Co-Learning in Hybrid Teams with Varying Robot Personalities

J.R. Dolfin

Delft University of Technology

TU Delft

Co-Learning in Hybrid Teams with Varying Robot Personalities

by

J.R. Dolfin

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuseday March 11, 2025 at 15:00.

Student number:	5560047	
Project duration:	March 1, 2024 – March 1, 2025	
Thesis committee:	Dr. ir. L. Peternel,	TU Delft, supervisor
	Ir. E.M. van Zoelen,	TU Delft, supervisor
	Dr. ir. J. Kober,	TU Delft

TU Delft

Preface

This thesis marks the culmination of my Master's research in Robotics at Delft University of Technology. Throughout this project, I have explored the influence of robot personality on co-learning dynamics in human-robot interaction. The journey has been both challenging and rewarding, providing me with valuable insights and practical experience that I will carry forward in my career.

I would like to express my gratitude to my supervisors, Luka Peternel and Emma van Zoelen, for their guidance and feedback during the development of this work. Their expertise has been valuable in helping me refine my research approach and methodology.

I also extend my appreciation to the PhD candidates at the lab for their technical assistance and collaborative input throughout the research process. Their contributions have enhanced the quality of this project in numerous ways.

I am particularly grateful to Nicky Mol for his significant assistance with implementing the robot controller, which was essential to this study's experimental setup. His technical knowledge and willingness to share expertise were crucial to successfully completing this work.

Finally, I extend my appreciation to all those who supported me during this research process. I hope the findings presented in this work will contribute meaningfully to the field of human-robot interaction and inspire further research in this exciting area.

J.R. Dolfin
Delft, March 2025

Contents

Co-Learning in Hybrid Teams with Varying Robot Personalities

Jesse Dolfin
Faculty: Mechanical Engineering
Department: Cognitive Robotics
TU Delft

***Abstract*—This study examines how fixed robot personalities (patient, impatient, leader, follower) influence co-learning in human-robot teams by answering the research question: *How do different robot personalities influence co-learning*. To do this, we implemented a reinforcement learning framework for a handover task where a robot and human participant co-learn to solve a task. The robot has personalities encoded along two axes: *patient/impatient* (via motion speed and stiffness) and *leader/follower* (via exploration rates and reward structures in phased Q-learning).**

Through a within-subject design, we analyze policy metrics and human perceptions. While task success rates remain stable, strategy and internal policy metrics vary significantly. This underpins the key finding: robot personality does not affect task performance since humans can adapt to overcome subtle differences in robot personality. However, robot personality significantly affects how the collaboration is performed as human-robot teams adopt different strategies for different robot personalities. Results demonstrate that robot personality is salient for differences in physical behaviour yet is unperceivable for modifications of internal parameters like exploration rate/decay and reward function for short interactions. This work bridges a critical gap in understanding how static robot traits shape collaborative adaptation, even when overt performance metrics remain unchanged.

I. Introduction

Human-robot collaboration has grown rapidly in recent years, driven by advancements in robotics and artificial intelligence (AI) [1], [2]. Robots are moving beyond repetitive tasks and are becoming collaborative agents capable of more dynamic and intuitive interactions with humans [3]. These systems, often called human-AI systems or hybrid teams, involve humans and robots working together on a shared task [4]. Hybrid teams are now widely used in manufacturing, healthcare, and service industries, where robots support humans in completing increasingly complex tasks [5].

Designing robots for effective collaboration requires a clear understanding of the dynamics in these hybrid teams. Co-learning studies this collaborative process, focusing on how both parties adapt to each other to achieve shared goals [6].

Since co-learning relies on team members adapting to each other and collaborating effectively, strong team cohesion is essential for successful outcomes [7]. The personality traits of team members, human or robot, play a critical role in building cohesion and shaping collaboration [8].

A critical but underexplored aspect of human-robot interaction (HRI) is the interplay between human and robot personality. While personality in HRI has been studied, most research has focused on human personality, with relatively few studies investigating how robot personality contributes to interaction outcomes. As noted in [9], there is an ongoing debate in the HRI community regarding whether human-robot personality matching leads to better interactions or whether strategic mismatches can enhance collaboration. However, these discussions remain inconclusive due to the limited number of studies explicitly examining robot personality. Addressing this gap is essential to developing robust design principles for hybrid teams.

One key step toward closing this gap is understanding how robots can exhibit personality in the first place. Studies show that, just like humans, robots can display distinct 'personalities' through their behaviours and decision-making patterns [10], [11]. These personalities influence how humans perceive and adapt to robots in collaborative settings. Since co-learning relies on mutual adaptation, a robot's personality may be crucial in shaping team cohesion and long-term learning dynamics.

Despite the growing body of co-learning research, studies have not yet directly examined the role of robot personality in these scenarios. Prior work has explored co-learning in simulated environments [4], [6], wizard-of-oz setups [12], and, more recently, in physically embodied environments [13]. However, these studies have not focused on the influence of robot personality.

This study aims to close this gap by examining how different robot personalities affect the co-learning process in hybrid teams. We focus on four robot personality types- patient, impatient, leader, and follower- implemented in a co-learning scenario. The study addresses the research question: *How do different robot personalities influence co-learning?* These personality types remain constant throughout the study, excluding dynamic personality adaptation to the human partner. The experiment involves one human and one robot performing a handover task based on the setup by Veldman-Loopik [13], where the robot hands over an object to a human engaged in

a secondary task. The robot can adapt and learn from the task through a novel q-learning approach that allows for the expression of personality through learning specific parameters and by coupling the algorithm with a finite state machine (FSM). The algorithm is explained in detail in section III-C.

The key takeaway of our results is that robot personality significantly impacts collaboration strategies and internal policy metrics (e.g., action consistency) even when task performance metrics are unaffected. This is likely due to human adaptability, closing the performance gap between personality types. Furthermore, we show that designing personality along physical vectors is salient, while personality modulated by internal metrics (e.g., epsilon decay) is not noticeable in short interactions. We provide the necessary background to support our method in Section II; here, we review co-learning, reinforcement learning, and personality modelling. To study the effects of personality, we created an experiment whose design is outlined in Sections III-IV; here, we detail our Q-learning framework and define our metrics. Lastly, we present our work in Section V and discuss these results in Section VI.

II. Background

To study how robot personalities affect co-learning, we need to understand three things: (1) how robots and humans adapt to each other (co-learning), (2) how robots can learn from experience (reinforcement learning), and (3) how to give robots consistent ”personalities” through their actions. This section explains what prior research tells us about these topics to support our design decisions and analyze how personality affects collaboration. We build on these ideas in our methods (Section III) and results (Section V).

A. Co-learning

We will build on the co-learning framework to understand how robots and humans adapt to one another. Van Zoelen [6] outlines two phases in this process: **Implicit Co-Adaptation**: Partners unconsciously adjust behaviors during interaction. For example, a robot might slow its movements to match a human’s pace, with neither party explicitly planning this change. 2. **Explicit Reinforcement**: Partners then formalize successful strategies through direct feedback (e.g., reward signals or verbal communication), making these behaviours repeatable in new contexts.

In their work, Van Zoelen [12] studied these adaptations. They identified four main categories of interaction patterns during the co-adaptation phase.

1) *Sudden adaptations* happen when the human or the robot quickly adapts their leader or follower role in response to an event in the task or a partner’s behaviour. This could include, for example, the human changing direction or the robot suddenly leading the task.
2) *Stable situations* are interactions *between* adaptations, where one of the partners (either human or robot) leads while the other follows. These are steady and recurring patterns of interaction without much change, like the human leading and pulling the robot along.
3) *Gradual adaptations* are slow transitions where the human or the robot gradually shifts their role from leader to follower (or vice versa). This process often occurs as they learn more about their partner’s behaviour.
4) *Active negotiations* involve a series of short, back-and-forth adaptations between humans and robots. They eventually transition to a new, stable situation through alternating minor adaptations.

These categories provide a basis for analyzing the various strategies that can emerge in hybrid teams, which will be discussed further in Section V.

Veldman-Loopik [13] expanded on co-learning by developing a physical co-learning setup and introducing several tools to evaluate co-learning dynamics in embodied environments:

- Performance Rate: Quantitative analysis of the team’s success rate based on task completion.
- Human Perception of Fluency: Participant questionnaires to assess subjective experiences of collaboration fluency, as per the works of G. Hoffman [14].
- Strategy Identification: Analysis of interaction patterns using video footage and Q-table data to identify emerging collaboration strategies.
- Robot Action Preferences: Analysis of action selection frequencies through the Q-table to observe adaptation trends over time.
- Qualitative Feedback: Post-experiment interviews or questionnaires to gather subjective insights from participants on their collaboration experience.

These tools offer valuable metrics for evaluating co-learning and will be used to assess how robot personality influences the co-learning process.

Veldman-Loopik also outlined five design requirements to ensure the presence of co-learning in a task [13], which inform our task design in Section III:

R1 The method ensures hard dependencies and allows for soft dependencies between humans and robots in both directions.
R2 The robot and human team members must learn at a comparable pace.
R3 Both the human and the robot are rewarded similarly based on their collaborative performance.
R4 The reinforcement learning algorithm can continuously adapt its behaviour during all stages of the learning process.
R5 The human and the robot must be able to observe each other’s state and actions, and neither should have any observability advantages.

To clarify the first requirement, a **hard dependency** exists when neither team member can complete the task alone, necessitating collaboration. A **soft dependency** is not essential for task completion but offers opportunities to enhance performance. The hard dependency is inherent in a handover task, as the handover cannot occur without both parties. Soft dependencies emerge in how the handover is executed; for example, misalignment in expectations about hand orientation can lead to failures, prompting both team members to adapt for more effective collaboration.

B. Reinforcement Learning

For robots to adapt dynamically in co-learning scenarios, especially when embodying distinct personalities, we employ reinforcement learning (RL) as the robot's adaptive mechanism. RL algorithms solve Markov Decision Processes (MDPs), which model interactions between an agent (robot) and its environment (including the human partner). An MDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r \rangle$, where:

- $\mathcal{S}$: Set of possible states.
- $\mathcal{A}$: Set of available actions.
- $\mathcal{T}(s, a)$: Transition function determining the next state s' given current state s and action a.
- $r(s, s')$: Immediate reward received when transitioning from state s to s'.

In practice, the reward structure and transition probabilities are usually unknown. RL enables agents to learn optimal or near-optimal policies by interacting with the environment and maximizing cumulative rewards. Agents start by taking random actions but gradually approximate the transition and reward functions from experience [15]. This process is illustrated in Figure 1.

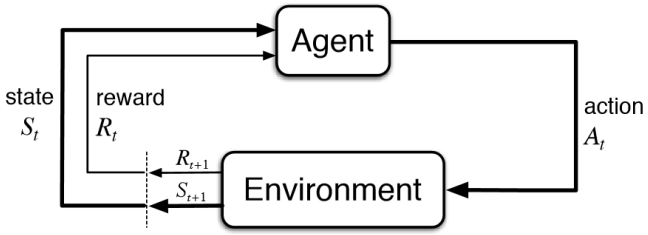

Fig. 1. Reinforcement Learning Diagram

In co-learning scenarios, adaptations need to happen quickly. Therefore, simple tabular methods like Q-learning are commonly used due to their efficiency and explainability. Van Zoelen [6] employed an adapted version of RL with options, where the concept of actions is extended to include temporally extended actions called options. This allows the agent to execute multi-step strategies that terminate upon reaching a subgoal, accelerating learning by identifying useful subgoals and creating policies to reach them [16].

Similarly, Veldman-Loopik [13] used an adapted MAXQ value decomposition algorithm, a hierarchical form of Q-learning. MAXQ decomposes an MDP into smaller sub-MDPs and breaks down the value function into an additive combination of subtask value functions. This decomposition enhances learning efficiency by allowing the algorithm to solve smaller sub-problems first [17].

These works highlight the benefits of using simple, decomposed Q-learning algorithms in co-learning scenarios. The speed of these methods allows for quick adaptations, which is crucial when time and data are limited. Moreover, the explainability of tabular methods enables a straightforward interpretation of the agent's policy by examining state-action values.

Complex methods like Deep Q Networks (DQNs) are impractical for co-learning due to their computational complexity and lack of explainability. Qureshi [18] demonstrated this limitation by requiring 14 days to train a DQN-based robot for a simple handshake, which is too slow for real-time human collaboration. This computational inefficiency motivates our approach of using lightweight, interpretable Q-learning, which provides precise control over the robot's decision-making process.

C. Embedding Robot Personality

By strategically modulating Q-learning parameters such as exploration rates and reward structures, we can encode distinct robot personalities that shape adaptation strategies during co-learning. To ground personality within our Q-learning framework, we draw on the well-established Big Five model [19]. Here, personality represents consistent behavioural tendencies—like patience or assertiveness—that remain stable across interactions. The Big Five traits provide our foundational model for understanding these behavioural patterns.

1) Openness to Experience: Traits like imagination, aesthetic sensitivity, attentiveness to inner feelings, preference for variety, intellectual curiosity, and challenging authority.
2) Conscientiousness: Includes competence, orderliness, dutifulness, striving for achievement, self-discipline, and deliberation.
3) Extraversion: Characterized by an interest in the external world, enjoyment of social interactions, enthusiasm, talkativeness, assertiveness, and sociability [20].
4) Agreeableness: Encompasses trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness.
5) Neuroticism: Associated with negative emotions such as anxiety, worry, fear, anger, frustration, envy, jealousy, pessimism, guilt, and depression [21].

The FFM is frequently used to describe robot personalities, sometimes extended in novel ways [22]. In robotics, personality embedding often focuses on one or a subset of these traits, with extraversion being the most commonly used [23].

For example, Mileounis et al. [24] embedded extraversion in an NAO robot using voice modulation to differentiate between extroversion and introversion, dominance, and submission. An extroverted, dominant robot spoke with a low pitch and assertiveness, performed many gestures, and spoke quickly with emotion. Conversely, an introverted, submissive robot had a higher pitch, spoke insecurely, used fewer gestures, and spoke slowly.
Since co-adaptation occurs implicitly in our research, we focus on expressing personality through the robot's actions rather than voice. Luo et al. [10] achieved implicit expression for extraversion, agreeableness, and conscientiousness by embedding gestures in a mechatronic face, such as nodding, head shaking, gaze aversion, and eye-rolling. For example, an agreeable robot smiled regardless of agreement or disagreement, while an extroverted robot shook its head when disagreeing.

III. Methods

With the background established, this section outlines the methodology used to investigate the influence of robot personality on co-learning. Specifically, we detail the co-learning task design, structured around Veldman-Loopik's requirements (Section II-A), and describe the experimental setup used to implement this task. Furthermore, we introduce a novel Q-learning implementation that decomposes the Q-table into distinct phases, facilitating structured decision-making. Finally, we present our approach to integrating the robot personality types within the reinforcement learning framework.

A. Task design

Since we are trying to evaluate personality types for co-learning, the task must allow co-learning to emerge. To ensure this is the case, we will follow the requirements outlined in section II-A by Veldman-Loopik.

Creating a coherent story for the experiment that makes sense to the human participant is important, as this helps to engage the human more naturally with the task. The story chosen for the design of this experiment is that the human participant needs to perform a teleoperated lumbar puncture.
A lumbar puncture, also known as a spinal tap, is a medical procedure where a needle is inserted into the lower spine to collect cerebrospinal fluid from the spinal canal [25].
A lumbar puncture is chosen because it is a sensitive operation that requires a great deal of attention. This is necessary to limit the observability of the human participant and reduce its learning speed to match that of the robot, contributing to requirement **R5** and **R2**.
The participant needs to perform this puncture and drain the epidural space. After the procedure, the human must stop the bleeding, stitch up the patient, cut the suture thread and bandage the wound. All of these handlings require different objects for the robot to deliver. This scenario naturally incorporates a handover task, making it a fitting context for evaluating co-learning in hybrid teams. To structure the task properly, we divided the different stages of the handover into distinct phases, which can be seen in table I. These phases provide clear decision points which allow soft dependencies to emerge, thereby contributing to design requirement **R1**.

Phase	Description
Phase 0	The robot is in its 'home' position. It moves to a neutral, upright position opposite the human participant and picks up the scissors.
Phase 1	The robot decides when to initiate the handover immediately as the human starts the draining process or after detecting a state change (e.g., completion of the draining process or a human request by holding up their hand).
Phase 2	The robot selects a handover orientation: 'serve' (palm facing up) or 'drop' (palm facing down).
Phase 3	The robot moves towards the human's hand and decides when to open its hand. It can close, partially open, or fully open its hand.
Phase 4	The handover is evaluated as either failed or successful. The robot updates its decision-making based on this experience, and the handover is reset, returning the robot to Phase 0.

TABLE I
Phases of the robot handover process

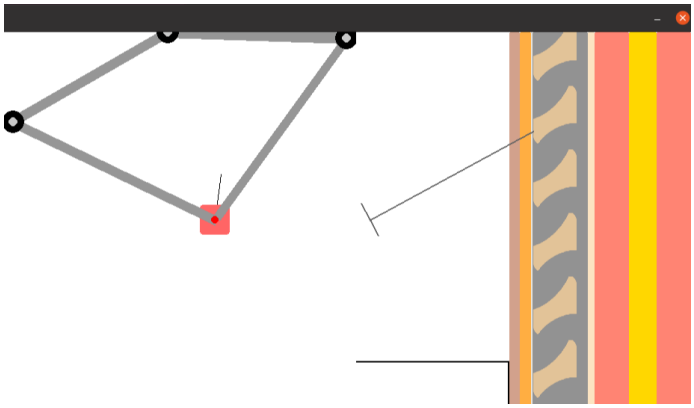

Fig. 2. the secondary task interface screenshot showing the teleoperated lumbar puncture simulation.

The lumbar puncture is implemented as a teleoperated simulation via pygame, shown in figure 2. The human controls the teleoperated arm on the left side of the screen, while the right side displays the teleoperation output. The right screen shows a needle and the multiple layers of tissue found in the human body. The task requires the human to push the needle through these tissue layers, providing visual feedback to the teleoperated arm on the left screen. Once the needle punctures the epidural space, the human must press the spacebar to initiate the draining process.

The damping and stiffness coefficients and the tissue layer sizes are modelled based on average values found

in the literature [26]. These parameters vary along a normal distribution, with new stiffness, damping, and size values sampled after each simulation reset. The implementation code is available on GitHub [27].

To reward the human complying with requirement **R3**, we introduced a negative chime and a red border when the handover failed. When the handover was successful, we introduced a positive chime and a green border. We also provided an end-screen message, either positive when the task was successful or negative when the task failed.

The task was adapted to better accommodate requirement **R2**, ensuring that the robot and human team members learn at a comparable pace. To achieve this, the feedback gain was decreased, and the tissue sizes were increased to make the puncture more manageable. Additionally, the code was modified to require the human to hold the spacebar during the draining process, with task failure resulting from premature release. This adjustment ensured a balanced interaction between the human and the robot. Finally, the code was integrated into the robot operating system (ROS), enabling communication between the simulation and the robot.

B. Software architecture

This communication is achieved by controlling the robot via ROS through the iiwa_ROS software package [28]. It uses an impedance controller to interact with the human participants safely.

On top of this software stack, we developed a python package [29] that implemented vision-related tasks, a finite state machine (FSM) to allow the robot to take the relevant actions, and a novel Q learning algorithm that enables the robot to perform complex tasks through the introduction of specific decision points that we call phases, which are directly coupled to a state in the FSM. The FSM handles the initialisation of different personality types and manages the main loop for testing multiple runs. It also interacts with the Q-learning agent to retrieve the current phase's appropriate action and connects it to the robot and hand controllers. A visual representation can be found in appendix D. The FSM also allows for secondary functions to be implemented for each phase, e.g., in phase 2, the robot decides on a handover orientation, 'serve' or 'drop', which is directly coupled to the state space. However, it also moves towards the participant's hand in this phase. By decoupling the hand-moving part from the RL mechanism, we significantly speed up the learning process. Encoding hand-location in the state space is costly regarding state-action values as it introduces a tuple (x_i, y_i, z_i) for some discrete workspace with resolution i. We decided to disconnect this from the RL mechanism since learning to move to the hand location is irrelevant to the co-learning process. Because the Q-learning algorithm is restricted to co-learing actions, we have little control over the expression of personality types. By introducing an FSM, we also increase the range of control over the robot's actions, extending beyond those necessary for task completion. For example, it allows us to change the robot and hand speeds, increasing our control over the expression of personality.

The vision-related tasks were achieved using a depth camera. This camera provides the location of the human hand so the robotic arm can move there accordingly. To obtain the location of the hand, we used MediaPipe [30], which offers ready-to-use Python implementations of a hand land-marker model. A handover is successful when the human has held the item for a second. The camera provides an RGB and a disparity image to obtain the location of the human hand. The RGB image is aligned with the disparity image, which means that each point in the RGB image corresponds to a depth point in the disparity image. The hand land-marker model is applied to the RGB image. Then, the point is deprojected from a pixel in the RGB image to a 3D point in the camera frame using the disparity image. This 3D point location is then published to a ROS topic.

C. Q-Learning

With the software architecture supporting the robot's interaction with its environment, we now detail the reinforcement learning algorithm that drives the robot's decision-making during the handover task. As discussed in Section II-B, we have chosen Q-learning as the most suitable algorithm for implementing the decision-making framework in our co-learning setup.

1) Algorithm Overview: Q-learning is a model-free reinforcement learning algorithm that updates its action-value function based on immediate rewards. However, the agent may struggle to associate actions with outcomes in scenarios with sparse or delayed rewards where feedback is infrequent, such as in our handover task, where success is determined only at the end of an episode. This can lead to inefficient exploration and slower learning, as the algorithm relies on consistent reward signals to adjust its policy effectively [31]. To address this challenge, we introduce an experience replay mechanism that modifies and distributes the final reward across the trajectory, facilitating better credit assignment over time.

2) State, Action, and Phase Spaces: We decompose the overall state-action space into a sequential structure consisting of distinct phases, each representing a specific stage of the handover process. The phase p is a function of the current state s. The phase transition function is then expressed as $p' = \Phi(s')$. This relationship allows the agent to update the phase based on the next state, ensuring coherent transitions.

Algorithm 1: Generalized Phase-Based State Transition Algorithm

Input: State $s \in \mathcal{S}$, Phase $p \in \mathcal{P}$, Action $a \in \mathcal{A}$
Output: Next state s', Next phase p'

Initialization: Define $\mathcal{A}_p \subset \mathcal{A}$ for each $p \in \mathcal{P}$

if $\mathit{Valid}(a, p) = 1$ **then**
  $s' \leftarrow \mathcal{T}(s, a)$
  $p' \leftarrow \Phi(s')$
else
  $s' \leftarrow s$
  $p' \leftarrow p$
return $s',\ p'$

Definition

$$\text{Valid}(a, p) = \begin{cases} 1, & \text{if } a \in \mathcal{A}_p \\ 0, & \text{if } a \notin \mathcal{A}_p \end{cases}$$

Each phase p defines a specific subset of possible actions $\mathcal{A}_p$ and valid states $\mathcal{S}_p$ that the robot can be in during that phase. This is captured concisely in algorithm 1.
To enforce phase-specific action constraints, we introduce a function $a_{\text{valid}} = A(p, a)$ that determines the validity of an action a in the current phase p. This function ensures that the agent only considers actions appropriate for the specific stage of the handover process, effectively filtering out invalid actions and reducing the state-action space.
For instance, in the initial phase, the robot's actions are limited to moving to its home position, while in the next phase, the robot is allowed to choose a hand orientation. By defining $\Phi(s)$ and $A(p, a)$, we create a structured environment where the agent's decisions are both state- and phase-dependent.

		Phase 1	Phase 2		Phase 3		
	Q-Table	**Action 1**	**Action 2**	**Action 3**	**Action 4**	**Action 5**	**Action 6**
Phase 1	**State 1**	**15.53**	0	0	0	0	0
	State 2	**6.24**	0	0	0	0	0
Phase 2	**State 3**	0	**-12.3**	**-1.21**	0	0	0
	State 4	0	**-6.53**	**1.1**	0	0	0
	State 5	0	**1.12**	**16.7**	0	0	0
	State 6	0	**18.13**	**12.19**	0	0	0
Phase 3	**State 7**	0	0	0	**20.23**	**13.13**	**-12.1**
	State 8	0	0	0	**12.19**	**13.9**	**-13.8**
	State 9	0	0	0	**16.23**	**19.7**	**-20.1**

Fig. 3. Visualisation of the phase decomposition of a Q-table

This decomposition is visualised in Figure 3. The Q-table $Q(s, a)$ is structured such that $Q(s, a) = 0$ for all invalid state-action pairs (s, a), as determined by the phase-specific constraints $A(p, a)$. For Phase 1 ($p = 1$), valid states are $\mathcal{S}_1 = \{1, 2\}$ and valid actions are $\mathcal{A}_1 = \{1\}$, resulting in $Q(s, a) = 0$ if $s \notin \mathcal{S}_1$ or $a \notin \mathcal{A}_1$. Similarly, for Phase 2 ($p = 2$), valid states are $\mathcal{S}_2 = \{3, 4, 5, 6\}$ and valid actions are $\mathcal{A}_2 = \{2, 3\}$. For Phase 3 ($p = 3$), valid states are $\mathcal{S}_3 = \{7, 8, 9\}$ and valid actions are $\mathcal{A}_3 = \{4, 5, 6\}$. Consequently, the algorithm only updates valid state and phase pairs, enforcing the constraint defined by $A(p, a)$.
By structuring the action and state spaces in this way, we effectively reduce the number of Q-values that need to be learned from $6 \times 9 = 54$ (states $\times$ actions) to 19. This is calculated as:

$$\sum_{p=1}^{|\mathcal{P}|} |\mathcal{A}_p| \cdot |\mathcal{S}_p| = 1 \times 2 + 2 \times 4 + 3 \times 3 = 2 + 8 + 9 = 19.$$

Detailed descriptions of the action and state spaces in our co-learning setup are provided in Appendices A and B, respectively.

3) Update Rule: We update the Q-values as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \delta_t,$$

Where α is the learning rate and δ_t is the temporal-difference error representing the difference between predicted and actual rewards.
The temporal-difference error is computed as:

$$\delta_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t),$$

Where r_t represents the reward received at time step t, and γ is the discount factor weighing the importance of future rewards.

4) Experience Replay Mechanism: One of the primary challenges in our setup is the delayed reward structure, as rewards are only given at the end of an episode when the success of the handover is evaluated. To address this, we implemented an experience replay mechanism inspired by Mnih et al. [32], enabling the agent to learn more efficiently from these sparse rewards. Each episode's experiences are stored as tuples $(s, a, s', r, \phi, \text{Valid})$, where ϕ denotes the phase and Valid indicates the validity of the action in the given phase, determined by the function $A(p, a)$.
Once the episode concludes, the experience replay mechanism reprocesses the collected experiences, allowing the algorithm to distribute the final reward across all valid state-action pairs in the trajectory. This approach ensures that the agent considers the entire sequence of decisions contributing to the handover's outcome rather than reinforcing only the final action.

5) Reward Function: In our setup, the agent receives rewards at the end of each episode based on the handover outcome. A successful handover earns a base reward of $+10$ per phase completed; a failed handover incurs a penalty of -10 per phase.
A time limit is imposed to encourage efficient completion. If the handover is completed within this limit, the agent receives an additional reward equal to twice the unused time. Failure to complete within the time results in an episode failure and the associated penalties.
A penalty of -2 is assigned if the robot attempts to close its hand when it is closed or partially open. This

discourages repetitive actions and prevents the robot from getting stuck in a loop.
Additionally, we augment the reward function to model the 'leader' personality type, characterised by a preference for the 'serve' orientation during handovers. If the agent selects the 'serve' orientation when appropriate, we grant an extra reward of $+10$.

6) Implementation Details: The primary implementation is the real-time training loop. This function trains the robot by selecting an appropriate action, outputting this to the FSM, and updating the Q-table when required. This function is explained in algorithm 2.

Algorithm 2: Real-Time Training Loop for Q-Learning Agent

Init: $Q(s,a) \sim 0.01 \cdot \mathit{Uniform}(0,1)$ `// Q-table`
Input: α, λ, *Env*`// Learning rate, Discount factor, RL environment`

Function `TrainStep(`ϵ, *Env*`)`**:**
  if *NewEpisode* **then**
    $s, \phi \leftarrow$ Env.Reset()
    $\mathcal{E} \leftarrow \emptyset$ `// Experience buffer`
    Terminated $\leftarrow$ **False**
    NewEpisode $\leftarrow$ **False**

  $\phi_{\text{current}} \leftarrow \phi$
  Valid $\leftarrow$ **False**
  while $\phi = \phi_{current}$ ***and not*** *Valid* **do**
    $a \leftarrow \text{EpsilonGreedy}(\epsilon, Q(s,a))$
    $s', r, \phi, \mathit{Terminated}, \text{Valid} \leftarrow \text{Env.Step}(a)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(s, a, s', r, \text{Valid})\}$
    if $r \neq 0$ ***and*** *Valid* **then**
      $\text{UpdateQTable}(s, a, r, s', \alpha, \gamma)$
    $s \leftarrow s'$

  if *Terminated* **then**
    NewEpisode $\leftarrow$ **True**
  return $a, \phi, \mathit{Terminated}$

The FSM calls the `TrainStep` function during each phase. This function initialises the state in the first phase by resetting the environment. `TrainStep` then repeatedly calls the environment's `step` function, updating the state and phase until a valid action is taken while recording the agent's experiences. If an action is valid (`Valid` parameter is **True**), the loop breaks, and the function returns the action, current phase, and termination flag, allowing the FSM to execute the action and proceed to the next phase. If a valid action yields a reward, the Q-table is immediately updated to prevent the agent from getting stuck in a loop.
At the end of each episode, we invoke the experience replay mechanism, which iterates over the stored experiences $\mathcal{E}$. For each valid action, it updates the Q-values by redistributing the final reward over the entire trajectory. Intermediate rewards are directly updated in the main training loop, so these need not be accounted for in this mechanism.
The training parameters are learning rate $\alpha = 0.15$, discount factor $\gamma = 0.8$. The exploration factor ϵ starts at 0.9 and decays by 5% per episode to a minimum of $\epsilon = 0.1$, ensuring the agent remains adaptable for co-learning. These values were determined through parameter tuning on a simulated version of the problem, optimising for the highest mean Q-value in the solved Q-table.

D. Robot Personality

Building on the Q-learning framework, we introduce distinct robot personality types embedded into the robot's decision-making processes through tailored adjustments in the learning algorithm.
The robot personalities are implemented through 2 vectors: the leader-follower axis is implemented through the reinforcement learning scheme, and the patient-impatient axis is implemented through the robot and hand controllers.
The leader personality type has a modified exploration scheme. The leader starts with a slightly lower exploration factor $\epsilon = 0.8$, which decays by 10% down to $\epsilon = 0.2$. The argument for this is that a leader should converge to a strategy quicker, hence the quicker decay rate, but should also be more open to change, thus the higher floor. Since a leader has a high extraversion, it will strongly prefer the 'serve' orientation, so it will receive an additional $+10$ each time it chooses this orientation, as explained in section III-C5. This preference for the serve orientation and faster convergence aligns with the FFM trait of extraversion, where proactive engagement and quick adaptability reflect high extraversion and openness to experience.
The follower personality type has a modified exploration scheme that starts with a lower exploration factor of $\epsilon = 0.6$, which decays by 20% per episode; it also has a much higher learning rate of $\alpha = 0.5$. These modifications make the follower personality type much quicker to converge to whatever strategy the human initially decides to follow. And since it has low extraversion, it will not try to change the strategy too much and will likely keep the current strategy. This design reflects low extraversion, aligning with agreeableness in the FFM, as the follower personality accommodates the human's strategy and shows a stable, cooperative approach.
The baseline time for the robot to complete a movement is 5 seconds. The impatient personality type modifies this to 2 seconds, significantly increasing the robot's speed. It also closes the hand in 1 second, where the baseline is 2 seconds. Furthermore, the joint stiffness slightly increases, making the movement more direct. These faster, more forceful movements reflect high neuroticism, in line with the FFM trait of impulsiveness, a lower tolerance for delay, and extraversion through its direct, fast-paced approach.

The patient personality type has a robot movement time of 7 seconds per movement and a hand closing speed of 3 seconds. It also has decreased joint stiffness, making the robot slightly sluggish at the start and allowing it to overshoot its target position slightly. It also has slightly reduced damping values, introducing a small damped oscillation around its setpoint value. This slower, more deliberate approach aligns with high agreeableness and low neuroticism in the FFM, showing a considerate, calm style that accommodates the human partner's pace.

IV. Experimental setup

Having outlined the task design and the Q-learning-based personality integration, the next step is to implement these elements in a physical experimental setup to test the robot's interaction with a human participant. The following section details the physical setup used to implement our methods, followed by an explanation of the experimental procedure and participant guidance. Finally, we outline the metrics used for result analysis.

A. Setup

The chosen hardware setup is visualised in Figure 4 and consists of the KUKA LBR iiwa7 R800 robotic arm with the qb-SoftHand as a gripper. The iiwa7 is a cobot arm, which means it is designed to be safe to work in environments alongside humans. It has torque sensors at each joint, which allow the robot to measure the physical interactions between the human and the robot, which improves the safety of the arm and allows for the interaction to be part of the co-learning process.

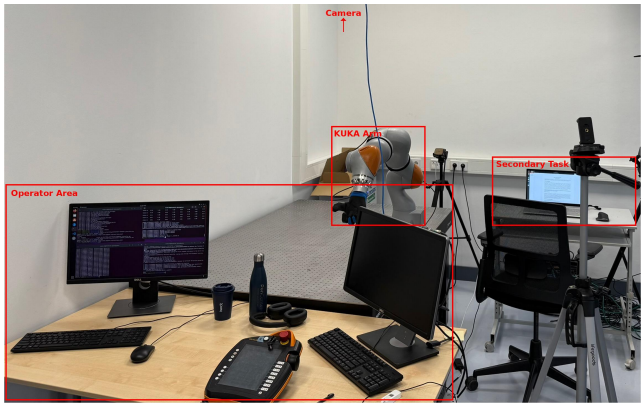

Fig. 4. Experimental setup for co-learning study. The operator area, KUKA arm, secondary task, and overhead camera are indicated

The SoftHand is a hand-like gripper that is tactile enough to pick up many objects. It is soft, so it can be safely used in rigid environments and alongside humans, as the hand will deform under load.
The chosen camera is the RealSense D455. This camera can be operated via the RealSense API, which provides standard functions for aligning the depth image to the RGB image and a deproject function. The distance from the ground to the camera is 3 meters. This camera is highly accurate for the full workspace, with a depth error of less than 2% at 4 meters.
The human participants sit at their desks next to the robot arm. They are given a keyboard, mouse, and screen to perform the secondary task.
The setup is filmed with a digital camera; the videos are analysed to determine the interaction patterns.

B. Procedure

Before the experiment begins, the participants are asked to complete an informed consent form. After the consent has been given, the participants will receive an introduction to the test setup and get time to familiarise themselves with the robot arm, which will be in compliant mode. This allows the participants to build trust in the arm and show that it is safe to work with.
The participants are then provided with a high-level explanation of the experiment. The robot's action space is not discussed, as this needs to be discovered during co-learning. The participants are given time to familiarise themselves with the secondary task and understand its dynamics.
Once the participants are familiarised with the setup, they will start the experiment with the baseline personality. Each run takes 11 episodes. The participants will get a form to fill out each time they complete a run. This form measures team fluency as described in section II-A and the perceived personality type. The personality types are randomised via a Latin square to balance the order effects of robot personality types on the co-learning process.
During the experiment, snapshots of the Q-table are taken to see how the robot strategies evolved over the episodes.

C. Metrics

With the experimental setup in place, we now define the key metrics used to evaluate the success and dynamics of human-robot collaboration during the handover task. We adopt metrics from Veldman-Loopik (introduced in Section II-A) to assess various aspects of co-learning, supplemented by measures tailored to our reinforcement-learning (RL) framework and personality embeddings.

1) Performance Metrics: To gauge high-level task success, we track the following:

- **Performance Rate**: the percentage of successful episodes per run:

$$M_{\text{PR}} = \frac{\text{Successful episodes}}{\text{Total episodes}} \times 100\%.$$

- **Cumulative Reward**: a traditional RL metric summing per-episode rewards over a run.

These metrics capture *task-level* outcomes, providing insight into whether different robot personalities lead to higher success or efficiency.

2) Human Perception of Fluency: The fluency metric describes the human experience of collaboration. Building on the works by Hoffman [14], we consider:

- Human-Robot Fluency, e.g., “The robot contributed to the fluency of the interaction”.
- Robot Relative Contribution, e.g., ”I was the most important team member on the team”.
- Trust in the robot, e.g., ”I trusted the robot to do the right thing at the right time”.
- Positive Teammate Traits, e.g., “The robot was intelligent”.
- Improvement, e.g., “The robot’s performance improved over time”.
- Working Alliance for Hybrid Teams, e.g., ”I am confident in the robot’s ability to help me”.

We compiled a selection of these questions into a questionnaire (Appendix C). Participants answered them after interacting with each personality type. This subjective measure of how different robot behaviours affect collaboration quality was aggregated into a composite **Fluency_Score** (1-7 Likert scale average).

3) Strategy Identification: Robots and humans may use various strategies because the task can be accomplished in multiple ways. Table II outlines the human strategies across the three primary handover phases. We combined the robot’s chosen strategy (linked to its action space) with the human’s strategy to form a *joint strategy*.

Phase	Human Strategy
Phase 1	**H1**: Human asks for the item as soon as draining starts. **H2**: The Human asks for the item after the draining process finishes. **H3**: Human does not ask for the item.
Phase 2	**H4**: Human chooses a ’serve’ orientation. **H5**: Human chooses a ’drop’ orientation.
Phase 3	**H6**: Human signals robot by pulling on end-effector. **H7**: Human waits for the robot to release the item.

TABLE II
HUMAN STRATEGIES ACROSS DIFFERENT PHASES.

We quantify strategy evolution using the following:

- **Total_Strategy_Changes**: Count of distinct joint strategies adopted per run
- **stability**: Percentage of consecutive episodes where the joint strategy remained unchanged

Video footage of the experiment further supports our identification of emergent collaborative patterns.

4) RL Metrics: The robot develops action preferences through Q-learning within the reinforcement learning (RL) framework. Analysing snapshots of the Q-table reveals how the robot’s *policy* evolves throughout multiple episodes. We adopt the following RL-specific metrics to evaluate the learning dynamics and policy stability:

a) Entropy.: Measures the randomness or level of exploration in the robot’s action choices [33]. For a given state s, the Q-values are converted into a probability distribution over actions using the softmax function:

$$p(a|s) \;=\; \frac{\exp(Q(s,a))}{\sum_b \exp(Q(s,b))},$$

where $Q(s,a)$ is the Q-value for state s and action a. The entropy for a state is then computed as:

$$H(s) \;=\; -\sum_a p(a|s)\log(p(a|s)),$$

where $p(a|s)$ is the probability of taking action a in state s. Finally, the overall **entropy** is the average entropy across all states:

$$H \;=\; \frac{1}{|S|}\sum_{s\in S} H(s),$$

Where S is the set of all states. Higher entropy values indicate greater randomness or exploration in the robot’s actions, while lower entropy suggests that the robot has learned a more deterministic policy.

b) ActionConsistency.: Quantifies how consistently the robot repeats specific actions once they are learned. Formally:

$$\text{ActionConsistency} \;=\; \frac{1}{N}\sum_{s=1}^{N}\mathbb{I}\big(L_i \;=\; L_{i+1}\big),$$

Where:

- N is the total number of states.
- $L_i = \arg\max_a Q_i(s,a)$
- $\mathbb{I}(\cdot)$ is the indicator function, returning one if the best actions match and zero otherwise.

c) QGap.: Measures how confidently the robot identifies its best action in each state. For a given state s, we define the gap as the difference between the highest and second-highest Q-values:

$$\text{gap}(s) \;=\; \max_a Q(s,a) \;-\; \max_{a\neq \arg\max_b Q(s,b)} Q(s,a),$$

Where:

- $Q(s,a)$ is the Q-value for state s and action a,
- $\arg\max_a Q(s,a)$ gives the action with the highest Q-value.

The overall **QGap** is then the average gap across all states:

$$QGap \;=\; \frac{1}{|S|}\sum_{s\in S}\text{gap}(s),$$

Where S is the set of all states. A larger $QGap$ indicates greater confidence in the policy, as the best action is much better than the alternatives. Conversely, a smaller $QGap$ reflects uncertainty or ties between actions.

d) Convergence.: Assesses the overall stability of the Q-table across episodes via the L1 distance between consecutive Q-tables. For two Q-tables $Q^{(i)}$ and $Q^{(i+1)}$ from consecutive episodes, we define:

$$d_{\mathsf{L1}} \;=\; \sum_{s}\sum_{a}\Big|\,Q^{(i)}(s,a)\;-\;Q^{(i+1)}(s,a)\Big|\,.$$

A smaller d_{L1} indicates fewer changes in Q-values, implying a more stable (converged) policy. If d_{L1} remains very low for multiple consecutive episodes, we infer that learning converges effectively.
These metrics provide insight into the learning process, showing how well the task is performed and how confidently and consistently the robot arrives at its decisions.

5) Perceived Personality Feedback: Building on our personality implementation, we asked participants to rate how *patient* or *impatient* the robot appeared (i.e., Patient_Impatient_Score). Similarly, they assessed whether the robot's behaviour seemed more *leader*-like or *follower*-like (Leader_Follower_Score). In Section V, we report how these subjective ratings correlate with each robot personality type.

- **Patient_Impatient_Score**: 1-7 rating (1=very patient, 7=very impatient)
- **Leader_Follower_Score**: 1-7 rating (1=strong follower, 7=strong leader)

V. Results

To investigate how different robot personalities (Follower, Impatient, Leader, Patient) influenced our core metrics, we conducted a series of repeated-measures ANOVAs, treating *personality* as a within-subject factor [34]. Each participant experienced all four conditions, enabling us to assess how each outcome measure changed based on the robot's assigned personality type.

TABLE III
Repeated-measures ANOVA results. Degrees of freedom (df) reflect Greenhouse-Geisser corrections (GG) where sphericity was unmet. Sig. indicates significance at $\alpha = 0.05$ (* for $p < 0.05$, ** for $p < 0.01$, ns for non-significant).

Metric	df (GG)	MSE	F	p	Sig.
Mean Performance Rate	(2.28, 31.98)	560.57	1.64	.208	ns
Cumulative Reward	(2.20, 30.77)	649280.32	1.54	.230	ns
Total Strategy Changes	(2.40, 33.58)	5.40	4.97	**.009**	**
Stability	(2.44, 34.13)	0.04	5.76	**.005**	**
Fluency Score	(2.31, 32.34)	1.21	0.46	.664	ns
Patient-Impatient Score	(2.16, 30.22)	1.84	4.94	**.012**	*
Leader-Follower Score	(2.04, 28.51)	1.11	2.06	.145	ns
Avg Entropy	(2.59, 36.22)	0.02	2.59	.075	ns
Avg QGap	(2.41, 33.74)	2.10	2.87	.061	ns
Avg Convergence	(1.61, 22.56)	93.99	0.30	.697	ns
Avg ActionConsistency	(2.41, 33.75)	0.00	3.35	**.039**	*

Table III highlights significant main effects of *Personality* ($p < 0.05$) on four metrics: *Total Strategy Changes*, *stablility*, *Patient-Impatient Score*, and *Avg ActionConsistency*. Greenhouse–Geisser corrections (GG) were applied to the degrees of freedom where sphericity assumptions were violated [35]. All other metrics were non-significant ($p > 0.05$), indicating that while high-level performance outcomes (e.g., success rate, total reward) showed minimal variation across personalities, the strategies used to interact with the personality types differed significantly.
This subsection focuses on the metrics that showed significant effects of robot personality: *Total Strategy Changes*, *stablility*, *Patient-Impatient Score*, and *Avg ActionConsistency*. These metrics highlight how variations in personality influenced the robot's internal policy characteristics and the chosen strategy, as well as the participants' perceptions, even as other performance-related outcomes remained unaffected. We treat total strategy changes and Stability as a single unified metric since Stability and total strategy changes follow a perfect negative correlation, as shown in appendix F, Figure 19.

A. Patient-Impatient Score

The repeated-measures ANOVA for the Patient-Impatient Score was significant ($p = 0.012$), indicating a difference in perceived personality across conditions. Post-hoc Tukey-corrected pairwise comparisons [36] revealed that only the difference between the impatient and patient personality types was significant ($p = 0.0174$). Figure 5 visually supports this result, showing a clear separation between patient and impatient personalities. In contrast, the leader and follower personalities remained closely clustered, with both having a leader-follow score of around 0. The Patient personality type was perceived as patient and slightly a follower.

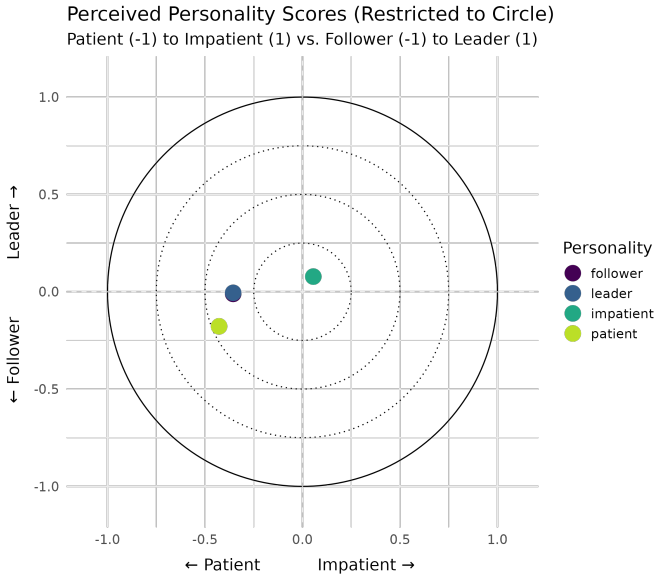

Fig. 5. Perceived personality scores. These scores indicate how distinct the personality types are from one another.

These results confirm that participants largely perceived the patient and impatient personalities as intended, while the leader and follower traits were less distinguishable.

On average, the robots were rated somewhat more patient, and even the impatient robot was not perceived as impatient overall. Video analysis shows that the impatient robot dropped the object most frequently (Appendix G), which did not result in higher impatient scores. This is likely because some participants found this personality type to be the most precisely timed and the quickest for their strategy. Conversations with participants after the experiment revealed that many preferred the impatient robot, as it allowed them to obtain the item quickly and shift focus away from the secondary task, reducing mental strain.

B. Strategy

(*Total Strategy Changes*/*Stability*) showed significance with $p = 0.009$ and $p = 0.005$, respectively.

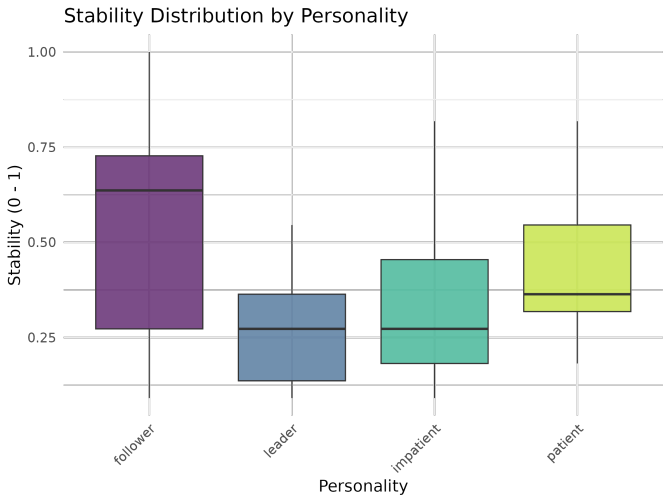

Fig. 6. Box plot showing stability distributions by personality type.

The Follower personality exhibits the highest median stability score, approximately 0.65, though it also has the largest variability among all personality types. Notably, it is the only personality type with a stability score of 1.00, indicating zero strategy changes. The Impatient and Leader personalities show similar median stability scores, with the Impatient personality exhibiting slightly greater variability.

Post-hoc Tukey-corrected pairwise comparisons confirmed that the Follower personality had significantly higher Stability than both Impatient ($p = 0.018$) and Leader ($p = 0.023$). However, no significant differences were found between the Patient and any other personality type. This suggests that while the Patient personality demonstrated relatively high median Stability, it did not differ enough from the others to reach statistical significance.

The Patient personality type is also the only one that does not show a stability score of 0. This means that no participants exhibited entirely random strategies with this personality. One possible explanation is that the Patient personality's slower, more predictable handovers provided a consistent interaction pattern, preventing erratic strategy shifts. Participants may have adapted to this personality by adjusting their behaviour in a structured manner, leading to greater overall Stability. In contrast, personality types with more aggressive or unpredictable handover timing may have induced more frequent and erratic adjustments in participant strategies.

While the stability analysis provides insight into overall trends in human-robot collaboration, it does not capture how humans adjust their behavior to different robot personalities. We conducted a video analysis to classify and quantify interaction patterns to better understand these adaptations. Following the framework in Section II-A, interactions were categorized as either stable situations, where behavior remained consistent over time, or sudden adaptations, where abrupt adjustments occurred in response to task demands. Table IV summarizes the observed interaction patterns.

TABLE IV
CATEGORIZED INTERACTION PATTERNS OBSERVED DURING HUMAN-ROBOT COLLABORATION.

Interaction Pattern	Category
Human holds item until the robot releases	Stable Situation
Human waits for robot to open hand	Stable Situation
Misalignment	Stable Situation
Robot and human meet in the middle	Sudden Adaptation
Human moves towards robot	Sudden Adaptation
Human takes object from robot with force	Sudden Adaptation
Human touches hand unnecessarily	Sudden Adaptation
Robot drops object	Sudden Adaptation
Robot moves away	Sudden Adaptation
Robot moves towards human	Sudden Adaptation

To quantify the influence of robot personality on interaction dynamics, we analyzed how frequently each pattern occurred for different robot personalities. Table V highlights the two interaction patterns with the most pronounced variation across personality types: "Human takes object from robot with force" and "Robot drops object."

TABLE V
INTERACTION PATTERNS WITH SIGNIFICANT VARIATION ACROSS PERSONALITY TYPES.

Interaction Pattern	Personality Type	Count
Human takes object from robot with force	Follower	17
	Leader	18
	Impatient	13
	Patient	**47**
Robot drops object	Follower	9
	Leader	14
	Impatient	**36**
	Patient	2

Participants were far more likely to take the object when interacting forcefully with the Patient personality. In these cases, the human had already drained the epidural space and was actively waiting for the object. Given the relatively high cognitive load of monitoring the screen, participants likely sought to reduce it by forcibly taking the object instead of waiting for the slower, patient robot to release it. In contrast, robots with more precisely timed handovers (Leader, Follower) synchronized better with the human's task rhythm, reducing the need for forceful retrieval.

Similarly, the Impatient robot dropped objects significantly more often than the other personality types. This occurred because the impatient robot was more likely to release the object prematurely, before the human had finished draining the epidural space. As a result, participants could not catch the object, leading to a failed handover. This aligns with the impatient robot's design, where faster actions prioritize speed over synchronization with the human partner's availability.

These findings suggest that Stability plays a crucial role in human-robot collaboration, but an important question remains: Does Stability correlate with actual performance? Figure 7 explores this relationship by illustrating the connection between mean performance rate (MPR) and Stability.

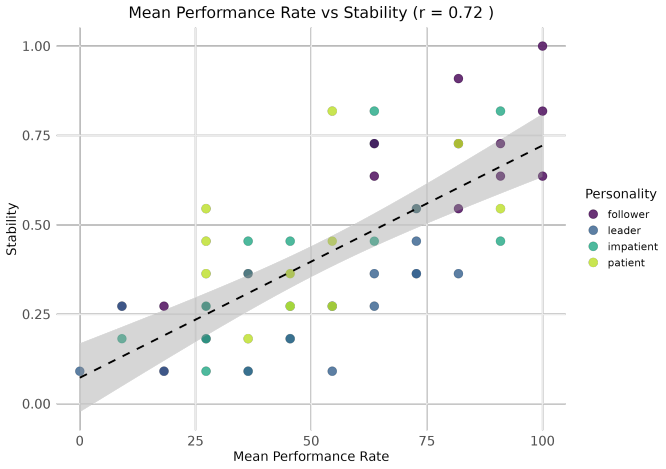

Fig. 7. Scatter plot of Mean Performance Rate (MPR) versus Stability across personality types. Each point represents an individual participant's data, color-coded by personality type. A clear upward trend is evident, with higher Stability generally coinciding with higher performance rates. The dashed line indicates a linear regression fit with a 95% confidence band (shaded region).

Statistical tests confirm this observation, with a Pearson correlation coefficient of $r = 0.724$ ($p < 0.0001$) and a Spearman rank correlation of $\rho = 0.728$ ($p < 0.0001$) [37]. Both coefficients indicate a robust correlation, suggesting that participants who maintained more stable interaction strategies (i.e., changed their approach less frequently) tended to achieve higher performance rates.

These findings align with reinforcement learning principles, wherein a consistently successful strategy is positively reinforced, making the algorithm more likely to converge on a policy.

Notably, although personality type influenced the degree of strategy stability, it did not significantly affect the mean performance rate. Instead, Stability itself emerges as the key factor: Participants with predictable, well-practiced interaction patterns realized more consistent and efficient handovers, thereby improving performance outcomes.

C. Avg ActionConsistency

With $p = 0.039$, robot personality significantly influenced how consistently it repeated selected actions. The *follower* condition had the highest average action consistency, while the *impatient* condition exhibited the lowest. The *Leader* and *patient* personalities fell in between.

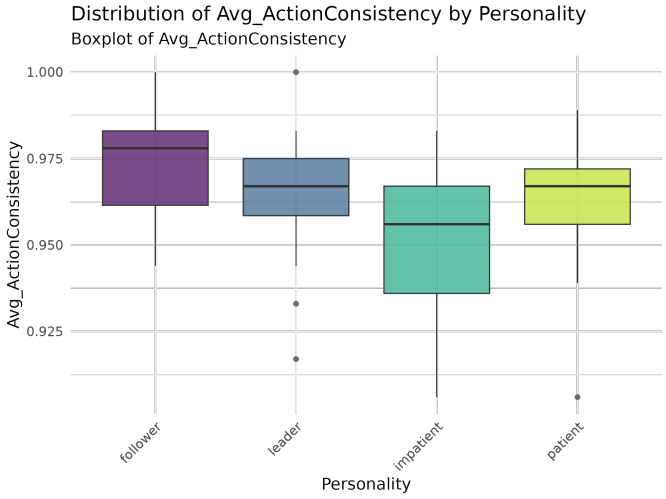

Fig. 8. Box plot showing the average action consistency for each personality type

Figure 8 shows that the follower personality type has the highest median action consistency. The impatient personality type has the lowest action consistency and the highest deviation out of all personality types. The leader and follower personality types show similar trends in personality type.

Post-hoc Tukey-corrected pairwise comparisons confirmed that the Follower personality had significantly higher action consistency than both Impatient ($p = 0.021$) and Leader ($p = 0.038$). However, no significant difference was found between the Patient and other conditions, suggesting that while Patient and Leader personalities demonstrated moderate action consistency, they were not statistically distinct from each other or the Impatient condition. This aligns with expectations, as the Follower personality is designed to be highly consistent, converging to a learned policy quickly with its quicker epsilon (ϵ) decay and higher learning rate (α).

Figure 9 illustrates how action consistency evolved over episodes. The Follower personality consistently demonstrated the highest action consistency, with a smooth increasing trend reaching near 1.00 in later episodes. In contrast, the Impatient personality exhibited the most erratic behaviour, with frequent fluctuations across

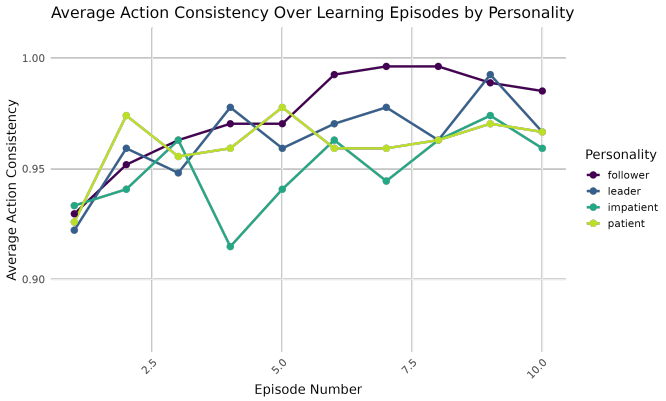

Fig. 9. Average action consistency over episodes, grouped by personality type.

episodes. The Leader and Patient personalities followed a similar trajectory, displaying moderate consistency but not as smooth as the Follower. By the final episodes, the action consistency of the Leader, Patient, and Impatient personalities converged to similar values. This suggests that while the Follower maintained strong repeatability throughout, the other personalities gradually stabilized, albeit at a lower consistency level than the Follower.

VI. Discussion

With the results laid out, we now reflect on their implications. This chapter aims to answer the research question: **"How do different robot personalities influence co-learning?"** Our findings reveal that while overall task performance remains stable across personality types, significant differences emerge in strategy adaptation and subjective perception. This suggests that robot personality does not directly impact performance but plays a key role in shaping how collaborative strategies evolve within human-robot teams.

A. Strategy Adaptation and Performance

Our results indicate that robot personality significantly influences strategy metrics, demonstrating that personality expression affects the stability of the strategies participants adopt.

Qualitative observations further support this finding: participants adjust their behaviour in response to the robot's personality. For example, when interacting with the impatient robot, they take the object more often with force (see Table V). This suggests that humans dynamically adapt to the robot's behaviour to establish more stable interaction patterns.

Moreover, we observe a strong positive correlation between strategy stability and performance ($r = 0.724, p < 0.0001$; see Fig. 7), indicating that as participants settle on a stable strategy, their performance improves.

Despite these adaptations, task performance metrics such as *Mean Performance Rate* and *Cumulative Reward* show no significant differences across personality types. This suggests that humans compensate for variations in robot behaviour by adjusting their strategies, effectively stabilizing performance across different robot personalities rather than improving overall task performance.

Interestingly, the robot's strategy also changed in response to its personality type. For instance, although the impatient robot initially dropped the object frequently, repeated failures led it to adjust by keeping its hand closed longer, ultimately giving participants enough time to grasp the object successfully.

This mutual adaptation demonstrates the flexibility of human-robot teams and reinforces that optimizing robot behaviour alone is insufficient for successful collaboration. Instead, it highlights the importance of designing robots that support strategy discovery and mutual adaptation rather than enforcing rigid interaction patterns [6].

B. Internal Policy Dynamics

Analyzing internal policy metrics reveals that robot personality significantly affects behavioural consistency. Specifically, the *Avg ActionConsistency* metric varied with personality type ($p = 0.039$), with the Follower condition exhibiting the highest consistency and the Impatient condition the lowest. This finding aligns with the strategy stability results reported in Figure 6, where the Follower personality type exhibited the highest median stability.

While *Avg ActionConsistency* reached significance, other reinforcement learning metrics did not. However, *Avg Entropy* ($p = 0.075$) and *Avg QGap* ($p = 0.061$) displayed notable trends. These trends align with theoretical expectations regarding reinforcement learning dynamics.

The *Avg QGap* metric reflects the difference in Q-values between the best and second-best action. A larger Q-gap indicates that a specific action is consistently reinforced with a positive reward, leading to more repeatable behaviour [38]. Hence, a higher Q-gap suggests a more stable strategy. This aligns with the earlier finding that stable strategies differ between personality types.

Similarly, *Avg Entropy* provides insight into action selection variability. In stable strategies, positive reinforcement strengthens preferred actions, reducing entropy as actions become more predictable [39]. In contrast, unstable strategies, where failed handovers frequently receive negative rewards, result in fluctuating action preferences. The policy fails to converge since, in our implementation, the agent repeatedly selects the action with the highest value, receives a negative reward, and then shifts its preference to another action.

Importantly, this instability persists regardless of ϵ. Actions are naturally more random when ϵ is high. When ϵ is low, the selection process becomes more deterministic, but instability remains because negative reinforcement pushes different actions to take precedence at different moments. This prevents a single action from being reinforced consistently, leading to a broader action distribution and higher entropy.

In contrast to these trends, *Avg Convergence* shows no significance for personality ($p = 0.697$). This suggests that, on average, participants settled on stable policies regardless of the robot's personality. The overall convergence is likely driven more by human adaptation than personality, leading to similar final policies across conditions. This further supports our finding that humans compensate for variations in robot behaviour to maintain task performance.

These findings contribute to understanding how human adaptation shapes strategy formation in human-robot collaboration. Our results suggest that even when initial conditions differ due to personality implementation, human adaptability helps establish stable interaction patterns over time, which may be leveraged in reinforcement learning-based approaches. [40].

C. Perception Metrics

The results show a significant difference in the *Patient-Impatient Score* ($p = 0.012$), indicating that designing robot personality along physical vectors is salient. In contrast, the *Leader-Follower Score* does not show any significance for personality type ($p = 0.145$), which shows that humans are less likely to perceive changes in the decision-making patterns of the robot. This could be because the RL parameters require time to decay and update the q-values meaningfully. Given the short duration of the runs, participants may not have experienced a noticeable difference in the robot's decision-making behaviour.

This finding has important implications for robot design. If the goal is to create robots with recognizable personalities, emphasis should be placed on physical manifestations of personality traits rather than internal decision-making parameters. Physical traits create immediate perceptual cues humans can recognize and adapt to, while decision-making patterns may only become apparent over extended interactions.

Similarly, perceived fluency of interaction did not vary significantly with personality type ($p = 0.664$), suggesting that, on average, the specific robot personality did not influence how smoothly participants experienced the collaboration. A likely explanation is that individual preferences differed enough to cancel out the effects of fluency on average. For example, some participants preferred the Impatient personality because it was faster, while others favoured the Patient personality because it allowed them to focus on the screen without feeling rushed.

Another possible explanation is that participants naturally adapted their expectations and behaviour to the robot's personality over time. For instance, those interacting with an Impatient robot may have unconsciously adjusted their timing to match its faster pace, while those with a Patient robot may have felt no urgency to change their approach. This adaptive behaviour could have minimized perceived differences in fluency across personality types.

These findings suggest that if the goal is to design robot personalities that are immediately perceptible, emphasis should be placed on physical vectors such as movement speed or timing, as these provide clear and recognizable cues for users. However, the aim is to shape team dynamics and influence collaboration patterns over time. Decision-making traits can still be integrated into personality expression as they do not significantly impact initial perception. This enables greater control over team strategy without distorting the robot's perceived personality.

D. Implications for Co-Learning in Hybrid Teams

Our results contribute to the ongoing debate in the HRI community on whether human-robot personality matching enhances interaction quality or whether mismatches might be more effective [9]. Our findings indicate that robot personality does not significantly impact perceived team fluency or task performance, suggesting strict personality matching is not required for effective collaboration. Instead, the ability to adapt interaction strategies appears to be the key factor in maintaining successful human-robot teamwork.

These findings have practical implications for designing collaborative robots across various domains, including manufacturing, healthcare, and service sectors. To enhance hybrid team performance, designers should consider the following:

1) **Adopting follower-type designs for stable policies**: Robots converging on strategies quickly lead to higher action consistency and more stable interactions, improving task performance.
2) **Promoting strategic flexibility**: Co-learning systems should support diverse interaction strategies, enabling teams to find suitable approaches that mitigate potential performance issues from personality-specific behaviours.
3) **Implementing personality through mixed vectors**: Prioritize physical vectors (e.g., movement speed, stiffness) for clear and immediate personality expression, especially in short-term interactions, while using decision-making parameters to shape long-term collaboration dynamics.

These design recommendations closely reflect van Zoelen et al.'s emphasis on "co-learning," wherein humans and robots iteratively adapt their behavior to each other rather than relying on rigid, predefined roles (Section 2 in their paper). In particular, van Zoelen et al.'s analysis of co-adaptive interaction patterns illustrates how successful collaboration emerges from recurring co-adaptive behaviors, such as dynamically adjusting who takes the lead or reacting to changes in the partner's actions. This approach supports flexible strategy discovery, ultimately improving collaborative fluency over time. Although van Zoelen et al. did not focus on personality traits, our results (Section VI-C, Table V, Figure 5) show that expressing personality through clear, perceivable behaviors (such as speed and timing) similarly supports mutual adaptation,

reinforcing van Zoelen et al.'s conclusion that transparent, dynamically evolving actions foster stronger co-learning and engagement.

E. Methodological Reflections and Limitations

a) Sample Size and Statistical Power.: A notable constraint of this study is the relatively small participant pool ($n = 15$) in a repeated measures design. While repeated measures can enhance statistical efficiency, such a modest sample size may limit the power to detect subtler effects of robot personality. Therefore, certain performance metrics, such as *Mean Performance Rate* and *Cumulative Reward*, could show undetected differences. This limitation is particularly relevant for metrics that approached but did not reach statistical significance, such as *Avg Entropy* ($p = 0.075$) and *Avg QGap* ($p = 0.061$), which might reveal significant differences with a larger sample. However, as this study is exploratory, the sample size is sufficient to identify key trends and provide insights into the influence of robot personality on co-learning. Future work can build upon these findings with larger participant pools to improve generalizability.

b) Time Constraints on Real-Time Reinforcement Learning.: Each condition involved a limited number of interaction episodes (11 per personality), restricting the time available for the human or robot to converge on optimal strategies. Reinforcement learning typically benefits from extended interaction horizons for policy stabilization. When participants must simultaneously learn a secondary task (the lumbar puncture simulation), this short timeframe could also limit their ability to adjust effectively. This limitation might particularly affect the perception of leader/follower traits, which rely more heavily on reinforcement learning parameters and require more interactions to become salient.

c) Control of Human Factors.: Individual variations in motor skills, adaptability, and cognitive strategies were not explicitly controlled. Some participants may have more difficulty navigating the teleoperation task, while others adapt readily. In addition, no direct measures of mental workload (e.g., NASA-TLX [41]) were collected. Different perceived task difficulty and stress levels could shape how participants respond to or perceive the robot's behaviours. Future work could incorporate these measures to understand better how cognitive load influences adaptation to different robot personalities.

d) System Implementation Constraints.: The phase-based Q-learning architecture and finite-state machine (FSM) streamline the action space, allowing a greater expression of personality at the cost of limiting emergent behaviours that might otherwise develop in a more flexible or continuous learning environment. While this approach enabled precise experimental control over the personality types, it may not fully capture the richness of human-robot co-adaptation that could emerge in less constrained interaction paradigms.

F. Future Directions

a) Long-Term Interactions and Statistical Power.: As noted in our methodological reflections, a small participant pool and limited interaction episodes may obscure the subtle effects of robot personality. Future studies should include more participants and extend the number of episodes per condition to allow humans and robots to reach more stable strategies. Larger-scale, longer-term experiments would boost statistical power and better capture the evolving nature of co-learning, especially where humans require time to fully develop strategies and the robot benefits from extended learning horizons. Longitudinal studies could reveal whether the leader/follower distinction becomes more salient over time as reinforcement learning parameters have more opportunity to influence observable behaviour.

b) Personality Design.: While this work focused on two primary axes of personality, incorporating additional Big Five traits or multi-dimensional models could reveal a wider spectrum of behaviours and further differentiate robot interactions. Designing personalities suited for low-modality robots where expressive capabilities are limited to movement speed, orientation, or force remains challenging. However, innovations in reward shaping and motion-parameter tuning could help produce more distinct behaviours and clearer human perceptions. Future work might explore how combinations of physical characteristics and decision-making parameters could create more recognizable and consistent personality expressions.

c) Adaptive Personalities.: While this study focused on static robot personalities, future work could explore adaptive personality models that adjust in real time based on human preferences, task performance, or contextual cues. Since individuals have different collaboration preferences, a robot capable of dynamically adjusting personality-related parameters could enhance engagement and improve interaction efficiency over time.
For example, a collaborative robot could initially adopt a consistent behavioural pattern, such as a follower role, to help the human form a mental model. Then, it could gradually adjust its behaviour based on observed user tendencies, potentially shifting toward a leader-patient role. Investigating the impact of such adaptations on team performance and engagement could provide valuable insights into optimizing long-term human-robot collaboration.

VII. Conclusion

This paper investigated how different robot personalities influence co-learning in human-robot collaboration.

The results demonstrate that while personality does not directly impact overall task performance, it significantly affects strategy adaptation and perceived interaction dynamics. Specifically, we found that stability in human-robot interaction strategies strongly correlates with performance, reinforcing that adaptability is a key factor in effective co-learning.

Our findings indicate that personality-driven behavioural differences primarily shape how hybrid teams adjust their strategies rather than determining final task outcomes. For instance, the impatient personality led to frequent object drops, yet the robot adapted by waiting longer before releasing, allowing participants to retrieve the item in time. Conversely, the patient personality increased forceful object retrievals, suggesting that participants compensated for the robot's slower release timing. These adaptations highlight the bidirectional nature of co-learning, where both the human and the robot adjust to form coherent strategies.

Moreover, our analysis of reinforcement learning dynamics showed that personality influences action consistency, with the follower personality exhibiting the highest repeatability and the impatient personality the lowest. Despite these differences, traditional reinforcement learning convergence metrics did not reveal significant effects of personality, suggesting that humans ultimately compensate for behavioural inconsistencies. This underscores the robustness of human adaptation in hybrid teams and raises questions about the necessity of strict personality matching in human-robot collaboration.

These findings have important implications for designing robots in collaborative settings. First, ensuring that robots quickly converge on a stable strategy enhances interaction consistency. Second, encouraging strategic flexibility allows teams to find suitable approaches that mitigate potential performance issues arising from personality-specific behaviours. Lastly, personality implementation should prioritize easily perceptible physical traits like movement speed and stiffness to enhance perception while using decision-making parameters to shape long-term collaboration dynamics.

References

[1] G. Kootstra, A. Bender, T. Perez, and E. van Henten, *Robotics in Agriculture*. Germany: Springer, Mar. 2020, pp. 1–19.

[2] N. G. Hockstein, C. Gourin, R. Faust, and D. J. Terris, "A history of robots: from science fiction to surgical robotics," *Journal of robotic surgery*, vol. 1, pp. 113–118, 2007.

[3] A. Hentout, M. Aouache, A. Maoudj, and I. Akli, "Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017," *Advanced Robotics*, vol. 33, no. 15-16, pp. 764–799, 2019.

[4] K. van den Bosch, T. Schoonderwoerd, R. Blankendaal, and M. Neerincx, "Six challenges for human-ai co-learning," in *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*. Springer, 2019, pp. 572–589.

[5] Futura Automation, "A history timeline of industrial robotics," 2019, accessed: 2024-09-26. [Online]. Available: https://futura-automation.com/2019/05/15/a-history-timeline-of-industrial-robotics/

[6] E. M. Van Zoelen, K. Van Den Bosch, and M. Neerincx, "Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning," *Frontiers in Robotics and AI*, vol. 8, p. 692811, 2021.

[7] M. R. Barrick, G. L. Stewart, M. J. Neubert, and M. K. Mount, "Relating member ability and personality to work-team processes and team effectiveness," *Journal of Applied Psychology*, vol. 83, no. 3, pp. 377–391, 1998.

[8] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[9] C. Esterwood and L. P. Robert, "A systematic review of human and robot personality in health care human-robot interaction," *Frontiers in Robotics and AI*, vol. 8, p. 748246, 2021.

[10] L. Luo, K. Ogawa, G. Peebles, and H. Ishiguro, "Towards a personality ai for robots: Potential colony capacity of a goal-shaped generative personality model when used for expressing personalities via non-verbal behaviour of humanoid robots," *Frontiers in Robotics and AI*, vol. 9, p. 728776, 2022.

[11] M. Y. Lim, J. D. A. Lopes, D. A. Robb, B. W. Wilson, M. Moujahid, E. De Pellegrin, and H. Hastie, "We are all individuals: The role of robot personality and human traits in trustworthy interaction," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Aug. 2022. [Online]. Available: http://dx.doi.org/10.1109/RO-MAN53752.2022.9900772

[12] E. M. van Zoelen, K. van den Bosch, M. Rauterberg, E. Barakova, and M. Neerincx, "Identifying interaction patterns of tangible co-adaptations in human-robot team behaviors," *Frontiers in Psychology*, vol. 12, p. 645545, 2021.

[13] H. W. Veldman-Loopik, "A method for embodied co-learning in interdependent human-robot teams," Master's thesis, Delft University of Technology, 2023, supervisors: Dr. Ir. Luka Peternel, Ir. Emma van Zoelen.

[14] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.

[15] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

[16] M. Stolle and D. Precup, "Learning options in reinforcement learning," in *Abstraction, Reformulation, and Approximation: 5th International Symposium, SARA 2002 Kananaskis, Alberta, Canada August 2–4, 2002 Proceedings 5*. Springer, 2002, pp. 212–223.

[17] T. G. Dietterich, "Hierarchical reinforcement learning with the maxq value function decomposition," *Journal of artificial intelligence research*, vol. 13, pp. 227–303, 2000.

[18] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Robot gains social intelligence through multimodal deep reinforcement learning," in *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)*. IEEE, 2016, pp. 745–751.

[19] S. Roccas, L. Sagiv, S. H. Schwartz, and A. Knafo, "The big five personality factors and personal values," *Personality and social psychology bulletin*, vol. 28, no. 6, pp. 789–801, 2002.

[20] Oxford University Press. (2023) Extroversion. *Oxford English Dictionary*. [Online; accessed 30-Sep-2024]. [Online]. Available: https://www.oed.com

[21] E. R. Thompson, "Development and validation of an international english big-five mini-markers," *Personality and individual differences*, vol. 45, no. 6, pp. 542–548, 2008.

[22] G. Matthews, P. A. Hancock, J. Lin, A. R. Panganiban, L. E. Reinerman-Jones, J. L. Szalma, and R. W. Wohleber, "Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems," *Personality and individual differences*, vol. 169, p. 109969, 2021.

[23] R. Alahmad, C. Esterwood, S. Kim, S. You, and Q. Zhang, "A review of personality in human–robot interactions," *Ann Arbor*, vol. 1001, pp. 48 109–1285, 2020.

[24] A. Mileounis, R. H. Cuijpers, and E. I. Barakova, "Creating robots with personality: The effect of personality on social intelligence," in *Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part I 6*. Springer, 2015, pp. 119–132.

[25] C. M. Doherty and R. B. Forbes, "Diagnostic lumbar puncture," *The Ulster medical journal*, vol. 83, no. 2, p. 93, 2014.

[26] I. Foundation, “Tissue properties database: Density,” 2024, accessed: 2024-10-02. [Online]. Available: https://itis.swiss/virtual-population/tissue-properties/database/density/

[27] J. Dolfin, “Teleoperated lumbar puncture simulation,” 2024, accessed: 2024-11-14. [Online]. Available: https://github.com/JesseDolfin/Teleoperated_LumbarPuncture_Simulation

[28] K. Chatzilygeroudis, M. Mayr, B. Fichera, and A. Billard, “iiwa_ros: A ROS stack for KUKA’s IIWA robots using the Fast Research Interface,” 2019, software available from GitHub. [Online]. Available: https://github.com/epfl-lasa/iiwa_ros

[29] J. Dolfin, “Co-learning robot personalities repository,” 2024, accessed: 2024-10-02. [Online]. Available: https://github.com/JesseDolfin/co_learning_robot_personalities

[30] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, “Mediapipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.

[31] J. Eschmann, “Reward function design in reinforcement learning,” in *Reinforcement Learning Algorithms: Analysis and Applications*, B. Belousov, H. Abdulsamad, P. Klink, S. Parisi, and J. Peters, Eds. Cham: Springer International Publishing, 2021, pp. 25–33. [Online]. Available: https://doi.org/10.1007/978-3-030-41188-6_3

[32] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[33] A. Juliani, “Maximum entropy policies in reinforcement learning & everyday life,” 2018. [Online]. Available: https://awjuliani.medium.com/maximum-entropy-policies-in-reinforcement-learning-everyday-life-f5a1cc18d32d

[34] L. Statistics, “Repeated measures anova - understanding a repeated measures anova,” https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php, accessed: 2025-01-21.

[35] S. W. Greenhouse and S. Geisser, “On methods in the analysis of profile data,” *Psychometrika*, vol. 24, no. 2, pp. 95–112, 1959.

[36] XLSTAT Help Center, “How to interpret contradictory results between anova and multiple pairwise comparisons,” 2023. [Online]. Available: https://help.xlstat.com/6741-how-interpret-contradictory-results-between-anova-and

[37] Melanie, “Pearson and spearman correlations: A guide to understanding and applying correlation methods,” *DataScientest*, January 2024. [Online]. Available: https://datascientest.com/en/pearson-and-spearman-correlations-a-guide-to-understanding-and-applying-correlation-methods

[38] M. G. Bellemare, G. Ostrovski, A. Guez, P. S. Thomas, and R. Munos, “Increasing the action gap: New operators for reinforcement learning,” 2015. [Online]. Available: https://arxiv.org/abs/1512.04860

[39] N. Team, “Entropy in machine learning — applications, examples, alternatives,” 2024, accessed: 2025-02-24. [Online]. Available: https://nebius.com/blog/posts/entropy-in-machine-learning

[40] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, “Game-theoretic modeling of human adaptation in human-robot collaboration,” in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 323–331.

[41] NASA Human Systems Integration Division, “NASA Task Load Index (TLX) Paper/Pencil Version,” https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf, accessed: 2025-02-20.

Appendix

A. Action Space Table

The full set of action that the algorithm can take per phase are shown in table VI

Phase	Action Description	Index
0	Move to the home position	0
1	Initiate the handover immediately	1
1	Wait for a state change	2
2	Go to the 'serve' handover orientation	3
2	Go to 'drop' handover orientation	4
3	Open the hand	5
3	Open the hand partially	6
3	Close the hand	7

TABLE VI
Actions with Phase and Index

B. State Space Table

Similarly, the states are defined in table VII. The state of the robot signifies the phase the experiment is in; for this reason, the phases are also added to the table.

Phase	State Description	Index
0	Home	0
1	Start handover immediately & the hand is in the workspace	1
1	Start handover immediately & the hand is not in the workspace	2
1	Wait for a state change & the hand is in the workspace	3
1	Wait for a state change & the hand is not in the workspace	4
2	Human hand orientation: serve & robot hand orientation: serve	5
2	Human hand orientation: serve & robot hand orientation: drop	6
2	Human hand orientation: drop & robot hand orientation: serve	7
2	Human hand orientation: drop & robot hand orientation: drop	8
2	Human hand orientation: unknown & robot hand orientation: serve	9
2	Human hand orientation: unknown & robot hand orientation: drop	10
3	Human input detected & Robot hand is open	11
3	Human input detected & Robot hand is partially open	12
3	Human input detected & Robot hand is closed	13
3	No human input detected & Robot hand is open	14
3	No human input detected & Robot hand is partially open	15
3	No human input detected & Robot hand is closed	16

TABLE VII
States with Phase and Index

C. Human Fluency Questionaire

This questionnaire aims to evaluate the perceived fluency and effectiveness of the collaboration between the human participant and the robot. The questions are designed to measure several key aspects of interaction, such as collaboration fluency, trust, shared goals, and the relative contribution of each party. Participants are asked to rate their level of agreement with each statement based on their experience during the experiment, using a Likert scale from 1 (Strongly Disagree) to 7 (Strongly Agree).

1	2	3	4	5	6	7
Strongly Disagree	Somewhat Disagree	Disagree	Neutral	Agree	Somewhat Agree	Strongly Agree

Questionnaire

Please rate the following statements based on your experience collaborating with the robot. Indicate your level of agreement by writing a number between 1 (Strongly Disagree) and 7 (Strongly Agree).

---

1. The robot improved over time. _____ (1-7)
2. The team worked fluently together. _____ (1-7)
3. The robot adapted to my input as the task progressed. _____ (1-7)
4. The robot contributed to the team's success. _____ (1-7)
5. I trusted the robot to act according to our shared goals. _____ (1-7)
6. The robot made decisions that aligned with my expectations for timing and handover actions. _____ (1-7)
7. The robot was committed to the task. _____ (1-7)
8. I felt like an equal partner in the team. _____ (1-7)
9. I had to guide the robot more than expected. *(reverse scored)* _____ (1-7)
10. The robot demonstrated an understanding of the shared task goals. _____ (1-7)
11. I adjusted my actions based on the robot's behaviour. _____ (1-7)
12. The robot made independent decisions when appropriate. _____ (1-7)
13. I had to constantly monitor the robot's actions to ensure task success. *(reverse scored)* _____ (1-7)
14. The robot and I shared a mutual understanding of the task requirements. _____ (1-7)
15. The robot made independent decisions when appropriate to support the task. _____ (1-7)

Categories

- **Collaboration Fluency**: 2, 3, 4
- **Relative Contribution**: 8, 9, 12
- **Trust in Robot**: 5, 13
- **Positive Teammate Traits**: 6, 7, 12, 15
- **Perception of Improvement**: 1, 3, 11
- **Perception of Shared Goal**: 5, 10, 14

D. Finite state machine

Figure 10 shows the FSM used to control the robotic arm during the handover task. It defines the task's phases, transitions, and actions based on sensory inputs, human signals, and reinforcement learning decisions.

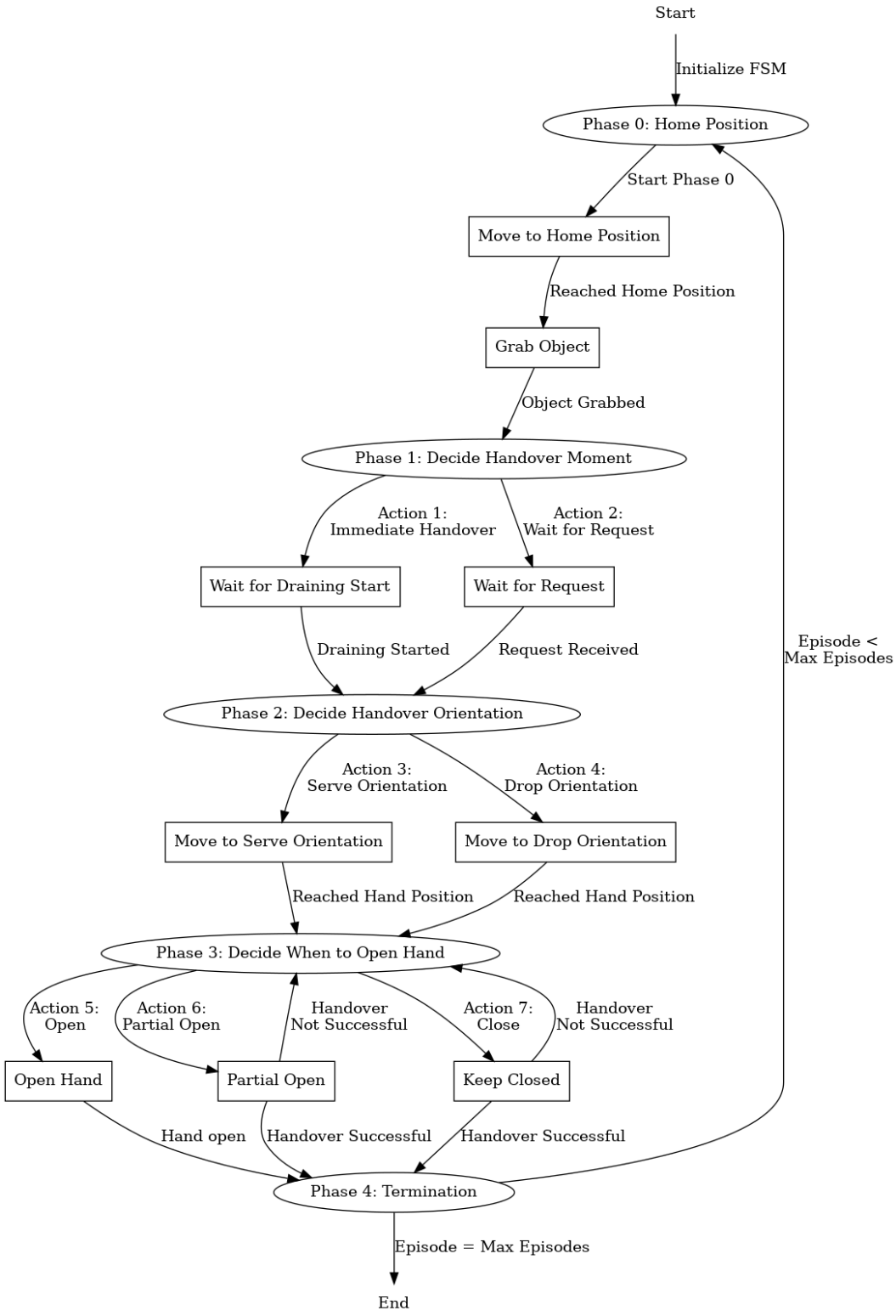

Fig. 10. Finite State Machine Diagram

E. Correlation Matrix

Figure 11 shows the full correlation matrix of all the key metrics used in the analysis of the results.

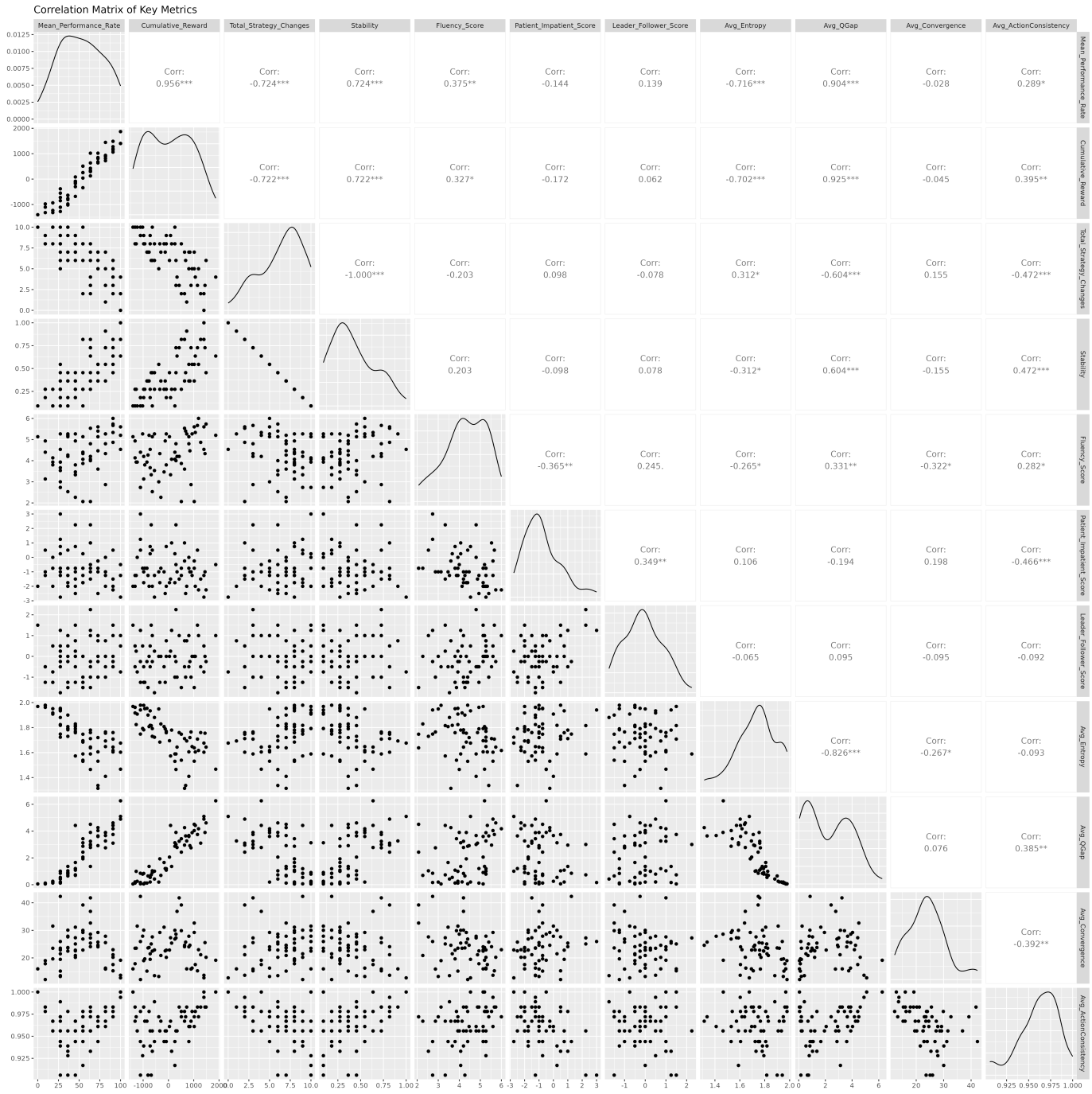

Fig. 11. Correlation Matrix

F. Extra Plots

In these appendices, we outline all the remaining plots we examined in our paper that are not significant and/or do not show any visually interesting effects.

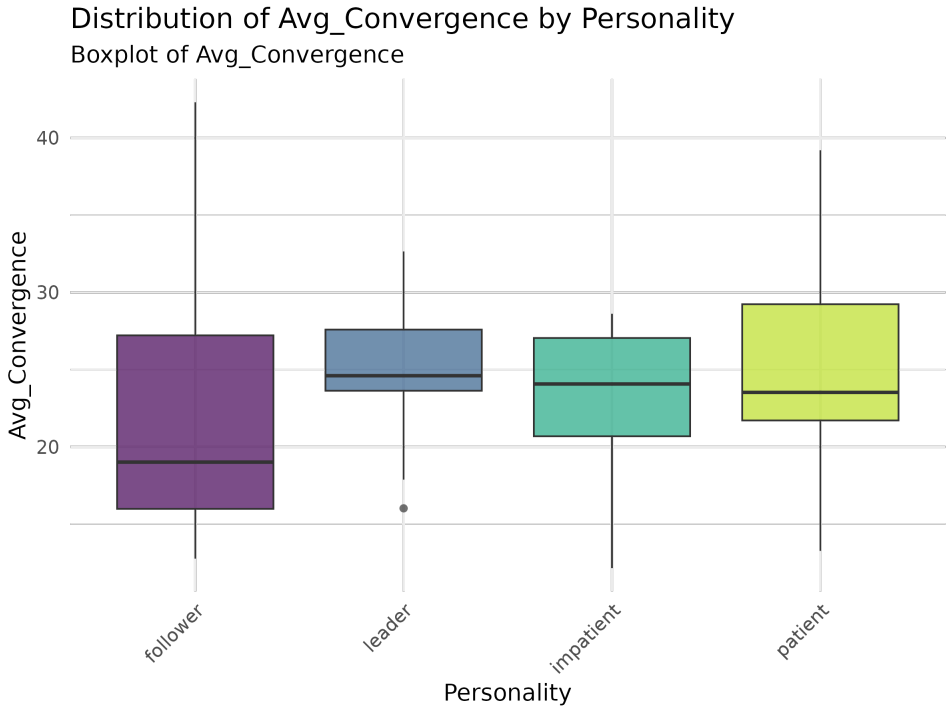

Fig. 12. Box plot showing the average Q-table convergence across different personality types.

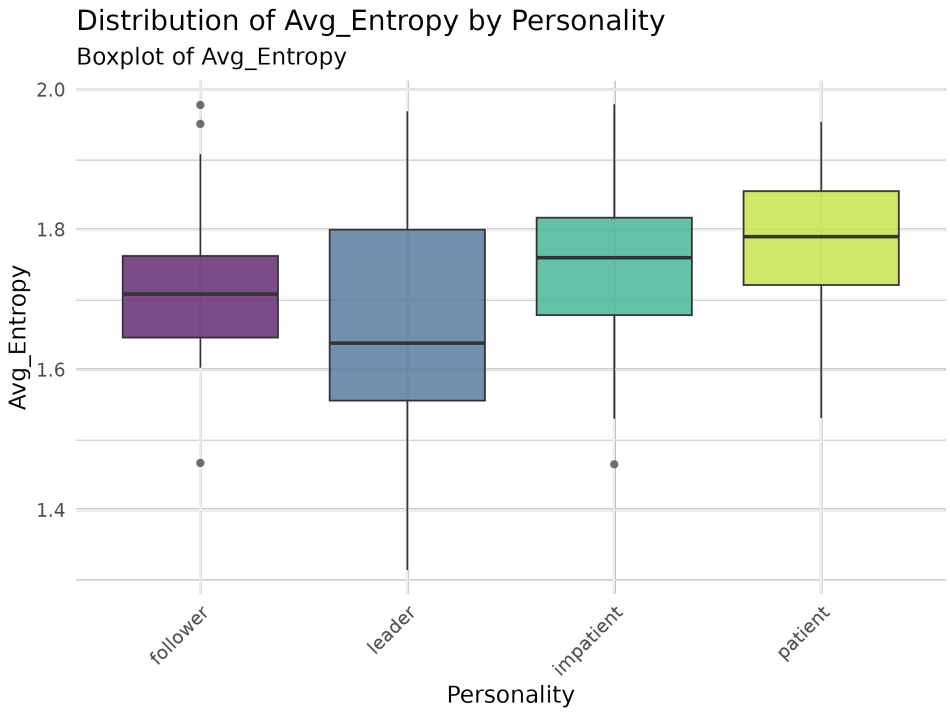

Fig. 13. Box plot showing the average entropy per personality type.

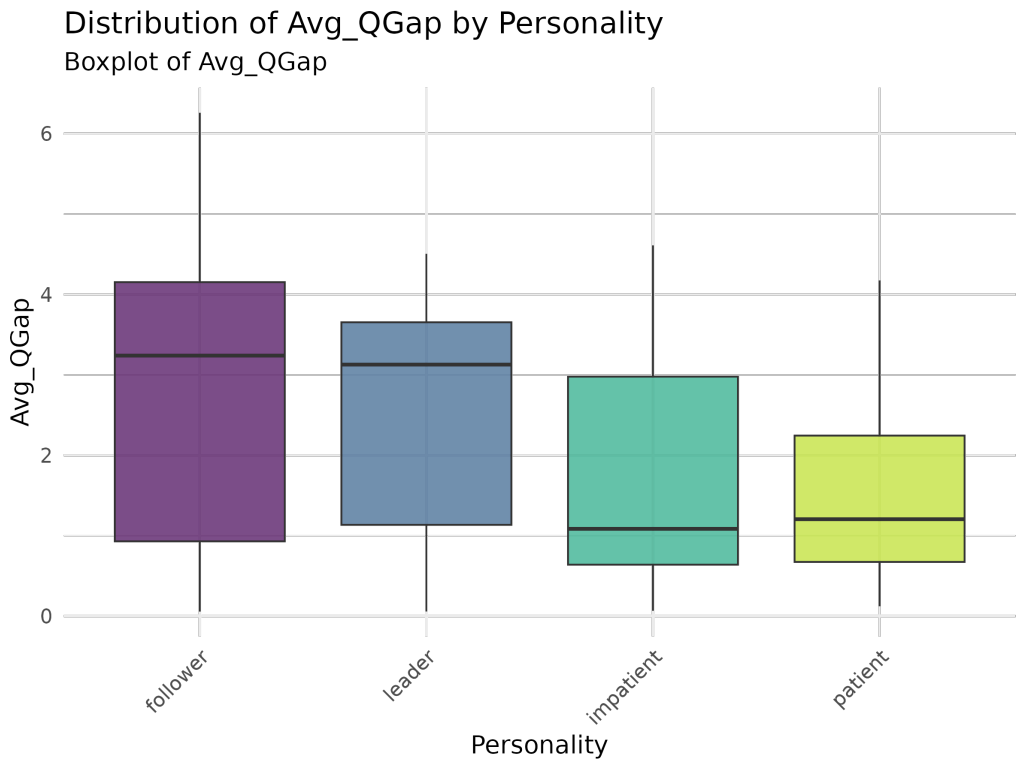

Fig. 14. Box plot showing the average Q-gap for each personality type.

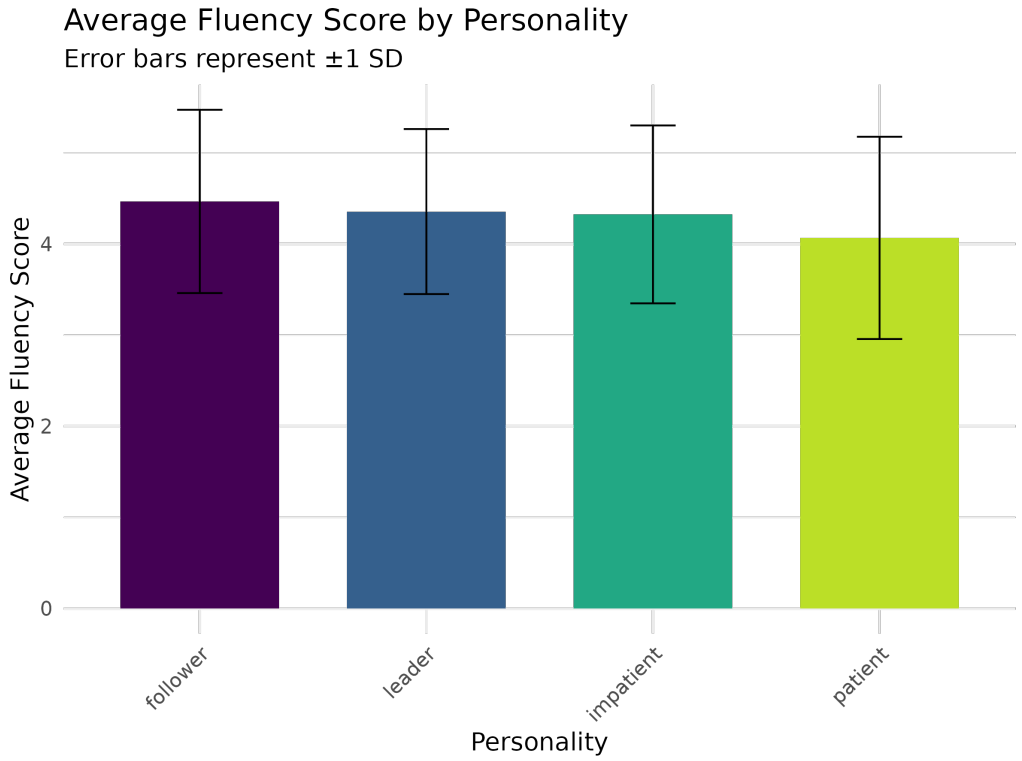

Fig. 15. Fluency score comparison across different personality types.

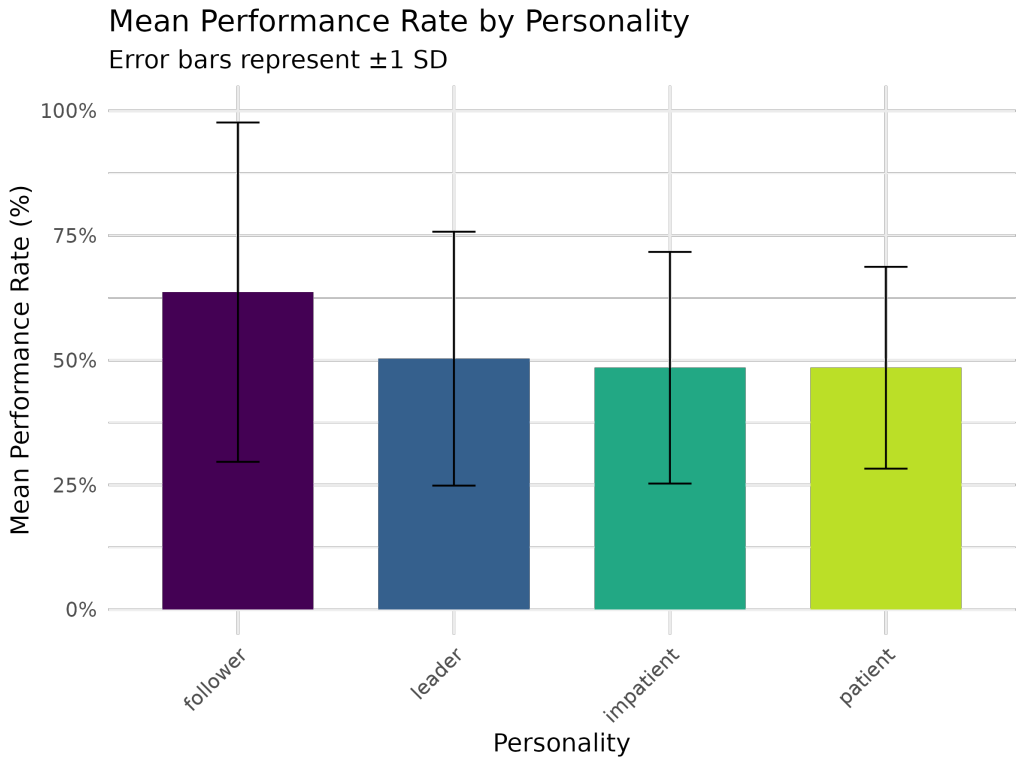

Fig. 16. Mean performance rate for each personality type.

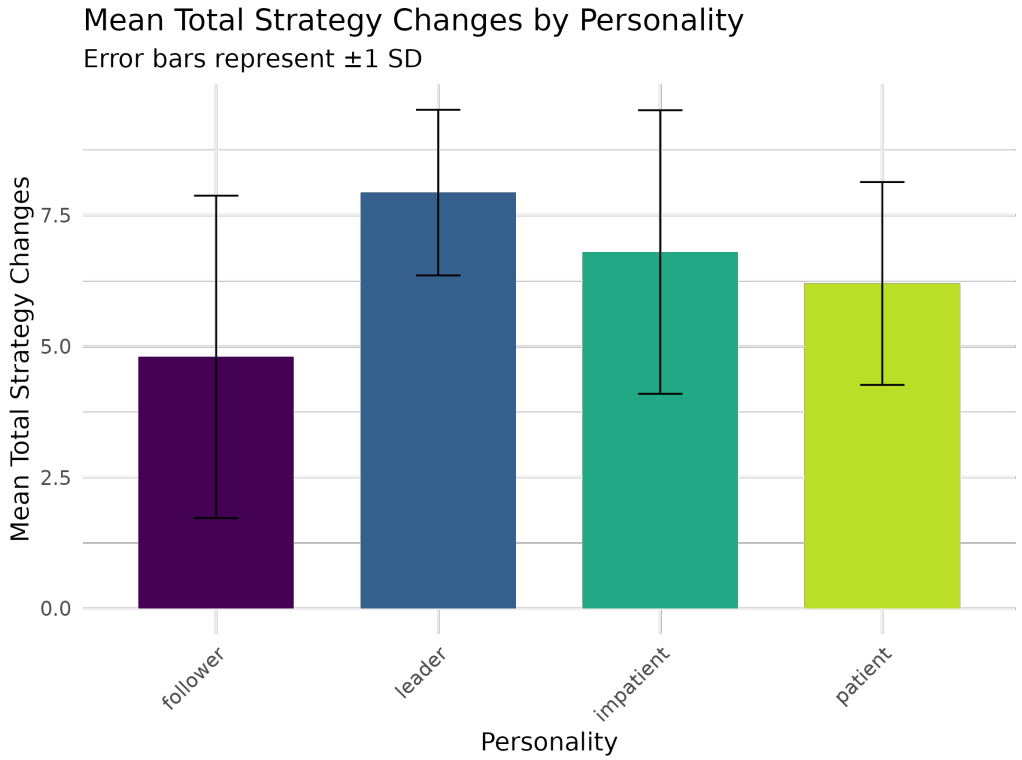

Fig. 17. Total number of strategy changes per personality type.

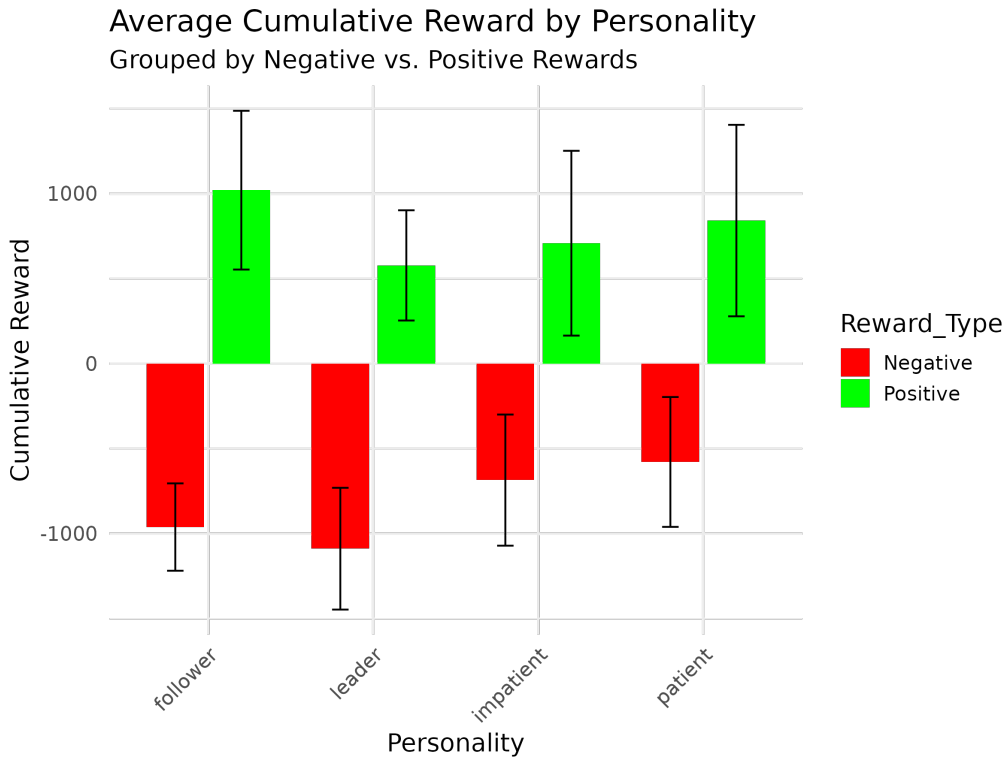

Fig. 18. Cumulative reward distribution across different personality types, separated by positive and negative rewards.

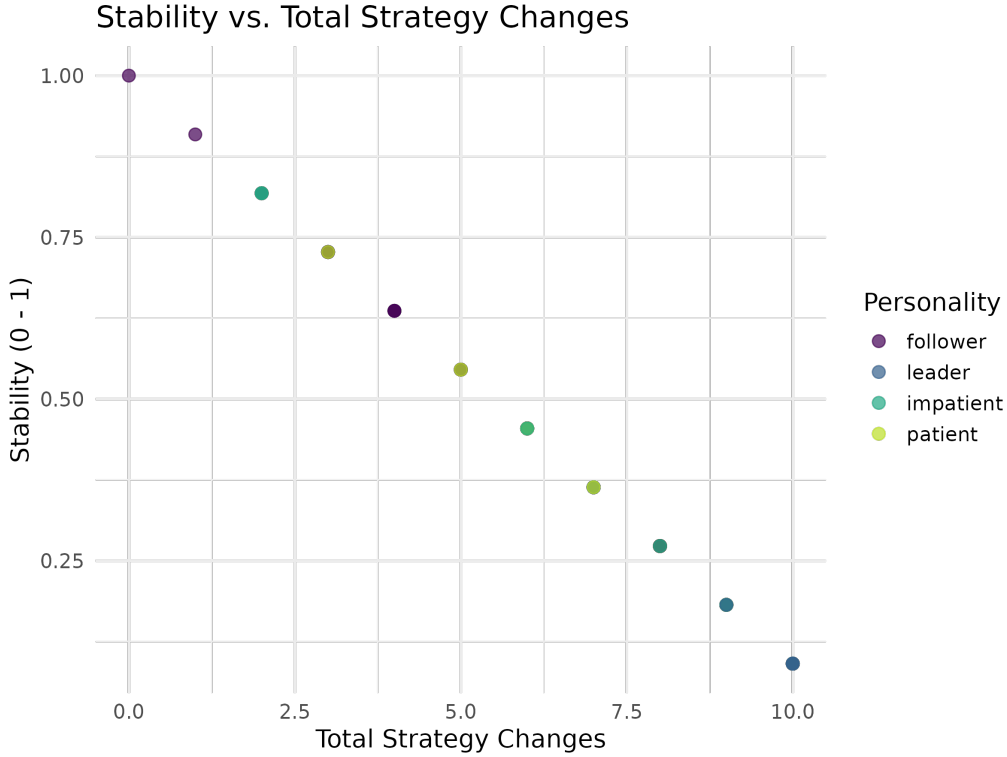

Fig. 19. Perfect negative correlation between stability and strategy changes.

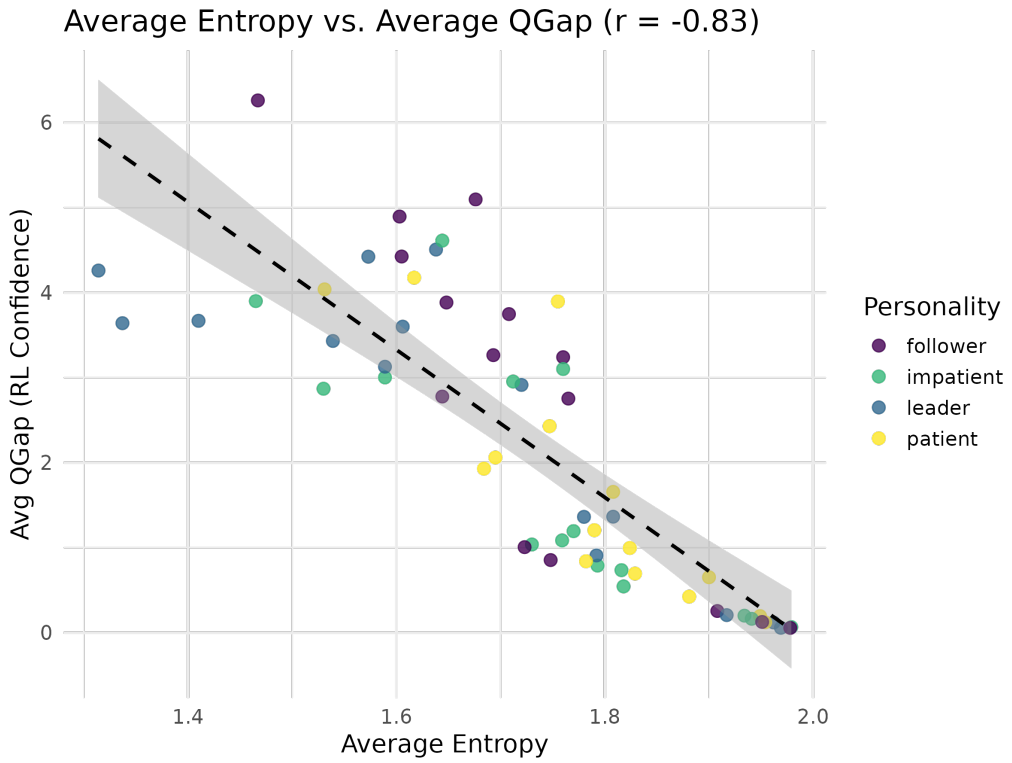

Fig. 20. Scatter plot of entropy vs. QGap, color-coded by personality type. The dashed line represents a linear regression fit, with the shaded region denoting the confidence interval.

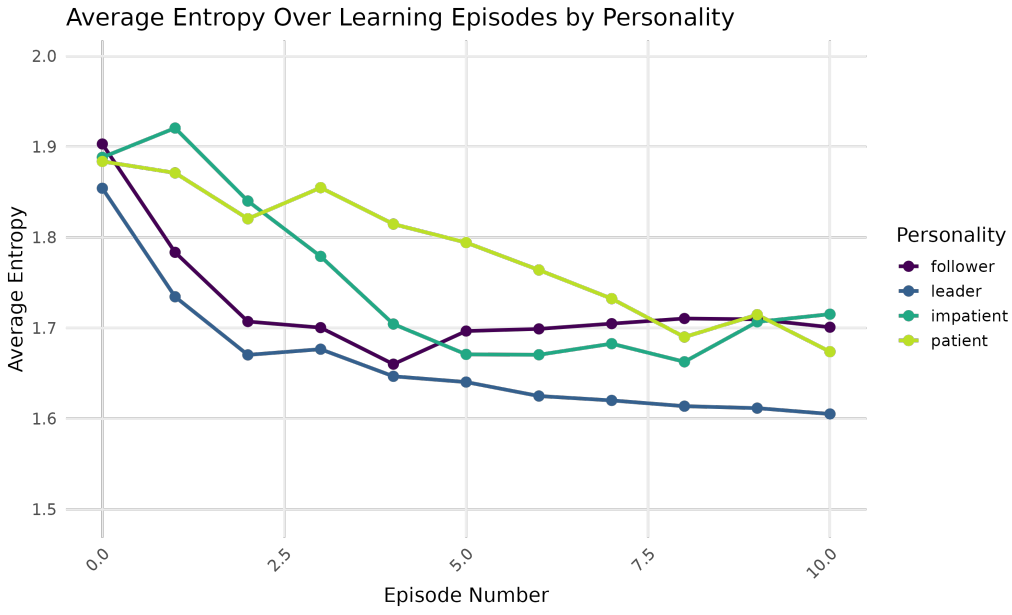

Fig. 21. Average entropy for each personality type plotted over the learning episodes

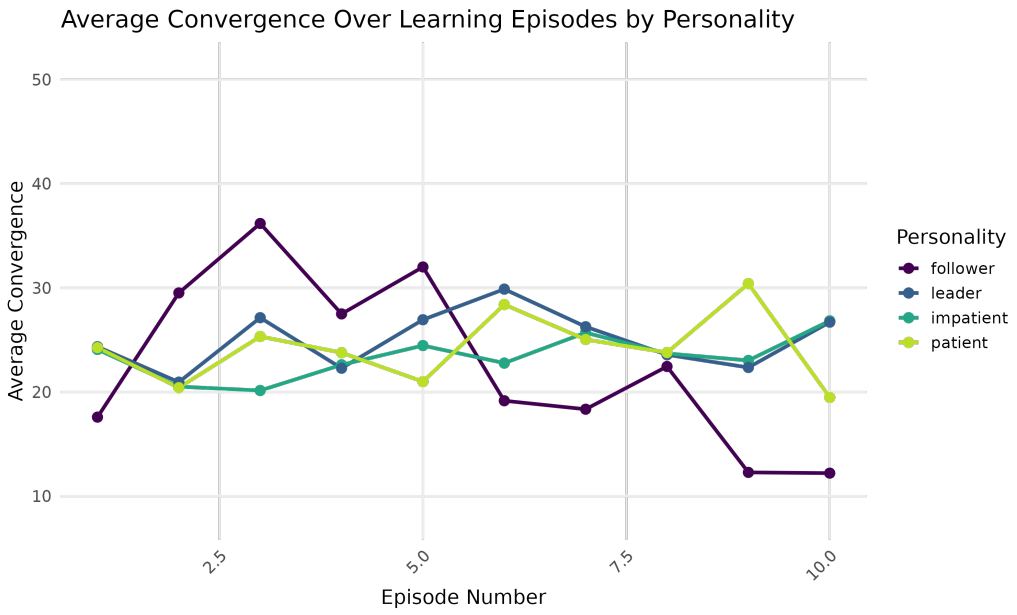

Fig. 22. Average Q-table convergence for each personality type plotted for the learning episodes

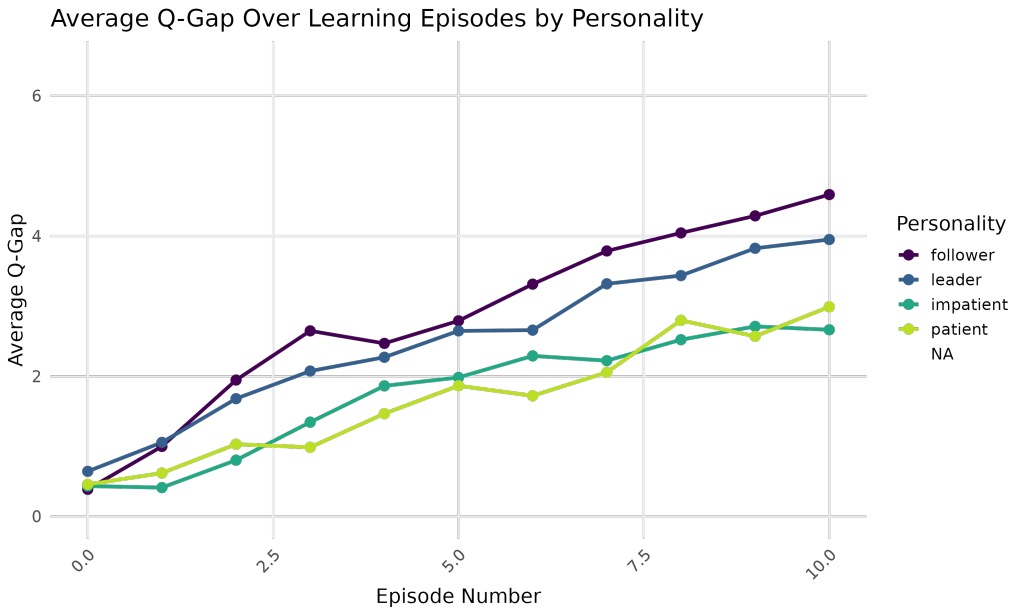

Fig. 23. QGap over episodes by personality type.

G. Interaction Patterns

TABLE VIII

CATEGORIZED INTERACTION PATTERNS OBSERVED DURING HUMAN-ROBOT COLLABORATION. INTERACTION PATTERNS ARE CLASSIFIED AS EITHER *STABLE SITUATIONS*, WHICH REPRESENT CONSISTENT AND RECURRING BEHAVIOURS, OR *SUDDEN ADAPTATIONS*, WHICH INVOLVE ABRUPT CHANGES IN BEHAVIOUR IN RESPONSE TO TASK DEMANDS OR PARTNER ACTIONS.

Interaction Pattern	Category	Personality Type	Count
Human holds item until the robot releases	Stable Situation	Patient	16
Human holds item until the robot releases	Stable Situation	Impatient	14
Human holds item until the robot releases	Stable Situation	Leader	25
Human holds item until the robot releases	Stable Situation	Follower	35
Human waits for robot to open hand	Stable Situation	Patient	49
Human waits for robot to open hand	Stable Situation	Impatient	50
Human waits for robot to open hand	Stable Situation	Leader	43
Human waits for robot to open hand	Stable Situation	Follower	48
misalignment	Stable Situation	Patient	6
misalignment	Stable Situation	Impatient	5
misalignment	Stable Situation	Leader	5
misalignment	Stable Situation	Follower	5
Robot and human meet in the middle	Sudden Adaptation	Patient	11
Robot and human meet in the middle	Sudden Adaptation	Impatient	9
Robot and human meet in the middle	Sudden Adaptation	Leader	14
Robot and human meet in the middle	Sudden Adaptation	Follower	1
Human moves towards robot	Sudden Adaptation	Patient	44
Human moves towards robot	Sudden Adaptation	Impatient	45
Human moves towards robot	Sudden Adaptation	Leader	31
Human moves towards robot	Sudden Adaptation	Follower	33
Human takes object from robot with force	Sudden Adaptation	Patient	47
Human takes object from robot with force	Sudden Adaptation	Impatient	13
Human takes object from robot with force	Sudden Adaptation	Leader	18
Human takes object from robot with force	Sudden Adaptation	Follower	17
Human touches hand unnecessarily	Sudden Adaptation	Patient	4
Human touches hand unnecessarily	Sudden Adaptation	Impatient	3
Human touches hand unnecessarily	Sudden Adaptation	Leader	1
Human touches hand unnecessarily	Sudden Adaptation	Follower	3
Robot drops object	Sudden Adaptation	Patient	2
Robot drops object	Sudden Adaptation	Impatient	36
Robot drops object	Sudden Adaptation	Leader	14
Robot drops object	Sudden Adaptation	Follower	9
Robot moves away	Sudden Adaptation	Patient	12
Robot moves away	Sudden Adaptation	Impatient	21
Robot moves away	Sudden Adaptation	Leader	24
Robot moves away	Sudden Adaptation	Follower	12
Robot moves towards human	Sudden Adaptation	Patient	56
Robot moves towards human	Sudden Adaptation	Impatient	66
Robot moves towards human	Sudden Adaptation	Leader	66
Robot moves towards human	Sudden Adaptation	Follower	73