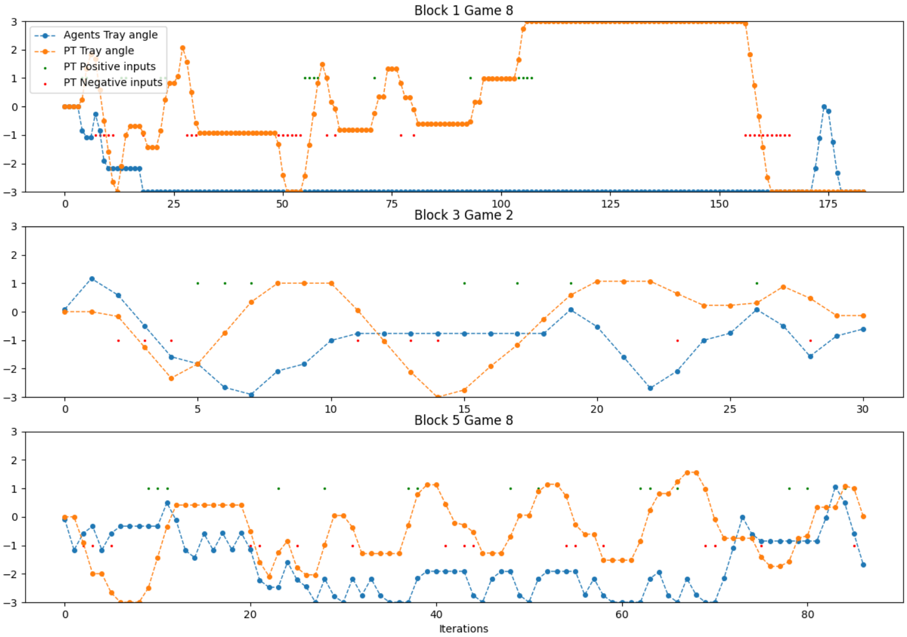


Figure 7.7 Tray angles and players controls during the P_NTL_2 collaboration.

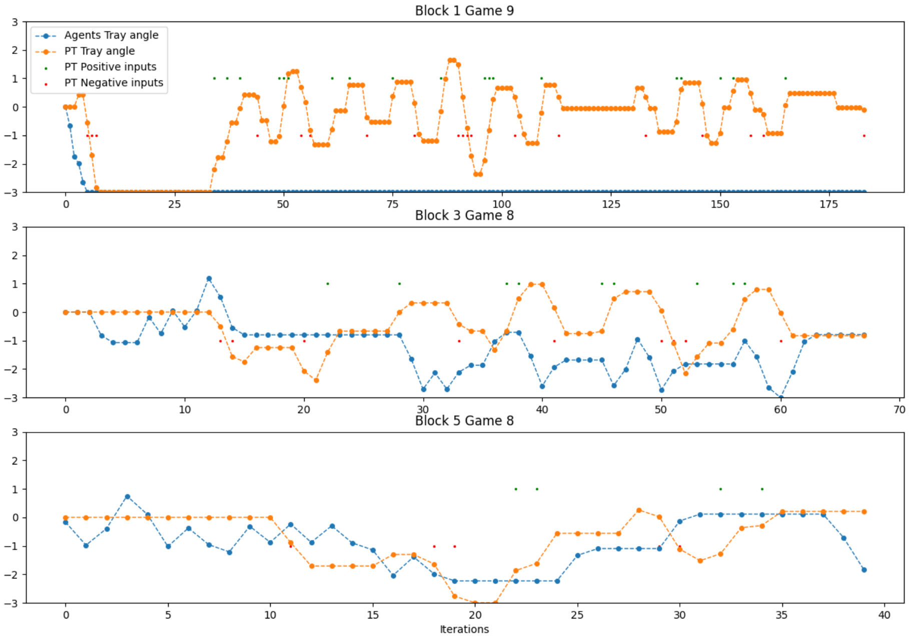


Figure 7.8 Tray angles and players controls during the P_NTL_4 collaboration.

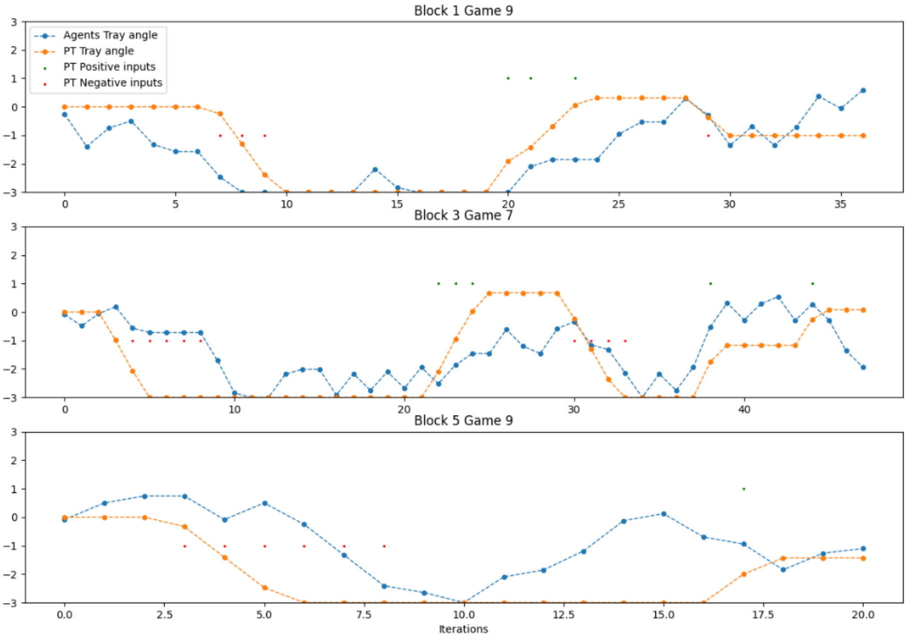


Figure 7.9 Tray angles and players controls during the P_TL_2 collaboration.

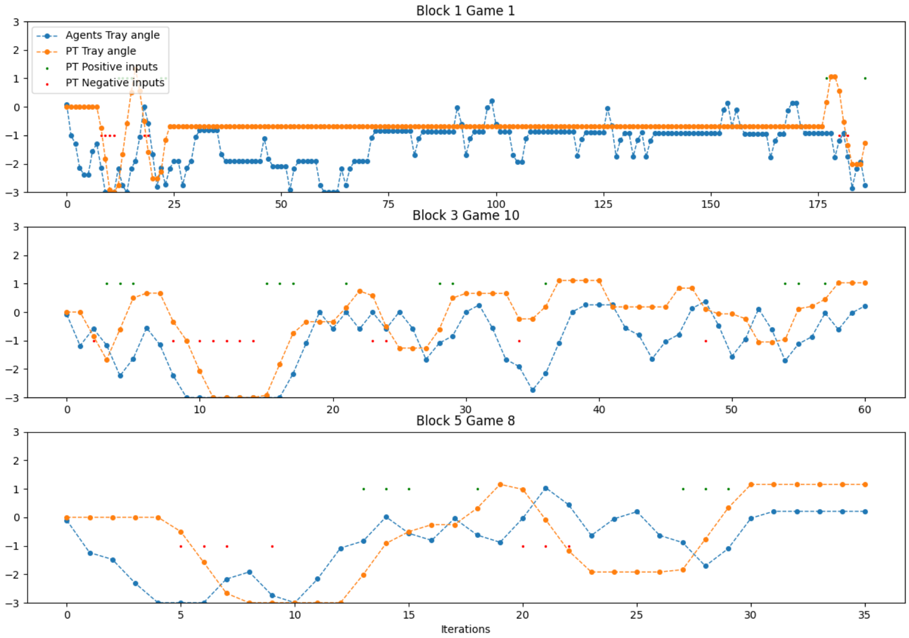


Figure 7.10 Tray angles and players controls during the P_TL_3 collaboration.

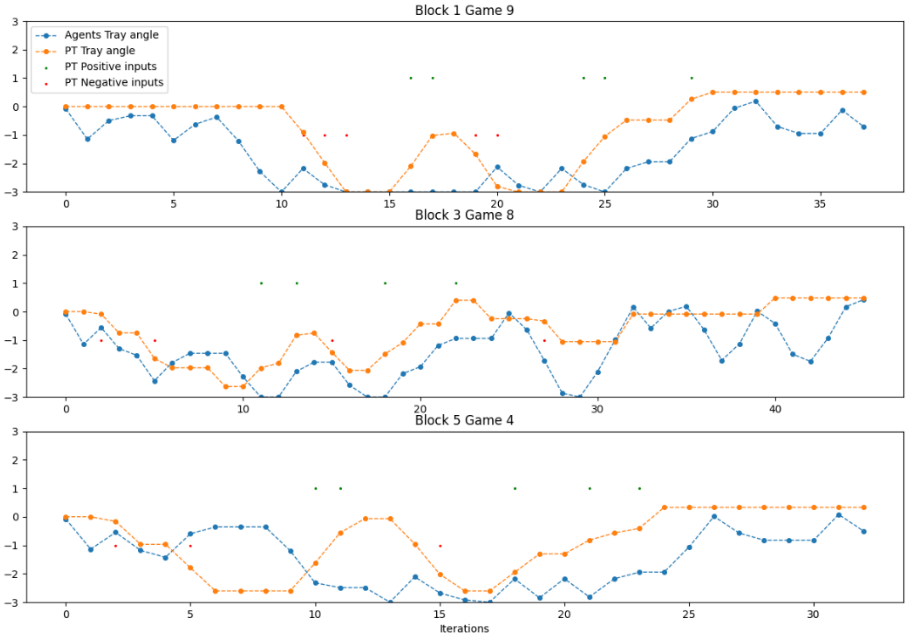


Figure 7.11 Tray angles and players controls during the P_TL_4 collaboration.