National Technical University of Athens

School of Electrical and Computer Engineering

Division of Signals, Control and Robotics

Transfer Learning exploiting Demonstrations in a Human-Robot Interactive Game

DIPLOMA THESIS of
Nikolaos Stavrou

Supervisor:
Assoc. Prof. Costas Tzafestas, NTUA

Co-Supervisor:
Dr. Maria Dagioglou, NCSR "Demokritos"

Athens, March 2024

National Technical University of Athens

School of Electrical and Computer Engineering

Division of Signals, Control and Robotics

Transfer Learning exploiting Demonstrations in a Human-Robot Interactive Game

DIPLOMA THESIS of

Nikolaos Stavrou

Supervisor:
Assoc. Prof. Costas Tzafestas, NTUA

Co-Supervisor:
Dr. Maria Dagioglou, NCSR "Demokritos"

Approved by the examination committee on 28 March, 2024.

Costas Tzafestas	Ioannis Kordonis	Haralambos Psilakis
Associate Professor NTUA	Assistant Professor NTUA	Lecturer NTUA
.....	.....	.....

Athens, March 2024

.....

Nikolaos Stavrou
Graduate of Electrical and Computer Engineering NTUA

Περίληψη

Η συνεργασία ανθρώπου-ρομπότ απαιτεί συστήματα που έχουν ικανότητα προσαρμογής αλλά και μάθησης από αλληλεπιδράσεις, ώστε να εκτελούν αποτελεσματικά αλληλοεξαρτώμενες εργασίες. Αυτή η εργασία επεκτείνει πρόσφατη έρευνα, εστιάζοντας στη δυναμική της συνεργασίας ανθρώπου-ρομπότ όπου οι άνθρωποι καλούνται να συνεργαστούν με έναν πράκτορα βαθιάς ενισχυτικής μάθησης για την επίτευξη ενός κοινού στόχου. Η απόδοση σε τέτοιου είδους αλληλεπιδράσεις εξαρτάται από την ικανότητα του πράκτορα βαθιάς ενισχυτικής μάθησης να προσαρμόζεται στον εκάστοτε χρήστη. Η μελέτη μας υλοποιεί μια προσέγγιση μεταφοράς μάθησης μέσω επιδείξεων, συγκεκριμένα μέσω της βαθιάς Q-μάθησης από επιδείξεις, με στόχο την βελτίωση της συνεργασίας ανθρώπου-ρομπότ. Σε διαφοροποίηση με την προηγούμενη έρευνα που χρησιμοποιήθηκε επαναχρησιμοποίηση πολιτικής, ως μέθοδος μεταφοράς μάθησης, η προσέγγισή μας, σε συνδυασμό με προσαρμογές στις ρυθμίσεις του αλγορίθμου Soft Actor Critic, επιδιώκει να ενισχύσει την προσαρμοστικότητα και την αποτελεσματικότητα της μάθησης. Πραγματοποιήθηκαν πειράματα με 24 συμμετέχοντες. Τα ευρήματά μας υποδηλώνουν ότι η μεταφορά μάθησης, μέσω της μάθησης από επιδείξεις, μπορεί να επηρεάσει σημαντικά τις επιδόσεις.

Λέξεις κλειδιά αλληλεπίδραση ανθρώπου-ρομπότ · συνεργατική μάθηση · βαθιά δίκτυα Q · βαθιά ενισχυτική μάθηση · μεταφοράς μάθηση · αλγόριθμος soft actor-critic· μάθηση από επιδείξεις

Abstract

Enhancing Human-Robot Collaboration requires robots that are not only socially aware but also proficient in adapting and learning from interactions to perform interdependent tasks effectively. This thesis extends recent research by focusing on the dynamics of Human-Robot Collaboration where humans collaborate with a Deep Reinforcement Learning agent to achieve a common goal. The performance in such collaborations depends on the Deep Reinforcement Learning agent's ability to adapt and learn from its human partner and vice versa. Our study implements an alternative Transfer Learning (TL) approach, Learning from Demonstrations, specifically through Deep Q-Learning from Demonstrations (DQfD), aimed at encouraging more efficient human-robot teamwork. In contrast to the foundational work that utilized Probabilistic Policy Reuse, our approach, coupled with adjustments to the Soft Actor Critic algorithm's settings, seeks to enhance adaptability and learning outcomes. We conducted experiments involving 24 participants to evaluate the impact of these changes. Our findings suggest that the direct transfer of expertise with Learning from Demonstrations, complemented by specific SAC algorithm settings can significantly influence the collaborative performance.

Keywords human-robot interaction · collaborative learning · deep Q-networks · deep reinforcement learning · transfer learning · soft actor-critic algorithm · learning from demonstrations

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο της ολοκλήρωσης των προπτυχιακών σπουδών μου στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου. Η πειραματική ανάπτυξη και μελέτη διεξήχθει στο εργαστήριο της δραστηριότητας ρομποτικής Roboskel , του Εργαστηρίου Τεχνολογίας Γνώσεων και Λογισμικού (SKEL | The AI lab) , Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών «ΔΗΜΟΚΡΙΤΟΣ».

Η επιτυχής ολοκλήρωση της διπλωματικής εργασίας αυτής, οφείλεται σε μεγάλο βαθμό στην συνεργασία την οποία ανέπτυξαν μαζί μου, επιστήμονες και ερευνητές τους οποίους και ευχαριστώ θερμά. Ειδικότερα:

Τον καθηγητή μου κ. Κωνσταντίνο Τζαφέστα τόσο για το ότι αποδέχθηκε την πρότασή μου για την εκπόνηση της εργασίας υπό επίβλεψή του αλλά και για τις συμβουλές που μου προσέφερε καθ' όλη την διάρκεια εργασίας μου αυτής.

Την μεταδιδακτορική ερευνήτρια του εργαστηρίου Roboskel κα Μαρία Δαγιόγλου για την εμπιστοσύνη που μου έδειξε, δίνοντας μου την ευκαιρία να εργαστώ με τον ρομποτικό βραχίονα του εργαστηρίου και κυρίως για τη συνεχή καθοδήγησή της σε κάθε βήμα της εργασίας και την άμεση ανταπόκριση σε κάθε πρόβλημα που ανέκυπτε.

Ακόμη ευχαριστώ τον κ. Γιώργο Σταυρινό για την σημαντική συνεισφορά του στις ρυθμίσεις που αφορούσαν τα τεχνικά χαρακτηριστικά των συστημάτων που χρησιμοποιήθηκαν και τον μεταδιδακτορικό ερευνητή κ. Χρήστο Σπαθάρη για τις συμβουλές του και την βοήθειά του στη συγγραφή και διάρθρωση του κειμένου της εργασίας μου.

Contents

1 Εκτενής Περίληψη **12**
- 1.1 Εισαγωγή . . . 12
- 1.2 Παρεμφερής έρευνα . . . 13
  - 1.2.1 Βαθιά ενισχυτική μάθηση στη συνεργασία ανθρώπου-ρομπότ . . . 13
  - 1.2.2 Περιορισμοί βαθιάς ενισχυτικής μάθησης και μεταφορά μάθησης . . . 14
  - 1.2.3 Συνεισφορά . . . 15
- 1.3 Μεθοδολογία . . . 16
  - 1.3.1 Επισκόπηση της συνεργατικής εργασίας . . . 17
  - 1.3.2 Ομάδα χωρίς μεταφορά μάθησης . . . 18
  - 1.3.3 Ομάδα μεταφοράς μάθησης . . . 18
  - 1.3.4 Πράκτορας βαθιάς ενισχυτικής μάθησης . . . 19
  - 1.3.5 Πειραματική διαδικασία . . . 21
  - 1.3.6 Μετρικές . . . 22
- 1.4 Αποτελέσματα κύριας μελέτης . . . 24
  - 1.4.1 Αντικειμενικά αποτελέσματα . . . 26
  - 1.4.2 Υποκειμενικά αποτελέσματα . . . 28
  - 1.4.3 Σύγκριση με προηγούμενη εργασία . . . 30
- 1.5 Αποτελέσματα Φάσης 2 . . . 31
  - 1.5.1 Αντικειμενικά αποτελέσματα . . . 32
  - 1.5.2 Υποκειμενικά αποτελέσματα . . . 35
- 1.6 Συμπεράσματα . . . 36

2 Introduction **37**

3 Background **39**
- 3.1 Machine learning . . . 39
  - 3.1.1 Artificial Neural Networks . . . 40
- 3.2 Reinforcement Learning . . . 43
  - 3.2.1 RL Formulation . . . 43
  - 3.2.2 Reinforcement Learning Algorithms . . . 45
- 3.3 Deep Reinforcement Learning . . . 48
  - 3.3.1 Key Algorithms in Deep Reinforcement Learning . . . 49
  - 3.3.2 Actor-Critics Variants Overview . . . 50
- 3.4 Soft Actor-Critic . . . 50
  - 3.4.1 Soft Actor-Critic formulation . . . 51
- 3.5 Transfer Learning . . . 54
  - 3.5.1 Deep Q-Learning from Demonstrations (DQfD) . . . 56

4 Related work **58**
4.1 Overview of Human-Robot Interaction (HRI) . . . 58
4.2 Deep Reinforcement Learning in Robotics . . . 59
4.2.1 Implementations of DRL in Robotic Tasks . . . 59
4.2.2 Limitations of DRL in Robotics . . . 61
4.3 Motivation and contribution . . . 62

5 Methodology **64**
5.1 Research Approach . . . 64
5.1.1 Overview of the Collaborative task . . . 64
5.1.2 Reinforcement Learning agent . . . 65
5.1.3 Robot Control . . . 67
5.2 Study design . . . 68
5.2.1 No transfer Learning Group . . . 68
5.2.2 Transfer Learning Group . . . 70
5.2.3 Experimental procedure . . . 72
5.3 Measures . . . 73
5.3.1 Objective Measures . . . 73
5.3.2 Subjective Measures . . . 74
5.3.3 Understanding Participant Personalities in Human-Robot Collaboration . . . 75

6 Results **78**
6.1 Results - Main Study . . . 78
6.1.1 Objective Results . . . 80
6.1.2 Subjective Results . . . 84
6.1.3 Comparison with previous work . . . 85
6.2 Follow-up study . . . 86
6.2.1 Objective Results . . . 87
6.2.2 Subjective Results . . . 91

7 Conclusion and Discussion **93**

A Appendix Title **95**
.1 Entropy Tuning . . . 95
.2 Questionnaire Appendices . . . 97
.2.1 Appendix A: Big Five Personality Traits Questionnaire . . . 97
.2.2 Appendix B: Schwartz Portrait Values Questionnaire . . . 97
.2.3 Appendix C: AI Attitude Scale . . . 98
.2.4 Appendix D: Additional Personal Questions . . . 98
.2.5 Appendix E: Questionnaire 1 . . . 98
.2.6 Appendix F: Questionnaire 2 . . . 98
.2.7 Appendix F: information letter and consent form . . . 99

List of Figures

1.1 Αρχικοποιημένη θέση. Οι κινήσεις του ρομπότ περιορίζονται στο τετράγωνο. Το τελικό στοιχείο δράσης τοποθετείται σε μία από τις τέσσερις αρχικές θέσεις (′o′) και η ομάδα ανθώπου-ρομπότ πρέπει να το φέρει στο κέντρο (⊗) του τετραγώνου. 17
1.2 Διάγραμμα παιχνιδιού χωρίς μεταφορά μάθησης 18
1.3 Διάγραμμα παιχνιδιού με μεταφορά μάθησης 19
1.4 Πειραματική διαδικασία 21
1.5 Η χωροθέτηση του πειράματος συνεργασίας ανθρώπου-ρομπότ. 22
1.6 Big 5 No TL group 25
1.7 Big 5 TL group 25
1.8 PVQ 21 No TL group 25
1.9 PVQ 21 TL group 25
1.10 Αποτελέσματα στα μπλοκ δοκιμών: Νίκες, Ανταμοιβές, και κανονικοποιημένη διανυθείσα απόσταση 26
1.11 Συμπεριφορά της ομάδας με τον ειδικό (1η σειρά), μια ομάδα με LfD(2η σειρά) και μια ομάδα χωρίς μεταφορά μάθησης (3η σειρά). Τα θερμοδιαγράμματα δείχνουν τη θέση της κουκκίδας λέιζερ στη βασικό μπλοκ, στο 3ο και στο τελευταίο μπλοκ δοκιμής. Οι αριθμοί υποδεικνύουν τη συχνότητα με την οποία η κουκκίδα καταλάμβανε κάθε κελί (1εκ. × 1εκ.) - δηλαδή ο κανονικοποιημένος αριθμός των ζευγών x, y που μετρήθηκαν μέσα στο κελί και στα δέκα παιχνίδια μιας παρτίδας. 27
1.12 Κρίση ελέγχου 29
1.13 Μετρήσεις συνεργασίας 29
1.14 Καμπύλες ανταμοιβών στην μέθοδο [13] 31
1.15 Νίκες, ανταμοιβές και κανονικοποιημένη διανυθείσα απόσταση (φάση 2) . 33
1.16 Συνδυασμένα στοιχεία για ανταμοιβές φάσης 2 33
1.17 Συμπεριφορά της ομάδας με τον ειδικό (1η σειρά), μια ομάδα με LfD(2η σειρά) και μια ομάδα χωρίς μεταφορά μάθησης (3η σειρά). Τα θερμοδιαγράμματα δείχνουν τη θέση της κουκκίδας λέιζερ στη βασικό μπλοκ, στο 3ο και στο τελευταίο μπλοκ δοκιμής. Οι αριθμοί υποδεικνύουν τη συχνότητα με την οποία η κουκκίδα καταλάμβανε κάθε κελί (1 εκ. × 1 εκ.) - δηλαδή ο κανονικοποιημένος αριθμός των ζευγών x, y που μετρήθηκαν μέσα στο κελί και στα δέκα παιχνίδια μιας παρτίδας. 34
1.18 Μπλοκ δοκιμών του ειδικού (φάση 2) 34
1.19 Κρίση ελέγχου 35
1.20 Μετρήσεις συνεργασίας (φάση 2) 36

3.1 Classification of the most common machine learning algorithms [43] . . 40

3.2 Biological neurons to Artificial neurons [44] . . . 41
3.3 Artificial Neural Network [46] . . . 42
3.4 The agent–environment interaction in a MDP [50] . . . 44
3.5 RL algorithms categorization [53] . . . 46
3.6 Discrete SAC pseudo code,[38] . . . 54
3.7 DQfD pseudocode,[33] . . . 57

4.1 The robotic setup used for Shafti's HRC experiment,[93] . . . 61

5.1 Initialized Position. The robot's movements are constrained within a $20cm \times 20cm$ area. The EE is placed in one of the four starting ('$\circ$') positions and the HR team has to bring the EE in the centre ($\otimes$) of the square. A laser pointer attached to the EE of the robot provides to the human visual feedback about the position of the EE that is controlled. 65
5.2 No Transfer learning Game Diagram . . . 68
5.3 Performance of Different Initialized Agents . . . 70
5.4 Transfer Learning Game Diagram . . . 72
5.5 Experimental procedure . . . 72
5.6 The Human-Robot Collaboration setup The robot is placed in the middle of 1m x 1m table. . . . 73
5.7 Subjective fluency metric scales and items used in our studies [39] . . . 74

6.1 Big 5 No TL group . . . 79
6.2 Big 5 TL group . . . 79
6.3 PVQ 21 no TL group . . . 79
6.4 PVQ 21 TL group . . . 79
6.5 Wins over the testing blocks . . . 81
6.6 Rewards over the testing blocks . . . 81
6.7 Normalized Traveled Distance over the testing blocks . . . 81
6.8 Testing blocks behaviour of the team with the expert human (1st row), a team in the LfD group (2nd row) and a team in the No TL group (3rd row). The heatmaps show the laser dot's position in the Baseline, 3th and last testing blocks. The numbers indicate the frequency with which the dot occupied each cell (1cm $\times$ 1cm) - that is the normilized number of x, y pairs counted within the cell in all ten games of a batch. . . . 82
6.9 Judgement of Control . . . 84
6.10 Collaboration Metrics . . . 85
6.11 Rewards of PPR method,[13] . . . 86
6.12 Wins, Rewards and Normalized Traveled Distance of follow-up study . 88
6.13 Combined Figures for of follow-up study Rewards . . . 89
6.14 Testing blocks (of follow-up study) behaviour of the a team in the LfD group (1st row) and a team in the No TL group (2nd row). The heatmaps show the laser dot's position in the Baseline, 3th and last testing blocks. The numbers indicate the frequency with which the dot occupied each cell (1cm $\times$ 1cm) - that is the number of x, y pairs counted within the cell in all ten games of a batch. . . . 89
6.15 Testing blocks of the expert of follow-up study) . . . 90
6.16 Judgement of Control . . . 91

6.17 Collaboration Metrics of follow-up study 92

1 Different constant target entropies performance,[112] 96
2 Rewards for different entropies 97
3 Information letter 1 112
4 Information letter 2 113
5 Consent form 113
6 Attitude towards AI 114

List of Tables

1.1 Ρυθμίσεις διακριτού SAC 20
1.2 Χαρακτηριστικά των συμμετεχόντων στη μελέτη 24
1.3 Άποψη για την Τεχτητή Νοημοσύνη 25
1.4 Αποτελέσματα στατιστικών δοκιμών για το πρώτο και το τελευταίο μπλοκ 27
1.5 Two-Way Mixed ANOVA στην κανονικοποιημένη διανυθείσα απόσταση . 28
1.6 Χαρακτηριστικά των συμμετεχόντων (φάση 2) 31
1.7 Άποψη για την Τεχνητή Νοημοσύνη (φάση 2) 32
1.8 Αποτελέσματα μικτής ANOVA δύο κατευθύνσεων σε κανονικοποιημένη διανυθείσα απόσταση 35

5.1 discrete SAC settings 67
5.2 Game Parameters 69

6.1 Characteristics of Study Participants 78
6.2 Attitude towards AI 80
6.3 Statistical Test Results for First and Last Blocks 83
6.4 Two-Way Mixed ANOVA results on normalized travelled distance . . . 83
6.5 Characteristics of follow-up study Participants 87
6.6 Attitude towards AI follow-up study 87
6.7 Two-Way Mixed ANOVA results on normalized travelled distance . . . 91

1 Big Five Questions - Openness to Experience 99
2 Big Five Questions - Conscientiousness 100
3 Big Five Questions - Extraversion 101
4 Big Five Questions - Agreeableness 102
5 Big Five Questions - Neuroticism 103
6 Personal Values Questionnaire (PVQ) Questions- part 1 104
7 Personal Values Questionnaire (PVQ) Questions- part 2 105
8 Personal Values Questionnaire (PVQ) Questions- part 3 106
9 Questions about Personal Information, Experience in Gaming, and Knowledge in AI 107
10 AI Attitude Scale Questions- part 1 108
11 AI Attitude Scale Questions-part 2 109
12 Subjective measures separated into each measure part 1. 110
13 Subjective measures separated into each measure-part 2. 111

Chapter 1

Εκτενής Περίληψη

1.1 Εισαγωγή

Η συνεργασία ανθρώπου-ρομπότ μπορεί να θεωρηθεί ως υποκατηγορία της αλληλεπίδρασης ανθρώπου-ρομπότ [1]. Ο διαχωρισμός της με τις άλλες κατηγορίες αλληλεπίδρασης έγκειται στο γεγονός ότι κατά τη συνεργασία, οι άνθρωποι καλλούνται να αλληλεπιδράσουν με τα ρομπότ για την επίτευξη κοινών στόχων. Αυτός ο τομέας είναι εγγενώς διεπιστημονικός, αντλώντας από τομείς όπως η ρομποτική, η τεχνητή νοημοσύνη, η αλληλεπίδραση ανθρώπου-υπολογιστή, η κοινωνιολογία, η ψυχολογία και άλλους. Η συνεργασία ανθρώπου-ρομπότ βρίσκει πρακτικές εφαρμογές σε διάφορους τομείς της καθημερινότητας, όπως η εκπαιδευση[2], οι θεραπευτικές παρεμβάσεις [3], η φροντίδα ηλικιωμένων [4], αλλά και σε βιομηχανικά περιβάλλοντα [5].

Οι προκλήσεις της συνεργασίας με τα ρομπότ αυξάνονται όταν οι εργασίες που πρέπει να διατελέσουν γίνονται πιο περίπλοκες. Τα ρομπότ θα πρέπει να είναι σε θέση να λαμβάνουν αποφάσεις μόνα τους, ειδικά όταν εργάζονται στενά με ανθρώπους, για παράδειγμα όπως όταν συνεργάζονται για να σηκώνουν και να μετακινούν μεγάλα αντικείμενα [6]. Είναι σημαντικό για τα ρομπότ όχι μόνο να κινούνται με ασφάλεια αλλά και να κατανοούν τις προθέσεις των ανθρώπων, ώστε να μπορούν να βοηθήσουν με το βέλτιστο τρόπο [7]. Για να γίνουν αυτές οι αλληλεπιδράσεις ομαλές και φυσικές, τα ρομπότ θα πρέπει να προγραμματίζονται ώστε να παρατηρούν και να αντιδρούν στις ανθρώπινες συμπεριφορές, να μαθαίνουν από τη συνεργασία και να προσαρμόζουν τις ενέργειές τους.

Παρά την πρόοδο στον τομέα, η δημιουργία ομαλής συνεργασίας μεταξύ ανθρώπων και ρομπότ εξακολουθεί να αποτελεί πρόκληση. Η ταυτόχρονη μάθηση, μια κρίσιμη πτυχή της συνεργασίας, πολλές φορές είναι αργή καθώς εξαρτάται από διάφορους παράγοντες όπως η απαιτούμενη σωματική και πνευματική προσπάθεια, οι δεξιότητες του ανθρώπινου συνεργάτη, οι τεχνικές μηχανικής μάθησης που χρησιμοποιούνται και οι υπολογιστικές απαιτήσεις της εργασίας. Τα συνεργατικά ρομπότ, αναμένεται να μαθαίνουν γρήγορα, να αναγνωρίζουν τις ικανότητες των ανθρώπινων συνεργατών τους και να προσαρμόζονται στα δυνατά και τα αδύνατα σημεία τους.

Οι αλγόριθμοι βαθιάς ενισχυτικής μάθησης βρίσκουν ποικίλες εφαρμογές στον τομέα, από φορητά ρομπότ [8] και ρομποτικούς βραχίονες [9] έως drones [10] και άλλα. Η επιτυχία της βαθιάς ενισχυτικής μάθησης στα προαναφερθέντα προέρχεται από την ικανότητά της να μαθαίνει περίπλοκες κινήσεις και συμπεριφορές που είναι δύσκολο να επιτευχθούν με τις παραδοσιακές μεθόδους ελέγχου.

Ωστόσο, ένας από τους περιορισμούς της βαθιάς ενισχυτικής μάθησης είναι η γενίκευ-

ση, της υπάρχουσας γνώσης σε νέα περιβάλλοντα, άγνωστες καταστάσεις ή στη συνεργασία με νέους χρήστες [11]. Η εκπαίδευση ενός ρομπότ από το μηδέν για κάθε νέα εργασία και χρήστη είναι χρονοβόρα, με αποτέλεσμα πολλές φορές να καθιστά τη συνεργασία μη αποδοτική.

Για την αντιμετώπιση αυτού του ζητήματος, μία διαδεδομένη αντιμετώπιση είναι η μεταφορά γνώσης εντός των πλαισίων της βαθιάς ενισχυτικής μάθησης [12]. Τέτοιες προσπάθειες στοχεύουν όχι μόνο στην ανεξαρτησία των ρομπότ αλλά και στη διευκόλυνση της απρόσκοπτης ενσωμάτωσής τους σε ανθρώπινες ομάδες. Ο απώτερος στόχος είναι να προωθηθεί μια πιο διαισθητική και αποτελεσματική συνεργασία μεταξύ των δύο ετάιρων, αυξάνοντας έτσι τη συνολική παραγωγικότητα και την αρμονία της αλληλεπίδρασης.

Με βάση την εργασία [13], η παρούσα διπλωματική εργασία στοχεύει να συμβάλει στον τομέα της συνεργασίας ανθρώπου-ρομπότ εφαρμόζοντας μια προσέγγιση μεταφοράς μάθησης, μάθηση από επιδείξεις, εντός του αλγόριθμου βαθιάς ενισχυτικής μάθησης Soft Actor-Critic (SAC). Η μελέτη περιλαμβάνει συμμετέχοντες σε ένα πείραμα συνεργασίας ανθρώπου-ρομπότ, σχεδιασμένο να αξιολογεί την αποτελεσματικότητα της βαθιάς ενισχυτικής μάθησης και της μεταφοράς μάθησης, στην ενίσχυση της αποτελεσματικότητας της συνεργασίας.

Οι βασικές συνεισφορές αυτής της έρευνας περιλαμβάνουν τη μετάβαση της μεθόδου μεταφοράς μάθησης από την επαναχρησιμοποίηση πολιτικής στη μάθηση από επιδείξεις και την εφαρμογή της σε πραγματικό βραχίονα με τη χρήση του Λειτουργικού Συστήματος ROS . Επιπλέον, αυτή η διατριβή θα συγκρίνει τις δύο μεθόδους μεταφοράς μάθησης μέσω της ανάλυσης ανάλυσης των αποτελεσμάτων των συμμετεχόντων από το πείραμα. Θα συζητηθούν επίσης οι διαφορές μεταξύ αυτών των μεθόδων μεταφοράς μάθησης, αξιολογώντας τις αντίστοιχες επιπτώσεις τους στη διαδικασία συνεργασίας. Επιπλέον, η μελέτη αξιολογεί την επίδραση διαφορετικών εντροπιών στόχων στον αλγόριθμο SACστο πλαίσιο της μεταφοράς μάθησης, παρέχοντας πληροφορίες για το πώς αυτές οι παράμετροι επηρεάζουν τη δυναμική της συνεργασίας. Αυτή η ανάλυση στοχεύει στην ενίσχυση του συνεργασίας ανθρώπου-ρομπότ βελτιστοποιώντας τη διαδικασία μάθησης και προσαρμογής των ρομπότ για τη βελτίωση της συνεργασίας τους με τους ανθρώπους.

1.2 Παρεμφερής έρευνα

1.2.1 Βαθιά ενισχυτική μάθηση στη συνεργασία ανθρώπου-ρομπότ

Η ενσωμάτωση της βαθιάς ενισχυτικής μάθησης στη ρομποτική έδωσε νέες δυνατότητες στα ρομπότ, καθιστώντας τα πιο προσαρμοστικά κατά τη συνεργασία με τους ανθρώπους.

Πρώτον, στην ασφάλεια και την εμπιστοσύνη, η οποία είναι ζωτικής σημασίας σε σενάρια συνεργασίας ρομπότ-ανθρώπων. Οι αλγόριθμοι βαθιάς ενισχυτικής μάθησης μπορούν να ρυθμιστούν ώστε να εκπαιδεύουν τα ρομπότ για να δίνουν προτεραιότητα στην ασφάλεια, όπως μέσω της κίνησης τους εντός προκαθορισμένων ορίων αλλά και της επιδέξειας ανταπόκρισης τους σε τυχών απρόβλεπτες ανθρώπινες ενέργειες, κερδίζοντας με αυτόν τον τρόπο την εμπιστοσύνη των ανθρώπων.[14].

Επίσης, πολλές εργασίες στα πλαίσια συνεργασίας ανθρώπου-ρομπότ απαιτούν περίπλοκη λήψη αποφάσεων και έλεγχο. Η ικανότητα της βαθιάς ενισχυτικής μάθησης στον χειρισμό αυτών των πολύπλοκων εργασιών, όπως η πλοήγηση σε ιδιόμορφα περιβάλλοντα,

ενισχύουν τη συμβολή της στη συνεργασία [15].

Ένα άλλο πλεονέκτημα της είναι ότι ¨εκπαιδεύει' τα ρομπότ να μαθαίνουν και να προσαρμόζονται συνεχώς στην ανθρώπινη τεχνογνωσία, βελτιώνοντας την απόδοσή τους με την πάροδο του χρόνου και μειώνοντας τον χρόνο μάθησης [16].

Επιπλέον, οι αλγόριθμοι βαθιάς ενισχυτικής μάθησης παρέχουν προσαρμοστικότητα σε πραγματικό χρόνο, διασφαλίζοντας ότι τα ρομπότ μπορούν να λαμβάνουν στιγμιαίες αποφάσεις, ευθυγραμμίζοντας τις ενέργειές τους με τις ανθρώπινες προθέσεις και στόχους, βελτιώνοντας την ποιότητα της αλληλεπίδρασης. [17].

1.2.2 Περιορισμοί βαθιάς ενισχυτικής μάθησης και μεταφορά μάθησης

Η βαθιά ενισχυτική μάθηση θέτει νέα σημεία αναφοράς στη ρομποτική, επιτρέποντας στα ρομπότ να αναλαμβάνουν όλο και πιο σημαντικούς ρόλους σε διάφορους τομείς. Ωστόσο ακόμα αντιμετωπίζονται σημαντικές προκλήσεις σε πρακτικές εφαρμογές της.

Ένα σημαντικό ζήτημα είναι η **ανεπάρκεια δειγμάτων** (sample insufficient), καθώς οι αλγόριθμοι βαθιάς ενισχυτικής μάθησης χρειάζονται πολλά δεδομένα για να αναπτύξουν βελτιστοποιημένες πολιτικές. Μια πιθανή αντιμετώπιση σε αυτό είναι η παράλληλη χρήση πολλαπλών ρομπότ για τη συλλογή δεδομένων, όπως φαίνεται στο [18], αν και αυτό μπορεί να είναι μη πρακτικά εφαρμόσιμο λόγω υψηλού κόστους.

Μια εναλλακτική λύση είναι η εκπαίδευση τους σε περιβάλλον προσομοίωσης, κάτι που είναι πιο γρήγορο αλλά και πιο οικονομικό, και στη συνέχεια η εφαρμογή των πειραματικών πολιτικών στον πραγματικό κόσμο. Τεχνικές μεταφοράς, όπως η ¨Μεταφορά μηδενικής βολής' [19], περιλαμβάνουν άμεση εφαρμογή πολιτικής από προσομοιώσεις σε πραγματικές καταστάσεις, υπό την προϋπόθεση ότι το προσομοιωμένο περιβάλλον αντικατοπτρίζει τις πραγματικές συνθήκες. Αυτή η μέθοδος αξιολογήθηκε στο [20] μέσω εργασιών όπως η ώθηση και η ολίσθηση που πραγματοποιήθηκαν από ένα ρομπότ. Ωστόσο, η αποτελεσματικότητα των πολιτικών που μαθαίνονται στις προσομοιώσεις μπορεί να ποικίλλει όταν εφαρμόζονται στον πραγματικό κόσμο λόγω εγγενών διαφορών μεταξύ του προσομοιωμένου και πραγματικού περιβάλλοντος ή της πολυπλοκότητας και της απρόβλεπτης κατάστασης του πραγματικού κόσμου.

Για να γεφυρωθεί αυτό το χάσμα, χρησιμοποιούνται μέθοδοι όπως η ¨τυχαιοποίηση χώρου' domain randomization, η οποία περιλαμβάνει τυχαιοποίηση παραμέτρων προσομοίωσης για να καλύψει μια σειρά από σενάρια πραγματικού κόσμου. Για παράδειγμα, το [21] περιγράφει την εκπαίδευση ενός ανιχνευτή αντικειμένων σε διάφορες προσομοιωμένες ρυθμίσεις, οι οποίες θα μπορούσαν στη συνέχεια να λειτουργήσουν αποτελεσματικά σε εφαρμογές πραγματικού κόσμου χωρίς περαιτέρω εκπαίδευση, για εργασίες όπως επιλογή και τοποθέτηση. Αυτή η προσέγγιση έχει επίσης χρησιμοποιηθεί σε άλλους τομείς, όπως η εκτίμηση θέσης και προσανατολισμού [22] και η τμηματοποίηση [23], ενισχύοντας την ευστάθεια και λειτουργικότητα των εφαρμογών βαθιάς ενισχυτικής μάθησης στη ρομποτική [24].

Ένας άλλος περιορισμός είναι η **εξισορρόπηση εξερεύνησης-εκμετάλλευσης**, ένα θεμελιώδες δίλημμα στην ενισχυτική μάθηση. Σε ρομποτικές εφαρμογές, η τυχαία εξερεύνηση που απαιτείται για τη μάθηση αυτών των αλγορίθμων μπορεί να οδηγήσει σε μη ασφαλείς ενέργειες, προκαλώντας δυνητικά πιθανή μηχανική βλάβη. Μια πρόσφατη εργασία που αντιμετωπίζει αυτό το ζήτημα είναι το [24], όπου οι συγγραφείς παρουσιάζουν μεθόδους, ώστε αρχές ασφάλειας να μπορούν να ενσωματωθούν στην ενισχυτική μάθηση.

Η ασφάλεια έχει επίσης ληφθεί υπόψη σε εφαρμογές πραγματικού κόσμου, όπως στο [25], όπου χρησιμοποιείται ένα νευρωνικό δίκτυο για την πρόβλεψη του αποτελέσματος μιας ενέργειας όσον αφορά την ασφάλεια. Ένας άλλος περιορισμός της τυχαίας εξερεύνησης είναι ότι μπορεί να είναι μια χρονοβόρα διαδικασία λόγω της υψηλής διάστασης του χώρου δράσης σε ρομποτικές εφαρμογές (π.χ. ένας πράκτορας βαθιάς ενισχυτικής μάθησης που ελέγχει την περιστροφή του τροχού ενός αυτοοδηγούμενου αυτοκινήτου). Ένα παράδειγμα εργασίας που αντιμετωπίζει αυτό το ζήτημα είναι το [26], όπου ενδεδειγμένες τροχιές έχουν χρησιμοποιηθεί στη διαδικασία εκμάθησης στα αρχικά στάδια.

Τέλος ένα σημαντικό ζήτημα είναι η **γενίκευση της γνώσης** (generalize knowledge) από τα ρομπότ, δηλαδή η λειτουργία σε νέο, άγνωστο περιβάλλον. Οι παραδοσιακές προσεγγίσεις εκπαίδευσης από μηδενική βάση, όπως έχουμε ήδη αναφέρει, μπορεί να είναι χρονοβόρες και συχνά μη ενδεδειγμένες . Η μεταφορά μάθησης προσφέρει μια λύση με τη επαναχρησιμοποίηση γνώσης μεταξύ παρόμοιων εργασιών.

Στο [27], οι συγγραφείς παρουσιάζουν μια εφαρμογή μέσω της διαμόρφωσης κόστους (reward shaping) προκειμένου να βελτιωθεί η εκπαίδευση ενός πράκτορα ενισχυτικής μάθησης που χρησιμοποιείται για πλοήγηση, μεταβάλλοντας τη συνάρτηση κόστους με βάση τις δυνατότητες που παρέχει ο αλγόριθμος SLAM.

Η μάθηση από επιδείξεις έχει εφαρμοστεί στο [26], όπου επιλύεται το ζήτημα της αραιής συνάρτησης ανταμοιβής σε ένα σενάριο επιλογής και τοποθέτησης (pick and place). Η χρήση αυτών των τροχιών επιτρέπει στον πράκτορα να επιλύσει το έργο, το οποίο μπορεί να ήταν ανέφικτο με τυχαία εξερεύνηση.

Επίσης χρησιμοποιούνται ¨προ-μαθημένες πολιτικές' για την εκπαίδευση αυτών των πρακτόρων. Για παράδειγμα, η επιλογή πολιτικών (policy distillation) έχει εφαρμοστεί στη ρομποτική σε ένα πρόβλημα μάθησης [28], όπου ο στόχος είναι ένας μόνο πράκτορας να ¨μάθει' τρεις διαφορετικές πολιτικές για τρεις διαφορετικές εργασίες πλοήγησης και να επιλέξει ποια πολιτική να χρησιμοποιήσει προσδιορίζοντας σε πραγματικό χρόνο την προς επίλυση εργασία.

Μία άλλη προσέγγιση είναι η άμεση επαναχρησιμοποίηση πολιτικής [29], όπου ο πράκτορας μπορεί να επιλέξει μια ενέργεια με βάση μίας προ-μαθημένης πολιτικής αντί της δικής του πολιτικής. Αυτή η ιδέα έχει χρησιμοποιηθεί στο [30], όπου οι συγγραφείς διδάσκουν ένα ανθρωποειδές ρομπότ πώς να περπατάει γρήγορα, εκμεταλλευόμενοι μια πολιτική που επιτρέπει στο ρομπότ να περπατάει με κανονική ταχύτητα. Η γνώση έχει επίσης μεταφερθεί μεταξύ μορφολογικά διαφορετικών ρομπότ, όπως στο [31], όπου οι συγγραφείς εκπαιδεύουν έναν ρομποτικό χειριστή 3 συνδέσμων σε τρεις διαφορετικές εργασίες και εκμεταλλεύονται τις πολιτικές προκειμένου να εκπαιδεύσουν ένα ρομπότ 4 συνδέσμων.

Τέλος, υπάρχει η μάθηση από αναπαράσταση (representation learning). Μία εφαρμογή της παρουσιάζεται στο [32], όπου οι συγγραφείς δείχνουν πώς η εξαγωγή σημαντικών χαρακτηριστικών από το περιβάλλον μπορεί να επιταχύνει τη διαδικασία μάθησης ενός πράκτορα ενισχυτικής μάθησης στην περίπτωση της πλοήγησης κινητών ρομπότ.

1.2.3 Συνεισφορά

Η εργασία βασίζεται στο προηγούμενο πείραμα [13], όπου ένας άνθρωπος είναι υπεύθυνος για τον έλεγχο του τελικού στοιχείου δράσης του ρομπότ στον άξονα (y), ενώ ένας πράκτορας βαθιάς ενισχυτικής μάθησης ελέγχει τον άξονα (x), με στόχο την εκμάθηση της λύσης μιας εργασίας σε πραγματικό χρόνο. Ο στόχος είναι να καθοριστεί εάν η μεταφορά γνώσης από έναν προεκπαιδευμένο έμπειρο πράκτορα μπορεί να βελτιώσει τη συνολική

απόδοση της ομάδας. Στην εργασία [13] χρησιμοποιήθηκε η πιθανολογική επαναχρησιμοποίηση πολιτικής [29], μια τεχνική μεταφοράς μάθησης, η οποία επιτρέπει στον πράκτορα να επιλέξει με πιθανότητα, μια ενέργεια που βασίζεται σε μια προηγούμενη πολιτική εμπειρογνωμόνων αντί για τη δική του.

Στην προσέγγισή μας, εισάγουμε ορισμένες τροποποιήσεις στη μέθοδο. Η κύρια διαφοροποίηση από την προηγούμενη εργασία [13] έγκειται στην εφαρμογή της βαθιάς Q μάθησης από επιδείξεις (Deep Q-learning from Demonstrations) [33], μιας τεχνικής μεταφοράς μάθησης που υπάγεται στην κατηγορία μάθησης από επιδείξεις.

Ένα από τα πλεονεκτήματα της μάθησης από επιδείξεις έναντι της επαναχρησιμοποίησης πολιτικής είναι η φύση της αλληλεπίδρασης μεταξύ του ανθρώπου και του πράκτορα. Στην πρώτη, ο συμμετέχων αλληλεπιδρά με τον δικό του πράκτορα βαθιάς ενισχυτικής μάθησης, ο οποίος μπορεί να ενσωματώνει μάθηση από επιδείξεις ειδικών. Αυτή η προσέγγιση διασφαλίζει ότι ο συμμετέχων είναι κεντρικός στη διαδικασία μάθησης, αναπτύσσοντας τη δική του πολιτική χωρίς να επηρεάζεται άμεσα από τις αποφάσεις ενός ειδικού πράκτορα. Αυτό ενθαρρύνει μια πιο ανθρωποκεντρική εμπειρία μάθησης, όπου ο άνθρωπος χρήστης επηρεάζει σημαντικά τη μαθησιακή τροχιά του ρομπότ, ευθυγραμμίζοντάς το με τη δικιά του μοναδική στρατηγική.

Ένας άλλος λόγος που επιλέξαμε τη συγκεκριμένη μέθοδο είναι λόγω της συμβατότητάς του με το πλαίσιο SAC [34]. Το DQfD συνδυάζει τη δύναμη του Q-learning με τη μάθηση από επιδείξεις, επιτρέποντας στον πράκτορα να επωφεληθεί από την καθοδήγηση των ειδικών. Αυτή η μέθοδος ενσωματώνεται με τον SAC , έναν αλγόριθμο βαθιάς ενισχυτικής μάθησης εκτός πολιτικής, γνωστό για τη σταθερότητα και την απόδοσή του. Αυτή η συμβατότητα μας παρέχει ένα ισχυρό πλαίσιο για την εργασία μας. Επιπλέον, το DQfD λειτουργεί με τρόπο εκτός σύνδεσης και εκτός πολιτικής, επιτρέποντας αποτελεσματική μάθηση από προηγούμενες εμπειρίες [35], βελτιώνοντας την προσαρμοστικότητα και την απόδοση του ρομπότ μας σε σενάρια συνεργασίας σε πραγματικό χρόνο.

Οι κύριες συνεισφορές της μελέτης μας είναι:

- Εφαρμογή DQfD ως μεθόδου μεταφοράς μάθησης στην εργασία.
- Διεξαγωγή συγκριτικής μελέτης με ομάδες ανθρώπου-ρομπότ για την αξιολόγηση του αντίκτυπου της μεθόδου.
- Σύγκριση της με τη μέθοδο PPR που εφαρμόστηκε σε προγενέστερες εργασίες.
- Επίδραση διαφορετικών εντροπιών στόχων και συναρτήσεων ενεργοποίησης στον αλγόριθμο SAC στην αποτελεσματικότητα της συνεργασίας.

1.3 Μεθοδολογία

Σε αυτό το κεφάλαιο, παρουσιάζεται η επισκόπηση των μεθοδολογιών και των πειραματικών ρυθμίσεων που χρησιμοποιήθηκαν στη μελέτη συνεργασίας ανθρώπου-ρομπότ. Επίσης, περιγράφεται λεπτομερώς η χρήση του αλγόριθμου Soft Actor Critic (SAC) για τον έλεγχο του ρομπότ στο πείραμα HRC . Επιπλέον, αυτή η ενότητα εισάγει την εφαρμογή της μεθόδου μεταφοράς μάθησης Deep Q-Learning from Demonstrations (DQfD) , στο πείραμα μας.

Στη συνέχεια, αναφέρονται τόσο τα αντικειμενικά όσο και τα υποκειμενικά κριτήρια που χρησιμοποιούνται για την αξιολόγηση της ποιότητας της συνεργασίας ανθρώπου-ρομπότ, όπως και η σημασία της κατανόησης των προσωπικοτήτων των συμμετεχόντων.

1.3.1 Επισκόπηση της συνεργατικής εργασίας

Η ομάδα αποτελείται από ένα ρομποτικό βραχίονα 6 βαθμών ελευθερίας και έναν άνθρωπο. Το ρομπότ τοποθετείται στη μέση του επιπέδου (επιφάνειας) εργασίας και το τελικό στοιχείο είναι κάθετο στο επίπεδο εργασίας και μπορεί να κινείται παράλληλα με αυτό σε ένα ορισμένο ύψος. Επίσης, ένα λέιζερ, που δείχνει προς το επίπεδο εργασίας, είναι προσαρτημένο στο τελικό στοιχείο δράσης. Ενώ ο άνθρωπος ελέγχει την κίνηση του τελικού στοιχείου δράσης στον έναν άξονα (y), χρησιμοποιώντας το πληκτρολόγιο, ένας παράγοντας βαθιάς ενισχυτικής μάθησης, είναι υπεύθυνος για την κίνηση στον άξονα (x). Με τον συνδυασμό των κινήσεων δύο εταίρων το ρομπότ μπορεί να κινηθεί στο επίπεδο xy , περιορισμένο σε ένα τετράγωνο διαστάσεων 20 εκ × 20 εκ.

Στην αρχή κάθε δοκιμής - ¨παιχνιδιού¨, το ρομπότ επιλέγει τυχαία μια αρχική θέση από τις τέσσερις δυνατές (στις γωνίες του τετραγώνου) και τοποθετείται από πάνω της, όπως απεικονίζεται στο σχ. 1.1. Τη στιγμή που φτάνει στην αρχική θέση, μια ακολουθία τριών σύντομων και ενός μεγάλου ηχητικού τόνου «βεεπ» σηματοδοτούν την έναρξη του παιχνιδιού. Όταν ξεκινά το παιχνίδι, ο στόχος της ομάδας είναι να φέρει την κουκκίδα λέιζερ μέσα στον κύκλο της θέσης του στόχου, που βρίσκεται στο κέντρο του τετραγώνου. Η ομάδα κερδίζει αν καταφέρει να φέρει την κουκίδα λέιζερ μέσα στον κύκλο της θέσης του στόχου, με ακτίνα=0,01 m/s , με ταχύτητα μικρότερη από 0,05 m/s, σε 30 δευτερόλεπτα από την αρχή του παιχνιδιού.

Σχήμα 1.1: Αρχικοποιημένη θέση. Οι κινήσεις του ρομπότ περιορίζονται στο τετράγωνο. Το τελικό στοιχείο δράσης τοποθετείται σε μία από τις τέσσερις αρχικές θέσεις (′o′) και η ομάδα ανθώπου-ρομπότ πρέπει να το φέρει στο κέντρο (⊗) του τετραγώνου.

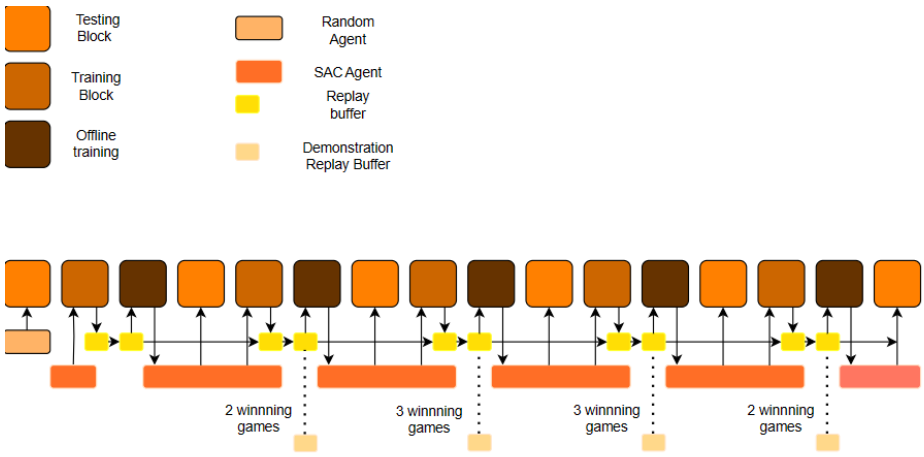
Υπάρχουν δύο ξεχωριστές ομάδες συμμετεχόντων:

1. **Ομάδα χωρίς μεταφορά μάθησης**: Οι συμμετέχοντες σε αυτήν την ομάδα αλληλεπιδρούν με έναν πράκτορα βαθιάς ενισχυτικής μάθησης 1.3.4 ξεκινώντας χωρίς προηγούμενη εκπαίδευση. Η διαδικασία μάθησης του πράκτορα βασίζεται αποκλειστικά στην αλληλεπίδρασή του με τον συμμετέχοντα κατά τη διάρκεια του πειράματος.

2. **Ομάδα με μεταφορά μάθησης**: Οι συμμετέχοντες αυτής της ομάδας συνεργάζονται με έναν πράκτορα βαθιάς ενισχυτικής μάθησης που μαθαίνει από τα δεδομένα επίδειξης του ειδικού αλλά και από την αλληλεπίδραση με τον συμμετέχοντα.

1.3.2 Ομάδα χωρίς μεταφορά μάθησης

Το πλήρες παιχνίδι, χωρίς μεταφορά μάθησης, αποτελείται από 110 δοκιμές και παρουσιάζεται στο σχήμα 1.2.



Σχήμα 1.2: Διάγραμμα παιχνιδιού χωρίς μεταφορά μάθησης

Δομή παιχνιδιού

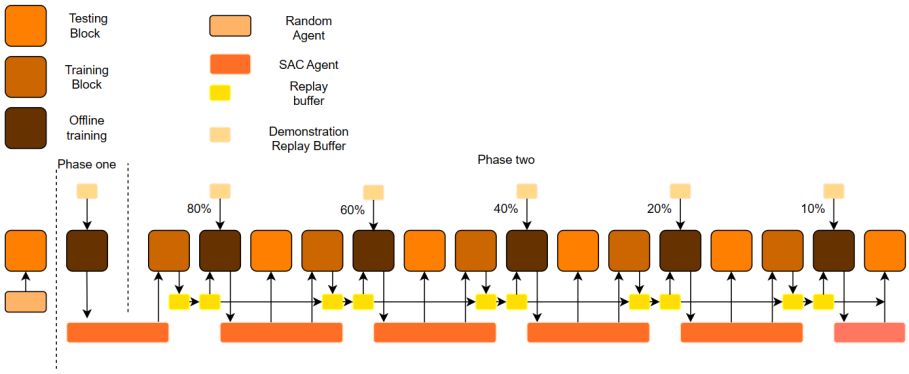
Κάθε μπλοκ στο πείραμά μας αποτελείται από 10 δοκιμές. Αναφερόμαστε σε μια ακολουθία ενός μπλοκ εκπαίδευσης, ακολουθούμενη από μία συνεδρία εκπαίδευσης εκτός σύνδεσης (offline G/U), και στη συνέχεια ένα μπλοκ δοκιμής, ως batch. Τα δεδομένα που συλλέγονται κατά τη διάρκεια των μπλοκ εκπαίδευσης αποθηκεύονται στο buffer επανάληψης. Αυτά τα δεδομένα χρησιμοποιούνται στη συνέχεια στις συνεδρίες εκπαίδευσης εκτός σύνδεσης. Η κύρια εστίασή μας για ανάλυση και αποτελέσματα θα είναι στα μπλοκ δοκιμών (test blocks). Είναι σημαντικό να σημειωθεί ότι ο πράκτορας εφαρμόζει την ίδια πολιτική τόσο στο μπλοκ εκπαίδευσης (train block) όσο και στο μπλοκ δοκιμών. Θα μπορούσαμε να χρησιμοποιήσουμε μόνο ένα μπλοκ μεταξύ κάθε εκπαίδευσης εκτός σύνδεσης, αλλά υιοθετήσαμε αυτήν την προσέγγιση για να διευκολύνουμε την άμεση σύγκριση με τα ευρήματα από την προηγούμενη εργασία [13].

1.3.3 Ομάδα μεταφοράς μάθησης

Στη μελέτη μας, ως «ειδικός» ορίζεται ένα άτομο με μεγάλη εμπειρία στο παιχνίδι, έως και 30-40 ώρες παιχνιδιού. Ο ειδικός συμμετέχει στην ίδια ρύθμιση παιχνιδιού σαν την ομάδα

χωρίς μεταφορά μάθησης. Ορισμένες από τις αλληλεπιδράσεις μεταξύ του ειδικού και του πράκτορα, συγκεκριμένα 10 επιτυχημένα παιχνίδια, καταγράφονται (όπως φαίνεται από τις διακεκομμένες γραμμές στην Εικόνα 5.2). Αυτή η συλλογή αλληλεπιδράσεων ειδικών σχηματίζει τον buffer δεδομένων επίδειξης, ο οποίος χρησιμοποιείται για τη μεταφορά γνώσης.

Η προσέγγισή για μεταφορά μάθησης είναι εμπνευσμένη από τη μέθοδο DQfD [33]. Αυτή η μέθοδος περιλαμβάνει δύο διακριτές φάσεις. Στην αρχική φάση, εστιάζουμε στην προ-εκπαίδευση, η οποία περιλαμβάνει εκπαίδευση εκτός σύνδεσης αποκλειστικά με το buffer επιδείξεων. Αυτή η φάση δίνει στον πράκτορα μια αρχική κατανόηση του περιβάλλοντος χρησιμοποιώντας δεδομένα από την αλληλεπίδραση με τον ειδικό. Η δεύτερη φάση σηματοδοτεί την έναρξη της αλληλεπίδρασης του πράκτορα με τον εκάστοτε συμμετέχοντα. Εδώ, η εκπαίδευση του πράκτορα κατά τη διάρκεια των συνεδριών εκτός σύνδεσης ενσωματώνει δεδομένα επίδειξης από τον ειδικό και δεδομένα που δημιουργούνται από τους ίδιους τους συμμετέχοντες. Συγκεκριμένα, μειώνουμε σταδιακά την αναλογία των δεδομένων επίδειξης σε κάθε συνεδρία εκπαίδευσης με την πάροδο των μπλοκς. Αυτή η στρατηγική επιτρέπει στον πράκτορα να προσαρμόζεται στη μοναδική στρατηγική κάθε συμμετέχοντα. Το σχήμα 1.3 απεικονίζει την διαδικασία του γκρούπ μεταφοράς μάθησης.



Σχήμα 1.3: Διάγραμμα παιχνιδιού με μεταφορά μάθησης

1.3.4 Πράκτορας βαθιάς ενισχυτικής μάθησης

Το πείραμα αλληλεπίδρασης ανθρώπου ρομπότ περιλαμβάνει έναν άνθρωπο που ελέγχει την επιτάχυνση στον άξονα (y) και έναν παράγοντα ενισχυτικής μάθησης που ελέγχει την επιτάχυνση του άξονα (x) του τελικού στοιχείου δράσης.

Επιλέξαμε το συγκεκριμένο αλγόριθμο για το πείραμα μας καθώς είναι ένας (model-free) αλγόριθμος, δηλαδή ο SAC δεν απαιτεί ένα προκαθορισμένο μοντέλο του περιβάλλοντος, καθιστώντας τον προσαρμόσιμο σε απρόβλεπτα περιβάλλοντα (αλληλεπίδραση ανρθώπου-ρομπότ). Επιπλέον ο SAC ως αλγόριθμος εκτός πολιτικής χρησιμοποιεί προηγούμενες εμπειρίες για την εκπαίδευσή του, ένα βασικό πλεονέκτημα σε σενάρια με περιορισμένα δεδομένα αλληλεπίδρασης σε πραγματικό χρόνο. Ένα άλλο πλεονέκτημά του είναι η δυνατότητα μάθησης εκτός σύνδεσης offline training . Αυτό το χαρακτηριστικό επιτρέπει στον SAC να βελτιστοποιεί την πολιτική του χρησιμοποιώντας δεδομένα που έχουν συλλεχθεί προηγουμένως, κάτι που είναι επωφελές σε περιβάλλοντα όπου η συνεχής online μάθηση δεν είναι εφικτή. Τέλος ο SAC προσαρμόζει την εξερεύνηση μέσω της κανονικοποίησης της εντροπίας, το οποίο είναι σημαντικό σε εργασίες αλληλεπίδρασης που απαιτούν

προσαρμοστικότητα και ανταπόκριση.

Υλοποίηση και παράμετρι του SAC

Ο αλγόριθμος βασίζεται στη μοντελοποίηση μέσω Μαρκοβιανής διαδικασίας λήψης αποφάσεων. Πιο συγκεκριμένα:

- **Κατάσταση (S):** Περιλαμβάνει τη θέση και τη ταχύτητα $\{eepos_{\mathrm{x}}, eepos_{\mathrm{y}}, eevel_{\mathrm{x}}, eevel_{\mathrm{y}}\}$ του τελικού στοιχίου δράσης.
- **Δράση (A):** Ορίζεται η επιτάχυνση στον άξονα x , με πιθανές τιμές -1, 0 ή 1.
- **Ανταμοιβή (R):** Χρησιμοποιήθηκε αραιή συνάρτηση κόστους με ποινή για καταστάσεις εκτός στόχου (-1) και ανταμοιβή για την επίτευξη του στόχου (10).

Στον Πίνακα 1.1, παρουσιάζονται οι ρυθμίσεις για τις υπερπαραμέτρους του μοντέλου, οι οποίες βασίζονται στο [13], με τις ακόλουθες δύο αλλαγές:

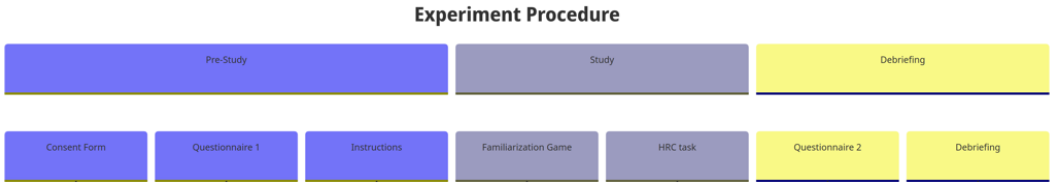
- **Συνάρτηση ενεργοποίησης:** Η συνάρτηση ενεργοποίησης των νευρωνικών άλλαξε από ReLU σε Tanh . Αυτή η τροποποίηση έγινε λόγω της συμβατότητας της Tanh με την αρχικοποίηση Xavier [36] και άλλες εργασίες όπως η [37] όπου χρησιμοποίησαν τον SAC με την Tanh για να καθορίσουν τη βέλτιστη γωνία διεύθυνσης για αυτόνομη οδήγηση σε μία πίστα.
- **Στόχος εντροπίας:** Ο στόχος εντροπίας τροποποιήθηκε από $0,36*(-\log(1/|A|))$ σε $0,98 * (-\log(1/|A|))$, όπου το A αντιπροσωπεύει το χώρο δράσης με 3 πιθανές ενέργειες. Αυτή η προσαρμογή ευθυγραμμίζεται με την προσέγγιση που εισήχθη στην εργασία του κ. Χριστοδούλου για το διακριτό SAC [38], όπου προτάθηκε αυτή η συγκεκριμένη διατύπωση στόχου εντροπίας.

Table 1.1: Ρυθμίσεις διακριτού SAC

Hyper-parameter	[13] value	our value
Layers	2 fully connected, 1 output	2 fully connected, 1 output
Fully connected layer units	32,32, moves available:3	32,32, moves available:3
Batch size	256	256
Replay buffer size	1000000	1000000
Discount rate	0.99	0.99
Learning rate Actor	0.0003	0.0003
Learning rate Critic	0.0003	0.0003
Learning rate Alpha temperature	0.001	0.001
Optimizer	Adam	Adam
Weight initializer	Xavier initialization	Xavier initialization
Activation function	Relu	**Tanh**
Networks update per off-line training	14.000	14.000
Loss function	Mean square error	Mean square error
Entropy target	0.36*(-log(1/|A|))	**0.98*(-log(1/|A|))**

1.3.5 Πειραματική διαδικασία

Οι συμμετέχοντες χωρίστηκαν σε 2 ομάδες, οι μισοί έπαιξαν το παιχνίδι χωρίς μεταφορά μάθησης ενώ οι άλλοι μισοί έπαιξαν με. Πριν από την έναρξη του πειράματος δόθηκε ένα έντυπο συγκατάθεσης στους συμμετέχοντες που τους ενημέρωνε ότι η εμπλοκή τους ήταν εθελοντική και ότι καμία προσωπική πληροφορία δεν θα χρησιμοποιηθεί χωρίς τη συναίνεσή τους. Η ενημερωτική επιστολή και το έντυπο συγκατάθεσης βρίσκονται στα Παραρτήματα 5. Το πρωτόκολλο μελέτης εγκρίθηκε από την Επιτροπή Ερευνών του Εθνικού Κέντρου Ερευνας Φυσικών Επιστημών «ΔΗΜΟΚΡΙΤΟΣ». Μετά το έντυπο συγκατάθεσης, οι συμμετέχοντες ζήτησαν συμπλήρωσαν το Ερωτηματολόγιο 1 5.3.3, το οποίο αφορά στοιχεία για τον χαρακτήρα κάθε συμμετέχοντα καθώς και την άποψη του για την Τεχνητή Νοημοσύνη.



Σχήμα 1.4: Πειραματική διαδικασία

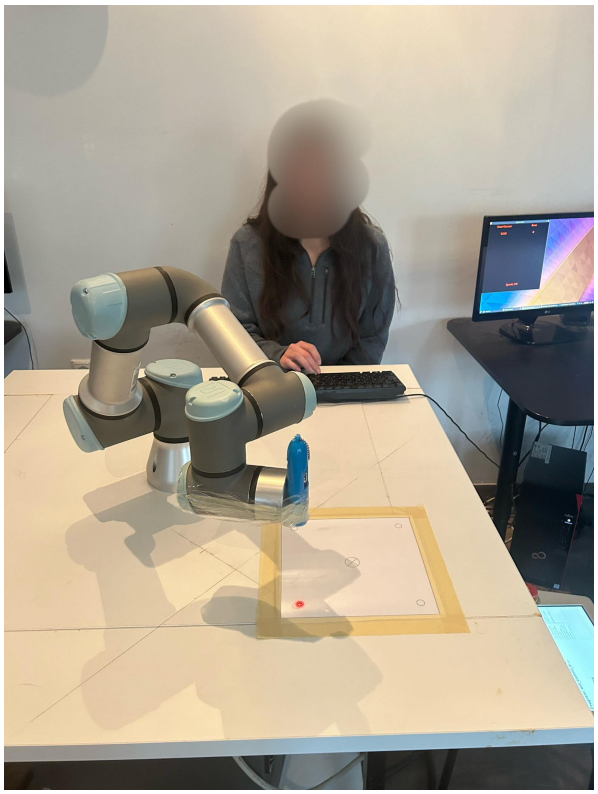
Μετά τη συμπλήρωση του Ερωτηματολογίου 1 δόθηκαν οδηγίες που δόθηκαν σε κάθε συμμετέχοντα, σχετικά με τη φύση της συνεργασίας με το ρομπότ, την εργασία που αναλαμβάνει ο καθένας τους, τον άξονα που ελέγχει κάθε μέλος της ομάδας, τον συνολικό αριθμό δοκιμών (110) καθώς και ότι κάθε παιχνίδι θα μπορούσε να τελειώσει με ¨νίκη' ή ¨ήττα¨, το καθένα με διακριτούς ήχους. Εξηγήθηκε ότι οι κινήσεις του ρομπότ περιορίζονταν σε μια καθορισμένη τερτραγωνική περιοχή, όπως φαίνεται στο 1.1. Επιπλέον, οι συμμετέχοντες ενημερώθηκαν για την ασφάλεια λόγω των κινηματικών περιορισμών του ρομπότ και του κουμπιού απενεργοποίησης έκτακτης ανάγκης. Οι συμμετέχοντες δεν ενημερώθηκαν για τον πράκτορα βαθιάς ενισχυτικής μάθησης που ελέγχει την κίνηση στον άξονα (x) ούτε για την εκπαίδευση εκτός σύνδεσης του πράκτορα κάθε 20 παιχνίδια, ούτε την ομάδα τους.

Οι οδηγίες ενημέρωσαν επίσης τους συμμετέχοντες για τον τύπο ελέγχου που είχαν στον άξονα (y). Συγκεκριμένα η οδηγία που δίνεται για το χειρισμό είναι:

- 'i¨: Ο συμμετέχων μπορεί να απομακρύνει το τελικό στοιχείο δράσης από αυτόν.
- 'm¨: Ο συμμετέχων μπορεί να μετακινήσει το τελικό στοιχείο δράσης προς το μέρος του.
- 'k¨: Όταν πατάει το κουμπί 'k¨, ο συμμετέχων δίνει εντολή στο τελικό στοιχείο δράσης να συνεχίσει να κινείται με τον ίδιο ακριβώς τρόπο που κινούνταν τη στιγμή που πάτησε το κουμπί.

Οι συμμετέχοντες έλαβαν θέση όπως απεικονίζεται στην Εικόνα 1.5 και ξεκίνησαν το παιχνίδι εξοικείωσης, όπως περιγράφεται στο 5.2.1, προκειμένου να αποκτήσουν σαφέστερη κατανόηση και έλεγχο της κίνησης του ρομπότ κατά μήκος ενός άξονα. Μόλις ολοκληρώθηκε η εξοικείωση, προχώρησαν στο κύριο παιχνίδι συνεργασίας, όπως περιγράφεται στο 5.1.1. Μετά το παιχνίδι, οι συμμετέχοντες συμπλήρωσαν το Ερωτηματολόγιο

2 5.3.2, σκοπός του οποίου είναι να αξιολογήσουν τα υποκειμενικά μέτρα της συνεργασίας. Στο τέλος πραγματοποιήθηκε ενημερωτική συνεδρία. Αυτή η συνεδρία έδωσε την ευκαιρία στους συμμετέχοντες να μοιραστούν τις εμπειρίες και τις σκέψεις τους σχετικά με την αλληλεπίδραση, προσφέροντας πολύτιμες πρωτογενείς πληροφορίες για τη συνεργασία ανθρώπου-ρομπότ.



Σχήμα 1.5: Η χωροθέτηση του πειράματος συνεργασίας ανθρώπου-ρομπότ.

1.3.6 Μετρικές

Αντικειμενικά Κριτήρια

Στη μελέτη μας, τα αντικειμενικά μέτρα χρησιμεύουν ως ποσοτικοποιήσιμοι δείκτες, παρέχοντας μια αντικειμενική αξιολόγηση της αλληλεπίδρασης μεταξύ των ανθρώπινων συμμετεχόντων και του ρομπότ. Αυτά τα μέτρα περιλαμβάνουν:

- **Συνολικός χρόνος αλληλεπίδρασης**: Το μέγεθος αυτό μετράει τη συνολική διάρκεια της ενεργής εμπλοκής μεταξύ του συμμετέχοντα και του ρομπότ κατά τη διάρκεια κάθε εργασίας.

- **Σκορ**: Ξεκινώντας από 150, το σκορ μειώνεται κατά ένα για κάθε πλαίσιο ελέγχου (control frame) κατά τη διάρκεια του παιχνιδιού. Αυτό ευθυγραμμίζεται με τη συχνότητα των πλαισίων ελέγχου, που εμφανίζονται κάθε 200 ms, εντός της διάρκειας των 30 δευτερολέπτων μιας δοκιμής. Συνεπώς έχουμε 150 πλαίσια ελέγχου ανά παιχνίδι.

- **Αριθμός νικών**: Το μέγεθος αυτό παρακολουθεί το συνολικό αριθμό των επιτυχημένων δοκιμών σε ένα μπλοκ (10 παιχνίδια).
- **Κανονικοποιημένη διανυθείσα απόσταση**: Υπολογίζεται ως η απόσταση που διανύθηκε, πολλαπλασιασμένη με το ποσοστό του συνολικού χρόνου που διαρκεί ένα παιχνίδι.
- **Θερμογράμματα**: Πρόκειται για χωρικές αναπαραστάσεις που δείχνουν τη συχνότητα των θέσεων του τελικού στοιχείου δράσης σε όλο το χώρο εργασίας παρέχοντας μια οπτική αναπαράσταση της αλληλεπίδρασης.

Για να αναλύσουμε αυτά τις αντικειμενικές μετρήσεις, χρησιμοποιήσαμε μεθόδους στατιστικής ανάλυσης, όπως η mixed ANOVA, για να συμπεράνουμε αν τα αποτελέσματα μας επιδυκνύουν στατιστικές σημαντικότητες.

Υποκειμενικά κριτήρια

Τα υποκειμενικά μέτρα είναι σημαντικά για την κατανόηση της ανθρώπινης αντίληψης κατά την αλληλεπίδραση με το ρομπότ. Αυτές οι καταγραφές βασίζονται στις προσωπικές απόψεις και εμπειρίες των συμμετεχόντων, αποτυπώνοντας πτυχές της αλληλεπίδρασης που είναι δύσκολο να ποσοτικοποιηθούν αντικειμενικά. Η μελέτη μας χρησιμοποίησε ένα ερωτηματολόγιο εμπνευσμένο από το έργο του Hoffman [39], εστιάζοντας σε έξι βασικές πτυχές της συνεργασίας: ευχέρεια στην αλληλεπίδραση ανθρώπου-AI, συμβολή της TN, βελτίωση της ομάδας, εμπιστοσύνη, εκπαίδευση και συμμαχία. Το ερωτηματολόγιο είναι προσαρμοσμένο στους στόχους της μελέτης και έχει σχεδιαστεί για να καταγράφει τις αντιλήψεις και εμπειρίες κάθε συμμετέχοντα καθ΄ όλη τη διάρκεια της αλληλεπίδρασης.

Κατανόηση των προσωπικοτήτων των συμμετεχόντων στην αλληλεπίδραση ανθρώπου-ρομπότ

Αναγνωρίζοντας τη σημασία των ατομικών διαφοροποιήσεων, η έρευνά μας δίνει επίσης έμφαση στην κατανόηση των προσωπικοτήτων των συμμετεχόντων στη συνεργασία ανθρώπου-ρομπότ. Η προσωπικότητα ενός ανθρώπου που συμμετέχει στο πείραμα μπορεί να επηρεάσει σημαντικά τη στρατηγική και την αλληλεπίδρασή του με το ρομπότ, απαιτώντας μια προσαρμοστική απόκριση από το ρομποτικό σύστημα. Για τον σκοπό αυτό, αναπτύξαμε το ¨Ερωτηματολόγιο 1¨, το οποίο οι συμμετέχοντες συμπληρώνουν πριν από την έναρξη του παιχνιδιού. Αυτό το ερωτηματολόγιο χωρίζεται σε τρία μέρη:

1. **Big Five (Χαρακτηριστικά της προσωπικότητας)**: Αυτό το τμήμα, περιλαμβάνει 50 ερωτήσεις που μας βοηθούν να κατανοήσουμε πέντε βασικές πτυχές της προσωπικότητας: Εξωστρέφεια, Ευπροσάρμοστικότητα, Ευσυνειδησία, Συναισθηματική σταθερότητα/Νευρωτισμός και Ανοιχτός σε νέες εμπειρίες.
2. **Schwartz Portrait Values Questionnaire (PVQ)**: Χρησιμοποιώντας το PVQ-21 αποτυπώνονται δέκα βασικές ανθρώπινες αξίες.
3. **AI Attitude Scale**: Αυτή η κλίμακα μας βοηθά να κατανοήσουμε τα συναισθήματα και τις σκέψεις των συμμετεχόντων σχετικά με την TN.

Μέσω αυτής της προσέγγισης, η μελέτη μας στοχεύει στη δημιουργία μιας κατανόησης τόσο των αντικειμενικών όσο και των υποκειμενικών πτυχών της αλληλεπίδρασης ανθρώπου-ρομπότ.

1.4 Αποτελέσματα κύριας μελέτης

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα της μελέτης. Στην αρχή αποτυπώνονται τα χαρακτηριστικά των συμμετεχόντων. Στη συνέχεια συγκρίνονται τα χαρακτηριστικά και οι αξίες της προσωπικότητας μεταξύ ομάδων χρησιμοποιώντας τα Big Five και Schwartz Portrait Values Questionnaire (PVQ). Εξετάζεται επίσης τη στάση των συμμετεχόντων απέναντι στην τεχνητή νοημοσύνη, η οποία είναι σημαντική για την κατανόηση της αλληλεπίδρασής τους με το ρομπότ.

Επίσης παρουσιάζονται τα αντικειμενικά αποτελέσματα, όπως οι χρόνοι αλληλεπίδρασης, οι καμπύλες μάθησης και οι μετρήσεις απόδοσης. Διενεργείται στατιστική ανάλυση για την επαλήθευση της ορθότητας των ευρημάτων, εστιάζοντας στις διαφορές απόδοσης μεταξύ των ομάδων. Επιπρόσθετα, το κεφάλαιο διερευνά τις υποκειμενικές εμπειρίες, συμπεριλαμβανομένων των αντιλήψεων των συμμετεχόντων για τον έλεγχο και την ποιότητα της συνεργασίας.

Το κεφάλαιο αντιπαραβάλλει επίσης αυτά τα ευρήματα με την προηγούμενη μελέτη [13] και εισάγει μια νέα πειραματική φάση (Φάση 2), με τροποποιημένες ρυθμίσεις.

Πίνακας 1.2: Χαρακτηριστικά των συμμετεχόντων στη μελέτη

Χαρακτηριστικό	**Λεπτομέρειες**
Κατανομή φύλου	5 γυναίκες, 11 άνδρες
Ηλικία	16 - 31 ετών
Επικρατές χέρι	13 δεξιόχειρες, 3 αριστερόχειρες
Εμπειρία παιχνιδιού	8 με ’5 χρόνια, 2 με 3-5 χρόνια, 1 με 0-1 έτος, 5 με κανένα
Προτιμώμενες συσκευές παιχνιδιών	10 φορητοί υπολογιστές, 3 κονσόλες, 2 κινητά, 1 κανένα

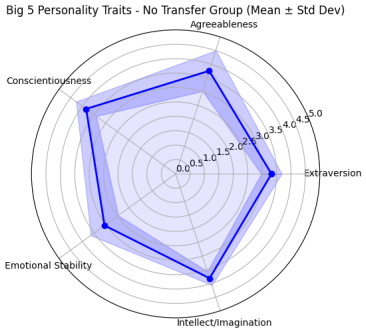
Big 5 και PVQ

Η σύγκριση των χαρακτηριστικών της προσωπικότητας μέσω του Big 5 των δύο ομάδων, παρουσιάζεται στο 1.6 και στο 1.7, δείχνει σύγκλιση ως προς τα χαρακτηριστικά της προσωπικότητάς τους. Αυτή η σύγκλιση υποδηλώνει ότι, όσον αφορά τη συνεργασία τους με το ρομπότ, και οι δύο ομάδες είναι πιθανό να εμφανίζουν παρόμοιες συμπεριφορές.

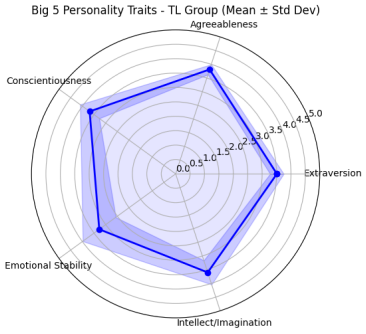
Κατά την εξέταση των αποτελεσμάτων PVQ, όπως απεικονίζεται στο 1.9 για την ομάδα με μεταφορά μάθησης και 1.8 για την ομάδα χωρίς αυτή, υπάρχουν διαφορές σε ορισμένα και συγκεκριμένα χαρακτηριστικά. Η πρώτη δείχνει οριακά υψηλότερες τιμές σε συγκεκριμένους τομείς, κάτι που πιθανώς καταδεικνύει περισσότερο ενθουσιώδη και δημιουργική προσέγγιση στις αλληλεπιδράσεις ανθρώπου-ρομπότ. Είναι ακόμη πιθανό να δείχνει μεγαλύτερη διάθεση για την απόκτηση νέων εμπειριών και ισχυρότερο κίνητρο για την επίτευξη στόχων κατά την επίτευξη των επιδιωκόμενων στόχων των εργασιών.

Πάντως, παρά τις διαφοροποιήσεις σε ορισμένες πτυχές των χαρακτηριστικών της προσωπική τους φυσιογνωμίας, οι δύο ομάδες παρουσιάζουν παρόμοια σύνθεση προσωπικότη-

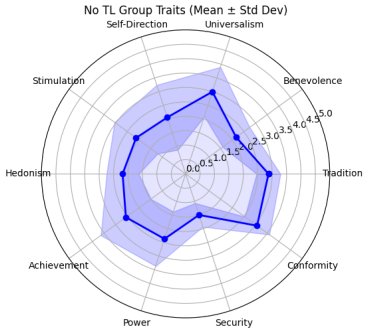
τας. Αυτή η σύγκλιση των προσωπικοτήτων οδηγεί στο συμπέρασμα ότι και οι δύο ομάδες είναι πιθανό να προσεγγίσουν τις αλληλεπιδράσεις ανθρώπου-ρομπότ με συγκρίσιμο τρόπο.



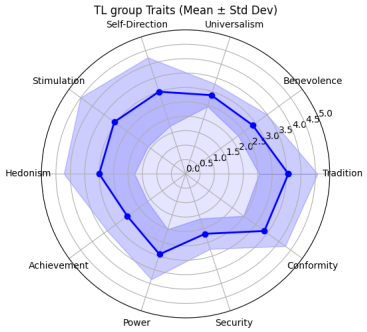
Σχήμα 1.6: Big 5 No TL group



Σχήμα 1.7: Big 5 TL group



Σχήμα 1.8: PVQ 21 No TL group



Σχήμα 1.9: PVQ 21 TL group

Άποψη για την Τεχνητή Νοημοσύνη

Και οι δύο ομάδες, έχουν θετική άποψη για την Τεχτητή Νοημοσύνη, όπως φαίνεται στον Πίνακα 1.3. Η ομάδα με μεταφορά μάθησης εμφανίζεται περισσότερο θετική σχετικά με την τεχνητή νοημοσύνη, στοιχείο που δείχνει ότι έχουν μεγαλύτερη εμπιστοσύνη στη συνεργασία με το ρομπότ κατά τη διάρκεια της εργασίας. Η άλλη ομάδα, αν και είναι επίσης θετική, μπορεί να είναι λίγο πιο προσεκτική και επιφυλακτική, στοιχείο που στοιχειοθετεί προσπάθεια ανίχνευσης νέων προοπτικών και αναζητήσεων της χρήσης και των εφαρμογών της τρέχουσας τεχνολογίας.

Πίνακας 1.3: Άποψη για την Τεχτητή Νοημοσύνη

Ομάδα	Μέσο	Τυπική απόκλιση
Χωρίς μεταφορά μάθησης	0.488	0.406
Με μεταφορά μάθησης	0.541	0.369

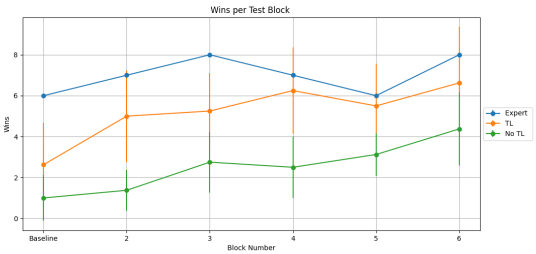
Συνολικά, με βάση τις παραπάνω απαντήσεις, οι δύο ομάδες δεν παρουσιάζουν σημαντικές διαφορές στα χαρακτηριστικά και τις συμπεριφορές τους.

1.4.1 Αντικειμενικά αποτελέσματα

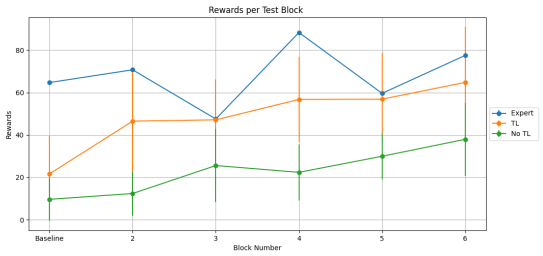
Στα παιχνίδια δοκιμών, η ομάδα με μεταφορά μάθησης εμφάνισε μέσο χρόνο αλληλεπίδρασης 21 λεπτών με τον πράκτορα, με τυπική απόκλιση 2,6 λεπτών. Αντίθετα, η ομάδα χωρίς είχε μέσο χρόνο αλληλεπίδρασης 26 λεπτά, με τυπική απόκλιση 1,4 λεπτά.

Οπτική αναπαράσταση της καμπύλης εκμάθησης στα 60 παιχνίδια δοκιμών, παρουσιάζεται στις Εικόνες 1.10α′ και 1.10β′, οι οποίες απεικονίζουν τις καμπύλες μάθησης για τις νίκες και τις ανταμοιβές, και στις δύο ομάδες, καθώς και απόδοση του ειδικού. Η ομάδα χωρίς μεταφορά μάθησης διατηρεί μια σχετικά σταθερή απόδοση, ενώ ο ειδικός δείχνει μια συνολική καλή απόδοση λαμβάνοντας υπόψη την όλη αλληλεπίδραση.

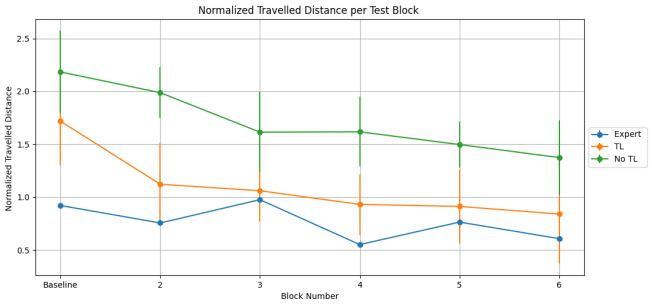
Επιπλέον, στο Σχήμα 1.10γ′, παρουσιάζουμε την κανονικοποιημένη διανυθείσα απόσταση που είναι η διανυθείσα απόσταση πολλαπλασιασμένη με το ποσοστό του συνολικού χρόνου που δαπανάται σε ένα παιχνίδι.



(α′) Νίκες στα μπλοκ δοκιμών



(β′) Ανταμοιβές στα μπλοκ δοκιμών

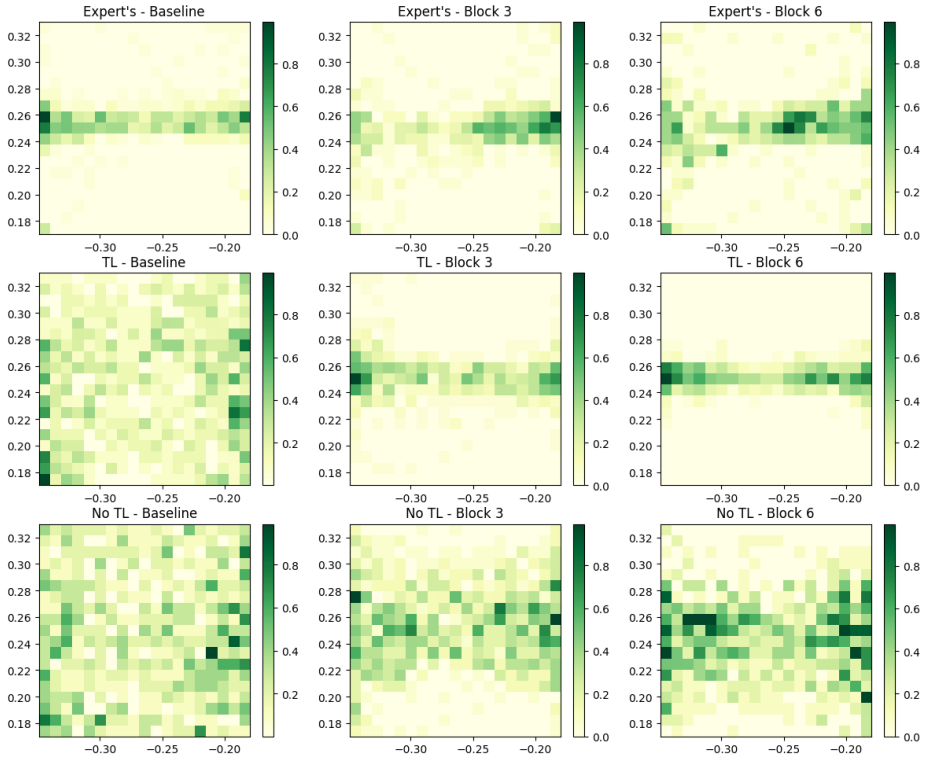


(γ′) Κανονικοποιημένη διανυθείσα απόσταση στα μπλοκ δοκιμών

Σχήμα 1.10: Αποτελέσματα στα μπλοκ δοκιμών: Νίκες, Ανταμοιβές, και κανονικοποιημένη διανυθείσα απόσταση

Ένα κρίσιμο εργαλείο για την κατανόηση της συμπεριφοράς των ομάδων είναι τα θερμογράμματα που παρουσιάζονται στο 6.8. Κατά τη διάρκεια του βασικού (baseline) μπλοκ (εδώ και τα 2 γκρουπ αλληλεπιδρούν με έναν τυχαίο πράκτορα), μια ευρεία κατάληψη στα κελιά υποδηλώνει μια διερευνητική φάση όπου οι ομάδες εξοικειώνονται με το περιβάλλον και την εργασία.

Από παιρετέρω εξέταση των μπλοκ δοκιμών, καταγράφονται ενδείξεις μεταφοράς μάθησης στην ομάδα. Αυτό είναι εμφανές από ένα μοτίβο όπου μια γραμμή υψηλής συχνότητας εμφανίζεται σταθερά σε ένα ορισμένο σημείο στον άξονα (x), δείχνοντας ότι η ομάδα επισκέπτεται συχνά μια συγκεκριμένη διαδρομή. Αυτό το μοτίβο υποδηλώνει ότι η ομάδα αξιοποιεί αποτελεσματικά τις επιδείξεις των ειδικών, αν και υποδηλώνει επίσης μια πιθανή εξάρτηση από αυτή τη στρατηγική.



Σχήμα 1.11: Συμπεριφορά της ομάδας με τον ειδικό (1η σειρά), μια ομάδα με LfD(2η σειρά) και μια ομάδα χωρίς μεταφορά μάθησης (3η σειρά). Τα θερμοδιαγράμματα δείχνουν τη θέση της κουκκίδας λέιζερ στη βασικό μπλοκ, στο 3ο και στο τελευταίο μπλοκ δοκιμής. Οι αριθμοί υποδεικνύουν τη συχνότητα με την οποία η κουκκίδα καταλάμβανε κάθε κελί (1εκ. × 1εκ.) - δηλαδή ο κανονικοποιημένος αριθμός των ζευγών x, y που μετρήθηκαν μέσα στο κελί και στα δέκα παιχνίδια μιας παρτίδας.

Στατιστικά αποτελέσματα

Σε όλη την ανάλυση, στοχεύουμε να παρατηρήσουμε τις διαφορές των 2 ομάδων και να αξιολογήσουμε τον αντίκτυπο της μεθόδου μεταφοράς μάθησης στην απόδοση των συμμετεχόντων. Στο πρώτο μπλοκ, το οποίο χρησιμεύει ως σύγκριση μεταξύ των δύο ομάδων, δεν εντοπίστηκαν στατιστικά σημαντικές διαφορές, είτε στις ανταμοιβές, είτε στις νίκες μεταξύ των ομάδων. Ωστόσο, στο τελευταίο μπλοκ, παρατηρήσαμε μια στατιστικά σημαντική διαφορά στις ανταμοιβές μεταξύ των δύο ομάδων, αποδεικνύοντας ότι η μεταφορά μάθησης είχε αντίκτυπο στις ανταμοιβές. Αυτά τα αποτελέσματα, συνοψίζονται στον πίνακα 1.4 που παρουσιάζει τα αποτελέσματα της δοκιμής Mann-Whitney Utόσο για το πρώτο όσο και για το τελευταίο μπλοκ.

Μπλοκ	**Μεταβλητή**	**Mann-Whitney U**	**Τιμή P**
Πρώτο (Baseline)	Ανταμοιβές	45,5	0,165
Πρώτο (Baseline)	Νίκες	47.0	0.119
Τελευταίο	Ανταμοιβές	52.0	0.038
Τελευταίο	Νίκες	51,5	0,043

Πίνακας 1.4: Αποτελέσματα στατιστικών δοκιμών για το πρώτο και το τελευταίο μπλοκ

Επιπλέον, έγινε χρήση της two-way mixed ANOVA μέσω της συνάρτησης `bwtrim` από το πακέτο `WRS2` [40] για να αξιολογήσουμε την αποτελεσματικότητα της μεταφοράς μάθησης σε διαφορετικά μπλοκ κατά τη διάρκεια του πειράματος. Αυτή η ανάλυση σχεδιάστηκε για να αξιολογήσει τόσο τις αλλαγές εντός της ομάδας (σε όλα τα μπλοκ-χρόνος) όσο και τις διαφορές μεταξύ των ομάδων. Τα αποτελέσματα παρουσιάζονται στον 6.4

Η ανάλυση αποκάλυψε:

Στατιστικά σημαντική διαφορά μεταξύ των ομάδων (p-value $< 0{,}0001$), που αποδεικνύει μια ουσιαστική διαφορά στην κανονικοποιημένη απόσταση μεταξύ των ομάδων. Αυτό το εύρημα υποστηρίζει την υπόθεση ότι η μεταφορά μάθησης επηρεάζει θετικά την απόδοση.

Στατιστικά σημαντική διαφορά μεταξύ των μπλοκ: (p-value $< 0{,}0001$), υποδηλώνοντας ότι η κανονικοποιημένη απόσταση αλλάζει σημαντικά σε διαφορετικά μπλοκ μέσα σε κάθε ομάδα. Αυτό το αποτέλεσμα είναι ενδεικτικό μιας διαδικασίας μάθησης ή προσαρμογής που συμβαίνει με την πάροδο του χρόνου.

Μη στατιστικα σημαντική επίδραση αλληλεπίδρασης: Η αλληλεπίδραση μεταξύ ομάδας και μπλοκ δεν ήταν στατιστικά σημαντική (p-value = 0,7168), καταδυκνύοντας ότι το μοτίβο της αλλαγής στα μπλοκ είναι αρκετά όμοιο μεταξύ των ομάδων. Και οι δύο ομάδες παρουσιάζουν παρόμοιες τροχιές βελτίωσης ή αλλαγής με την πάροδο του χρόνου.

Μεταβλητή	**Στατιστικά**	**τιμή p**
μπλοκ	8.0571	0.0010
ομάδα	78,0173	$< 0{,}001$
μπλοκ:ομάδα	0,5774	0,7168

Πίνακας 1.5: Two-Way Mixed ANOVA στην κανονικοποιημένη διανυθείσα απόσταση

Ενώ υπάρχει στατιστική σημαντικότητα μεταξύ των ομάδων, η έλλειψη στατιστικής σημαντικότητας μεταξύ μπλοκ και ομάδων υποδηλώνει ότι η τροχιά βελτίωσης, ενώ υπάρχει, δεν διαφέρει μεταξύ των ομάδων κατά τη διάρκεια του πειράματος.

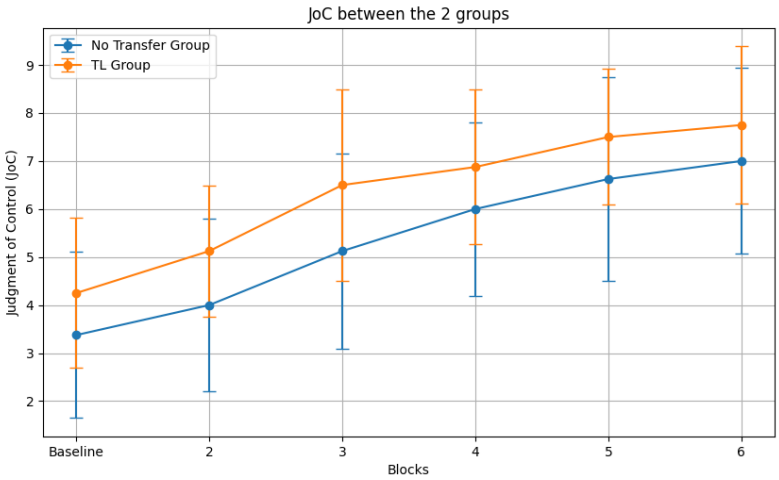
1.4.2 Υποκειμενικά αποτελέσματα

Κρίση ελέγχου

Στην ανάλυση του ελέγχου του ρομποτικού βραχίονα, οι συμμετέχοντες ρωτήθηκαν για την ικανότητά τους να κρίνουν τον έλεγχό τους, σε έξι διαφορετικά χρονικά διαστήματα (μετά το μπλοκ βάσης και σε κάθε εκπαίδευση εκτός σύνδεσης). Το ερώτημα ήταν:

Πώς θα βαθμολογούσατε την ικανότητά σας να ελέγχετε την κίνηση των χεριών στα τελευταία 10 παιχνίδια από 1 (χωρίς έλεγχο) έως 9 (πλήρης έλεγχος)·΄

Η γραφική παράσταση στο Σχήμα 1.12 καταγράφει οπτικά βελτιώσεις στην κρίση ελέγχου (JOC) με την πάροδο του χρόνου και για τις 2 ομάδες. Τα αποτελέσματα δείχνουν στατιστική σημασία μόνο στη βελτίωση των βαθμολογιών με την πάροδο του χρόνου (p=0,0001), όχι μεταξύ των ομάδων (p=0,246) ή στην αλληλεπίδραση μεταξύ ομάδας και χρόνου (p=0,53).

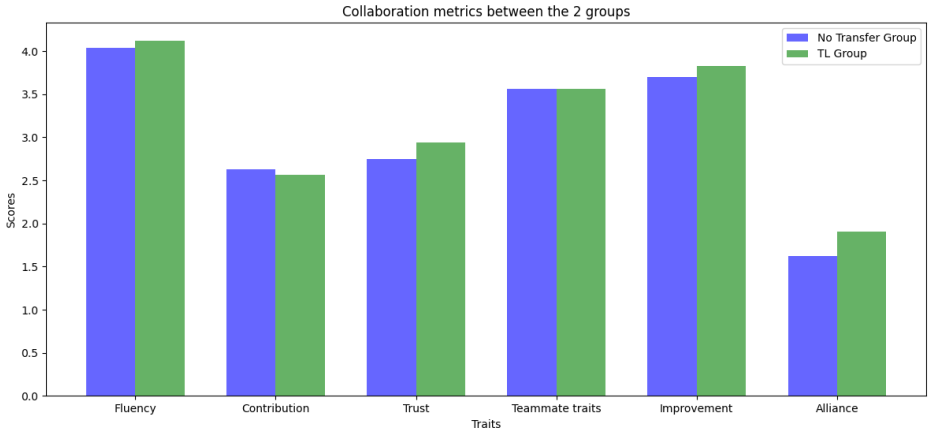


Σχήμα 1.12: Κρίση ελέγχου

Μετρήσεις συνεργασίας

Οι μετρήσεις συνεργασίας που παρουσιάζονται στο Σχήμα 1.13 χρησιμοποιούν μια κλίμακα Likert για να αξιολογήσουν την απόδοση των δύο ομάδων σε διάφορες διαστάσεις της ομαδικής εργασίας, Ερωτηματολόγιο 2 όπως παρουσιάζεται στο .2.6.

Με βαθμολογίες που ξεπερνούν το 3 στην ευχέρεια, τα χαρακτηριστικά συμπαίκτη και τη βελτίωση, και οι δύο ομάδες φαίνεται να έχουν ευνοϊκά αποτελέσματα, υποδηλώνοντας καλή ποιότητα συνεργασίας σε αυτούς τους τομείς. Οι μετρήσεις συνεισφοράς και εμπιστοσύνης κυμαίνονται γύρω από το μέσο της κλίμακας, αντανακλώντας μια ουδέτερη ή μέτρια στάση, όπου οι συμμετέχοντες ούτε συμφώνησαν ούτε διαφώνησαν έντονα με τις δηλώσεις που σχετίζονται με τα επίπεδα δέσμευσής τους ή την εμπιστοσύνη τους στους συμπαίκτες τους. Από αυτό συνεπάγεται ότι ενώ η συνεργασία ήταν λειτουργική, υπάρχει περιθώριο βελτίωσης στον τρόπο με τον οποίο αισθάνονται τα άτομα για τη συμβολή του πράκτορα και την αξιοπιστία του.



Σχήμα 1.13: Μετρήσεις συνεργασίας

1.4.3 Σύγκριση με προηγούμενη εργασία

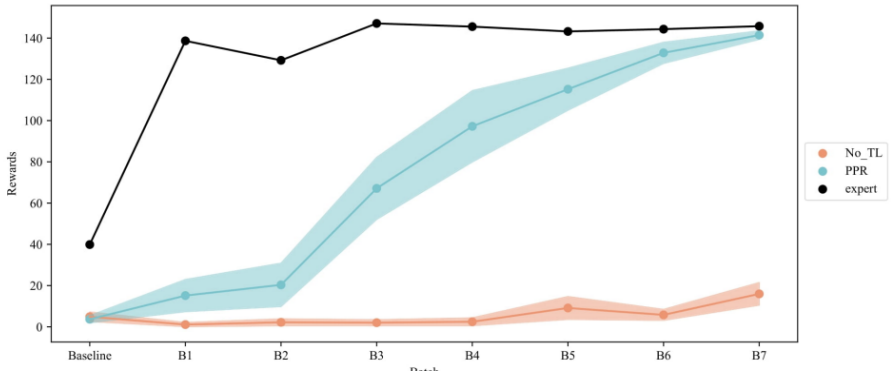
Αυτή η ενότητα παρουσιάζει μια συγκριτική ανάλυση της έρευνάς μας, η οποία χρησιμοποιεί μια προσέγγιση εκμάθησης μεταφοράς μάθησης από επιδείξεις, με τη μελέτη [13], στην οποία χρησιμοποίηθηκε μια μέθοδος επαναχρησιμοποίησης πολιτικής, στην ίδια διαδικασία.

Υπάρχουν 4 βασικές διαφορές μεταξύ των 2:

- **Παράμετρος στόχου εντροπίας:** Ορίσαμε την παράμετρο εντροπίας στόχου στο $0,98 \times (-\log(\frac{1}{|A|}))$, ενώ το [13] εφάρμοσε μια μικρότερη εντροπία στόχο $0,36 \times (-\log(\frac{1}{|A|}))$ 5.1.2.
- **Συνάρτηση ενεργοποίησης των νευρωνικών δικτύων του SAC :** Στην προσέγγισή μας, η συνάρτηση ενεργοποίησης ήταν Tanh, ενώ το [13] χρησιμοποιούσε ReLU 5.1.2.
- **Αριθμός παιχνιδιών:** Στο [13] ο συνολικός αριθμός παιχνιδιών ήταν 150 ενώ εμείς είχαμε 110 5.1.1.
- **Αρχικοποίηση του πράκτορα:** Στην προσέγγισή μας, οι συμμετέχοντες σε κάθε ομάδα ξεκινούν το πρώτο τους εκπαιδευτικό μπλοκ με έναν αρχικοποιημένο πράκτορα, ενώ στην εργασία [13], οι συμμετέχοντες αλληλεπιδρούσαν με έναν τυχαίο πράκτορα στην πρώτη μπλοκ εκπαίδευσης 5.2.1.

Από το σχήμα 1.14, μπορούμε να συγκρίνουμε τις ανταμοιβές σε μπλοκ δοκιμής για τη μέθοδο PPR με τις δικές μας στο 1.10β′. Αυτή η σύγκριση αποκαλύπτει τρεις βασικές διαφορές:

1. **Απόδοση ειδικών:** Η μελέτη μας δείχνει ότι οι ανταμοιβές του ειδικού διατηρούν μια μέση τιμή περίπου 70. Αντίθετα, το [13] δείχνει ότι ο ειδικός ξεκινά με χαμηλότερες βαθμολογίες αλλά φτάνει γρήγορα τη βαθμολογία του 140 από το δεύτερο μπλοκ και συγκλίνει εκεί.
2. **Απόδοση ομάδας μεταφοράς μάθησης:** Στο [13], η ομάδα εμφανίζει μια σταδιακή καμπύλη εκμάθησης, φτάνοντας την απόδοση σε επίπεδο ειδικού μέχρι το έβδομο μπλοκ. Η μέθοδός μας αποκαλύπτει μια πιο γρήγορη αρχική βελτίωση αλλά σύγκλιση σε χαμηλότερο επίπεδο ανταμοιβής, γύρω στο 70.
3. **Απόδοση ομάδας χωρίς μεταφορά μάθησης:** Το [13] αναφέρει ελάχιστη βελτίωση στην ομάδα, με μέγιστη ανταμοιβή 10 στο τελευταίο μπλοκ. Στη μελέτη μας, η ομάδα ξεκινά με μια μέση ανταμοιβή 10 και επιδεικνύει μάθηση, φτάνοντας τη μέση ανταμοιβή 40 στο τελικό μπλοκ.



Σχήμα 1.14: Καμπύλες ανταμοιβών στην μέθοδο [13]

Αρχικά, θεωρήσαμε ότι αυτές οι διαφορές θα μπορούσαν εν μέρει να αποδοθούν στους διαφορετικούς στόχους της εντροπίας μεταξύ των δύο μελετών. Για να το αξιολογήσουμε αυτό, πραγματοποιήσαμε πειράματα με διαφορετικές εντροπίες-στόχους κατά τη διάρκεια της συνεργασίας εμπειρογνωμόνων-πρακτόρων, μερικά από τα οποία παρουσιάζονται στο Παράρτημα .1.

Συνοπτικά, τα αποτελέσματα αποκάλυψαν ασταθή μάθηση κάτω από διαφορετικές ρυθμίσεις εντροπίας κατά τη διάρκεια αλληλεπιδράσεων με τον ειδικό. Ο πράκτορας αντιμετώπισε προκλήσεις στο να μάθει με συνέπεια μια αποτελεσματική πολιτική. Μια ρύθμιση υψηλής εντροπίας $0,98 \times (-\log(\frac{1}{|A|}))$, που ενθάρρυνε την εξερεύνηση, κατέστησε πιο δύσκολο τον εντοπισμό του υποκείμενου προβλήματος. Με βάση αυτά τα ευρήματα, προκύπτει ότι η αλλαγή στη συνάρτηση ενεργοποίησης Tanh μπορεί να οδήγησε δυνητικά σε ένα πρόβλημα εξαφανιζόμενης παραγώγου [41]. Ως αποτέλεσμα, δεν συνιστούμε τη χρήση του SAC με ενεργοποίηση Tanh , ειδικά σε αυτήν την μέθοδο.

1.5 Αποτελέσματα Φάσης 2

Πραγματοποιήθηκαν πρόσθετα πειράματα με τη συνάρτηση ενεργοποίησης στην αρχική της ρύθμιση (από Tanh σε Relu) και εντροπία στόχο $0,36 \times (-\log(\frac{1}{|A|}))$. Επιπλέον, προσθέσαμε 3 ακόμη batches για να είναι σε πλήρη συνάφεια με το [13]. Θα αναφερθούμε σε αυτά τα πειράματα ως φάση 2. Αποφασίσαμε να μην συμπεριλάβουμε τα ερωτηματολόγια Big 5 και το PVQ στα πειράματα της φάσης 2.

Αυτά τα πειράματα περιλάμβαναν 8 συμμετέχοντες με τα χαρακτηριστικά τους που παρουσιάζονται στο 1.6:

Πίνακας 1.6: Χαρακτηριστικά των συμμετεχόντων (φάση 2)

Χαρακτηριστικό	**Λεπτομέρειες**
Κατανομή φύλου	4 γυναίκες, 4 άνδρες
Ηλικία	22-24 ετών
Επικρατές χέρι	8 δεξιόχειρες
Εμπειρία παιχνιδιού	4 με ˜5 χρόνια, 1 με 3-5 χρόνια, 1 με ˚1 έτος, 2 με κανένα
Προτιμώμενες συσκευές παιχνιδιού	2 φορητοί υπολογιστές, 2 κονσόλες, 2 κινητά, 2 κανένα

Άποψη για την Τεχνητή Νοημοσύνη

Και οι δύο ομάδες, έχουν θετική άποψη για την Τεχνητή νοημοσύνη, όπως φαίνεται στον Πίνακα 1.7. Η στάση απέναντι στην τεχνητή νοημοσύνη έδειξε παρόμοια αποτελέσματα με την ομάδα μεταφοράς μάθησης να υπερισχύει, αλλά χωρίς σημαντικές διαφορές.

Πίνακας 1.7: Άποψη για την Τεχνητή Νοημοσύνη (φάση 2)

Ομάδα	Μέσο	Τυπική απόκλιση
Χωρίς μεταφορά μάθησης	0.448	0.621
Με μεταφορά μάθησης	0.666	0.660

1.5.1 Αντικειμενικά αποτελέσματα

Στο μέσο χρόνο αλληλεπίδρασης των δοκιμαστικών παιχνιδιών, οι δύο ομάδες είχαν μεγάλη διαφορά σε σχέση με την προηγούμενη. Συγκεκριμένα, η ομάδα με μεταφορά μάθησης εμφάνισε μέσο χρόνο αλληλεπίδρασης 11,45 λεπτών με τον παράγοντα, με τυπική απόκλιση 0,76 λεπτά. Αντίθετα, η ομάδα χωρίς είχε μέσο χρόνο αλληλεπίδρασης 29,7 λεπτά, με τυπική απόκλιση 6,35 λεπτά.

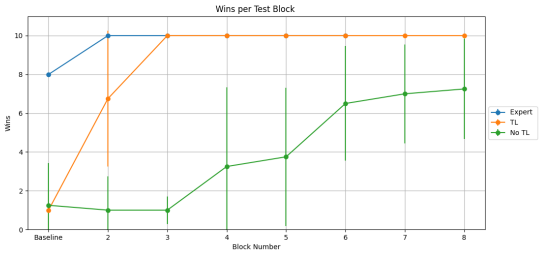
Μια οπτική αναπαράσταση της μαθησιακής προόδου στα 80 παιχνίδια δοκιμών των πειραμάτων φάσης 2, παρουσιάζεται στις Εικόνες 1.15α′ και 1.15β′, οι οποίες απεικονίζουν τις καμπύλες μάθησης για νίκες και ανταμοιβές, για τις δύο ομάδες, καθώς και την απόδοση του εμπειρογνώμονα. Επιπλέον, στο Σχήμα 1.15γ′, παρουσιάζουμε την κανονικοποιημένη διανυθείσα απόσταση.

Εστιάζοντας στην καμπύλη εκμάθησης ανταμοιβών 1.15β′ μπορούμε να παρατηρήσουμε τα παρακάτω:

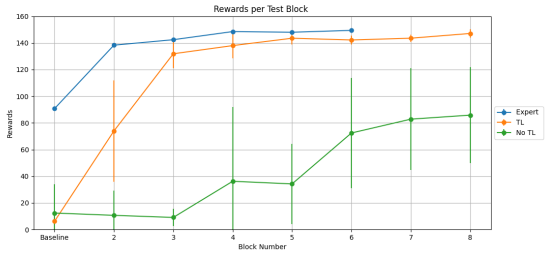
Δεδομένου ότι ο ειδικός πέτυχε τη μέγιστη ανταμοιβή από το δεύτερο μπλοκ δοκιμής επιλέξαμε να διατηρήσουμε τη συνέπεια με τα προηγούμενα πειράματα ζητώντας από τον ειδικό να συμμετέχει για τον ίδιο αριθμό μπλοκ. Επιπλέον, θα διατηρήσουμε τις ίδιες αναλογίες δεδομένων επίδειξης από τις αλληλεπιδράσεις τους, αντικατοπτρίζοντας τη μεθοδολογία της προηγούμενης μελέτης.

Για την ομάδα με μεταφορά μάθησης, τα διαγράμματα δείχνουν ότι παρόλο που υπάρχει διακύμανση στο δεύτερο μπλοκ δοκιμής, η ομάδα επιτυγχάνει σταθερά μια ανταμοιβή 140 από το τρίτο μπλοκ δοκιμής και μετά. Αυτή η απόδοση έρχεται σε αντίθεση με την προηγούμενη μελέτη, μας, στην οποία η μέγιστη ανταμοιβή που επιτεύχθηκε ήταν 60 από το 6ο μπλοκ. Επιπλέον, αυτό το αποτέλεσμα διαφέρει από το [13], όπου η ίδια ομάδα που χρησιμοποιεί PPR TL , έφτασε σε απόδοση συγκρίσιμη με αυτή ενός ειδικού στο τελευταίο (6ο) μπλοκ.

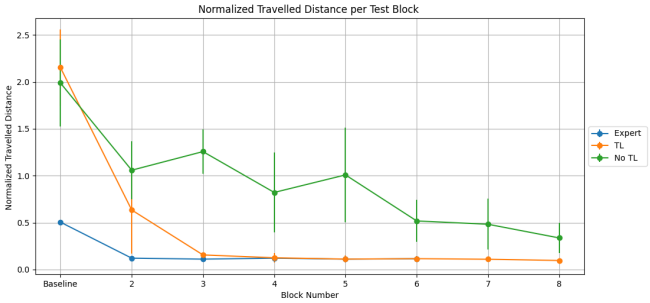
Η ομάδα χωρίς μεταφορά μάθησης πέτυχε ανταμοιβή 80 στο 6ο μπλοκ, που είναι σημαντική βελτίωση από τα 40 που είχαν επιτευχθεί υπό προηγούμενες συνθήκες στο ίδιο στάδιο. Γίνεται μια αξιοσημείωτη σύγκριση με την απόδοση της ομάδας κατά τη φάση 2, η οποία έφτασε σε ανταμοιβή 80 στο 6ο μπλοκ, σε αντίθεση με τα ευρήματα από το [13], όπου η μέση ανταμοιβή σε αυτό το στάδιο ήταν μόνο 20, καθώς φαίνεται στο Σχ. 1.14. Αυτή η βελτίωση μπορεί πιθανότατα να συνδεθεί με τη χρήση ενός αρχικοποιημένου πράκτορα 5.2.1 κατά τη διάρκεια του πρώτου μπλοκ εκπαίδευσης, σημειώνοντας τη μόνη διαφορά μεταξύ των δύο προσεγγίσεων των ομάδων χωρίς μεταφορά μάθησης.



(α′) Νίκες στα μπλοκ δοκιμών (φάση 2)



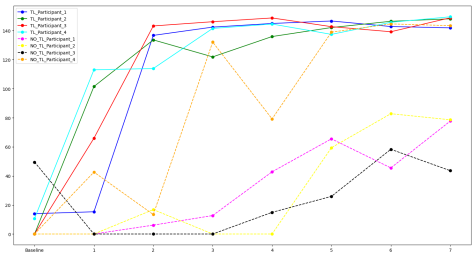
(β′) Ανταμοιβές στα μπλοκ δοκιμών (φάση 2)



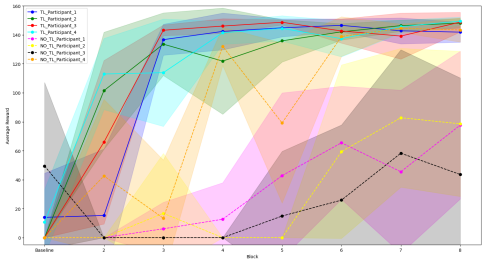
(γ′) Κανονικοποιημένη διανυθείσα απόσταση στα μπλοκ δοκιμών (φάση 2)

Σχήμα 1.15: Νίκες, ανταμοιβές και κανονικοποιημένη διανυθείσα απόσταση (φάση 2)

Στις Εικόνες 1.16α′ και 1.16β′, η ατομική απόδοση κάθε συμμετέχοντα αναπαρίσταται οπτικά. Μια ενδιαφέρουσα παρατήρηση είναι ότι ο συμμετέχων No TL 4 φαίνεται να συγκλίνει προς ένα μοτίβο ανταμοιβής παρόμοιο με αυτό της ομάδας μεταφοράς μάθησης, φτάνοντας σε επίπεδο ανταμοιβής 140. Αυτή η σύγκλιση στην απόδοση θυμίζει παρόμοια συμπεριφορά που παρατηρήθηκε σε έναν συμμετέχοντα στο [13], ο οποίος τελικά αποκλείστηκε από τη μελέτη.



(α′) Ανταμοιβές για όλους τους συμμετέχοντες (χωρίς τυπική απόκλιση)

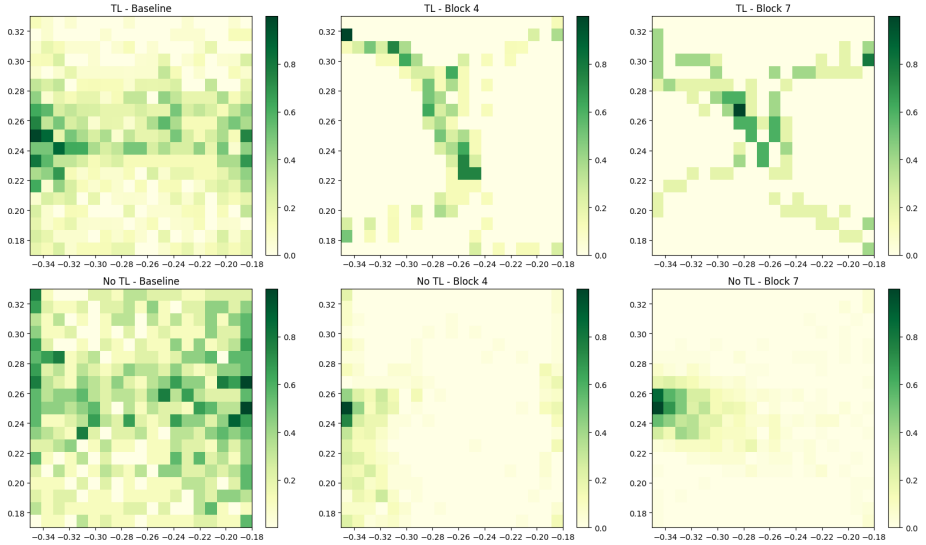


(β′) Ανταμοιβές για όλους τους συμμετέχοντες (με τυπική απόκλιση)

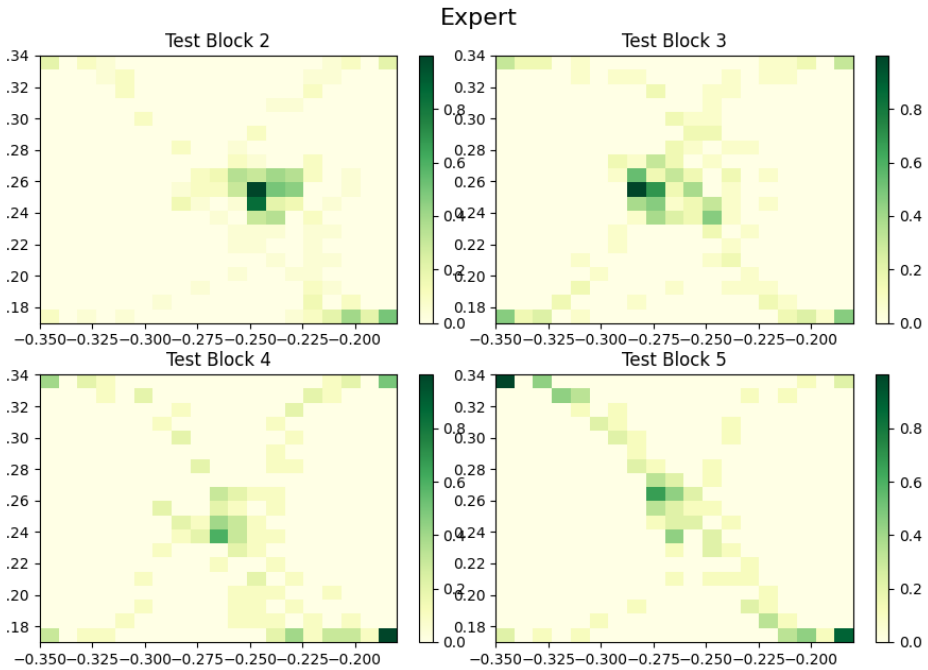
Σχήμα 1.16: Συνδυασμένα στοιχεία για ανταμοιβές φάσης 2

Τα θερμογράμματα για τη φάση 2, που απεικονίζονται στο Σχήμα 1.17, υποδεικνύοουν τα αποτελέσματα της μάθησης μεταφοράς εντός της ομάδας με μεταφορά μάθησης. Συγκεκριμένα, ένα μοτίβο σε σχήμα «X» εμφανίζεται στο τελικό μπλοκ, το οποίο όχι μόνο ευθυγραμμίζεται στενά με τα μοτίβα που παρατηρούνται στα θερμογράμματα του ειδικού

(Εικόνα 1.18), αλλά αντιπροσωπεύει επίσης τη βέλτιστη στρατηγική αλληλεπίδρασης που ελαχιστοποιεί τον χρόνο ολοκλήρωσης.



Σχήμα 1.17: Συμπεριφορά της ομάδας με τον ειδικό (1η σειρά), μια ομάδα με LfD(2η σειρά) και μια ομάδα χωρίς μεταφορά μάθησης (3η σειρά). Τα θερμοδιαγράμματα δείχνουν τη θέση της κουκκίδας λέιζερ στη βασικό μπλοκ, στο 3ο και στο τελευταίο μπλοκ δοκιμής. Οι αριθμοί υποδεικνύουν τη συχνότητα με την οποία η κουκκίδα καταλάμβανε κάθε κελί (1 εκ. $\times$ 1 εκ.) - δηλαδή ο κανονικοποιημένος αριθμός των ζευγών x, y που μετρήθηκαν μέσα στο κελί και στα δέκα παιχνίδια μιας παρτίδας.



Σχήμα 1.18: Μπλοκ δοκιμών του ειδικού (φάση 2)

(H)

Εφέ	Στατιστικά δοκιμών	p-value
μπλοκ	24.042	< 0.001
ομάδα	26.135	< 0.001
μπλοκ:ομαδα	3.539	0.00382

Πίνακας 1.8: Αποτελέσματα μικτής ANOVA δύο κατευθύνσεων σε κανονικοποιημένη διανυθείσα απόσταση

Στατιστικά αποτελέσματα

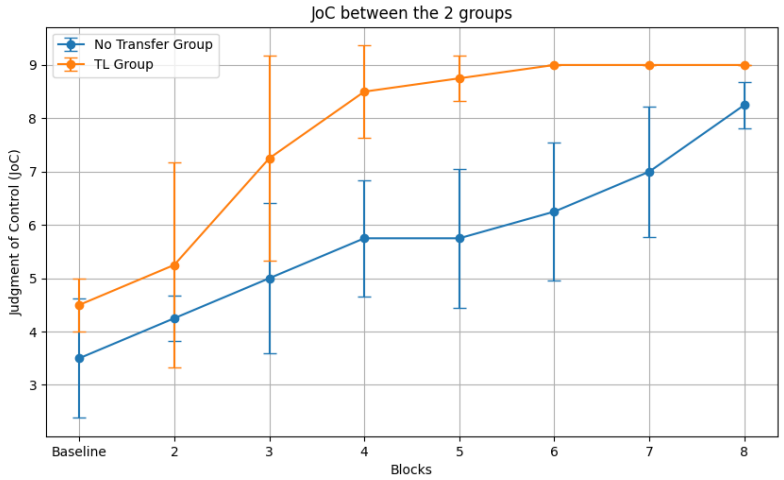
Η ίδια στατιστική ανάλυση με την πρώτη φάση του πειράματος χρησιμοποιήθηκε και στο σημείο αυτό. Τα αποτελέσματα παρουσιάζονται στον 1.8

Τα στατιστικά σημαντικά αποτελέσματα που παρατηρήθηκαν για όλες τις μεταβλητές μας υπογραμμίζουν την αποτελεσματικότητα της μεθόδου LfD , ειδικά με τις νέες ρυθμίσεις που ενσωματώνουν τις προσαρμογές του SAC (συνάρτηση ενεργοποίησης ReLU σε συνδυασμό με $0,36 \times (-\log(\frac{1}{|A|}))$ ρύθμιση εντροπίας).

1.5.2 Υποκειμενικά αποτελέσματα

Κρίση ελέγχου

Η γραφική παράσταση στο Σχήμα 1.19 υποδηλώνει βελτιώσεις στην κρίση ελέγχου. Υπάρχει στατιστική σημαντικότητα με την πάροδο του χρόνου (p-value $< 0,001$), μεταξύ των ομάδων (p-value $< 0,001$) αλλά όχι στην αλληλεπίδραση μεταξύ ομάδας και χρόνου (p-value=0.53).

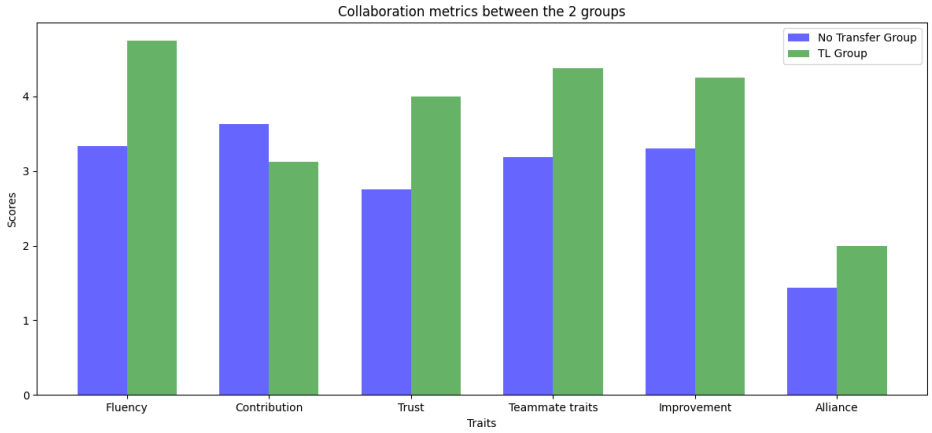


Σχήμα 1.19: Κρίση ελέγχου

Μετρήσεις συνεργασίας

Στη Φάση 2 της μελέτης, παρατηρήσαμε ότι η ομάδα με μεταφορά μάθησης πέτυχε σταθερά υψηλότερες βαθμολογίεςσυγκριτικά με την ομάδα χωρίς μεταφορά μάθησης σε πολλαπλές

μετρήσεις συνεργασίας ;;. Συγκεκριμένα, η ομάδα μεταφοράς μάθησης ξεπέρασε την άλλη ομάδα στους δείκτες: ευφράδεια, σημάδια συμπαίκτη, βελτίωση και εμπιστοσύνη.



Σχήμα 1.20: Μετρήσεις συνεργασίας (φάση 2)

1.6 Συμπεράσματα

Η παρούσα διπλωματική επεκτείνει το έργο [13] εξερευνώντας μια διαφορετική προσέγγιση μεταφοράς μάθησης, συγκεκριμένα τη βαθιά Q μάθηση από επιδείξεις. Η μελέτη μας διαφοροποιείται από αυτή τη μέθοδο με την προηγούμενη πιθανολογική επαναχρησιμοποίηση πολιτικής και εξετάζει τις επιπτώσεις των αλγοριθμικών τροποποιήσεων στην αποτελεσματικότητα της συνεργασίας. Παρά τις αρχικές προκλήσεις με τις προσαρμογές παραμέτρων του αλγόριθμου Soft Actor-Critic (SAC), η επαναφορά στις αρχικές ρυθμίσεις κατά την εφαρμογή της νέας μεθόδου μεταφοράς μάθησης έδειξε βελτιώσεις στη συνεργατική απόδοση, όπως αποδεικνύεται από τον συνολικό χρόνο αλληλεπίδρασης στα παιχνίδια δοκιμών και από την πρώτερη σύγκλιση της ομάδας μεταφοράς μάθησης.

Αναγνωρίζοντας τους περιορισμούς της μελέτης, όπως η μικρή ομάδα συμμετεχόντων και η ιδιαιτερότητα της εργασίας, η οποία δεν ενσωματώνει τις πολυπλοκότητες του πραγματικού κόσμου, οι μελλοντικές ερευνητικές κατευθύνσεις πρέπει να περιλαμβάνουν περαιτέρω διερεύνηση της μεθοδολογίας μεταφοράς μάθησης με υψηλότερη τιμή της παραμέτρου εντροπίας στόχου που πιθανώς να συμβάλλει στην καλύτερη εξατομίκευση του αλγορίθμου σε κάθε συμμετέχοντα. Επίσης μία άλλη κατεύθυνση προς την εξατομίκευση είναι η διερεύνηση του βέλτιστου ποσοστού μεταφοράς γνώσης από τον ειδικό.

Chapter 2

Introduction

Human-Robot Collaboration (HRC) can be considered a sub-field of Human Robot Interaction (HRI)[1]. HRC explores how humans and robots can work together to accomplish mutual objectives. This field is inherently multidisciplinary, drawing from areas such as robotics, artificial intelligence, human-computer interaction, sociology, psychology, and more, to develop systems where humans and robots can complement each other's capabilities. HRC finds practical applications in everyday scenarios including educational support [2], therapeutic interventions [3], and companionship [4], as well as in industrial settings [5].

The challenges of collaborating with robots get bigger when the tasks become more complicated. Robots need to be intellogent enough to make autonomous decisions, especially when working closely with people, like when lifting and moving big items together [6]. It is important for robots to not only move around safely but also to understand what people are trying to do so they can help in the right way [7]. To make these interactions smooth and natural, robots should be built to notice and react to human behaviors, learn from working together, and adjust their actions.

Despite the progress in robotics, creating smooth and fully independent teamwork between humans and robots is still challenging. Learning together, a crucial aspect of collaboration, can be slow and depends on various factors such as the physical and mental effort required, the human partner's skills, the machine learning techniques used, and the task's computational demands. Robots, or cobots, are expected to learn quickly, recognize their human partners' abilities, and adjust to their strengths and weaknesses.

Recent developments in deep reinforcement learning (DRL) have opened new ways for examining HRC in real-time across various applications, from mobile robots [8] and robotic arms [9] to drones [10] and more. DRL's success in these areas comes from its ability to learn complex motions and behaviors that are hard to achieve with traditional control methods.

However, a limitation in DRL for robotics is the challenge of generalizing learned knowledge to new, unfamiliar situations or when working with new partners [11]. The standard practice of training a robot from scratch for each new task is time-consuming and inefficient, affecting the team's productivity and leading to fatigue.

Addressing this issue, transferring knowledge within DRL frameworks [12] presents a solution, offering several strategies for enhancing learning efficiency. Such efforts are directed not only towards enhancing the independence of robots but also towards

facilitating their seamless integration into human teams. The ultimate goal is to foster a more intuitive and effective partnership between humans and robots, thereby elevating the overall productivity and harmony of collaborative endeavors.

Building upon the work in [13], this thesis aims to contribute to the field of HRC by implementing a Transfer Learning approach, Learning from Demonstrations (LfD), within the Soft Actor-Critic (SAC) DRL algorithm. The study involves human participants in an HRC experiment, designed to assess the effectiveness of DRL and transfer learning in enhancing the efficiency of HRC.

The key contributions of this research include transitioning the Transfer Learning method from policy reuse to learning from demonstrations, and implementing it within the Robot Operating System (ROS). Furthermore, this thesis will compare the two TL methods based on the analysis of participant outcomes from the HRC experiment. We will also delve into the differences between these TL methods, assessing their respective impacts on the collaborative process. Additionally, the study evaluates the effect of different target entropies on the SAC algorithm in the context of TL, providing insights into how these parameters influence the dynamics of HRC. This analysis aims at enhancing HRC by optimizing the learning and adaptation process of robots to improve their collaboration with humans.

The structure of this thesis is organized as follows: Chapter 3, we briefly review the fundamentals of Machine Learning (ML) and Reinforcement Learning (RL), with a particular emphasis on DRL and the SAC algorithm. We also introduce the concept of TL, and the main and its significance in enhancing learning efficiency. Chapter 4 explores existing research in HRI, presenting applications of DRL in robotics for HRC scenarios, discussing the limitations of current methodologies, and highlighting how TL offers promising solutions for overcoming these challenges. In Chapter 5, the approach and methods employed in this research are outlined, including the formulation of the DRL agent and the implementation of TL. This Chapter also discusses the objective and subjective measures that are used to evaluate the effectiveness of TL. Chapter 6 presents the experimental outcomes, showcasing the impact of the chosen TL method on improving HRC. This section highlights how adjustments to the SAC parameters, specifically target entropy and the activation function, modulate the effectiveness of the collaborative tasks and also includes a comparison with previous work [13] that employed a different TL method. Additionally, it discusses subsequent experiments that further explore modifications to the SAC settings, offering a comprehensive analysis of how these changes influence the dynamics of human-robot collaboration.

Chapter 3

Background

In this chapter, the discussion begins with a brief review of the primary categories of machine learning: supervised, unsupervised, and semi-supervised learning. Following this review, there is an exploration into the concept of Artificial Neural Networks, providing a crucial foundation for understanding more sophisticated algorithms. The focus then shifts to reinforcement learning, characterized by a learning process through feedback and adjustments and expands to deep reinforcement learning, renowned for its ability to handle more complex tasks. The Soft Actor-Critic algorithm, a key technique in deep reinforcement learning, is also introduced. Furthermore, the chapter introduces the fundamentals of Transfer Learning, emphasizing its role in enhancing learning processes by utilizing knowledge from previously encountered similar tasks. A special focus is placed on Learning from Demonstrations, an approach that effectively demonstrates the implementation of transfer learning strategies.

3.1 Machine learning

Machine learning (ML) is a field where computers learn to solve problems without being explicitly told how. It is something humans interact with on a daily basis, often without realizing it, influencing choices from the products we browse online to the movies we watch. Based on the analysis in [42], there are several ML categories and algorithms, as shown in Figure 3.1.

There are four primary learning approaches: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning The selection of the learning method and algorithm depends on the type of the problem and the availability of the data.

- **Supervised Learning:** In supervised learning, a dataset with pairs of inputs and their corresponding outputs, is collected. The goal is to use an optimization algorithm to build a function that not only fits this data well, but also predicts accurately on new data it has not seen before. Supervised learning is often used for two main tasks: regression, where continuous values are predicted, and classification, where the data is categorized into different groups. The process typically involves minimizing a loss function, which measures "how far off" the model's predictions are from the actual results in the dataset.

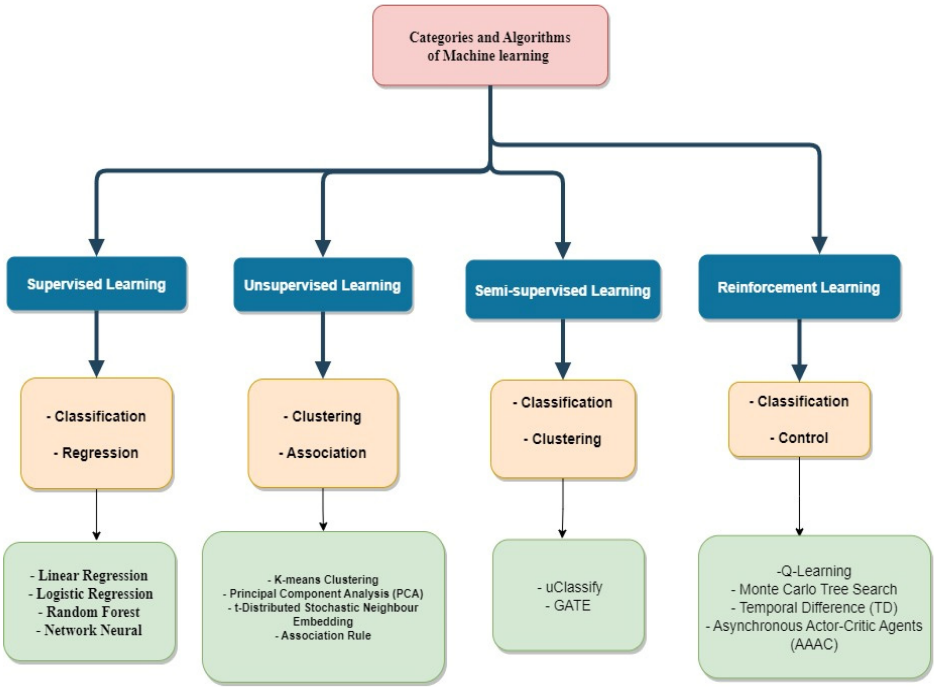


Figure 3.1: Classification of the most common machine learning algorithms [43]

- **Unsupervised Learning:** In unsupervised learning, the dataset does not have labeled input-output pairs. Instead, the emphasis is upon discovering hidden patterns and structures within it. This approach is useful in tasks like clustering, where similar data points are grouped together, dimensionality reduction, which simplifies data without losing important information, and density estimation, where the distribution of data points is determined. Unsupervised learning focuses more on uncovering insights and inherent features of the data.

- **Semi-Supervised Learning:** In semi-supervised learning, the dataset includes both labeled and unlabeled data. This method is useful when fully labeled data is scarce or costly. It combines elements of supervised and unsupervised learning, using the labeled portion to enhance learning and make predictions about the unlabeled part. It is effective for tasks like classification and regression when complete data labeling is impractical.

- **Reinforcement Learning (RL):** Reinforcement Learning is a type of machine learning where an agent learns to make decisions by performing actions in an environment to achieve a goal. The agent receives rewards based on its actions and learns to maximize cumulative rewards over time. RL is distinct for its focus on sequential decision-making and interaction with a dynamic environment. A more detailed exploration of Reinforcement Learning will be provided in Section 3.2.

3.1.1 Artificial Neural Networks

Traditional ML methods often rely on selecting specific features from data, which can be challenging or even impossible when dealing with large or complex datasets. That is

why in these occasions we prefer to use neural networks who are inspired by the human brain which consists of billions interconnected units (neurons) organized in layers. Just as neurons in the brain take inputs through our senses and transmit signals through dendrites and axons, artificial neurons in these networks receive and process data in a similar interconnected way. This structure enables neural networks to process large or complex datasets without the need for manual feature selection. Figure 3.2 illustrates the comparison between biological neurons and their artificial counterparts.

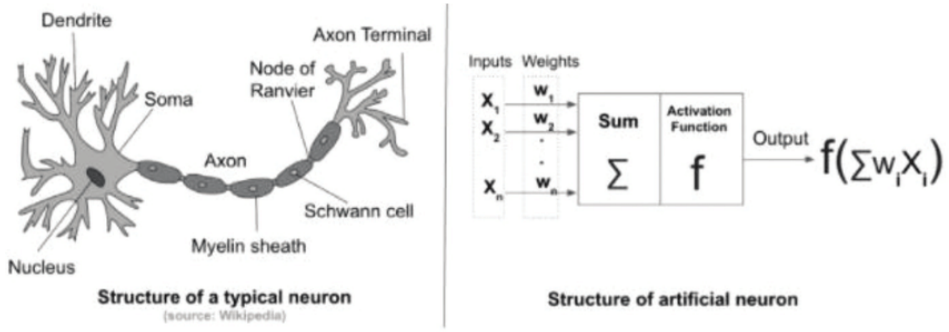


Figure 3.2: Biological neurons to Artificial neurons [44]

Each layer in a neural network consists of multiple neurons. These neurons receive input, process it, and pass the output to the next layer. The process begins with the input layer, where each neuron represents a feature of the input data. The input to each neuron is weighted, reflecting the importance or relevance of that feature with the context of the task. Additionally, a bias term is added to the input before it undergoes a transformation via an activation function. This function determines whether and to what extent the signal should be passed further along the network. The output of each neuron, after being processed by the activation function, is then passed on to the next layer. This process continues until the final layer, typically known as the output layer, which produces the final result of the neural network. The structure and depth of these layers can vary greatly, depending on the complexity of the task and the design of the network. Learning in neural networks occurs through a process known as backpropagation. This involves adjusting the weights of the connections between neurons based on the error of the network's output compared to the expected result [45]. The error is calculated using a loss function, which measures the difference between the network's prediction and the actual target values. During training, the network performs a forward pass to make predictions, and then a backward pass to propagate the error back through the network, updating the weights. This iterative process of forward and backward passes allows the network to learn from the data, gradually improving its predictions over time. The structure of a neural network is illustrated in Figure 3.3

- **Loss Functions:** Loss functions play a crucial role in neural network training. They quantify the error between predicted values ($\hat{y}$) and actual target values (y). A common one is Mean Squared Error (MSE) which measures the average squared difference between predicted and actual values, making it suitable for regression tasks.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{3.1}$$

- **Activation Functions:** Activation functions introduce non-linearity into the neural network, enabling it to capture complex patterns and relationships in the data. Common activation functions include:

  ☐ **Sigmoid:**

  The sigmoid function transforms input into a range between 0 and 1, suitable for binary classification problems.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

  ☐ **ReLU (Rectified Linear Unit):**

  ReLU is widely used due to its simplicity and effectiveness. It outputs x for positive inputs and 0 for negative inputs.

$$\text{ReLU}(x) = \max(0, x) \tag{3.3}$$

  ☐ **tanh (Hyperbolic Tangent):**

  Tanh is similar to the sigmoid but maps inputs to a range between -1 and 1, making it useful for centered data.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.4}$$

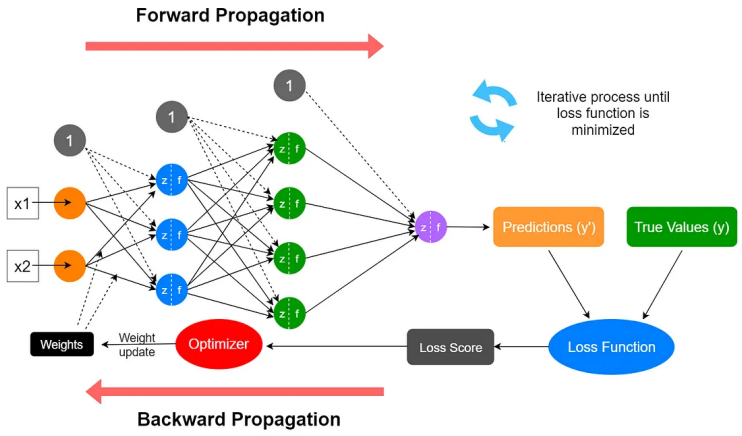


Figure 3.3: Artificial Neural Network [46]

There are several types of Artificial Neural Networks (ANNs) designed for specific tasks. For example **Feedforward Neural Networks (FNNs)** FNNs, also known as multi-layer perceptrons (MLPs), are the foundation of most neural networks. **Convolutional Neural Networks (CNNs)** are suitable for image-related tasks, such as

image classification and object detection. They use convolutional layers to automatically learn features from images. **Recurrent Neural Networks (RNNs)** are used for sequential data tasks like natural language processing and time series prediction. They have recurrent connections that allow them to maintain a hidden state and process sequences.

3.2 Reinforcement Learning

Reinforcement Learning (RL), often used to solve complex decision-making problems, involves an agent interacting with an environment over multiple time steps. RL differs in an important way from supervised and unsupervised learning. It does not use a dataset as a starting point. Instead, it generates data on-line or off-line as dictated by the needs of the optimization algorithm it uses.

In this section, we analyze the fundamental concepts of RL, drawing from the fundamental works of Dimitri P. Bertsekas [47], [48], and Richard S. Sutton and Andrew G. Barto [49].

3.2.1 RL Formulation

A basic tool of RL is the concept of Markov Decision Processes (MDPs). MDPs provide a mathematical framework to model decision-making scenarios where outcomes are influenced by both random factors and the decisions of the agent. This framework offers a structured approach for defining and formalizing an environment in RL. An MDP encompasses several key components:

- **States (S)**: The set of all possible states in the environment.
- **Actions (A)**: For each state s, $A(s)$ represents the set of possible actions.
- **Reward Function (r)**: A function $r : S \times A \to \mathbb{R}$, where $r(s, a)$ is the immediate reward received after taking action a in state s.
- **Transition Probability Function (P)**: A function $P : S \times A \times S \to [0, 1]$, where $P(s, a, s')$ gives the probability of transitioning to state s' after taking action a in state s.
- **Discount Factor (γ)**: A factor between 0 and 1 that discounts future rewards, reflecting the preference for immediate rewards over future rewards.

In the context of MDPs and Bellman equations, s' represents a generic subsequent state in theoretical discussions, while s_t and s_{t+1} denote the current and next states in a temporal sequence during agent-environment interactions.

In practice, at each timestep $t = 0, 1, 2, \ldots$, the agent interacts with its environment by observing the current state s_t, selecting an action a_t according to its policy π, and transitioning to the next state s_{t+1}, as illustrated in Figure 3.4. The policy π, is a function that maps states to probabilities of selecting each possible action. Every action taken by the agent results in a reward r_t and a transition to a new state s_{t+1}.

This continuous cycle of observation, action, and reaction forms the core of the learning and decision-making process in Reinforcement Learning.

The primary objective of the agent is to discover a policy π that maximizes the expected total discounted return. Under policy π, this return, denoted as $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, is the cumulative sum of rewards received over time, each discounted by the factor γ.

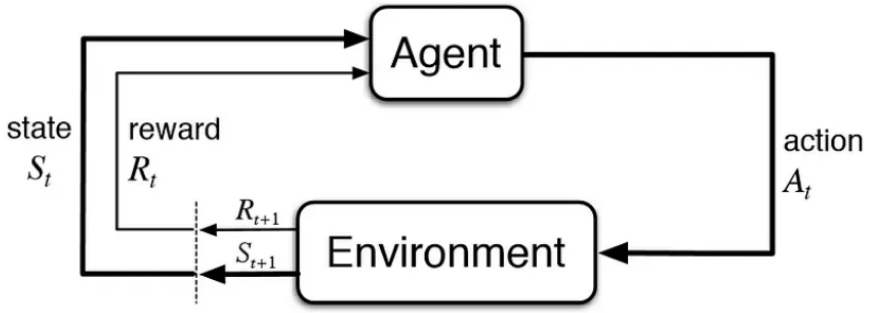


Figure 3.4: The agent–environment interaction in a MDP [50]

To achieve this goal, the agent relies on several key concepts:

- **State Value Function** $V^\pi(s)$: This function indicates the expected total discounted return when starting from state s and following policy π thereafter. It is given by:

$$V^\pi(s) = \mathbb{E}_\pi\{R_t \mid s_t = s\} = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right] \tag{3.5}$$

  It assesses the overall potential of being in a particular state under the policy.

- **Action-Value function (Q-value)** $Q^\pi(s, a)$: This function predicts the expected total discounted return from taking action a_t in state s_t, under policy π. The Q-value is crucial for determining the effectiveness of actions in specific states. It's defined as:

$$Q^\pi(s, a) = \mathbb{E}_\pi\{R_t \mid s_t = s, a_t = a\} = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right] \tag{3.6}$$

Bellman Equations: According to Bellman equations, the value function can be decomposed into two parts: the immediate reward (r) plus the discounted value function of the next state. Specifically, for a policy π, the Bellman equation for the value function is expressed as:

$$V^\pi(s) = \mathbb{E}_\pi\left[r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s')\right] \tag{3.7}$$

Similarly, the Bellman equation for action-value function $Q^\pi(s,a)$ under policy π is given by:

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) Q^\pi(s', \pi(s')) \right] \tag{3.8}$$

These Bellman equations play a crucial role in reinforcement learning by breaking down the problem into smaller sub-problems, making it possible to derive optimal policies. The ultimate goal is to find the **optimal policy** π^* and the associated **optimal value functions** $V^*(s)$ and $Q^*(s,a)$, which maximize the expected return. The optimal policy, π^*, selects actions that yield the highest expected cumulative reward, and the optimal value functions reflect the maximum achievable returns.

Exploration-Exploitation trade-off

In RL, the agent faces a fundamental dilemma known as the 'exploration-exploitation trade-off'. This dilemma arises from the need to balance two conflicting objectives: **exploration** and **exploitation**. Exploration involves the agent exploring its environment to gather information about the rewards associated with different actions and states. Without exploration, the agent might settle for suboptimal actions based on limited knowledge. Exploitation, on the other hand, involves choosing actions that are known to be effective based on the available information to maximize expected cumulative reward.

3.2.2 Reinforcement Learning Algorithms

In RL, there are mainly two different approaches for training an agent:

On-Policy Learning: In on-policy learning, the agent learns the value of the policy that it is currently using to make decisions. This includes learning from the exploration steps it takes. The policy being improved is the same policy used to make decisions. A typical example of an on-policy method is SARSA (State-Action-Reward-State-Action) [51].

Off-Policy Learning: Off-policy learning, in contrast, allows the agent to learn a policy different from the one it is executing. This approach enables the agent to learn from actions it has not taken, broadening its understanding of the environment. Q-learning [52] is a well-known off-policy method.

Both on-policy and off-policy learning have unique advantages and are suitable for different types of problems in Reinforcement Learning. Understanding these strategies is crucial for designing effective RL algorithms.

RL algorithms are also divided into **Model-Free** and **Model-Based** methods. Model-Free RL learns directly from interactions with the environment without explicitly modeling its dynamics, focusing on trial-and-error learning. In contrast, Model-Based RL uses or learns a model of the environment to plan actions, enabling more informed decision-making. Each approach offers distinct methodologies and is suitable for different types of problems. A summary of these approaches is presented in Figure 3.5.

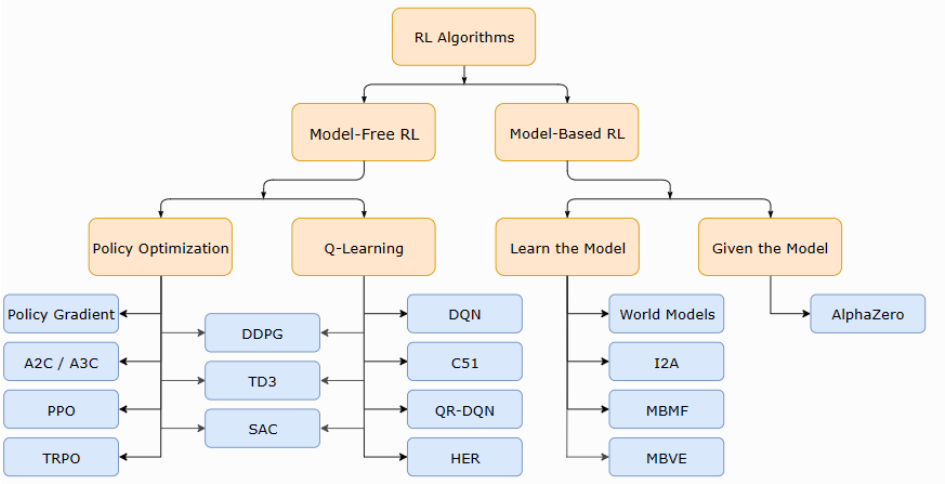


Figure 3.5: RL algorithms categorization [53]

Model-Free Reinforcement Learning

In Model-Free RL, the agent learns to make decisions based solely on the experiences it gains from interacting with the environment. It does not build an explicit model of the environment's dynamics. Instead, it focuses on estimating the value of states and actions to determine the best course of action. The learning of these algorithms comes with a trial and error procedure. Model-Free RL can be further divided into value-based, policy-based, and actor-critic methods.

Value-Based Algorithms These algorithms focus on estimating the value function to determine the best policy. They are characterized by a longer computation time due to the extensive state and action spaces.

- **Monte Carlo (MC) Methods:** In RL, MC methods are used to estimate the value function and policy based on averaging sample returns. These methods operate on complete episodes and do not require a model of the environment. The value of a state s under policy π is estimated by:

$$V(s) = \frac{1}{N(s)} \sum_{i=1}^{N(s)} R_{t,i} \tag{3.9}$$

  where $N(s)$ is the number of times state s is visited, and $R_{t,i}$ is the return following the i-th visit to state s. Monte Carlo methods are effective in episodic tasks where all episodes eventually terminate. They are suitable for episodic tasks with high variance and zero bias.

- **Temporal Difference (TD) Learning:** TD represents a class of model-free reinforcement learning methods that learn by bootstrapping from the current estimate of the value function. TD Learning is significant as it allows for learning directly from raw experience without a model of the environment's dynamics. Unlike Monte Carlo methods, TD learning updates estimates based partly on

other learned estimates, without waiting for a final outcome. A fundamental TD method is TD(0), represented by the formula:

$$V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)] \tag{3.10}$$

Here, α is the learning rate, γ is the discount factor, r is the reward received after transitioning from state s to s', $V(s)$ represents the estimated value of state s, and $V(s')$ is the estimated value of the next state s'. This method combines the sampling of Monte Carlo with the bootstrapping of dynamic programming. An extension of TD(0) is TD(λ), which considers multiple steps in the update process, providing a more flexible approach to learning suitable for continuous tasks.

Within the TD Learning framework, Q-Learning is a prominent off-policy, model-free algorithm used to find the maximum Q-value for state-action pairs. It updates Q-values for each state-action pair iteratively using the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{3.11}$$

Here, s and s' are the current and next states, a is the current action, r is the received reward, α is the learning rate, γ is the discount factor, and $\max_{a'} Q(s',a')$ estimates the best action for the next state s' according to the action value function. Q-Learning aims to learn a policy that maximizes the total reward over a trajectory and is effective in environments with discrete state and action spaces.

Policy-Based Algorithms These algorithms directly learn the policy based on the actions chosen for specific states, suitable for large state-space applications. Traditional policy-based methods are particularly effective in environments with high-dimensional or continuous action spaces and are known for their ability to learn stochastic policies. Most of the Policy-Based algorithms make use of the Policy Gradient Theorem.

- **Policy Gradient Theorem:** Let θ parametrize this policy as π_θ. The objective is to maximize the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau)]$, where $R(\tau)$ is the return of a trajectory τ generated by policy π_θ.

  The policy gradient theorem states that the gradient of the expected return with respect to the policy parameters θ can be expressed as an expectation:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)\right] \tag{3.12}$$

  Here, $Q^{\pi_\theta}(s,a)$ is the action-value function under policy π_θ, and the expectation is over the state and action space according to the policy π_θ. A standard approach to solve the maximization problem, is to use Gradient Ascent. In the gradient ascent update step, we adjust the policy parameters θ in the direction that increases the expected return:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

  Here, α is the learning rate, and this update is repeated iteratively to improve the policy.

Actor-Critic Algorithms Actor-Critic methods in Model-Free RL combine the concepts of value-based and policy-based approaches. The 'actor' is responsible for selecting actions based on a policy, while the 'critic' assesses these actions using a value function. This combination allows the agent to balance the direct policy approach with value-based learning, leading to efficient policy improvement.

The actor updates the policy in the direction suggested by the critic's value function. This framework integrates the strengths of both value-based and policy-based methods, providing more stability and efficiency in learning than using either approach alone. We will elaborate further in Actor Critic algorithms in the next subsection

Model-Based Reinforcement Learning

Model-Based RL employs models of the environment for more informed planning and decision-making. These models can either be given or learned. There are two main methods in Model-Based RL.

In **Learning the model** a common method is World Models [54], which creates internal representations of the environment, often using techniques like Variational Autoencoders and Recurrent Neural Networks. Effective in complex scenarios, they enable understanding and predicting dynamics, particularly in robotics and computer vision.

Given model methods utilize a pre-existing, known model of the environment. Common dynamic programming techniques include Policy Iteration and Value Iteration. These methods solve problems by decomposing them into simpler sub-problems using the environment's known model.

3.3 Deep Reinforcement Learning

As we explored the foundational concepts of RL, it is important to acknowledge its limitations, particularly in complex, high-dimensional environments. Traditional RL methods, while powerful, have limitations. Some of these include:

- **Scalability**: Traditional RL methods face computational challenges as the size of the state and action spaces increases.
- **Sample Efficiency**: These algorithms often require a large number of samples to learn effective policies, which can be impractical in real-world scenarios.
- **Generalization**: Traditional RL methods struggle with applying their learned policies to new, unseen environments.

These limitations have led to the emergence of Deep Reinforcement Learning (DRL), which combines the decision-making framework of RL with the representational power of deep neural networks.

Function Approximation is used in RL to estimate value functions and policies in large or continuous state spaces. Neural networks are often used in DRL for function approximation due to their ability to model complex, non-linear relationships. This integration addresses the scalability and generalization issues of traditional RL, enabling applications in more complex environments.

The field of DRL has achieved many breakthroughs, like the success of DeepMind's AlphaGo and its ability to master the game of Go, a task previously thought to be beyond the reach of computer algorithms [55],[56], or QT-Net [57], a novel adaptive trading model that uses DRL to autonomously develop quantitative trading strategies. These successes show the potential of DRL in solving complex, real-world problems.

3.3.1 Key Algorithms in Deep Reinforcement Learning

DRL encompasses a variety of algorithms, each designed to deal with different aspects of learning and decision-making in complex environments.

Deep Q-Networks (DQN) [58], firstly introduced by DeepMind, marked a significant advancement in DRL. Using only the pixel data from the Atari game screens, the DQN agent learned to play these games effectively, demonstrating its ability to process and act on complex visual inputs.

DQN uses a neural network to approximate the Q-value function, typically denoted as $Q(s, a; \theta)$ in this context. This differs from $Q^{\pi}(s, a; \theta)$, as DQN, an off-policy method, aims to directly approximate the optimal Q-value function, rather than one for a specific policy π. It employs two neural networks to enhance stability and performance: the main Q-network and a copy of it named the Target network. The main Q-network, parameterized by θ, is responsible for learning the Q-values. It is updated frequently and directly learns from the interactions with the environment. The Target network, parameterized by θ^{-}, is a more stable version of the main Q-network. It is used to generate the Q-value targets in the Q-learning update rule. The key reason for using the Target network is to provide consistent targets for a while, as frequent updates of Q-values can lead to instability due to the moving targets problem. The parameters of the Target network, θ^{-}, are periodically updated to match those of the main Q-network, θ, but this happens less frequently to maintain stability in the learning process. DQN also use experience replay. This technique involves storing experiences (s_t, a, r, s_{t+1}) in a replay buffer D to minimize correlations between consecutive learning samples. The loss function in DQN, $L(\theta)$, is defined as:

$$L(\theta) = \mathbb{E}_{(s_t,a,r,s_{t+1})\sim U(D)} \left[\left(Q(s_t, a; \theta) - y \right)^2 \right] \tag{3.13}$$

where $y = r + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^{-})$.

Here, θ represents the parameters of the main Q-network, and θ^{-} represents the parameters of the target network. The term $U(D)$ denotes uniform sampling from a replay buffer D, which is a collection of past experiences (s_t, a, r, s_{t+1}). The loss function helps the network learn the optimal values for state-action pairs by minimizing the difference between the current Q-value estimates, $Q(s, a; \theta)$, and the target Q-values, y, which are calculated using the more stable target network parameters, θ^{-}.

Policy Gradient Methods optimize the policy function directly, offering a more flexible approach in environments with continuous action spaces. Most of these methods use the policu gradient theorem, we discussed in Section 3.2.2.

Deep Q-Networks (DQN) and **Policy Gradient Methods** have shown us different ways to approach learning in complex environments. DQN focuses on learning value functions, while Policy Gradient Methods concentrate on directly learning the best actions. These methods set the stage for a more advanced type of algorithms: **Actor-Critic Methods**.

3.3.2 Actor-Critics Variants Overview

Actor-Critic Methods are a significant advancement in both DRL and traditional RL, combining the aspects of policy optimization (actor) with value function approximation (critic). This hybrid approach combines the benefits of both methodologies, leading to more efficient and stable learning processes. In the this subsection, we will mention some of the actor-critic methods, with a special focus on the Soft Actor-Critic algorithm, which we will be using in this thesis.

Some basic actor-critic variants are:

- **Advantage Actor Critic (A2C) / Asynchronous Advantage Actor Critic (A3C)** [59]: A3C enables simultaneous exploration by multiple agents in different environments, enhancing the learning speed. A2C provides a synchronous variant, contributing to more stable updates. Both methods utilize Temporal Difference (TD) error for action evaluation.

- **Deep Deterministic Policy Gradient (DDPG)**[60]: DDPG merges Q-learning with policy gradient methods and is well-suited for continuous action spaces. It uses deterministic policies for efficient exploration and incorporates a replay buffer and target networks, aiding stability in complex environments.

- **TRPO (Trust Region Policy Optimization)**[61] : TRPO uses a trust region to constrain policy updates, enhancing stability. Its key advantages include more stable policy updates, better sample efficiency, and improved convergence compared to standard policy gradient methods.

- **PPO AC (Proximal Policy Optimization Actor Critic)**[62]: PPO AC alternates between sampling data and optimizing a 'surrogate' objective function. Its strengths include combining policy gradient methods with minibatch updates, stability and simplicity in implementation, and competitive performance among actor-critic algorithms.

- **Soft Actor Critic (SAC)**: SAC, which will be elaborated further in the next subsection, integrates actor-critic architecture with entropy regularization.

3.4 Soft Actor-Critic

The Soft Actor-Critic SAC algorithm emerges as a solution to some of the limitations inherent in traditional actor-critic methods. These traditional approaches often suffer from instability and a lack of sufficient exploration, leading to premature convergence to suboptimal policies and difficulty in fully solving complex tasks. To address these issues, Haarnoja et al. [63] [64] propose a stochastic policy, emphasizing the role of entropy maximization as outlined in the next paragraph. Such an approach ensures that the policy not only seeks optimal rewards but also retains a level of randomness in its actions, a crucial element for discovering more effective strategies in both discrete and continuous state spaces.

Maximum Entropy RL

Maximum Entropy RL [65] introduced a novel approach to address the Exploration-Exploitation dilemma, Section 3.2.1 in RL. It integrated a measure of randomness, or entropy, into the decision-making policy of an agent. This technique encourages the policy to explore a broader range of actions, thus preventing the agent from converging too quickly to suboptimal strategies. By maximizing entropy along with the expected returns, the policy not only seeks immediate rewards but also maintains a degree of unpredictability in its actions, which is key to discovering more effective strategies and understanding the environment comprehensively.

The concept of entropy, originally adapted from physics [66], describes the measure of disorder or randomness in a system. In information theory, entropy represents the level of uncertainty or the informational content in a dataset. In the context of RL, entropy quantifies the unpredictability in an agent's action selection process. High entropy in RL implies that the agent is exploring its environment by trying out a variety of actions, which is particularly crucial during the initial stages of learning. This prevents the agent from prematurely locking onto a narrow set of potentially suboptimal actions.

3.4.1 Soft Actor-Critic formulation

In this section, we will adopt the following notation for consistency with SAC's paper: s_t represents the current state, s_{t+1} denotes the next state, a_t signifies the action taken at time t, and r_t the reward received at time t . This notation aligns with the conventions used in the Soft Actor-Critic (SAC) formulation as presented in the SAC paper.

SAC, embodying the principles of maximum entropy RL, optimizes a policy π to simultaneously maximize expected rewards and entropy. As illustrated in (eq. 3.14), the entropy H of a policy π at a state s_t is calculated as the negative logarithm of the probability of selecting any given action at that state

$$H(\pi(\cdot|s_t)) = -\log \pi(\cdot|s_t) \tag{3.14}$$

In SAC, this entropy serves as a critical component in the policy optimization process. It ensures that the policy does not become overly deterministic and remains capable of exploring new and potentially rewarding actions.

Therefore, the objective of SAC is formalized as:

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{T} \mathbb{E}_{\pi} \left[\gamma^t \left(r_t + \alpha H(\pi(\cdot|s_t)) \right) \right] \tag{3.15}$$

Here, π^* is the optimal policy, T denotes the number of timesteps, γ is the discount rate, α is the temperature parameter that determines the relative importance of the entropy term controlling the stochasticity of the optimal policy, and $H(\pi(\cdot|s_t))$ represents the entropy of the policy at state s. The notation τ_π denotes the trajectory distribution under policy π, which refers to the probability distribution of state-action pairs when the agent follows policy π over time.

The soft state-value function $V^\pi(s)$ (eq. 3.16) and soft action-value function $Q^\pi(s, a)$ (eq. 3.17) in SAC extend the concepts of the traditional state-value function (eq. 3.5

and action-value function (eq. 3.6) by incorporating maximum entropy regularization ($\alpha H(\pi(\cdot|s_t))$).

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t + \alpha H(\pi(\cdot|s_t)) \middle| s_0 = s\right] \tag{3.16}$$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t + \alpha \sum_{t=1}^{\infty} \gamma^t H(\pi(\cdot|s_t)) \middle| s_0 = s, a_0 = a\right] \tag{3.17}$$

From equations (3.16), (3.17) it is possible to derive the connection between soft state-value and soft action-value function (3.18) and the Bellman equation (3.19)

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[Q^{\pi}(s_t, a_t)\right] + \alpha H(\pi(\cdot|s_t)) \tag{3.18}$$

$$\begin{aligned} Q^{\pi}(s_t, a_t) &= \mathbb{E}_{\pi}\left[r_t + \gamma(Q^{\pi}(s_{t+1}, a_{t+1}) + \alpha H(\pi(\cdot|s_{t+1})))\right] \\ &= \mathbb{E}_{\pi}\left[r_t + \gamma V_{\pi}(s_{t+1})\right] \end{aligned} \tag{3.19}$$

Learning procedure

Dual Q networks In SAC algorithm, two separate Q-networks, as proposed in [67], are employed, denoted as Q_{θ_1} and Q_{θ_2}. This dual network architecture is designed to address the overestimation bias commonly observed in Q-learning algorithms. During training, for each action, the minimum value between these two networks is selected.

SAC learns through a process of soft policy iteration, which includes policy evaluation and policy improvement steps. In policy evaluation, SAC trains the Q-function to estimate expected future rewards, while in policy improvement, it adjusts the policy parameters to maximize expected rewards.

Policy evaluation Q-function is parameterized as $Q_\theta(s_t, a_t)$ using a neural network with parameters θ. The training process focuses on minimizing the Mean Squared Bellman Error, using the value network to form the Bellman backups using (eq. 3.20).

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}\left[\frac{1}{2}\left(Q_\theta(s_t, a_t) - \left(r_t + \gamma \mathbb{E}_{s_{t+1}\sim p(s_t,a_t)}[V_{\bar{\theta}}(s_{t+1})]\right)\right)^2\right] \tag{3.20}$$

Then the state-value function can be implicitly parameterised thought (eq. 3.21) which derives from (eq. 3.18) .

$$V^{\pi}(s_t) := \mathbb{E}_{\pi}[Q^{\pi}(s_t, a_t) - \alpha \log(\pi(a_t|s_t))] \tag{3.21}$$

Policy improvement The policy parameters can be learned by directly minimizing the expected KL-divergence:

$$J_{\pi}(\phi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi_\phi(\cdot|s_t) \,\middle\|\, \frac{\exp(Q_\theta(s_t, \cdot))}{Z_\theta(s_t)}\right)\right]. \tag{3.22}$$

The authors in [63] state that there are several options to minimize $J_\pi(\phi)$, but in this case it is convenient to apply the reparameterization trick as the Q-function ,which is represented by a neural network, can be differentiated. This works by reparameterizing the policy using a neural network which outputs the parameters of a probability distribution over actions. In the case of SAC, this distribution is often modeled as a Gaussian distribution with a mean and a covariance. The reparameterization is:

$$a_t = f_\phi(e_t; s_t) \tag{3.23}$$

Here e_t is a noise vector sampled from some fixed distribution.
Therefore, it is possible to rewrite (eq. 3.22) as (eq. 3.24).

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, e_t \sim \mathcal{N}} \left[\log \pi_\phi(f_\phi(e_t; s_t)|s_t) - Q_\theta(s_t, f_\phi(e_t; s_t)) \right], \tag{3.24}$$

Here π_ϕ is defined implicitly in terms of f_ϕ. We can approximate the gradient of (eq. 3.24) with:

$$\begin{aligned} \hat{\nabla}\phi J_\pi(\phi) = {} & \nabla_\phi \log \pi_\phi(a_t|s_t) \\ & + (\nabla_{a_t} \log \pi_\phi(a_t|s_t) - \nabla_{a_t} Q(s_t, a_t)) \nabla_\phi f_\phi(e_t; s_t), \end{aligned} \tag{3.25}$$

where a_t is evaluated at $f_\phi(e_t; s_t)$.

Entropy tuning SAC algorithm uses a stochastic policy enhanced with entropy regularization. The key component, the entropy coefficient α, is essential for balancing exploration and exploitation. Higher values of α encourage exploration, while lower values focus on exploitation. The optimal setting of α is non-trivial and varies across environments, requiring careful adjustment to achieve the most stable and rewarding learning outcomes. In [64], the authors propose a dynamic entropy approach, allowing the policy to explore more in states of uncertainty and exploit in states where the optimal action is clearer. The temperature loss gradient is formalized in (eq. 3.26):

$$J(\alpha) = \mathbb{E}_{a_t \sim \pi} \left[-\alpha \log \pi(a_t|s_t) - \alpha \bar{H} \right] \tag{3.26}$$

SAC adaptation for Discrete Action Settings

At [38] Christodoulou proposed an adaptation of the SAC algorithm to discrete action spaces. The steps in deriving the objectives are still valid, meaning (eq. 3.20, 3.24 and 3.26) still hold, with some modifications.

- The policy no longer outputs the mean and covariance of the action distribution, but directly outputs the action distribution, changing from $\pi : S \to \mathbb{R}^{2|A|}$ to $\pi : S \to [0, 1]^{|A|}$, using a softmax function in the final layer.
- The soft Q-function now outputs the Q-value for each possible action instead of just the input action, transitioning from $Q : S \times A \to \mathbb{R}$ to $Q : S \to \mathbb{R}^{|A|}$.

- In discrete action setting there is no need for an estimate in order to minimize the soft Q-function cost (eq. 3.20), since the action distribution is now completely determined, in contrast to continuous settings where the action probabilities are not explicitly defined. As a result, the soft state-value calculation changes from (eq. 3.21) to:

$$V^{\pi}(s_t) = \pi(s_t)^T \left[Q^{\pi}(s_t) - \alpha \log(\pi(s_t)) \right] \tag{3.27}$$

- The calculation of the temperature loss is adjusted similarly, with the temperature objective changing from (3.26) to (3.28).

$$J(\alpha) = \pi_t(s_t)^T \left[-\alpha \left(\log \pi_t(s_t) + \bar{H} \right) \right] \tag{3.28}$$

- With the policy outputting the exact action distribution, the reparameterization trick is no longer needed, allowing for a direct calculation in the policy objective (eq. 3.29).

$$J_{\pi}(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}} \left[\pi_t(s_t)^T \left[\alpha \log \pi_{\phi}(s_t) - Q_{\theta}(s_t) \right] \right] \tag{3.29}$$

The SAC algorithm with discrete actions is presented in 3.6.

Algorithm 1 Soft Actor-Critic with Discrete Actions (SAC-Discrete)

Initialise $Q_{\theta_1} : S \to \mathbb{R}^{|A|}$, $Q_{\theta_2} : S \to \mathbb{R}^{|A|}$, $\pi_{\phi} : S \to [0,1]^{|A|}$ ▷ Initialise local networks
Initialise $\bar{Q}_{\theta_1} : S \to \mathbb{R}^{|A|}$, $\bar{Q}_{\theta_2} : S \to \mathbb{R}^{|A|}$ ▷ Initialise target networks
$\bar{\theta}_1 \leftarrow \theta_1$, $\bar{\theta}_2 \leftarrow \theta_2$ ▷ Equalise target and local network weights
$\mathcal{D} \leftarrow \emptyset$ ▷ Initialize an empty replay buffer
for each iteration **do**
  for each environment step **do**
    $a_t \sim \pi_{\phi}(a_t|s_t)$ ▷ Sample action from the policy
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ ▷ Sample transition from the environment
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$ ▷ Store the transition in the replay buffer
  for each gradient step **do**
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J(\theta_i)$ for $i \in \{1, 2\}$ ▷ Update the Q-function parameters
    $\phi \leftarrow \phi - \lambda_{\pi} \hat{\nabla}_{\phi} J_{\pi}(\phi)$ ▷ Update policy weights
    $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_{\alpha} J(\alpha)$ ▷ Update temperature
    $\bar{Q}_i \leftarrow \tau Q_i + (1 - \tau)\bar{Q}_i$ for $i \in \{1, 2\}$ ▷ Update target network weights
Output θ_1, θ_2, ϕ ▷ Optimized parameters

Figure 3.6: Discrete SAC pseudo code,[38]

3.5 Transfer Learning

Transfer learning (TL) is a technique where a model or policy developed for one task is repurposed on a different but related task. In RL, TL involves using knowledge gained from one environment to enhance performance in another. This is particularly useful in HRC, where an agent must learn both the task and efficient cooperation with a human partner. TL can boost the learning efficiency of such teams, but it is crucial to avoid negative transfer, which can hinder problem-solving in new contexts.

TL can be implemented in various forms, each impacting training and co-learning differently. Zhungadi [12] categorizes TL in 5 approaches based on the type of information transferred:

- **Reward Shaping (RS)** [68]: This method involves modifying the reward structure of the learning environment based on external knowledge, such as insights from domain experts. The goal is to guide the agent towards more desirable behaviors by enhancing or diminishing rewards for certain actions. This can significantly speed up the learning process by providing more informative feedback to the agent. One example of RS in HRI is [69] in which, a RS framework is applied in a simulation scenario where a human-robot team performs a search-and-rescue mission. The results showed that the proposed framework successfully modifies the robot's optimal policy, enabling it to increase human trust with a minimal task performance cost.

- **Learning from Demonstration (LfD)**: LfD involves the agent learning from demonstrations provided by either a human expert or a pre-trained model. These demonstrations act as a guide, showing the agent effective strategies and actions in various states. This method is particularly useful in scenarios where the agent needs to learn complex tasks that are difficult to discover through trial and error alone. In the survey [70], the state of the art in LfD for collaborative robots is reviewed, with a focus on improving HRC by reducing complexity for human operators and aligning solutions with smart manufacturing. This approach empowers non-experts to teach robots new knowledge and enhances their collaboration with humans in various tasks [71].

- **Policy Transfer (PT)**: In PT, the agent leverages pre-trained policies from related tasks. This method is based on the assumption that certain aspects of the decision-making process in one task are applicable to another. By reusing these policies, the agent can bypass part of the learning process, adapting the existing knowledge to the new task. One approach to achieve this is thought policy distillation [72], which means that the agent will select an action by minimizing the the divergence of action distributions between the source policies and the target policies. Another way is direct policy reuse [29], where the agent selects, with a probability, an action based on a previous learned policy instead of his own policy. An example of a transfer policy is demonstrated in [73], where a robot learns both a task-focused policy and a safety-focused policy. The safety estimator model, helps determine which policy to use during execution. This approach enables robots to adapt to new environments while prioritizing safety

- **Inter-Task Mapping (ITM)**: ITM involves creating mapping functions between the source state space and target state space, the source action space to the target action space [74]. These functions help in translating the knowledge from one task to another, especially when the tasks are similar but not identical. Thought these mappings one of the previously mentioned TF methods can be applied. In [75], the results show that using multiple mappings significantly enhances transfer learning performance compared to single mapping or non-transfer methods.

- **Representation Transfer**: Representation learning aims at extracting features of the source problem which exist in the target problem as well. This is achieved by disentangling the state space, the action space or the reward space into task-invariant sub-spaces which are shared by both source and target domains. A work

of reusing representations is [76], which proposed the progressive neural network structure to enable knowledge transfer across multiple RL tasks in a progressive way.

Each of these methods offers unique advantages and can be chosen based on the specific requirements of the task. When implementing TL in RL, it is essential to consider the compatibility of these methods with the chosen RL framework and the nature of the tasks involved.

3.5.1 Deep Q-Learning from Demonstrations (DQfD)

DQfD, a variant of Learning from Demonstrations TL, is an innovative approach in RL, developed by Google DeepMind [33]

Algorithmic Pipeline

1. **Pre-training from Demonstrations**: Initially, the model undergoes pre-training using a dataset of expert demonstrations. This phase enables the model to acquire an understanding of desired behaviors, essential for effective learning.

2. **Combining Q-Learning and Demonstrations**: After pre-training, DQfD integrates traditional Q-learning updates with specialized loss functions derived from these expert demonstrations. This combination ensures the model learns effectively from both its experiences and the expert demonstrations.

3. **Regularization and Optimization**: To prevent overfitting to the demonstration data, DQfD employs regularization techniques. Additionally, it utilizes optimization methods like gradient descent to enhance the model's performance.

Mathematical Formulation

The DQfD loss function is a composite of several components, each serving a specific purpose:

- **Temporal Difference (TD) Loss**:

$$\text{TD Loss} = \left(Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right)^2$$

  This loss is fundamental in Q-learning, capturing the difference between predicted and actual rewards, thus guiding the model to learn the optimal policy.

- **Supervised Learning Loss from Demonstrations**:

$$\text{Supervised Loss} = \sum_{(s,a)\in\text{demonstrations}} \left(Q(s,a) - Q^*(s,a)\right)^2$$

  This component focuses on learning from expert demonstrations, aligning the model's actions with those of the expert.

- **Large Margin Classification Loss**:

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

This loss ensures that the values of non-expert actions are lower than the expert's actions, grounding the Q-values and aligning the policy with the expert's behavior.

- **L2 Regularization Loss**: This loss is applied to the network's weights and biases, reducing overfitting especially important when working with smaller datasets.

- **Overall Loss Function**:

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

The overall loss is a weighted sum of these components, balancing learning from expert behavior, adherence to the Bellman equation, prevention of overfitting, and policy alignment.

Algorithm 1 Deep Q-learning from Demonstrations.

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

Figure 3.7: DQfD pseudocode,[33]

The DQfD algorithm is depicted in 3.7. DQfD is applicable in various domains, particularly where expert demonstrations can bootstrap learning, such as in robotics and strategy games. There are other variants of DQfD, like [77], which involves simultaneous pretraining of policy functions and state-action value estimators using expert demonstrations. These are applied to algorithms like DDPG and ACER, demonstrating enhanced performance compared to traditional RL methods.

Chapter 4

Related work

This chapter talks about the history and development of Human-Robot Interaction (HRI). It starts from the first industrial robot made in the 1950s and goes to modern times where robots use advanced Artificial Intelligence (AI) and machine learning. It highlights the transformative role of Deep Reinforcement Learning (DRL) in robotics, enabling robots to undertake intricate tasks autonomously. The significance of DRL in facilitating effective Human-Robot Collaboration (HRC) is also examined, alongside its inherent challenges, such as the scarcity of learning data and the complexities involved in transferring knowledge. Finally, this chapter outlines the principal contributions of this study, shedding light on its impact on the field of HRC.

4.1 Overview of Human-Robot Interaction (HRI)

HRI has a significant evolution and advancement. It began in the late 1950s with the introduction of Unimate, the first industrial robot. This marked a new phase in HRI, known as 'indirect interaction' where humans programmed robots to execute specific tasks, effectively setting the stage for future advancements in robotics.

In the following decades, as outlined in [78], HRI transitioned from its initial directive nature, where robots were just executors of human commands, to a more nuanced and collaborative interaction. Sheridan's work analyze early challenges in HRI, such as the intricacies of human supervisory control and the emergence of social robotics.

The recent integration of advanced AI and machine learning has significantly revolutionized HRI. This evolution has broadened the scope of HRI, incorporating things like communication and ethical decision-making into robotic capabilities. Today, robots are integrating into daily human activities, not just enhancing human capabilities and experiences, but also make this interaction user friendly and anthropocentric.

As the field grows, it is important to understand the different types of interactions that occur between us and robots. Drawing from the categorization in [1], HRI can be broadly classified into three types, each representing a different level of interaction and autonomy:

- **Instruction**: We can view it as an one-way communication, this form involves humans issuing commands that robots execute. Here, robots act primarily as tools, carrying out tasks set by humans.

- **Cooperation**: This involves more dynamic interaction, where humans and robots work on separate tasks but contribute towards a shared objective. It requires a certain level of autonomous functioning from the robot, guided by human input.
- **Collaboration**: Here there is a sequence of interdependent actions meaning that each participant affect the actions of the other partner.

The applications of HRI are diverse, impacting various aspects of our lives. In healthcare, robots assist in complex surgeries [79]. Also rehabilitation robotics represents a significant advancement, particularly in providing therapy for stroke patients and those with motor disorders [80]. The manufacturing sector has been transformed by collaborative robots (cobots), enhancing performance and flexibility in assembly lines [5]. These robots are designed to work safely alongside humans, per-forming tasks that are either hazardous or repetitive. Social robots [81] are designed for direct interaction with humans, often with anthropomorphic designs, even for educational purposes [82]. Last but not least in entertainment and gaming, HRI introduces new forms of interaction, as explored in studies like [83].

The integration of robots in these sectors demonstrates the need of robots to be able to adapt in different functions and highlights the importance of continual advancements in this field. a key focus of current research in the field of HRI is to enhancing human-robot collaboration, especially in tasks requiring precision and adaptability, but at the same time keeping human as the center of the design.

4.2 Deep Reinforcement Learning in Robotics

4.2.1 Implementations of DRL in Robotic Tasks

In the domain of robotics, DRL has brought changes, as it is applied to a wide array of robotic systems to address tasks that require sophisticated decision-making and control. DRL offers a way for robots to learn and perform complex actions that would be difficult to achieve with traditional programming methods due to the complexity of the tasks.

For instance, mobile robots have utilized DRL to navigate autonomously [84],[85]. For example in [8] the authors use the Asynchronous Advantage Actor-Critic algorithm to enable a mobile robot to navigate without a map or a path planner but only using data from a 2D laser scan and a RGB-D camera. The goal is for the robot to get to a predefined goal pose while avoiding static obstacles and is achieved by training the robot to a simulated environment and then deploying it to the real world. Similarly, in robotic arm control [9], [86]. An example of robot arm motion control is [87] where a Deep Q network has been used in to enable a 3 DoF robotic arm to reach target configurations without prior knowledge of the goal and using only raw visual pixels as input to the network. Another DRL application to robotics is at grasping [88]. In [89] the authors propose a Q-learning based network architecture for improving the grasping capabilities of a robotic manipulator using raw visual data input from a multi-camera setup. By employing advanced learning algorithms, robots can learn to pick up and manipulate objects with a precision that mimics the dexterity of the human hand [18]. This learning process can be so detailed that it can even be based on raw visual data, allowing the robot to adapt to different tasks without being explicitly programmed for

each one. In drones [10], equipped with a stereo-vision front camera, learn to avoid obstacles within a geo-fenced area and reach their destination, demonstrating improved performance and consistency in their tasks. Other examples can be found [90] , [91]. In [92], the authors present a DRL algorithm based on maximum entropy RL in order to teach a quadrupedal how to walk. An additionally interesting result is that the robot learns without having access to his dynamic model.

DRL in HRC

The integration of DRL into robotics has had a profound impact on the capabilities of robots, making them more adept at collaborating with humans. This integration is particularly relevant in the context of HRC, where robots and humans work together on tasks, often in close proximity.

Firstly, safety and trust are crucial in HRC scenarios. DRL algorithms can be tuned to train robots to prioritize safety, operate within predefined boundaries, and respond adeptly to unforeseen human actions, thereby gaining confidence in their collaboration [14].

Also HRC often involves tasks that require intricate decision-making and control. DRL's proficiency in handling complex tasks, such as precision grasping and navigating through cluttered environments, enhance the robot's contribution to the collaboration [15].

Another advantage of DRL is that enables robots to continuously learn and adapt from human expertise, improving their performance over time and reducing the learning curve for new tasks [16].

Furthermore, DRL algorithms provide real-time adaptability, ensuring that robots can make instantaneous decisions and adjustments, aligning their actions with human intentions and goals, thereby augmenting the quality of HRC interactions [17].

A study in this area is presented in [93], where Shafti developed a HRC framework involving a robotic manipulator and a human user. The objective involves shifting a ball from a starting corner to a designated target location by manipulating a tilting platform, as depicted in 4.1. Within this setup, one axis of the platform is under human control, while the other axis is managed by the robot. Additionally, the platform features various obstacles that necessitate collaborative efforts from both the human and robot to navigate and achieve the goal. Lygerakis [94] created a visual simulation of this work in order to evaluate different training approaches. Then, in his work Koutrintzes [34] applied a TL methodology, called Deep Q-learning from Demonstrations, in this environment to examine how this can enhance the Human-Agent collaboration.

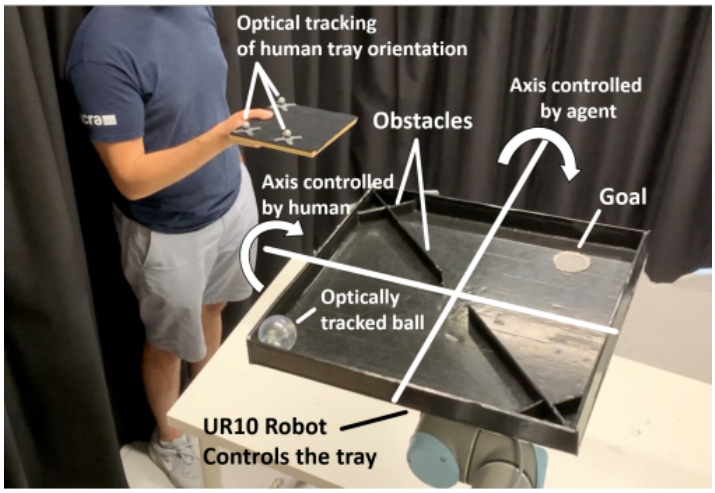


Figure 4.1: The robotic setup used for Shafti's HRC experiment,[93]

4.2.2 Limitations of DRL in Robotics

DRL might advance robotics in various aspects, making them more intelligent and autonomous but it comes with its limitations when it comes to practical robotics applications.

DRL algorithms often face **sample insufficiency** problems, as they need large datasets for the algorithm to develop policies close to the optimal. One way to overcome this issue is through the parallel utilization of multiple robots to gather data, as seen in [18] where fourteen robots were deployed for collecting data to train a model for predicting grasp. This strategy, while effective, may not be practical due to the high costs and potential lack of hardware resources.

An alternative solution is to train robots within simulated environments, which is not only quicker but also more cost-efficient, and then apply the learned policies in the real world. Tranfer Learning techniques like "Zero-shot Transfer" [19], involves direct policy application from simulations to real-life situations, provided the simulated environment closely mirrors the actual conditions. This method was evaluated in [20] through tasks such as reaching, pushing, and sliding conducted by a robotic manipulator. Yet, the efficacy of policies learned in simulations may vary when applied to the real world due to inherent differences between the simulated and real environments or the complexity and unpredictability of real-world conditions

To bridge this gap, "Domain Randomization" is employed, which involves randomizing simulation parameters to span a range of real-world scenarios. For instance, [21] describes training an object detector in diverse simulated settings, which could then operate effectively in real-world applications without further training, for tasks like pick-and-place operations. This approach has also been utilized in other areas such as pose estimation [22] and semantic segmentation [23], enhancing the robustness of DRL applications in robotics [24] .

Another limitation is the **exploration-exploitation trade-off**, a fundamental dilemma in reinforcement learning. In robotic applications, random exploration needed for learning can lead to unsafe actions, potentially causing mechanical damage. A recent work which addresses this issue is [24], where the authors present methods so that

safety principles can be incorporated in reinforcement learning. Safety has also been taken into account in real-world applications, like [25], where a neural network is used to predict the outcome of an action in terms of safety. Another limitation of random exploration is that it can be a time consuming procedure due to the high dimensionality of the action space in robotic applications (e.g. an agent controlling the rotation of the wheel of a self-driving car). An example work which addresses this issue is [26], where demonstrated trajectories have been utilised as a bias that governs the learning procedure in early stages.

One more issue of DRL in robotics is the ability of the robots to **generalize knowledge** in order to operate in new, unknown circumstances and environments. Most works try to solve a RL task by training the robot from scratch. However, this approach is non-optimal mainly because it is time consuming. One way to overcome this issue is by transferring knowledge among conceptually similar tasks, just as humans do. The methods for applying transfer learning in DRL have been presented in 3.5. In [27], the authors present an implementation of reward shaping in robotics in order to improve the training of an RL agent used for mobile navigation by altering its reward function based on the knowledge about the map provided by the SLAM algorithm. Learning from demonstration has been applied in [26] where the issue of sparse reward function is a pick-and-place scenario is addressed. The use of demonstrated trajectories for efficient exploration enables the agent to solve the task, which might have been infeasible with random exploration. Pre-learned policies have also been used for training DRL agents. For example, policy distillation has been applied in robotics in a continual learning problem [28], where the goal is for a single RL agent to learn three different policies for three different navigation tasks and learn which policy to use by identifying in real time the task to be solved. The second approach is direct policy reuse [29] , and the agent can select an action based on the pre-learned policy instead of his own policy. This idea has been used in [30], where the authors teach a humanoid robot how to walk fast by exploiting a policy that allows the robot to walk in a normal speed. Knowledge has also been transferred between morphological different robots, like in [31], where the authors train a 3-link robotic manipulator in three different tasks (target reaching, peg insertion, and block moving) and exploit the policies in order to train a 4-link one. Finally, in the context of representation learning, some works such as [95] show how to reuse the extracted features for transferring knowledge, while other such as [96] focus on the feature extraction. An application of representation learning in robotics is presented in [32], where the authors show how extracting important features from the environment can accelerate the learning procedure of an RL agent in the case of slot car racing and mobile robot navigation.

4.3 Motivation and contribution

Our work is based on Tsitos experiment [13], where a human, controlling the y-axis of the robot's end effector collaborates with a DRL agent controlling the x-axis, in order to learn how to solve a task in real-time. The experimental setup will be detailed in the next chapter 5.1. It focus is to determine whether knowledge transfer from a pre-trained expert agent can improve the overall team performance. In his work, Tsitos used Probabilistic Policy Reuse (PPR), a direct policy reuse TL technique 3.5, which

allows the agent to select, with a probability, an action based on a previous learned expert policy instead of his own.

In our approach, we introduce some modifications, which will be discussed in detail in the next chapter 5.1. However, the primary deviation from the previous work [13] is in our implementation of Deep Q-learning from Demonstrations (DQfD) [33], a TL technique based on Learning from Demonstration (LfD), instead of PPR.

One of the advantages of LfD over PPR is the nature of the interaction between the human participant and the agent. In LfD, the participant interacts with their own DRL agent, which may incorporate learning from expert demonstrations. This approach ensures that the participant is central to the learning process, developing their own policy rather than being directly influenced by an expert agent's decisions. This fosters a more human-centric learning experience, where the human user significantly influences the robot's learning trajectory, aligning it with their unique style.

Another reason we chose DQfD is because of its compatibility with the SAC framework [34]. DQfD combines the power of Q-learning with learning from demonstrations, allowing the agent to benefit from expert guidance. This method integrates with SAC, an off-policy RL algorithm known for its stability and performance. This compatibility provides us with a robust framework for our HRC task. Additionally, DQfD operates in an offline and off-policy manner, enabling efficient learning from past experiences [35], enhancing the adaptability and performance of our robot in real-time HRC scenarios.

The main contributions of our study are:

- Implementing LfD as a TL method to the task.
- Conducting a comparative study with human-robot teams to assess the impact of LfD.
- Comparison of LfD to the previously implemented PPR method.
- Discussing the influence of different target entropies and activation functions in the SAC algorithm on the effectiveness of the collaboration.

Chapter 5

Methodology

In this chapter, we present the overview of the methodologies and experimental setup used in the Human-Robot Collaboration (HRC) study.

The first section begins with an overview of the collaborative task. The game structure, familiarization process, and baseline and initialization phases for both groups are explained in detail. It also details the use of the Soft Actor Critic (SAC) algorithm for robot control in the HRC experiment. Furthermore, this section introduces the application of Deep Q-Learning from Demonstrations (DQfD), a Transfer Learning (TL) method, to enhance the SAC agent's performance by integrating learning from expert demonstrations, into our methodology.

The second section, outlines both objective and subjective measures used to evaluate quality of human-robot collaboration. It also emphasizes in understanding participant personalities through comprehensive questionnaires, an aspect critical to exploring individual interactions with AI and robots.

5.1 Research Approach

5.1.1 Overview of the Collaborative task

The human-robot collaborative task was taken from [13]. The team consists of a Universal Robot UR3, which is a non-redundant 6-DoF robotic arm and a human. The robot is placed to the middle of a 1m x 1m table and its end effector (EE) perpendicular to the table and can move parallel to it a certain height. Also, a laser is attached to the EE pointing to the table. While the human controls the movement of the EE in one axis (y-axis), using the keyboard, a DRL agent, as explained in Section 5.1.2, is responsible for the movement in the (x-axis). With the combination of the motions of two partners the robot can move in the xy pane, constrained in a 20cm X 20cm square.

At the beginning of each game, the robot chooses randomly an initial position out of the four possible (at the corners of the square) and placed on top of it, as it is depicted in Figure 5.1. At the time it reaches the starting position a sequence of three short and one long "beeps" indicates the start of the game. When the game starts the task of the human-DRL agent team is to bring the laser dot inside a the circle of the goal position, located at the center of the square. The goal position has a position and velocity tolerance which are 0.01m and 0.05 m/s respectively. The team wins if it manages to bring the laser dot inside the circle of the goal position, with radius=0.01,

with velocity lower than 0.05, at 30 secs from the beginning of its game, else it timeouts. The player will be informed with a different sound in both cases.

Figure 5.1: Initialized Position. The robot's movements are constrained within a $20cm \times 20cm$ area. The EE is placed in one of the four starting ('$\circ$') positions and the HR team has to bring the EE in the centre ($\otimes$) of the square. A laser pointer attached to the EE of the robot provides to the human visual feedback about the position of the EE that is controlled.

5.1.2 Reinforcement Learning agent

The HRC task, as detailed in Section 5.1.1 consists of a human controlling the y-axis and a DRL agent controlling the acceleration in x-axis of the EE. A discrete Soft Actor Critic algorithm, Section 3.4, is responsible for the movement of the EE in the x-axis.

For our task, we chose discrete SAC due to its features that align well with the requirements of this complex and dynamic task. Firstly, discrete SAC is a model-free algorithm, meaning it does not require a predefined model of the environment which is beneficial in HRC settings where the environment dynamics can be unpredictable and influenced by human actions. Being an off-policy algorithm, discrete SAC learns from past experiences, allowing it to leverage previously collected data. This is crucial in scenarios where real-time interaction data is valuable and limited. Additionally, discrete SAC incorporates offline training, which enables the algorithm to optimize its policy using previously gathered data. This aspect is particularly useful in HRC, where continuous online learning might be impractical. Moreover, discrete SAC's emphasis on entropy maximization ensures a balance between exploration and exploitation, a crucial limitation of DRL in HRC tasks as discussed in Section 4.2.2 where adaptability and responsiveness to new situations are essential. The algorithm's ability to adjust its exploration strategy dynamically through entropy regularization makes it well-suited

for HRC tasks. Overall, the combination of these features makes discrete SAC a suitable choice for our HRC task, providing a robust and flexible framework for effective HRC.

Discrete SAC employs a stochastic policy, using a softmax function to generate a probability distribution for action selection. Actions are chosen based on this distribution, ensuring a degree of randomness in decision-making. The SAC's exploration strategy is achieved by a soft policy, where all actions have a calculable chance of being selected. This probability is influenced by the entropy of the action probabilities and the alpha temperature, a parameter adjusted during offline training to align with the target entropy.

The MDP formulation for our problem is defined as follows:

- $\mathbf{S} = \{eepos_{\mathrm{x}}, eepos_{\mathrm{y}}, eevel_{\mathrm{x}}, eevel_{\mathrm{y}}\}$. Position and velocity of the EE
- $\mathbf{A} = \{-1, 0, 1\}$. Acceleration in the x-axis.
- $\mathbf{R} = \begin{cases} -1 & \text{non-goal state at each timestep (200ms)} \\ 10 & \text{goal} \end{cases}$

The selected states - position and velocity of the EE are crucial for the task, as the goal involves reaching a specific position with a velocity constraint. The inclusion of both position and velocity in the state representation allows the agent to learn not only how to reach the goal but also to control its approach speed, adhering to the defined tolerances.

The sparse reward function, taken by [93] means that the agent does not have an explicit knowledge of the goal position, so reaching goal is crucial to it forming a representation of state values. This simplifies the learning process and implicitly penalizes the time taken to reach the target. This approach emphasizes the importance of reaching the goal efficiently, as each timestep without success incurs a penalty, thereby encouraging the agent to find quicker paths to the goal.

In Table 5.1, the settings for the hyper-parameters of the model are presented, which are based on [13], with the following two changes:

- **Activation Function:** The activation function was changed from Relu to Tanh. This modification was made due to Tanh's compatibility with Xavier initialization [36] and other papers like [37] where they used SAC with Tanh to determine the optimal steering angle for autonomous race track.
- **Entropy Target:** The entropy target was modified from $0.36 * (-\log(1/|A|))$ to $0.98 * (-\log(1/|A|))$, where A represents the action space with 3 possible actions. This adjustment aligns with the approach introduced in Christodoulou's paper on discrete SAC [38], where this specific entropy target formulation was proposed.

Table 5.1: discrete SAC settings

Hyper-parameter	[13] value	this thesis value
Layers	2 fully connected, 1 output	2 fully connected, 1 output
Fully connected layer units	32,32, moves available:3	32,32, moves available:3
Batch size	256	256
Replay buffer size	1000000	1000000
Discount rate	0.99	0.99
Learning rate Actor	0.0003	0.0003
Learning rate Critic	0.0003	0.0003
Learning rate Alpha temperature	0.001	0.001
Optimizer	Adam	Adam
Weight initializer	Xavier initialization	Xavier initialization
Activation function	Relu	**Tanh**
Networks update per off-line training	14.000	14.000
Loss function	Mean square error	Mean square error
Entropy target	0.36*(-log(1/|A|))	**0.98*(-log(1/|A|))**

5.1.3 Robot Control

Human and RL Control

Both human operators and RL agent are in charge of the EE trajectory in their axes through specified accelerations. These accelerations are integrated over time to derive the velocities commanded to the robot. Thus, the control strategy, taken from [13] includes a feedforward term on the acceleration as delineated by the following equations:

$$\dot{x}_{\text{com}} = u \tag{5.1}$$

$$\dot{x}_{\text{com}} = \dot{x}_{\text{com}} + T_c \cdot \dot{x}_{\text{com}} \tag{5.2}$$

Here, u symbolizes the acceleration determined by either the human or the RL agent, $\dot{x}_{\text{com}}$ is the velocity commanded, and T_c represents the control cycle, set to 0.008 seconds corresponding to the robot controllers' operation at a frequency of 125Hz.

During the HRC task, both entities are permitted to command a set of three distinct accelerations, which includes:

- A positive acceleration of $+a\,\mathrm{m/s^2}$
- A negative acceleration of $-a\,\mathrm{m/s^2}$
- Zero acceleration, or $0\,\mathrm{m/s^2}$

where $a > 0$ is a predetermined positive constant. This control schema introduces an inherent challenge within the game, reinforcing the notion of mutual learning and adaptation as the human participant masters the dynamics of robot motion in collaboration with the RL agent.

Reset Mechanism

Prior to the initiation of each HRC game, a feedback control law, concerning the EE's position is employed to navigate the EE to a predefined starting location. This preparatory control law, taken by [13], is formulated as:

$$\dot{x}_{\text{com}} = K_p \cdot (x_{\text{des}} - x_{\text{curr}}) \tag{5.3}$$

In this context, x_{des} signifies the desired EE position, x_{curr} is its current position, and K_p is a gain matrix constituting the proportional control element, specifically a 2×1 matrix in this scenario. This mechanism ensures the robot's motion is appropriately calibrated, setting the stage for the forthcoming HRC interactions.

5.2 Study design

There are two distinct participant groups:

1. **No Transfer Learning Group**: Participants in this group interact with a DRL agent 5.1.2 starting without prior training. The learning process of the agent is solely based on its interaction with the participant during the experiment.

2. **Transfer Learning Group**: This group's participants collaborate with a DRL agent that learns both thought the participant's and the expert's demonstration data.

5.2.1 No transfer Learning Group

The complete game consists of 110 trials, following the pipeline of the game without TL as presented in Figure 5.2.

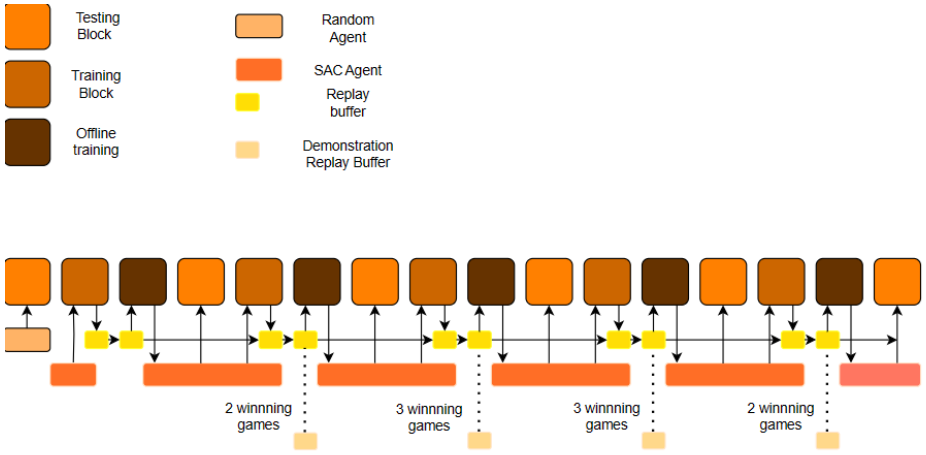


Figure 5.2: No Transfer learning Game Diagram

Game Structure

Each block in our experiment consists of 10 trials. We refer to a sequence of a training block, followed by offline gradient updates (G/U), and then a testing block, as a 'batch'. The data collected during the training blocks are stored in the replay buffer. This data is then utilized in the offline training sessions. Our primary focus for analysis and results will be on the testing blocks. It's important to note that the agent employs the same policy in both the training and testing blocks that immediately follow each offline training session. We could use only one block between each offline training but we adopted this approach to facilitate a direct comparison with the findings from Tsitos' previous work [13]. The parameters of the game are gathered in Table 5.2.

Table 5.2: Game Parameters

Game Parameters	**Values**
training games	50
Test games	60
Maximum games duration	30 secs
duration of RL agent	0.2 secs
Total update cycles	70000
win reward	10
non-win penalty	-1
goal position tolerance	0.01 m
goal velocity tolerance	0.05 m/s
maximum velocity in one axis	0.2

Familiarization

To provide participants the opportunity to assess their control over the robot, we conducted a preliminary test before the main 110 games. This test involved 10 games where the EE, could only move in the direction controlled by the human player. At the start of each game, the EE was placed in a starting position at the same height with the goal at the agnet's axis, so his contribution was not necessary. The player then had 10 seconds to move a red dot to a target area at a slow speed. This setup helped us assess the participant's skill in guiding the robot's movement. The target area and allowed position error were the same as in the main HRC game, but we set a stricter speed limit of 0.02m/s for this test. This approach was important because the robot's movement in the main games could influence the human's actions, and we wanted to understand each participant's basic ability to control the robot.

Baseline Block We refer to the initial testing block as the "baseline". In this block, all participants interact with a randomly acting agent, where each action is equally likely. The baseline block is the only shared phase between the two groups and serves to assess participant behavior and differences before the application of TL.

Initialized Agent

One of the changes we implemented is that, following the Baseline block, participants in each group now commence their first training block with an initialized agent, as opposed to Tsitos' work [13] where they interacted with a random agent.This approach was adopted due to the significant variance observed in the performance of randomly initialized agents. To address this, we conducted an experiment where an expert interacted with 15 different initialized agents, as illustrated in Figure 5.3. The performance of these agents varied greatly, with some nearly matching the efficiency of trained policies, while others struggled significantly.

In our study, we aimed to minimize the impact of this variance on participant performance and experience. Ideally, a larger sample size would have been used to absorb this variance, but due to practical constraints, we were limited to typical sample sizes of 10-20 participants.

Another potential solution was to refine the initialization process to achieve more consistent performance across agents. However, such a change could fundamentally alter the co-learning experience and limit our ability to compare our findings with similar studies. Therefore, we opted for a uniform initialization across all participants.

We selected the median-performing agent from our initial test (Run 4) as the standard initialization for all participants. This agent was used for the No TL group, ensuring a consistent starting experience for all participants and highlighting the influence of individual collaboration approaches on the learning process. At Section 5.2.2 we discuss about the Transfer Learning group's initialized agent.

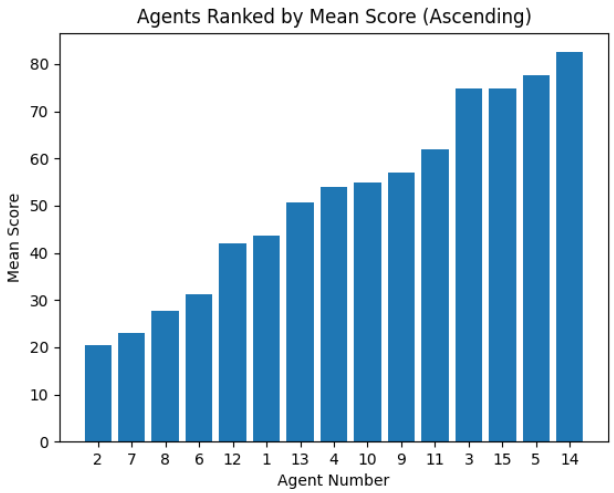


Figure 5.3: Performance of Different Initialized Agents

5.2.2 Transfer Learning Group

In our study, an "expert" is defined as an individual with extensive experience in the game, up to 30-40 hours of gameplay. The expert engages in the same game setup as a participant in the non-transfer learning group. Some of the interactions between the expert and the agent, particularly 10 successful games, are recorded (as illustrated

by the dotted lines in Figure 5.2). This collection of expert interactions formed the demonstration data buffer, which is used to transfer knowledge to the learning agent.

Our approach to TL in this game is inspired by the DQfD method [33], as detailed in Section 3.5.1. The application of DQfD in our HRC task is particularly beneficial for several reasons. Firstly, the expert demonstrations used in DQfD are directly relevant to our task, as both the experts and participants are involved in the same game environment. This direct applicability ensures that the knowledge transferred is useful. Secondly, there's a natural compatibility with the SAC framework, particularly through the Q Learning aspect of the SAC's critic network. This compatibility facilitates a smoother integration of the transfer learning process.

This method involves two distinct phases. In the initial phase, we focus on pre-training, which includes offline training exclusively with the demonstration replay buffer. This phase gives the agent an initial understanding of the environment by utilizing expert data. The second phase marks the beginning of the agent's interaction with human participants. Here, the agent's training during offline sessions incorporates a mix of expert demonstration data and data generated by the participants themselves. Notably, we gradually reduce the proportion of demonstration data in each subsequent offline training session. This strategy allows the agent to progressively adapt and its responses to the unique strategy of each participant.

Figure 5.4 illustrates the TL pipeline in detail. To address the issue of variance in agent initialization, as discussed in Section5.2.1, we employ the same initialized agent. The initial offline training session is conducted using this agent. The agent that emerges from this initial training session, having been exposed to the demonstration data, is then uniformly employed for all participants within the TL group.

Divergent from the original DQfD

In our methodology, we have chosen not to employ the two loss functions used in off-line training sessions in [33]: the Large Margin Classification Loss and the N-step Double Q-learning Loss with L2 Regularization. The Large Margin Classification Loss is designed to emphasize demonstration actions over other possibilities, creating a distinction between preferred and less optimal actions. This approach, could potentially restrict the agent's ability to adapt and personalize its policy to new users. On the other hand, the N-step Double Q-learning Loss, combined with L2 Regularization, is typically used to enhance generalization to new environments and improve sample efficiency. However, given the demonstrated capabilities of SAC algorithm in handling unseen scenarios and its sample efficiency, as evidenced in the works of Haarnoja [64] and Christodoulou [38], we decided instead of using these additional loss functions, a more straightforward approach for the off-line training session. The new agent is trained using only the demonstration replay buffer, without the modifications introduced by the additional loss functions. This approach is based on the confidence in the SAC's abilities and is aimed at ensuring that the agent can adapt more flexibly and personally to new users.

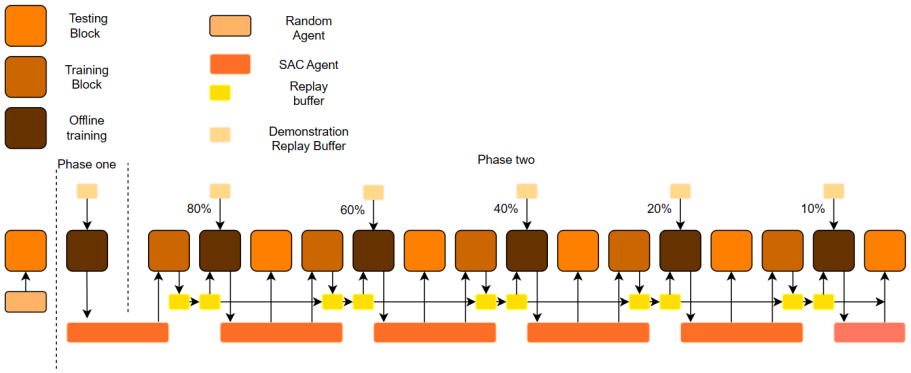


Figure 5.4: Transfer Learning Game Diagram

5.2.3 Experimental procedure

The participants were divided in 2 groups, half of them played the game without transfer learning while the other half played with LfD.

Before the beginning of the experiment a consent form where given to the participants informing them that their involvement was voluntary and that no personal information will be used against their will. The information letter and the consent form can be found in Appendices 5 The study protocol was approved by the Research Committee (REC) of NSCR "Democritus". After the consent form the participants requested to complete Questionnaire 1 which will be discussed in Section 5.3.3.

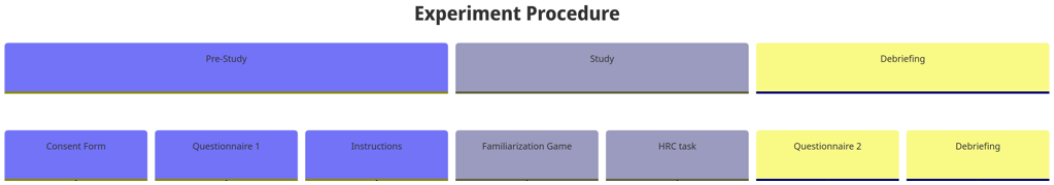


Figure 5.5: Experimental procedure

By the completion of Questionnaire 1 instructions where given to each participant, about the nature of the collaboration with the robot, the task they are sharing, the axis each member of the team controls, the total number of trials (110), Participants learned that each game could end in a "win" or "lose," each with distinct sounds. They were also briefed on the visualization module and that the games were divided into sets of 20. It was explained that the robot's movements were restricted to a designated rectangular area, as shown in Figure 5.1. Additionally, participants were assured of safety due to the robot's kinematic constraints and the emergency shutdown button. The participants where not informed about the DRL agent controlling the EE motion in the perpendicular (x-axis) were controlling neither the offline training of the agent every 20 games, neither their group (Transfer Learning, no Transfer Learning).

The instructions also informed the participants about the type of control they had at the (y-axis). Specifically the instruction given where:

- "i": The participant can move the EE away from him.
- "m": The participant can move the EE towards him

- "k": When pressing the "k" button, the participant commands the EE to continue moving the exact same way it was moving the moment he pressed the button.

The participants were directed to position themselves as shown in Figure 5.6, and started the familiarization game, as detailed in 5.2.1 in order to gain a clearer understanding and control over the robot's movement along one axis. Once the familiarization was concluded, they proceeded to the main HRC game, as described in 5.1.1. Following the HRC game, the participants requested to complete Questionnaire 2 5.3.2, which purpose is to evaluate the subjective measures of the collaboration. At the end a debriefing session was conducted. This session provided an opportunity for participants to share their experiences and thoughts on the collaborative effort, offering insights into the human-robot collaboration.

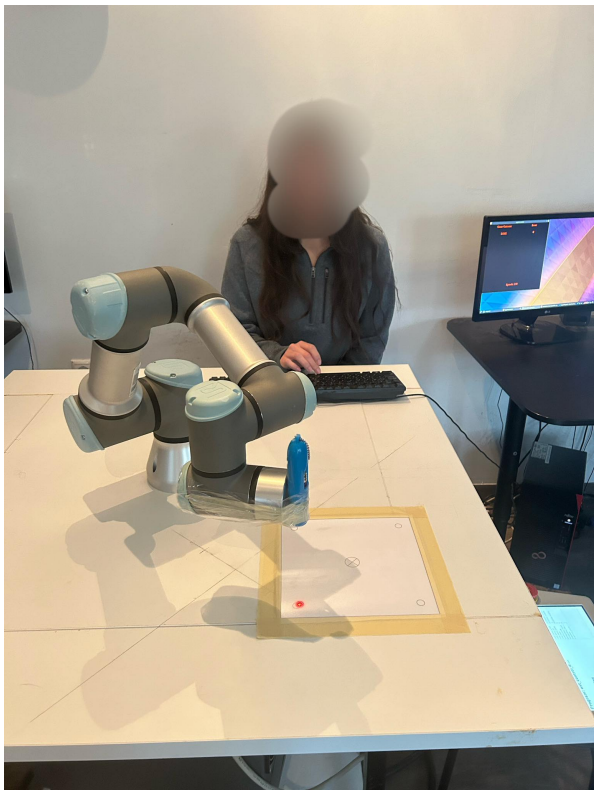


Figure 5.6: The Human-Robot Collaboration setup The robot is placed in the middle of 1m x 1m table.

5.3 Measures

5.3.1 Objective Measures

Objective measures are quantifiable indicators that allow for an impartial assessment. In our study, the following metrics were utilized to gauge performance:

- **Total Interaction Time:** The complete duration of active engagement between the human participant and the robot during each task.

- **Score:** Initiated from a base score of 150, the score decreases by one for every control frame during the game, aligning with the frequency of control frames (one every 200 ms) within the 30-second duration of a trial, resulting in 150 control frames per game.
- **Number of Wins:** The total count of successful trials within a series of game sessions, known as a game block.
- **Normalized Travel Distance:** This is the travelled distance multiplied by the percentage of the total time spent in a game.
- **Heatmaps:** Spatial representations showing the frequency of the end-effector's positions across the workspace during each block.

Statistical Approach for Objective Measures

Our analysis approach involved applying mixed ANOVA to analyze the data across different conditions and blocks, which allowed us to investigate the interaction effects comprehensively. We predetermined the alpha level for statistical significance at 0.05.

5.3.2 Subjective Measures

Subjective measures are essential in understanding how people perceive and interact with AI in collaborative tasks. These measures, based on each participant personal opinions and experience, help capture aspects which are challenging to quantify objectively. Many studies have presented subjective measures to evaluate better the experience of the human. Works in human-robot dialogue systems [97], conversation agents [98], explainable Ai [99] are some examples of focus for using subjective measures.

Our focus is on the validation of human-robot teaming fluency, so we focused on the proposed questionnaire by Hoffman. Our questionnaire 2, inspired by the work of Hoffman [39], focuses on six key aspects of collaboration: fluency in human-AI interaction, AI's contribution, team improvement, trust, training, and alliance, Figure 5.7.

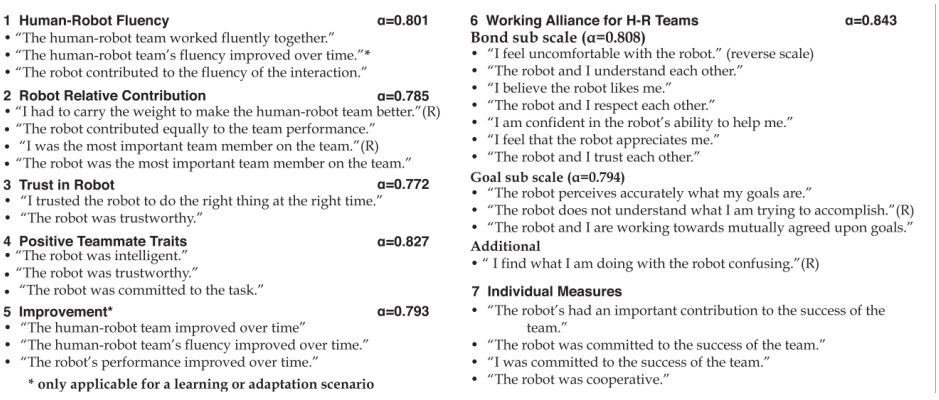


Figure 5.7: Subjective fluency metric scales and items used in our studies [39]

In each study the questionnaire changes to fit the objective. In [100] some of these measures are used to evaluate the relationship between human- robot interaction fluency

in job performance and job satisfaction. Another example is [101] where the classification in a human to robot grasping handover task is evaluated. Tsitos [13] used six of these measures, excluding the individual measure, in order to evaluate the difference in experience between the 2 groups.

In our study we developed Questionnaire 2 which is administered throughout and at the end of the interaction (the exact structure of the whole procedure will be explained at the next chapter). It is designed to capture the aspects of collaboration, reflecting the participant's evolving perceptions and experiences. The final Questionnaire 2 with [13] modifications can be found in Appendix .2.6

5.3.3 Understanding Participant Personalities in Human-Robot Collaboration

In research where humans are interacting with robots or artificial intelligence, the personalities of the human participants are of significant interest. More importantly in Human-Robot Collaboration (HRC), as an individual's personality forms their strategy during the interaction, making the robot adapt to it.

Previous works [102], [103] have utilized the Big Five personality trait questionnaire to better understand human experiences in Human Agent Collaboration. Additionally, a custom scale was used in [104] to explore future directions for personality research, focusing on how advances in information technology, such as AI and robotics, will require an understanding of individual traits.

That's why we developed 'Questionnaire 1', a set of questions helps us get a full picture of each person's personality and what they think about AI. We ask the the participants to answer this before they start the game.

Questionnaire 1

'Questionnaire 1' is made up of four different parts, each looking at different characteristics about a person:

1. **Big Five Personality Traits**
2. **Schwartz Portrait Values Questionnaire (PVQ)**
3. **AI Attitude Scale**
4. **Additional Personal Questions**

Big Five Personality Traits

The Big Five Personality Traits questionnaire, which can be found in Appendix .2.1, includes 50 questions that help us understand five key aspects of personality. This questionnaire focus on five main characteristics:

- **Extraversion**: This trait measures how outgoing and social a person is.
- **Agreeableness**: This trait looks at how kind, cooperative, and compassionate a person is. .

- **Conscientiousness**: This trait assesses how organized, responsible, and hard-working a person is. Highly conscientious individuals are usually very reliable and well-organized.
- **Emotional Stability/Neuroticism**: This trait measures how calm and emotionally stable a person is..
- **Openness to Experience**: This trait evaluates how open-minded, imaginative, and curious a person is.

The Big Five model, is a well-established framework in psychology. It suggests that a personality can be described using these five basic dimensions.

The questionnaire we use is based on the work of Goldberg [105] and translated by Tsaousi [106]. It consists of 50 items, with 10 questions for each personality dimension. These questions, divided by each trait, are provided in Appendix .2.1.

Schwartz Portrait Values Questionnaire

The Schwartz Portrait Values Questionnaire (PVQ) [107], developed by Shalom H. Schwartz, is a version of the original Schwartz Value Survey (SVS) [108]. Our study utilizes the PVQ-21, a shorter variant with 21 items. This version was selected for its ability to maintain participant engagement without loosing the depth of insight. The PVQ-21 captures Schwartz's ten basic human values, each representing unique motivational goals:

- **Self-Direction**: Emphasizing independent thought, creativity, and personal freedom.
- **Stimulation**: Focused on seeking excitement, novelty, and adventure.
- **Hedonism**: The pursuit of pleasure and enjoyment.
- **Achievement**: Aiming for personal success and demonstrating competence.
- **Power**: The desire for control, influence, and social status.
- **Security**: Prioritizing safety, stability, and order.
- **Conformity**: Adherence to social norms and expectations.
- **Tradition**: Respect for customs, cultural heritage, and traditional values.
- **Benevolence**: Concern for the welfare of others, showing empathy and compassion.
- **Universalism**: Valuing social justice, equality, and environmental sustainability.

Each question in the PVQ-21, in both English and Greek, are presented in .2.2

Attitudes towards AI Scale

The AI Attitude Scale helps us understand how people feel and think about AI. Schepman's at [109] developed 3 questionnaires :

- The first had two types of questions about AI. Some questions were positive, like "Artificial intelligence has many good uses." Others were negative, like "Artificial intelligence could take away people's jobs." There were 32 questions in total.
- The second asked 42 questions about how comfortable people are with AI doing things like "Translating languages in real-time" or "Helping police decide if someone might do a crime again."
- The third was similar to the second but compared what AI can do against what humans can do.

Out of the first part, 20 questions were kept. 7 were removed due to high association, and 5 were taken out because of Exploratory Factor Analysis. This final set of 20 questions, with 12 positive and 8 negative ones, is what we call the AI Attitude Scale.

Schepman tested this questionnaire with 100 people to see if it really shows how people generally feel about AI. He compared it with the second and third parts, which were more specific. In another study [110], Schepman looked at whether things like personality traits can be linked to how people feel about AI. For this, he used the Big Five personality traits and asked 300 people. The AI Attitude scale questionnaire can be found in Appendix .2.3

Additional Personal Questions

This section gathers more specific information about the participants, including: **Personal Information:** Age, gender, dominant hand, and any eye/neurological problems (5 questions). **Gaming Experience:** Assessing the participant's familiarity and experience with gaming (2 questions). **Knowledge about AI:** Questions to gauge the participant's understanding and awareness of AI technologies (3 questions). These questions can be found at Appendix .2.4.

Chapter 6

Results

In this chapter, the results of the HRC study are presented. The chapter covers the participants' characteristics, including demographics and personal attributes like gender, age, handedness, gaming experience, and preferred devices.

The analysis compares personality traits and values between groups using the Big Five personality traits (Big 5) and Schwartz Portrait Values Questionnaire (PVQ). It also examines participants' attitudes towards AI.

Objective results such as interaction times, learning curves, and performance metrics are detailed. This section highlights the differences and similarities in strategies and behaviors between groups over the course of the experiment.

A statistical analysis is conducted to validate the findings, focusing on performance differences between groups. Finally, the chapter explores subjective experiences, including participants' perceptions of control and collaboration quality.

The chapter also contrasts these findings with the previous study [13] and introduces a new follow-up study, with modified settings. This analysis explores the impact of these changes on the experiment's outcomes, offering insights into how alterations in the setup influence participant responses and interaction dynamics.

6.1 Results - Main Study

These experiments included 16 participants, 8 for each group, with their characteristics presented in Table 6.1:

Table 6.1: Characteristics of Study Participants

Characteristic	**Details**
Gender Distribution	5 women, 11 men
Age Range	16 - 31 years
Handedness	13 right-handed, 3 left-handed
Gaming Experience	8 with >5 years, 2 with 3-5 years, 1 with <1 year, 5 with none
Preferred Gaming Devices	10 laptops, 3 consoles, 2 mobiles, 1 none

Big 5 and PVQ

The comparison of the Big Five personality traits between the two groups, as shown in 6.1 and 6.2, shows a similarity in their personality profiles. This resemblance suggests that, in terms of HRC, both groups are likely to exhibit similar behaviors and attitudes towards the robot.

When examining the PVQ 21 results, as depicted in 6.4 for the TL group and 6.3 for the No TL group, there are differences in certain traits like universalism, stimulation, hedonism, and achievement. The TL group shows marginally higher values in these areas, which could imply a slightly more enthusiastic and creative approach to human-robot interactions. They might exhibit a greater openness to new experiences and a stronger motivation to achieve goals during tasks. However, it's important to note that these differences are relatively minor and do not significantly deviate from the overall similarity in personality composition between the two groups.

While there are some differences in certain aspects of their personality traits, the two groups exhibit a similar personality composition. This similarity indicates that both groups are likely to approach human-robot interactions in comparable way.

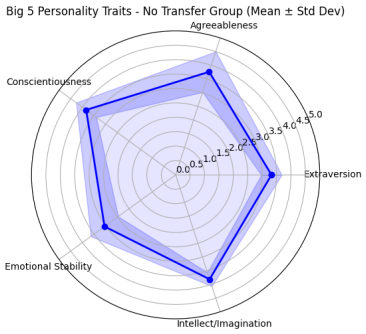


Figure 6.1: Big 5 No TL group

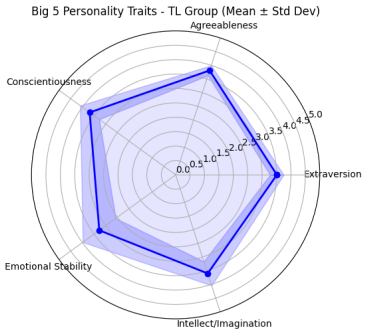


Figure 6.2: Big 5 TL group

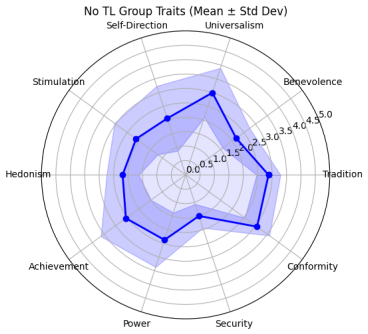


Figure 6.3: PVQ 21 no TL group

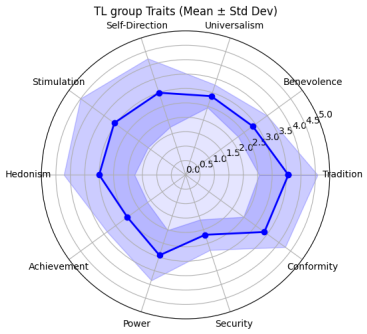


Figure 6.4: PVQ 21 TL group

Attitude towards AI

Both groups, No TL and TL, generally view AI positively, as shown in Table 6.2. The TL group seems a bit more positive about AI, which might make them more trusting

in working with the robot during the task. The No TL group, while also positive, might be a bit more cautious and curious, which could lead them to try out new ways thought the interaction. The assessment of attitudes towards AI concentrates solely in the difference between the positive and negative mean values.

Table 6.2: Attitude towards AI

Group	Mean score	Std
No TL	0.488	0.406
TL	0.541	0.369

Overall, based on the above answers, the two groups, do not exhibit significant differences in their characteristics and attitudes. Both groups show similar profiles in terms of the Big Five personality traits, suggesting comparable behavioral tendencies in their interactions. While there are some variations in the PVQ 21 results, with the TL group displaying slightly higher values in certain traits, these differences do not appear to be substantial enough to indicate a distinct divergence in their approach to AI and robotics. The attitudes towards AI, as reflected in the mean scores, are generally positive in both groups, with the TL group showing a little more positive outlook. However, this difference is not important, indicating that both groups are likely to engage with AI and robotics in similar ways, with perhaps minor variations in their levels of enthusiasm and trust.

6.1.1 Objective Results

In the test games, the TL group exhibited an average interaction time of 21 minutes with the agent, with a standard deviation of 2.6 minutes. In contrast, the NO TL group had a mean interaction time of 26 minutes, with a standard deviation of 1.4 minutes.

A visual representation of the learning progress over the 60 testing games, is presented to Figures 6.5 and 6.6, which depict the learning curves for Wins and Rewards, across both groups, as well as the expert's performance. These figures illustrate the strategies and performance levels of the TL and No TL groups, giving an insight on how they adapt and improve over time. As the experiment progresses, the TL group performance is close to the expert's. The No TL group maintains a relatively stable performance, while the expert shows an overall good performance thought the whole interaction.

Additionally, in Figure 6.7, we present the normalized traveled distance which is the travelled distance multiplied by the percentage of the total time spent in a game.

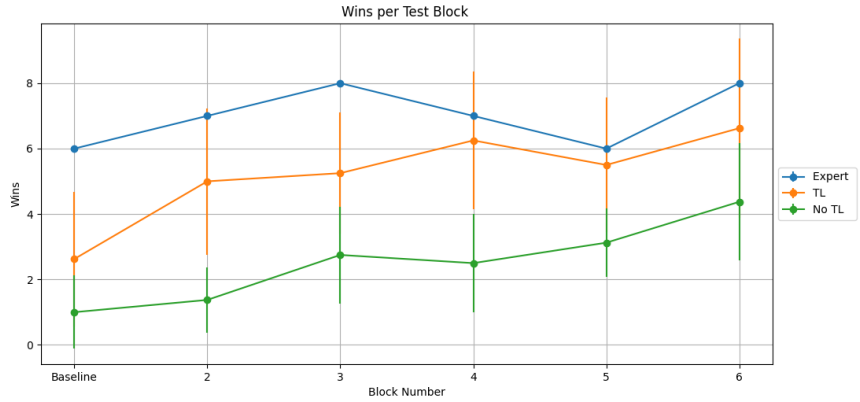


Figure 6.5: Wins over the testing blocks

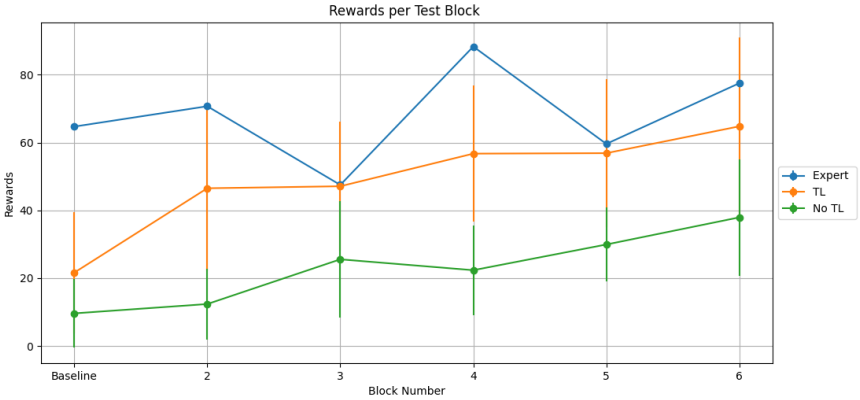


Figure 6.6: Rewards over the testing blocks

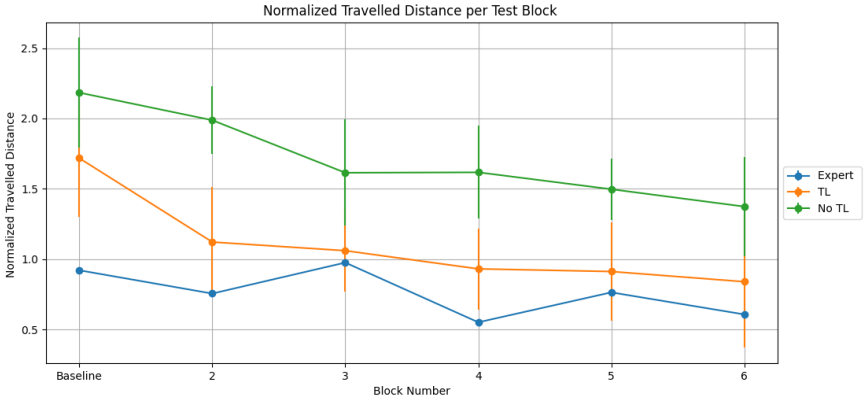


Figure 6.7: Normalized Traveled Distance over the testing blocks

Heatmaps are a key tool for analyzing group behavior, as depicted in Figure 6.8. The expert, who has extensive experience with the game, aims to teach the goal to the

agent from the start, unlike the baseline games. of the participants, where interactions with a random agent result in a widespread occupancy across the grid. This widespread distribution signifies an exploratory phase, where teams are getting accustomed to the environment and the task, indicating an absence of targeted learning initially.

Moving through the experiment, in the TL group, a noticeable pattern emerges. A consistently appearing high-frequency line on a specific point of the x-axis demonstrates the group's repeated engagement with a certain path. This pattern points to the transfer of knowledge from the expert demonstrations, signaling effective learning within the TL group. However, the expert's strategy seems to constrain personal strategies. In contrast, the No TL group, which does not benefit from expert data, exhibits a broader spectrum of strategies. This suggests that while the expert's strategy aids in focused learning for the TL group, it might limit the development of individual strategies, as evidenced by the more diverse approaches in the No TL group.

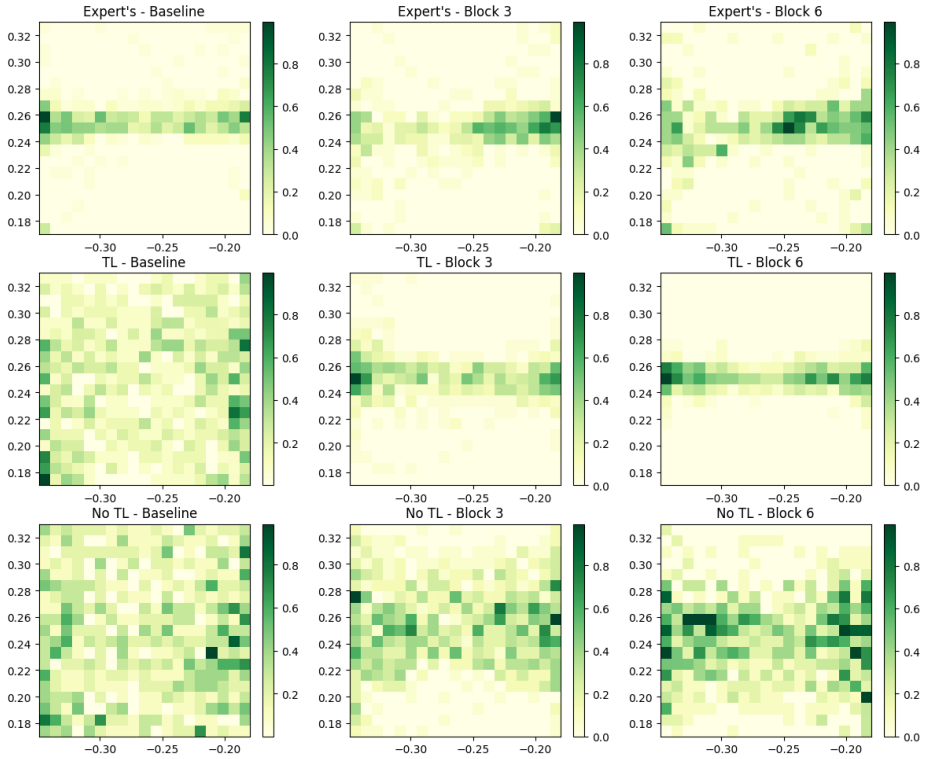


Figure 6.8: Testing blocks behaviour of the team with the expert human (1st row), a team in the LfD group (2nd row) and a team in the No TL group (3rd row). The heatmaps show the laser dot's position in the Baseline, 3th and last testing blocks. The numbers indicate the frequency with which the dot occupied each cell (1cm × 1cm) - that is the normilized number of x, y pairs counted within the cell in all ten games of a batch.

Statistical results

Throughout our analysis, we aimed to assess the differences of the two groups and evaluate the impact of TL on participant performance. Our initial normality tests revealed that neither rewards nor wins data followed a normal distribution in both the first and last blocks of 10 games, so we proceeded with non-parametric statistical tests.

In the first block, which serves as a baseline comparison, no statistically significant differences were detected in either rewards or wins between the TL and NO TL groups. However, in the last block, we observed a statistically significant difference in rewards and rewards between the two groups, indicating that TL had an impact on cumulative rewards. To summarize these results, we have created a Table 6.3 presenting the Mann-Whitney U test results for both the first and last blocks.

Block	Variable	Mann-Whitney U Statistic	P-value
First (Baseline)	Rewards	45.5	0.165
First (Baseline)	Wins	47.0	0.119
Last	Rewards	52.0	0.038
Last	Wins	51.5	0.043

Table 6.3: Statistical Test Results for First and Last Blocks

Furthermore, we employed a robust two-way mixed ANOVA using the `bwtrim` function from the `WRS2` package [40] to assess the efficacy of TL across different blocks of the experiment. This analysis was designed to evaluate both within-group changes (across blocks-time) and between-group differences (TL vs. No TL groups). The results are presented in Table6.4

The analysis revealed:

A **highly significant group effect** (p-value < 0.0001), indicating a substantial difference in normalized distance between the TL and No TL groups. This finding supports the hypothesis that Transfer Learning positively impacts performance, with the TL group showing more favorable results compared to the No TL group.

A **significant block effect :** (p-value $= 0.0010$) was observed, suggesting that normalized distance changes meaningfully across different blocks within each group. This effect is indicative of a learning or adaptation process occurring over time.

Non-Significant Interaction Effect: The interaction between group and block was not significant (p-value $= 0.7168$), implying that the pattern of change across blocks is consistent between the TL and No TL groups. Both groups exhibit similar trajectories of improvement or change over time.

Effect	Test Statistic	p-value
Block	8.0571	0.0010
Group	78.0173	< 0.001
Block:Group	0.5774	0.7168

Table 6.4: Two-Way Mixed ANOVA results on normalized travelled distance

The significant group effect underlines the effectiveness of Transfer Learning in enhancing performance metrics, as measured by normalized distance. The lack of a significant interaction effect suggests that the improvement trajectory, while present, does not differ between the TL and No TL groups across the experiment's duration.

6.1.2 Subjective Results

Judgement of Control

In the analysis of perceived control over the robotic arm, participants were asked for their ability to judge their control, across six different intervals (after the baseline block and at each offline training session). The question was:

"How would you rate your ability to control arm movement in the last 10 games from 1 (no control) to 9 (complete control)?"

The plot in Figure 6.9 visually suggests improvements in Judgment of Control (JOC) over time for both the TL Group and the No TL, with the TL Group appearing to have higher scores. This might implies a greater sense of control mastery in the TL Group.

However, a detailed statistical analysis using a robust two-way mixed ANOVA reveals a different aspect. The results show statistical significance only in the improvement of JOC scores over time (p=0.0001), not between the groups (p=0.246) or in the interaction between group and time (p=0.53). This indicates that both groups improved their perceived control ability over time, regardless of the transfer learning aspect.

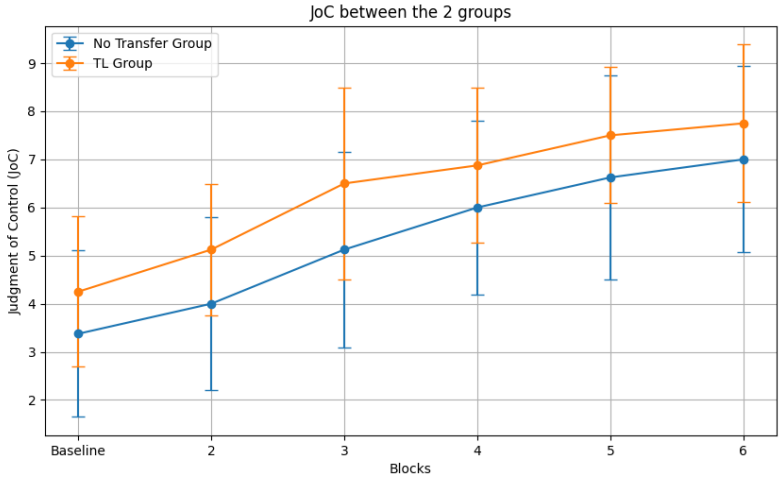


Figure 6.9: Judgement of Control

Collaboration Metrics

The collaboration metrics presented in Figure 6.10 utilize a Likert scale to evaluate the performance of the two groups—TL Group and No TL—across various dimensions of teamwork, Questionnaire 2 as presented in .2.6.

With scores exceeding 3 in Fluency, Teammate Traits, and Improvement, both groups are seen to have favorable outcomes, indicating good collaboration quality in these areas. These higher scores suggest that participants found the interactions smooth, valued the qualities of their teammates positively, and recognized notable progress in their collaborative efforts throughout the task.

Contribution and Trust metrics hover around the midpoint of 3 on the scale, reflecting a neutral or moderate stance, where participants neither strongly agreed nor

disagreed with the statements related to their engagement levels or their trust in teammates. This might imply that while the collaboration was functional, there is room for improvement in how individuals feel about the agent's input and the reliability of their team members. Nevertheless, this is not surprising given the overall objective performance.

Notably, the Alliance metric scores near 2 for both groups, indicating a less favorable perception of team unity. This lower score may point to potential issues in forming a strong partnership or bond within the teams, which could be a concern for tasks requiring high levels of cooperation and shared goals.

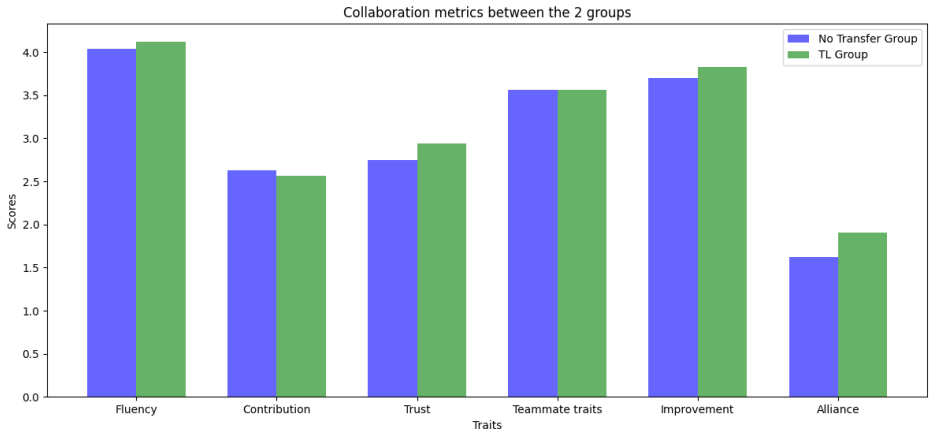


Figure 6.10: Collaboration Metrics

6.1.3 Comparison with previous work

This section presents a comparative analysis of our research, which employs a Learning from Demonstrations transfer learning approach, with the study by Tsitos [13], which utilized a Policy Reuse method, in the same task. There are 4 main differences between the 2 works:

- **Entropy Target Parameter:** We had the target entropy parameter set at $0.98 \times (-\log(\frac{1}{|A|}))$, while [13] applied a smaller target entropy of $0.36 \times (-\log(\frac{1}{|A|}))$ 5.1.2.
- **Activation Function of SAC's Neural Networks:** In our approach, the activation function was Tanh, while [13] used ReLU 5.1.2.
- **Number of Batches:** [13] had 7 batches in his study, while we had 5 5.1.1.
- **Initialization of Agent in First Training Block:** In our approach, participants in each group begin their first training block with an initialized agent, while in [13] work, participants interacted with a random agent at the first training block 5.2.1.

In Figure 6.11, we compare the rewards over testing blocks for the PPR method with ours in Figure6.6. This comparison reveals three main differences:

1. **Expert Performance:** Our study shows the expert's rewards maintaining a mean value of around 70. In contrast, [13] shows the expert starting with lower scores but quickly reaching the score of 140 from the second block and converges there.

2. **TL Group Performance:** In [13], the TL group shows a gradual learning curve, reaching expert-level performance by the seventh block. Our method reveals a more rapid initial improvement but a plateau at a lower reward level, around 70.

3. **No TL Group Performance:** [13] reports minimal improvement in the No TL group, with a maximum reward of 10 in the last block. In our study, the No TL group starts at a mean reward of 10 and shows observable learning, reaching a mean reward of 40 in the final block.

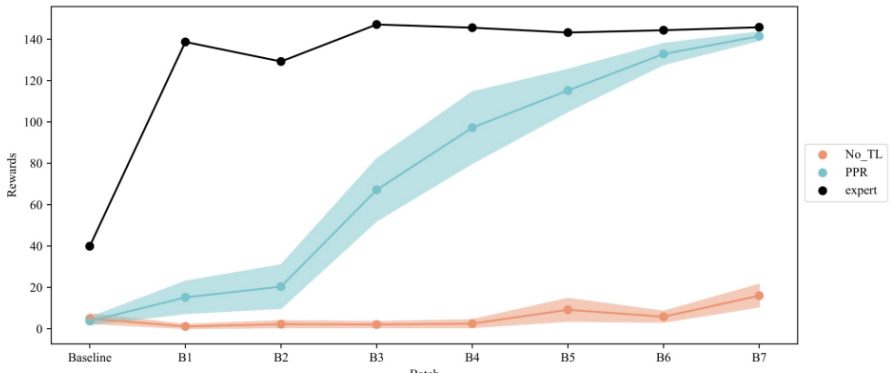


Figure 6.11: Rewards of PPR method,[13]

At first, we considered that these differences could be partly attributed to variations in the entropy target objectives between the two studies. To assess this, we conducted experiments with different target entropies during the expert-agent collaboration, some of which are presented in Appendix .1.

In summary, the results revealed unstable learning under different entropy settings during interactions with the expert. The agent faced challenges in consistently learning an effective policy. A high entropy setting of $0.98 \times (-\log(\frac{1}{|A|}))$, which encouraged exploration,made it harder to identify the underlying problem. Based on these findings, it is suggested that the change to the Tanh activation function may have contributed, potentially leading to a vanishing gradient problem [41]. Given these observations, our analysis shifted towards examining the impact of the activation function change. Therefore, in light of these results, we advise against employing SAC with Tanh activation within this particular experimental framework, as the focus on activation function alterations emerged as a critical area based on the challenges encountered with entropy settings.

6.2 Follow-up study

The activation function was reverted to its original setting (from Tanh to ReLU), and additional experiments were conducted using a target entropy of $0.36 \times (-\log(\frac{1}{|A|}))$

along with the ReLU activation function. Furthermore, three more batches were added to align with [13], and these experiments are referred to follow-up study. The Big 5 and PVQ were not included in the of follow-up study experiments.

These experiments included 8 participants with their characteristics presented in 6.5:

Table 6.5: Characteristics of follow-up study Participants

Characteristic	**Details**
Gender Distribution	4 women, 4 men
Age Range	22-24 years
Handedness	8 right-handed
Gaming Experience	4 with >5 years, 1 with 3-5 years, 1 with <1 year, 2 with none
Preferred Gaming Devices	2 laptops, 2 consoles, 2 mobiles, 2 none

Attitude towards AI

Both groups, No TL and TL, generally view AI positively, as shown in Table 6.6. The Attitude towards AI exhibited similar results, with the TL group having a more positive view of AI. However, based on the statistical test result showing a p-value of approximately 0.141, there were no significant differences between the groups' attitudes towards AI.

Table 6.6: Attitude towards AI follow-up study

Group	Mean score	Std
No TL	0.448	0.621
TL	0.666	0.660

6.2.1 Objective Results

In the test games average interaction time, the two groups had a big difference compared with the previous. Specifically the TL group exhibited an average interaction time of 11.45 minutes with the agent, with a standard deviation of 0.76 minutes. In contrast, the NO TL group had a mean interaction time of 29.7 minutes, with a standard deviation of 6.35 minutes.

A visual representation of the learning progress over the 80 testing games of follow-up experiments, is presented to Figures 6.12a and 6.12b, which depict the learning curves for Wins and Rewards, across both groups, as well as the expert's performance. Additionally, in Figure 6.12c, we present the normalized traveled distance which is the travelled distance multiplied by the percentage of the total time spent in a game.

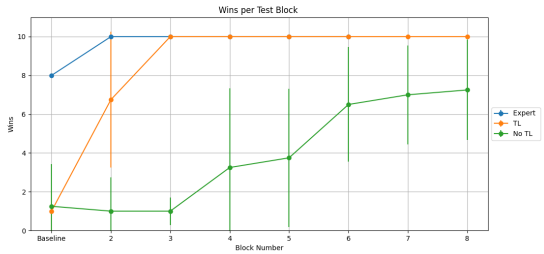
Focusing on the rewards learning curve, Figure 6.12b we can observe:

Given that the expert reached the maximum reward from the second test block (similar to [13]), we have chosen to maintain consistency with previous experiments by having the expert participate for an identical number of blocks. Additionally, we will

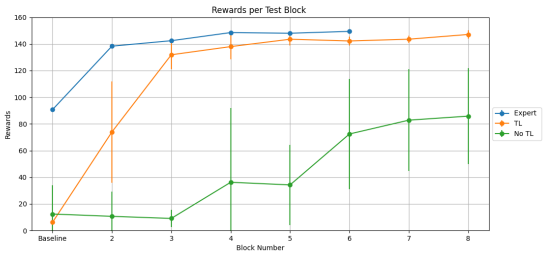
preserve the same proportions of demonstration data from their interactions, mirroring the methodology of our earlier study.

For the TL group, observations indicate that although there is variance in the second testing block, the group consistently achieves a reward of 140 from the third testing block onwards. This performance contrasts with our previous study, in which the maximum reward reached was 60 by the 6th block. Additionally, this outcome differs from [13], where the TL group, employing PPR TL, reached a performance comparable to that of an expert by the final (6th) block.

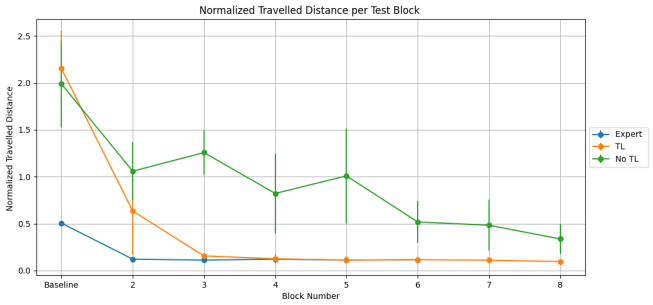
The No TL group achieved a reward of 80 at the 6th block, which is a significant improvement from the 40 achieved under previous conditions at the same stage. A notable comparison is drawn with the "No TL" group's performance during the follow-up study, which reached a reward of 80 at the 6th block, contrasting with findings from [13], where the mean reward at this stage was only 20, as shown in Fig. 6.11. This improvement can likely be linked to the use of an initialized agent Section 5.2.1 during the first training block, marking the only difference between the two "No TL" groups' approaches. We should mention that this might not be the case as it could be a result of better participants in the TL group and also less participants than [13].



(a) Wins over the testing blocks of follow-up study



(b) Rewards over the testing blocks of follow-up study

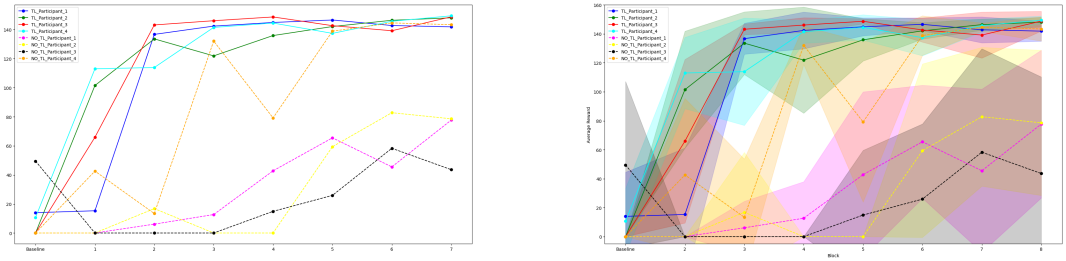


(c) Normalized Traveled Distance over the testing blocks of follow-up study

Figure 6.12: Wins, Rewards and Normalized Traveled Distance of follow-up study

In Figures 6.13a and 6.13b, the individual performance of each participant is visually represented. An interesting observation can be made from these plots. Specifically, it is notable that Participant No TL 4 appears to be converging towards a reward pattern similar to that of the TL group, particularly reaching a reward level of 140. This convergence in performance is reminiscent of a similar behavior observed in a

participant in [13], indicating that with this entropy setting, participants can reach the optimal behaviour but not as fast TL participants.



(a) Testing blocks rewards for all participants (without standard deviation)

(b) Testing blocks rewards for all participants (with standard deviation)

Figure 6.13: Combined Figures for of follow-up study Rewards

The heatmaps for the of follow-up study, depicted in Figure 6.14, distinctly highlight the effects of transfer learning within the TL group. Notably, an 'X' shaped pattern emerges in the final block, which not only aligns closely with the patterns observed in the expert's heatmaps (Figure 6.15) but also represents the optimal interaction strategy that minimizes completion time. This pattern underscores the successful transfer of strategic knowledge, as the TL group's learning buffer was effectively enriched with interactions drawn from the expert's demonstrations.

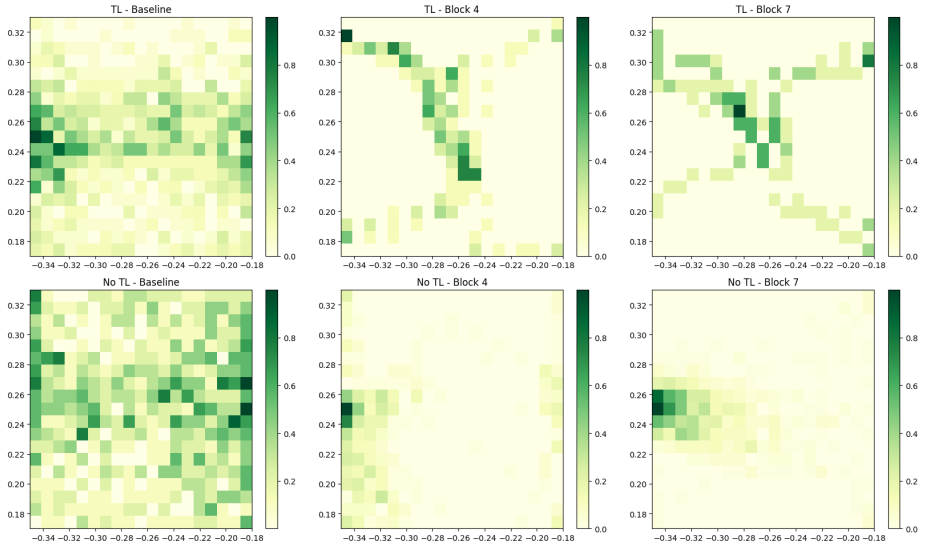


Figure 6.14: Testing blocks (of follow-up study) behaviour of the a team in the LfD group (1st row) and a team in the No TL group (2nd row). The heatmaps show the laser dot's position in the Baseline, 3th and last testing blocks. The numbers indicate the frequency with which the dot occupied each cell (1cm × 1cm) - that is the number of x, y pairs counted within the cell in all ten games of a batch.

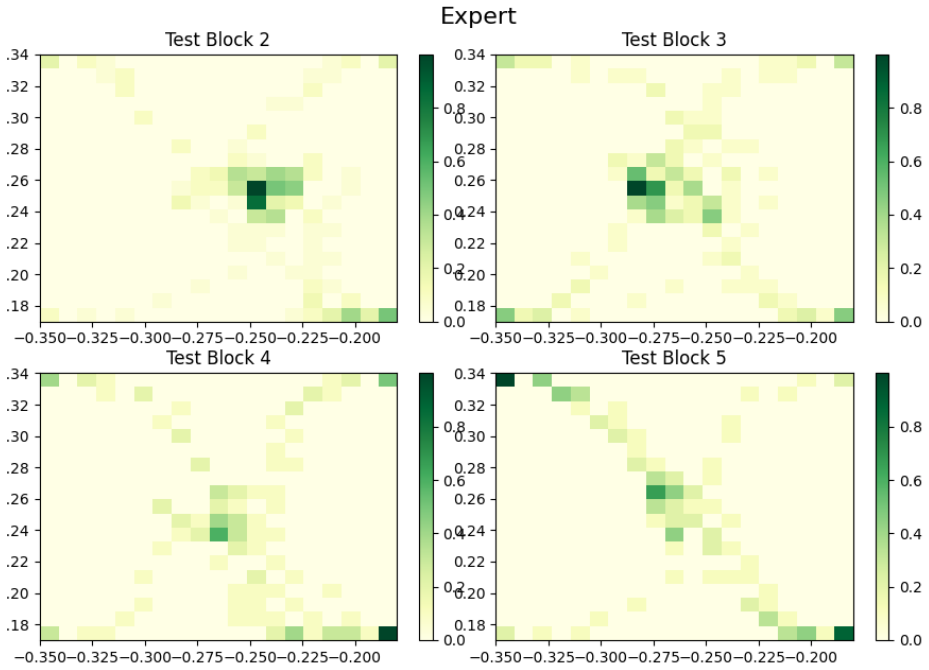


Figure 6.15: Testing blocks of the expert of follow-up study)

Statistical results

Furthermore, we employed a robust two-way mixed ANOVA using the `bwtrim` function from the `WRS2` package [40] to assess the efficacy of TL across different blocks of the experiment. This analysis was designed to evaluate both within-group changes (across blocks-time) and between-group differences (TL vs. No TL groups). The results are presented in 6.7

The analysis revealed:

Effect of Group: The analysis revealed a highly significant group effect (p-value < 0.001), indicating a substantial difference in normalized distance between the TL and No TL groups. This finding supports the hypothesis that Transfer Learning positively impacts performance, with the TL group showing more favorable results compared to the No TL group.

Effect of Block: A significant block effect was also observed (p-value $= 0.0010$), suggesting that normalized distance changes meaningfully across different blocks within each group. This effect is indicative of a learning or adaptation process occurring over time.

Interaction Effect: The interaction between group and block was significant (p-value $= 0.00382$), implying that the pattern of change across blocks is consistent between the TL and No TL groups. Both groups exhibit similar trajectories of improvement or change over time.

The statistically significant results observed for all our variables underscore the effectiveness of the LfD method, especially with the new settings incorporating the SAC's adjustments (ReLU activation function combined with a $0.36 \times (-\log(\frac{1}{|A|}))$ entropy setting).

Effect	Test Statistic	p-value
Block	24.042	< 0.001
Group	26.135	< 0.001
Block:Group	3.539	0.00382

Table 6.7: Two-Way Mixed ANOVA results on normalized travelled distance

6.2.2 Subjective Results

Judgement of Control

The plot in Figure 6.16 visually suggests improvements in Judgment of Control (JOC) over the end of each testing block (8) for both the TL Group and the No TL, with the TL Group appearing to have higher scores. This might implies a greater sense of control mastery in the TL Group.

The results show statistical significance only in the improvement of JOC scores over time (p-value < 0.001), between the groups (p-value < 0.001) but not in the interaction between group and time (p=0.53).

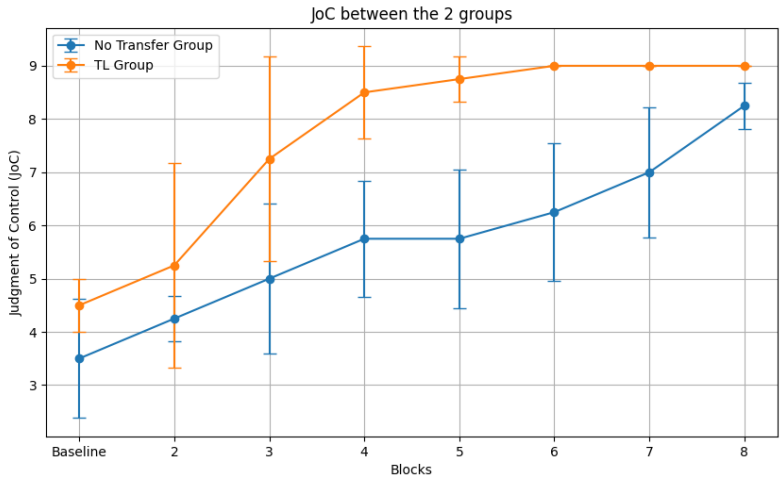


Figure 6.16: Judgement of Control

Collaboration Metrics

In the follow-up study, we observed that the TL Group consistently achieved higher scores than the No TL across multiple collaboration metrics 6.17. Specifically, the TL Group outperformed the No TL in the dimensions of Fluency, Teammate Traits, Improvement and Trust. These higher scores suggest that participants in the TL Group experienced smoother interactions, held more positive perceptions of their teammates, demonstrated notable progress in their collaborative efforts, and reported a better sense of team compared to the No TL.

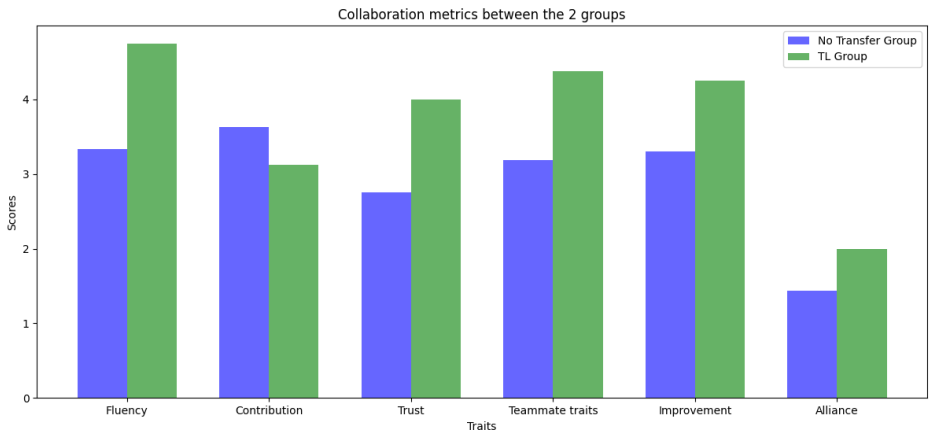


Figure 6.17: Collaboration Metrics of follow-up study

Chapter 7

Conclusion and Discussion

In [13], a human-agent collaboration game was developed to assess the interaction dynamics between humans and a Soft Actor-Critic agent. This game had a human and an agent with the joint control of a robotic manipulator's end-effector movement in the xy-plane, where the human was responsible for the y-axis and the agent for the x-axis. The objective was to move the end-effector to a specified target location at a controlled with low speed. This setup used as a method to explore the efficacy of leveraging an expert-trained agent's policy (probabilistic policy reuse method) for Transfer Learning in facilitating the collaboration between a RL agent and a human. The investigation aimed to understand the impact of policy reuse on the learning curve and performance in collaborative tasks.

Building upon this work [13], this thesis explored another approach by adopting a Learning from Demonstrations Transfer learning method called Deep Q-learning by Demonstrations, with a focus on the direct transfer of expertise through demonstrations. This adjustment aimed to assess and compare the relative effectiveness of policy reuse with Learning from Demonstrations in enhancing the human robot collaboration. Furthermore, our study introduced modifications to the SAC algorithm's parameters, including an increase in target entropy from $0.36 \times (-\log(\frac{1}{|A|}))$ to $0.98 \times (-\log(\frac{1}{|A|}))$ to encourage exploration and a switch from ReLU to the tanh activation function.

The results from 16 participants indicated that these modifications did not successfully promote an efficient transfer of expert knowledge. Subsequent analyses with varying target entropy values suggested that the employment of the tanh activation function led to unstable learning results. Follow-up study, with 8 participants, reverting to the original settings while implementing our Learning from Demonstrations method, yielded more promising results. The overall interaction time for the Transfer Learning group was significantly reduced.

The Transfer Learning group achieved earlier convergence compared to the original study's outcomes using policy reuse [13]. Remarkably, our findings indicate that convergence was attained as early as the second training block, a significant acceleration in learning pace compared to the previous study, which only saw convergence by the seventh block. This advancement underscores the efficiency of the Transfer Learning approach in facilitating rapid adaptation and learning within the given tasks, highlighting its potential to significantly reduce the time required for the human-agent team to achieve optimal performance levels.

Additionally, we observed that initializing the agent in the first training block posi-

tively impacted the rewards in the No TL group, surpassing those in the original study [13] which commenced with an agent taking random actions. However, this improvement is not conclusively attributed solely to the methodological changes. It might also be due to the relatively small sample size of participants involved in the study or potentially "better" participants in the TL group. These factors introduce variability that could skew the observed outcomes, suggesting that further research with a larger and more diverse participant pool is necessary to validate these findings definitively.

In acknowledging the limitations of our study, it's important to note several factors that could influence the outcomes and interpretations of our findings. Firstly, the sample size of our participant pool was relatively small, with 16 participants in the initial phase and 8 in the second. This limitation restricts the generalizability of our results to broader contexts. Additionally, the specificity of the task may not fully encapsulate the complexities and variabilities found in real-world human-robot collaboration scenarios. Our study's reliance on a single learning method, Learning from Demonstrations, for Transfer Learning, also presents a limitation. While providing valuable insights, this focus may overlook the potential benefits of other TL methods.

For future work, we recommend further exploration into the implementation of Learning from Demonstrations with a higher target entropy parameter, giving more room for exploration to the agent, which could facilitate more personalized adjustments for each participant. Additionally, it would be valuable to conduct experiments to determine the optimal percentage of expert knowledge transfer, also allowing for personalization in the learning process.

Appendix A

Appendix Title

.1 Entropy Tuning

Based on the results from our comparative analysis with Tsitos' study [13], this section explores the critical role of entropy tuning in SAC algorithm. The differences observed in the performance underscore the significance of the entropy target parameter in determining learning outcomes. Our objective is to examine how adjustments in entropy levels can influence the learning curve and strategy development in transfer learning scenarios.

We conducted a series of experiments involving an expert with consistent performance in the task, focusing on testing different entropy targets. These experiments are aimed at gaining a deeper understanding of the impact of entropy parameter adjustments on the learning process.

The importance of entropy tuning in SAC is highlighted by recent research in the field. For instance, in [111] Meta-SAC is introduced, which builds upon the SAC algorithm [64] and utilize metagradient along with a novel meta loss objective to automatically tune the entropy temperature in SAC. This approach has shown promising results, outperforming previous methods like SAC-v2 [64] in complex tasks like humanoid-v2.

In this study [112] the authors introduce the Target Entropy Scheduled SAC (TES-SAC) method, applied to the SAC algorithm in discrete action space settings, particularly focusing on a range of Atari 2600 games and classical control tasks. They examined the effects of different constant target entropy values on the SAC algorithm, specifically testing $H = C \cdot \log |A|$ where C varied among 0.98, 0.5, and 0.01 and their method TES-SAC 1. This analysis showed the influence of entropy settings on learning efficacy and overall performance .

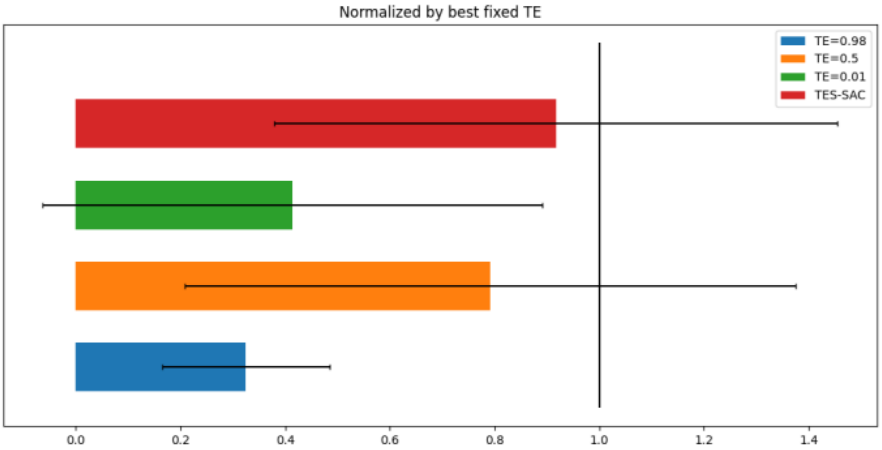
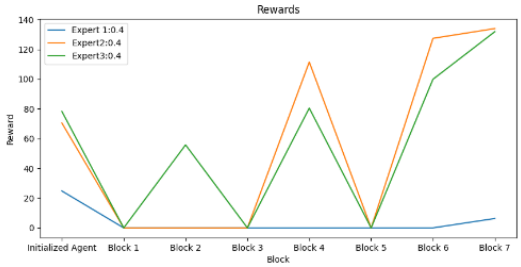


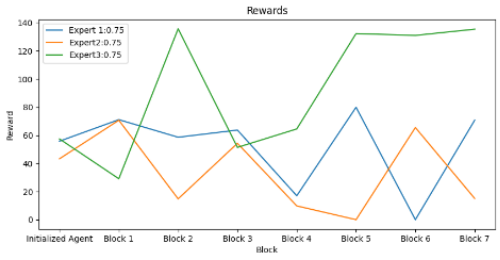
Figure 1: Different constant target entropies performance,[112]

Another observation was the dynamic response of the temperature variable α and policy entropy under these entropy configurations. For example, in environments like MsPacman, a low target entropy such as $H = 0.01 \cdot \log |A|$ led to an exponential decrease in α, due to the policy entropy's inability to meet such low targets, influenced by the minimum entropy level of the environment. In contrast, environments like BattleZone, where the minimum entropy level is near zero, did not exhibit a rapid decline in α even with low target entropy settings. These results highlight that the effectiveness of target entropy in SAC can significantly differ across various environments.

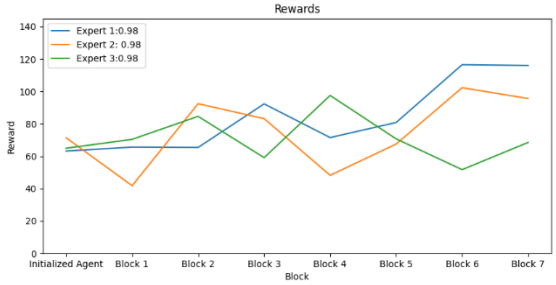
In our experiment, we examined the behavior of the agent in the agent-expert interaction for various constant target entropy values on the SAC algorithm. We present results for three of them $H = C \cdot \log |A|$ where C varied among 0.98, 0.75, and 0.36 and 2 illustrates the performance of these three settings, based on the rewards of each block.



(a) Rewards for C=0.36



(b) Rewards for C=0.75



(c) Rewards for C=0.98

Figure 2: Rewards for different entropies

Figures 2 display the rewards obtained under different entropy settings while the agent interacting with the expert. Interestingly, all the entropy settings revealed signs of unstable learning. In each case, the agent exhibited a pattern of transitioning from high rewards to lower rewards and then returning to a similar performance level, without being able to consistently learn a good policy. The performance instability was less pronounced with a high entropy setting of C=0.98, which encourages more exploration. This increased exploration appeared to mask the underlying issues with the SAC algorithm, making it challenging to identify the root cause of the problem. This led us to consider the change of the activation function to Tanh, might have caused the problem in the learning procedure. One possible reason is the vanishing gradient problem [41].

.2 Questionnaire Appendices

.2.1 Appendix A: Big Five Personality Traits Questionnaire

This appendix contains the 50 questions of the Big Five Personality Traits questionnaire. The questionnaire assesses five key personality characteristics: Extraversion3, Agreeableness 4, Conscientiousness 2, Emotional Stability/Neuroticism 5, and Openness to Experience1.

.2.2 Appendix B: Schwartz Portrait Values Questionnaire

Appendix B presents the 21 questions of the Schwartz Portrait Values Questionnaire, detailed in Tables 6, 7, 8. This questionnaire is designed to evaluate the participant's value system.

.2.3 Appendix C: AI Attitude Scale

In Schepman's study on AI attitudes, the refined 20-item AI Attitude Scale, which includes 12 positive and 8 negative questions, is presented in 6 and is translated to Greece in 10,11. This scale, developed after rigorous analysis and validation, effectively captures the general public's perception of AI."

.2.4 Appendix D: Additional Personal Questions

This section includes the additional personal questions that were asked to gain insights into the participants' backgrounds and experiences. It covers personal information, gaming experience, and knowledge about AI 9.

.2.5 Appendix E: Questionnaire 1

Questionnaire 1 combines all the questionnaires from .2.1,.2.2,.2.3,.2.4, . This questionnaire was used to assess personality traits, values, attitudes towards AI, and personal background information of the participants before the start of the collaborative process.

.2.6 Appendix F: Questionnaire 2

We utilized a questionnaire to assess subjective experiences in human-AI collaboration. Our approach is based on the methodology of Tsitos, with initial concepts are based inHoffman and with Tsitos feedback we improved measures to yield more meaningful results.

A key observation from Tsitos was the inconsistency in responses regarding the robot's contribution, primarily due to ambiguous framing of questions related to performance levels. To address this, we divided the Human-Agent Contribution category into two segments: four questions about the last block's performance and two about the overall process. This structure aims to unify participants' mindset while capturing both the training process and the final outcome contributions. The first four questions, focusing on the final block's performance, are:

1. Assessment of the team's performance in the last ten tests.
2. Personal responsibility for this performance.
3. The performance as a joint team effort.
4. The AI system's primary responsibility for the performance.

We altered the grouping of questions 2-4, pairing questions 2 and 4 with a reverse scale for question 2. This grouping provides insights into the perceived AI contribution in the final block. However, we believe that a narrative approach, interpreting individual responses, is more insightful than presenting grouped results. This method helps understand the context of team effort and responsibility, whether it's about assigning blame or recognizing contributions.

Post these four questions, participants responded to two additional questions about the human agent's contribution throughout the entire interaction:

5. Personal importance as a team member.

6. The AI system's importance as a team member.

Though these questions aim to clarify the perceived importance of each member, we find their internal consistency limited. They are better used individually for a comprehensive narrative.

In the Improvement category, we added two new questions to evaluate each member's role in team enhancement. These questions are crucial for understanding participants' perceptions of each member's importance in the co-learning experience. All questions were translated into Greek for our participants and are detailed in Questionnaire 2 12. 13

.2.7 Appendix F: information letter and consent form

Information letter can be found in tables 3, 4and the consent form at 5

Table 1: Big Five Questions - Openness to Experience

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

Table 2: Big Five Questions - Conscientiousness

Greek	**English**	**Pos/Neg**
Είμαι πάντοτε προετοιμασμένος	Am always prepared	Pos
Αφήνω τα πράγματά μου ολόγυρα	Leave my belongings around	Neg
Δίνω προσοχή στις λεπτομέρειες	Pay attention to details	Pos
Τα κάνω άνω κάτω	Make a mess of things	Neg
Κάνω τις «αγγαρείες» αμέσως	Get chores done right away	Pos
Συχνά ξεχνώ να βάζω τα πράγματα πίσω στη σωστή τους θέση	Often forget to put things back in their proper place	Neg
Μου αρέσει η τάξη	Like order	Pos
Αποφεύγω αυτά που πρέπει να κάνω	Shirk my duties	Neg
Ακολουθώ ένα πρόγραμμα	Follow a schedule	Pos
Είμαι ακριβής στη δουλειά μου	Am exacting in my work	Pos

Table 3: Big Five Questions - Extraversion

Greek	**English**	**Pos/Neg**
Είμαι η ζωή σε ένα πάρτι	Am the life of the party	Pos
Δεν μιλώ πολύ	Don't talk a lot	Neg
Αισθάνομαι άνετα όταν βρίσκομαι ανάμεσα σε ανθρώπους	Feel comfortable around people	Pos
Προτιμώ να μένω στο παρασκήνιο	Keep in the background	Neg
Αρχίζω συζητήσεις	Start conversations	Pos
Έχω ελάχιστα πράγματα να πω	Have little to say	Neg
Μιλώ με πολλούς διαφορετικούς ανθρώπους στα πάρτι	Talk to a lot of different people at parties	Pos
Δεν μου αρέσει να προσελκύω την προσοχή πάνω μου	Don't like to draw attention to myself	Neg
Δεν με ενοχλεί να είμαι το επίκεντρο της προσοχής	Don't mind being the centre of attention	Pos
Είμαι ήσυχος/η όταν βρίσκομαι ανάμεσα σε ξένους	Am quiet around strangers	Neg

Table 4: Big Five Questions - Agreeableness

Greek	**English**	**Pos/Neg**
Αισθάνομαι μικρό ενδιαφέρον για τους άλλους	Feel little concern for others	Neg
Ενδιαφέρομαι για τους ανθρώπους	Am interested in people	Pos
Προσβάλλω τους άλλους	Insult people	Neg
Συμπάσχω με τα συναισθήματα των άλλων	Sympathize with others' feelings	Pos
Δεν ενδιαφέρομαι για τα προβλήματα των άλλων	Am not interested in other people's problems	Neg
Έχω μαλακή καρδιά	Have a soft heart	Pos
Δεν ενδιαφέρομαι πραγματικά για τους άλλους ανθρώπους	Am not really interested in others	Neg
Βρίσκω χρόνο για τους άλλους	Take time out for others	Pos
Αισθάνομαι τα συναισθήματα των άλλων	Feel others' emotions	Pos
Κάνω τους ανθρώπους να αισθάνονται άνετα	Make people feel at ease	Pos

Table 5: Big Five Questions - Neuroticism

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

Table 6: Personal Values Questionnaire (PVQ) Questions- part 1

English Question (Male Version)	**Greek Question (Combine Male and Female Version)**
BENEVOLENCE	
It's very important to him to help the people around him. He wants to care for other people.	Είναι πολύ σημαντικό για αυτήν/όν να βοηθά τους ανθρώπους που την/τον περιβάλλουν. Ενδιαφέρεται για το καλό των άλλων.
It is important to him to be loyal to his friends. He wants to devote himself to people close to him.	Είναι σημαντικό για αυτήν/όν να είναι πιστή/ος στους φίλους της/του. Θέλει να αφοσιώνεται στους ανθρώπους που βρίσκονται κοντά της/του
UNIVERSALISM	
He thinks it is important that every person in the world be treated equally. He wants justice for everybody, even for people he doesn't know.	Πιστεύει πως είναι σημαντικό όλοι οι άνθρωποι στον κόσμο να αντιμετωπίζονται ισότιμα. Πιστεύει ότι όλοι πρέπει να έχουν ίδιες ευκαιρίες στη ζωή
It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.	Της/Του είναι σημαντικό, να ακούει ανθρώπους με διαφορετικές απόψεις από τις δικές της/του. Ακόμα και όταν διαφωνεί θέλει να μπορεί να τους κατανοεί.
He strongly believes that people should care for nature. Looking after the environment is important to him.	Πιστεύει ακράδαντα ότι οι άνθρωποι πρέπει να προστατεύουν τη φύση. Η προστασία του περιβάλλοντος είναι πολύ σημαντική για αυτήν/όν.
SELF-DIRECTION	
Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.	Είναι πολύ σημαντικό για αυτήν/όν να έχει καινούργιες ιδέες και να είναι δημιουργική/ος. Τον/Την αρέσει να κάνει πράγματα με τον δικό της/του πρωτότυπο τρόπο.
It is important to him to make his own decisions about what he does. He likes to be free to plan and to choose his activities for himself.	Είναι σημαντικό για αυτήν/ον να λαμβάνει τις δικές της/του αποφάσεις για ότι πρόκειται να κάνει. Θέλει να είναι ελεύθερη/ος και να μην εξαρτάται από άλλους.

Table 7: Personal Values Questionnaire (PVQ) Questions- part 2

STIMULATION	
He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.	Της/Του αρέσουν οι εκπλήξεις και θέλει να κάνει πάντα καινούρια πράγματα. Πιστεύει ότι στη ζωή είναι σημαντικό να κάνεις πολλά διαφορετικά πράγματα
He looks for adventures and likes to take risks. He wants to have an exciting life.	Αναζητεί την περιπέτεια και είναι ριψοκίνδυνος. Θέλει η ζωή της/του να είναι συναρπαστική
HEDONISM	
Having a good time is important to him. He likes to "spoil" himself.	Η καλοπέραση είναι σημαντική για αυτήν/όν. Της/Του αρέσει να καλομαθαίνει τον εαυτό της/του.
He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.	Πάντα ψάχνει ευκαιρία για γλέντι. Είναι σημαντικό για αυτήν/όν να κάνει πράγματα που την/τον ευχαριστούν.
ACHIEVEMENT	
It is very important to him to show his abilities. He wants people to admire what he does.	Είναι πολύ σημαντικό γι' αυτήν/όν να δείχνει τις ικανότητές της/του. Θέλει ο κόσμος να θαυμάζει αυτό που κάνει.
Being very successful is important to him. He likes to impress other people.	Η επιτυχία της/του, είναι πολύ σημαντική για την/τον ίδια/ιο. Ελπίζει ότι ο κόσμος θα αναγνωρίσει τα επιτεύγματά της/του.
POWER	
It is important to him to be rich. He wants to have a lot of money and expensive things.	Είναι σημαντικό γι' αυτήν/όν να είναι πλούσια/ιος. Θέλει να έχει πολλά λεφτά και ακριβά πράγματα.
It is important to him to be in charge and tell others what to do. He wants people to do what he says.	Είναι σημαντικό για αυτήν/όν να την/τον σέβονται οι άλλοι. Θέλει οι άλλοι να κάνουν αυτό που τους λέει.

Table 8: Personal Values Questionnaire (PVQ) Questions- part 3

SECURITY	
It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.	Είναι πολύ σημαντικό για αυτήν/όν να ζει σε ένα ασφαλές περιβάλλον. Αποφεύγει οτιδήποτε θα μπορούσε να θέσει σε κίνδυνο την ασφάλειά της/του.
It is very important to him that his country be safe from threats from within and without. He is concerned that social order be protected.	Είναι πολύ σημαντικό για αυτήν/ον η κυβέρνηση να μπορεί να εγγυηθεί για την ασφάλειά της/του. Θέλει ένα κράτος ισχυρό, ικανό να προστατεύσει τους πολίτες του.
CONFORMITY	
He believes that people should do what they're told. He thinks people should follow rules at all times, even when no-one is watching.	Πιστεύει ότι οι άνθρωποι πρέπει να κάνουν αυτό που τους λένε. Πιστεύει ότι οι άνθρωποι πρέπει πάντα να τηρούν τους κανόνες, ακόμα και όταν κανείς δεν τους βλέπει.
It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.	Είναι σημαντικό για αυτήν/όν να συμπεριφέρεται πάντα σωστά. Θέλει να αποφεύγει να κάνει πράγματα που οι άλλοι θα έλεγαν ότι είναι λάθος.
TRADITION	
He thinks it's important not to ask for more than what you have. He believes that people should be satisfied with what they have.	Είναι σημαντικό γι' αυτήν/όν να είναι ταπεινή/ός και μετριόφρων. Προσπαθεί να μην τραβά την προσοχή. Πιστεύει ότι οι άνθρωποι πρέπει πάντα να τηρούν τους κανόνες, ακόμα και όταν κανείς δεν τους βλέπει.
Religious belief is important to him. He tries hard to do what his religion requires.	Η παράδοση είναι κάτι πολύ σημαντικό για αυτήν/όν. Προσπαθεί να τηρεί τα ήθη και τα έθιμα.

English	Greek
Personal Information	
Gender	Φύλο
Age	Ηλικία
Dominant Hand	Επικρατές χέρι
Diagnosed Neurological Condition	Διαγνωσμένη νευρολογική πάθηση
Use of Myopia Glasses/Lenses	Χρήση γυαλιών/φακών μυωπίας
Experience in Gaming	
What experience do you have with gaming?	Τι εμπειρία έχεις με παιχνίδια (γαμινγ)·
What is your preferred platform	Ποιά είναι η προτιμώμενη πλατφόρμα σας ·
Knowledge about AI	
How would you describe your relationship with AI?	Πως θα χαρακτηρίζατε τη σχέση σας με την ΤΝ·
Do you come into contact with AI applications in your daily life?	Έρχεστε σε επαφή με εφαρμογές ΤΝ στην καθημερινότητά σας·
What is the main source of information on developments around AI issues?	Ποια είναι η κύρια πηγή ενημέρωσης των εξελίξεων γύρω από θέματα ΤΝ·

Table 9: Questions about Personal Information, Experience in Gaming, and Knowledge in AI

Table 10: AI Attitude Scale Questions- part 1

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

Table 11: AI Attitude Scale Questions-part 2

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

Table 12: Subjective measures separated into each measure part 1.

Greek	English	Pos/Neg
FLUENCY		
Η ομάδα ανθρώπου - ρομπότ συνεργάστηκε απρόσκοπτα	The human-robot team worked together seamlessly	Pos
Η συνεργασία της ομάδας έγινε πιο εύρυθμη με τη πάροδο του χρόνου.	The team's cooperation has become more fluid over time.	Pos
Το ρομπότ συνεισέφερε στην εύρυθμη συνεργασία της ομάδας.	The robot contributed to the fluid collaboration of the team.	Pos
CONTRIBUTION		
Εγώ είχα την κύρια ευθύνη γι' αυτήν την επίδοση.	I had the main responsibility for this performance.	Neg
Το ρομπότ είχε την κύρια ευθύνη γι' αυτή την επίδοση.	The robot was primarily responsible for this performance.	Pos
TRUST		
Είχα εμπιστοσύνη στο ρομπότ ότι θα έκανε το σωστό πράγμα τη σωστή στιγμή.	I had confidence in the robot that it would do the right thing at the right time.	Pos
Υπήρχε αμοιβαία εμπιστοσύνη ανάμεσα σε μένα και το ρομπότ.	There was mutual trust between me and the robot.	Pos
TEAMMATE TRAITS		
Το ρομπότ ήταν ευφυές.	The robot was intelligent.	Pos
Το ρομπότ ήταν αξιόπιστο.	The robot was trustworthy.	Pos
Το ρομπότ ήταν αφοσιωμένο στην επίτευξη του στόχου.	The AI system was dedicated to achieving the goal.	Pos
Το ρομπότ ήταν συνεργάσιμο.	The robot was cooperative.	Pos

Table 13: Subjective measures separated into each measure-part 2.

IMPROVEMENT		
Η ομάδα ανθρώπου - ρομπότ βελτιώθηκε με την πάροδο του χρόνου.	The human-robot team improved over time.	Pos
Η επίδοσή μου βελτιώθηκε κατά τη διάρκεια του πειράματος.	My performance improved during the experiment.	Pos
Η επίδοση του ρομπότ βελτιώθηκε κατά τη διάρκεια του πειράματος.	The performance of the robot improved during the experiment.	Pos
Εγώ είχα την κύρια ευθύνη για την βελτίωση της ομάδας.	I had the main responsibility for the improvement of the team.	Pos
Το ρομπότ είχε την κύρια ευθύνη για την βελτίωση της ομάδας	The robot had the main responsibility for the improvement of the team	Pos
ALLIANCE		
Πίστευα ότι το ρομπότ μπορούσε να με βοηθήσει.	I believed that the robot could help me.	Pos
Το ρομπότ μπορούσε να αντιληφθεί τις προθέσεις μου.	The robot could perceive my intentions.	Pos
Το ρομπότ δεν καταλάβαινε τι προσπαθούσα να πετύχω.	The robot didn't understand what I was trying to achieve.	Neg
Θεωρώ ότι η συνεργασία με το ρομποτ ήταν μπερδευτική	I think working with the robot was confusing	Neg
EXTRA ITEMS		
Πως κρίνετε την επίδοση της ομάδας στις τελευταίες δέκα δοκιμές·	How do you judge the team's performance in the last ten tests?	Pos/Neg
Η ευθύνη για αυτή την επίδοση ήταν: .	The responsibility for this performance was:.	Pos/Neg
Πως σου φάνηκε η συνεργασία με το ρομπότ· Έχεις κάποια σχόλια για το πείραμα·	What did you think of working with the robot? Do you have any comments on the experiment?	Pos/Neg

Συνοδευτική Ενημερωτική Επιστολή

Έρευνα για τη συνεργασία ανθρώπου-Τεχνητής Νοημοσύνης

Έχετε εκδηλώσει ενδιαφέρον να λάβετε μέρος στη διεξαγωγή έρευνας που γίνεται στα πλαίσια εκπόνησης διπλωματικής εργασίας του Νικόλα Σταύρου, σχολή Ηλεκτρολόγων Μηχανικών Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, σε συνεργασία με το ΕΚΕΦΕ "Δημόκριτος"[1]. Η έρευνα έχει ως στόχο την ανάπτυξη μεθόδων για την εύρυθμη συνεργασία ενός ανθρώπου με ένα ρομπότ, σε ένα πραγματικό περιβάλλον όπου απαιτείται συνδυασμός κινήσεων από τους δύο συνεργάτες.

Πρακτικές πληροφορίες

Η διεξαγωγή της έρευνας γίνεται στο χώρο του εργαστηρίου Roboskel (κτίριο Κεντρικής Βιβλιοθήκης, ΕΚΕΦΕ Δημόκριτος). Η διαδικασία συλλογής δεδομένων ολοκληρώνεται σε **μια επίσκεψη** που θα διαρκέσει περίπου 1.5 ώρα. Η συμμετοχή σας γίνεται σε εθελοντική βάση και δεν έχει κανένα όφελος για σας, οικονομικό, ή οποιασδήποτε άλλης φύσης.

Η διαδικασία / Μπορώ να διακόψω την διαδικασία;

Στα πλαίσια της διεξαγωγής του πειράματος καλείστε να συνεργαστείτε με ένα σύστημα ΤΝ για να ελέγξετε από κοινού τη θέση ενός βραχίονα και να τον μεταφέρετε σε μία θέση-στόχο. Εσείς ελέγχετε την κίνηση του βραχίονα σε έναν άξονα μέσω ενός πληκτρολογίου. Το σύστημα ΤΝ είναι υπεύθυνο να ελέγχει την κίνηση του άλλου άξονα. Συνολικά θα εκτελέσετε 100 δοκιμές. Κάθε δοκιμή ολοκληρώνεται είτε όταν επιτευχθεί ο στόχος είτε αν παρέλθουν 30 δευτερόλεπτα. Στο τέλος κάθε δοκιμής θα ενημερώνεστε για την επίδοση της ομάδας σας. Ανά τακτά χρονικά διαστήματα θα γίνονται διαλείμματα ώστε να αποφευχθεί οποιαδήποτε κόπωση.

Η συμμετοχή σας είναι εθελοντική. Μπορείτε να εγκαταλείψετε τη διαδικασία οποιαδήποτε στιγμή.

Τι είδους δεδομένα θα συλλεχθούν και πως θα χρησιμοποιηθούν;

Στην αρχή της διαδικασίας θα σας ζητηθεί να συμπληρώσετε ένα ερωτηματολόγιο με κάποια δημογραφικά στοιχεία και άλλες ερωτήσεις που αφορούν τις γνώσεις σας, την εμπειρία σας και την

Figure 3: Information letter 1

προσωπικότητά σας. Κατά τη διάρκεια και στο τέλος του πειράματος, θα κληθείτε να σχολιάσετε διάφορες πτυχές της συνεργασίας σας με την ΤΝ.

Επίσης, άλλα δεδομένα που θα συλλεχθούν είναι , στοιχεία σχετικά με την επίδοση του συστήματος ΤΝς που χρησιμοποιείται και οι ενέργειες του ανθρώπου, δηλαδή τα πλήκτρα που πατάει κατά την εξαγωγή των πειραμάτων. Τα δεδομένα θα χρησιμοποιηθούν για ερευνητικούς σκοπούς.

Η συλλογή όλων των δεδομένων θα γίνει ανώνυμα με τη χρήση κωδικού ονόματος συμμετέχοντα (π.χ. Υ1Ζ, κτλ.)

Τι είδους πληροφορίες θα είναι διαθέσιμες δημόσια;

Δημόσια σε περίπτωση δημοσίευσης θα γίνουν διαθέσιμα τα αποτελέσματα στατιστικής ανάλυσης επί των συλλεχθέντων δεδομένων καθώς και μεμονωμένες απαντήσεις σε ανοιχτού τύπου ερωτήσεις χωρίς **να συναφθεί καμία πληροφορία που αφορά κωδικό όνομα ή άλλα δημογραφικά στοιχεία του ατόμου**.

Οι φόρμες συγκατάθεσης θα φυλαχθούν εμπιστευτικά από το ΙΠΤ.

Υπάρχουν κίνδυνοι και ενοχλήσεις κατά τη διάρκεια της συλλογής δεδομένων;

Η συμμετοχή σας στο πείραμα δεν έχει κανένα κίνδυνο για εσάς, ούτε θα αισθανθείτε κάποια ενόχληση κατά τη διάρκειά του. Επίσης δεν χρειάζεται καμία προφύλαξη κατά τη διάρκειά του.

Αν έχετε κάποιες περαιτέρω απορίες για τα προαναφερθέντα, μη διστάσετε να μας ρωτήσετε.

Η συμμετοχή σας είναι εθελοντική. Η άρνηση παροχής συγκατάθεσης δεν επιφέρει καμία αρνητική συνέπεια σε εσάς. Διατηρείτε επίσης το δικαίωμα αναίρεσης της συγκατάθεσής σας οποιαδήποτε στιγμή κατά τη διάρκεια της συμμετοχής σας στη μελέτη ή και μετά από αυτή. Στην περίπτωση τέτοιας αναίρεσης, τα συγκεντρωμένα δεδομένα σας θα διαγραφούν άμεσα.

Αν έχετε άλλες ερωτήσεις, μπορείτε να απευθυνθείτε στους:

Δρ. Μαρία Δαγιόγλου, τηλ.: 2106503201, email: mdagiogl@iit.demokritos.gr.

Figure 4: Information letter 2

Αντίγραφο Συμμετέχοντα

Δήλωση συγκατάθεσης

Αριθμός συμμετέχοντα:_____

Δηλώνω υπεύθυνα ότι έχω διαβάσει τη δήλωση συγκατάθεσης και τη συνοδευτική ενημερωτική επιστολή. Η φύση, ο σκοπός και οι πιθανές επιπτώσεις της διαδικασίας μου έχουν εξηγηθεί επαρκώς. Γνωρίζω ότι έχω δικαίωμα να εγκαταλείψω τη διαδικασία οποιαδήποτε στιγμή. Δηλώνω ότι συμφωνώ στη συμμετοχή της παρούσας έρευνας. Αγ. Παρασκευή, __ / __ / 202.. _____ Υπογραφή _____ Ονοματεπώνυμο [με κεφαλαία] Υπεύθυνος Ερευνητής Δρ. Βαγγέλης Καρκαλέτσης Ινστ. Πληροφορικής & Τηλεπικοινωνιών,Ε.Κ.Ε.Φ.Ε. «Δημόκριτος»	Έχετε διαβάσει την ενημερωτική συνοδευτική επιστολή; | ΝΑΙ | ΟΧΙ | Είχατε την ευκαιρία να απευθύνετε διευκρινιστικές ερωτήσεις; | ΝΑΙ | ΟΧΙ | Λάβατε ικανοποιητικές απαντήσεις στις ερωτήσεις σας; | ΝΑΙ | ΟΧΙ | Λάβατε επαρκείς πληροφορίες για τη διαδικασία; | ΝΑΙ | ΟΧΙ | Από ποιόν ενημερωθήκατε; Όνομα: …….. Γνωρίζετε ότι έχετε το δικαίωμα να αποχωρήσετε οποιαδήποτε στιγμή πριν ή κατά τη διάρκεια της διαδικασία, χωρίς να αιτιολογήσετε τους λόγους της απόφασής σας; | ΝΑΙ | ΟΧΙ |

Figure 5: Consent form

Subscale (not for display)	Number (not for display)	Item
Positive	1	For routine transactions, I would rather interact with an artificially intelligent system than with a human.
Positive	2	Artificial Intelligence can provide new economic opportunities for this country.
Negative	3	Organisations use Artificial Intelligence unethically.
Positive	4	Artificially intelligent systems can help people feel happier.
Positive	5	I am impressed by what Artificial Intelligence can do.
Negative	6	I think artificially intelligent systems make many errors.
Positive	7	I am interested in using artificially intelligent systems in my daily life.
Negative	8	I find Artificial Intelligence sinister.
Negative	9	Artificial Intelligence might take control of people.
Negative	10	I think Artificial Intelligence is dangerous.
Positive	11	Artificial Intelligence can have positive impacts on people's wellbeing.
Positive	12	Artificial Intelligence is exciting.
Attention Check	**A**	I would be grateful if you could select Strongly agree.
Positive	13	An artificially intelligent agent would be better than an employee in many routine jobs.
Positive	14	There are many beneficial applications of Artificial Intelligence.
Negative	15	I shiver with discomfort when I think about future uses of Artificial Intelligence.
Positive	16	Artificially intelligent systems can perform better than humans.
Positive	17	Much of society will benefit from a future full of Artificial Intelligence
Positive	18	I would like to use Artificial Intelligence in my own job.
Negative	19	People like me will suffer if Artificial Intelligence is used more and more.
Negative	20	Artificial Intelligence is used to spy on people

Scoring: Check compliance with the Attention Check, then discount it from the scoring. Score items marked "Positive" as Strongly disagree = 1; (Somewhat) disgree = 2; Neutral = 3; (Somewhat) agree = 4; and Strongly agree = 5. Score the items marked "Negative" in reverse so that Strongly disagree = 5; (Somewhat) disgree = 4; Neutral = 3; (Somewhat) agree = 2; and Strongly agree = 1. Then take the mean of the positive items to form an overall score for the positive subscale, and the mean of the negative items to form the negative subscale. The higher the score on each subscale, the more positive the attitude. We do not recommend calculating an overall scale mean.

Figure 6: Attitude towards AI

Bibliography

[1] Judith Bütepage and Danica Kragic. *Human-Robot Collaboration: From Psychology to Social Robotics*. 2017. arXiv: 1705.10146 [cs.RO].

[2] Rinat Rosenberg-Kima et al. "Human-Robot-Collaboration (HRC): Social Robots as Teaching Assistants for Training Activities in Small Groups". In: *Name of the Conference Proceedings*. Research output: Chapter in Book/Report/Conference proceeding › Conference contribution › peer-review.

[3] Hoang-Long Cao et al. "Robot-Assisted Joint Attention: A Comparative Study Between Children With Autism Spectrum Disorder and Typically Developing Children in Interaction With NAO". In: *IEEE Access* (2020). Received October 30, 2020; accepted December 2, 2020; date of publication December 14, 2020; date of current version December 28, 2020. DOI: 10.1109/ACCESS.2020.3044483.

[4] M. Jeon. "Turning HART into HEART: Human Emotional AI/Robot Teaming". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 65. 1. 2021, pp. 1044–1048. DOI: 10.1177/1071181321651136. URL: https://doi.org/10.1177/1071181321651136.

[5] Eloise Matheson et al. "Human–Robot Collaboration in Manufacturing Applications: A Review". In: *Robotics* 8.4 (2019). ISSN: 2218-6581. URL: https://www.mdpi.com/2218-6581/8/4/100.

[6] Hongjun Xing et al. "Human-Robot Collaboration for Heavy Object Manipulation: Kinesthetic Teaching of the Role of Wheeled Mobile Manipulator". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2021, pp. 2962–2969. DOI: 10.1109/IROS51168.2021.9635910.

[7] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. "Human-robot mutual adaptation in collaborative tasks: Models and experiments". In: *The International Journal of Robotics Research* 36.5-7 (2017). PMID: 32855581, pp. 618–634. DOI: 10.1177/0278364917690593. eprint: https://doi.org/10.1177/0278364917690593. URL: https://doi.org/10.1177/0278364917690593.

[8] Hartmut Surmann et al. *Deep Reinforcement learning for real autonomous mobile robot navigation in indoor environments*. 2020. arXiv: 2005.13857 [cs.RO].

[9] Tobias Johannink et al. *Residual Reinforcement Learning for Robot Control*. 2018. arXiv: 1812.03201 [cs.RO].

[10] Guillem Muñoz et al. "Deep Reinforcement Learning for Drone Delivery". In: *Drones* 3.3 (2019). ISSN: 2504-446X. DOI: 10.3390/drones3030072. URL: https://www.mdpi.com/2504-446X/3/3/72.

[11] Hai Nguyen and Hung La. "Review of Deep Reinforcement Learning for Robot Manipulation". In: *2019 Third IEEE International Conference on Robotic Computing (IRC)*. 2019, pp. 590–595. DOI: 10.1109/IRC.2019.00120.

[12] Zhuangdi Zhu et al. "Transfer learning in deep reinforcement learning: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[13] Athanasios C. Tsitos and Maria Dagioglou. *Enhancing team performance with transfer-learning during real-world human-robot collaboration*. 2022. arXiv: 2211.13070 [cs.RO].

[14] Gal Dalal et al. *Safe Exploration in Continuous Action Spaces*. 2018. arXiv: 1801.08757 [cs.AI].

[15] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518 (2015), pp. 529–533. URL: https://api.semanticscholar.org/CorpusID:205242740.

[16] Timothy P. Lillicrap et al. *Continuous control with deep reinforcement learning*. 2019. arXiv: 1509.02971 [cs.LG].

[17] Sergey Levine et al. *End-to-End Training of Deep Visuomotor Policies*. 2016. arXiv: 1504.00702 [cs.LG].

[18] Sergey Levine et al. *Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection*. 2016. arXiv: 1603.02199 [cs.LG].

[19] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. "Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (2020), pp. 737–744. URL: https://api.semanticscholar.org/CorpusID:221971078.

[20] Eugene Valassakis, Zihan Ding, and Edward Johns. *Crossing The Gap: A Deep Dive into Zero-Shot Sim-to-Real Transfer for Dynamics*. 2020. arXiv: 2008.06686 [cs.RO].

[21] Josh Tobin et al. *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World*. 2017. arXiv: 1703.06907 [cs.RO].

[22] Martin Sundermeyer et al. *Implicit 3D Orientation Learning for 6D Object Detection from RGB Images*. 2019. arXiv: 1902.01275 [cs.CV].

[23] Xiangyu Yue et al. *Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data*. 2022. arXiv: 1909.00889 [cs.CV].

[24] Lukas Brunke et al. *Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning*. 2021. arXiv: 2108.06266 [cs.RO].

[25] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. *OptLayer - Practical Constrained Optimization for Deep Reinforcement Learning in the Real World*. 2018. arXiv: 1709.07643 [cs.RO].

[26] Ashvin Nair et al. *Overcoming Exploration in Reinforcement Learning with Demonstrations*. 2018. arXiv: 1709.10089 [cs.LG].

[27] Nicolò Botteghi et al. *On Reward Shaping for Mobile Robot Navigation: A Reinforcement Learning and SLAM Based Approach*. 2020. arXiv: 2002.04109 [cs.RO].

[28] René Traoré et al. *DisCoRL: Continual Reinforcement Learning via Policy Distillation*. 2019. arXiv: 1907.05855 [cs.LG].

[29] Fernando Fernández and Manuela Veloso. "Probabilistic Policy Reuse in a Reinforcement Learning Agent". In: New York, NY, USA: Association for Computing Machinery, 2006. ISBN: 1595933034. DOI: 10.1145/1160633.1160762. URL: https://doi.org/10.1145/1160633.1160762.

[30] Javier García and Diogo Shafie. "Teaching a humanoid robot to walk faster through Safe Reinforcement Learning". In: *Engineering Applications of Artificial Intelligence* 88 (2020), p. 103360. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2019.103360. URL: https://www.sciencedirect.com/science/article/pii/S0952197619302921.

[31] Abhishek Gupta et al. *Learning Invariant Feature Spaces to Transfer Skills with Reinforcement Learning*. 2017. arXiv: 1703.02949 [cs.AI].

[32] Rico Jonschkowski and Oliver Brock. "State Representation Learning in Robotics: Using Prior Knowledge about Physical Interaction". In: *Robotics: Science and Systems*. 2014. URL: https://api.semanticscholar.org/CorpusID:5467016.

[33] Todd Hester et al. *Deep Q-learning from Demonstrations*. 2017. arXiv: 1704.03732 [cs.AI].

[34] Dimitrios Koutrintzes. "Knowledge transfer in human-artificial intelligence collaboration". MA thesis. 2023. DOI: 10.26267/unipi_dione/3273. URL: https://dione.lib.unipi.gr/xmlui/handle/unipi/15851?locale-attribute=en.

[35] Haochen Liu et al. *Improved Deep Reinforcement Learning with Expert Demonstrations for Urban Autonomous Driving*. 2022. arXiv: 2102.09243 [cs.RO].

[36] Stanford University. *CS230 Deep Learning*. https://cs230.stanford.edu/section/4/. Accessed: 2024-01-30. 2023.

[37] Harvey Merton et al. *Deep Reinforcement Learning for Local Path Following of an Autonomous Formula SAE Vehicle*. 2024. arXiv: 2401.02903 [cs.RO].

[38] Petros Christodoulou. *Soft Actor-Critic for Discrete Action Settings*. 2019. arXiv: 1910.07207 [cs.LG].

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

[40] Patrick Mair and Rand Wilcox. "Robust statistical methods in R using the WRS2 package". In: *Behavior research methods* 52 (2020), pp. 464–488.

[41] Wikipedia contributors. *Vanishing Gradient Problem – Wikipedia, The Free Encyclopedia*. [Online; accessed 30-January-2024]. 2022. URL: https://en.wikipedia.org/wiki/Vanishing_gradient_problem.

[42] Batta Mahesh. "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR).[Internet]* 9.1 (2020), pp. 381–386.

[43] Mohammad Mustafa Taye. "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions". In: *Computers* 12.5 (2023). ISSN: 2073-431X. DOI: 10.3390/computers12050091. URL: https://www.mdpi.com/2073-431X/12/5/91.

[44] Adel El-Shahat. "Introductory Chapter: Artificial Neural Networks". In: Feb. 2018. ISBN: 978-953-51-3780-1. DOI: 10.5772/intechopen.73530.

[45] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536. URL: https://api.semanticscholar.org/CorpusID:205001834.

[46] Rukshan Pramoditha. *Overview of a Neural Network's Learning Process*. https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa. Accessed: date-of-access. 2002.

[47] Dimitri P. Bertsekas. *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. Athena Scientific, 2020, p. 376. ISBN: 978-1-886529-07-6.

[48] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019, p. 388. ISBN: 978-1-886529-39-7.

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

[50] Shweta Bhatt. *Reinforcement Learning 101*. [Online; accessed Date]. 2018. URL: https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292.

[51] G. Rummery and Mahesan Niranjan. "On-Line Q-Learning Using Connectionist Systems". In: *Technical Report CUED/F-INFENG/TR 166* (Nov. 1994).

[52] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine learning* 8 (1992), pp. 279–292.

[53] OpenAI. *Spinning Up in Deep Reinforcement Learning*. Year of Access. URL: https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html.

[54] David Ha and Jürgen Schmidhuber. "World Models". In: (2018). DOI: 10.5281/ZENODO.1207631. URL: https://zenodo.org/record/1207631.

[55] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

[56] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550 (2017), pp. 354–359. URL: https://api.semanticscholar.org/CorpusID:205261034.

[57] Maochun Xu et al. *Deep Reinforcement Learning for Quantitative Trading*. 2023. arXiv: 2312.15730 [q-fin.TR].

[58] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518 (2015), pp. 529–533. URL: https://api.semanticscholar.org/CorpusID:205242740.

[59] Volodymyr Mnih et al. *Asynchronous Methods for Deep Reinforcement Learning*. 2016. arXiv: 1602.01783 [cs.LG].

[60] Timothy P. Lillicrap et al. *Continuous control with deep reinforcement learning*. 2019. arXiv: 1509.02971 [cs.LG].

[61] John Schulman et al. "Trust region policy optimization". In: *International conference on machine learning*. PMLR. 2015, pp. 1889–1897.

[62] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[63] Tuomas Haarnoja et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 1861–1870. URL: https://proceedings.mlr.press/v80/haarnoja18b.html.

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

[65] Brian D Ziebart et al. "Maximum entropy inverse reinforcement learning." In: *Aaai*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438.

[66] Wikipedia contributors. *Entropy*. https://en.wikipedia.org/wiki/Entropy. Accessed: January 17, 2024. 2024.

[67] Scott Fujimoto, Herke van Hoof, and David Meger. *Addressing Function Approximation Error in Actor-Critic Methods*. 2018. arXiv: 1802.09477 [cs.AI].

[68] Tim Brys et al. "Policy Transfer using Reward Shaping." In: *AAMAS*. 2015, pp. 181–188.

[69] Yaohui Guo, X. Jessie Yang, and Cong Shi. *Reward Shaping for Building Trustworthy Robots in Sequential Human-Robot Interaction*. 2023. arXiv: 2308.00945 [cs.RO].

[70] Arturo Daniel Sosa-Ceron, Hugo Gustavo Gonzalez-Hernandez, and Jorge Antonio Reyes-Avendaño. "Learning from Demonstrations in Humanndash;Robot Collaborative Scenarios: A Survey". In: *Robotics* 11.6 (2022). ISSN: 2218-6581. DOI: 10.3390/robotics11060126. URL: https://www.mdpi.com/2218-6581/11/6/126.

[71] Jangwon Lee. *A survey of robot learning from demonstrations for Human-Robot Collaboration*. 2017. arXiv: 1710.08789 [cs.RO].

[72] Andrei A. Rusu et al. *Policy Distillation*. 2016. arXiv: 1511.06295 [cs.LG].

[73] Wenhao Yu, C. Karen Liu, and Greg Turk. *Protective Policy Transfer*. 2020. arXiv: 2012.06662 [cs.RO].

[74] Matthew E. Taylor, Peter Stone, and Yaxin Liu. "Transfer Learning via Inter-Task Mappings for Temporal Difference Learning". In: *Journal of Machine Learning Research* 8.1 (2007), pp. 2125–2167.

[75] Anestis Fachantidis et al. "Transfer Learning via Multiple Inter-task Mappings". In: Sept. 2011, pp. 225–236. ISBN: 978-3-642-29945-2. DOI: 10.1007/978-3-642-29946-9_23.

[76] Andrei A. Rusu et al. *Progressive Neural Networks*. 2022. arXiv: 1606.04671 [cs.LG].

[77] Xiaoqin Zhang and Huimin Ma. *Pretraining Deep Actor-Critic Reinforcement Learning Algorithms With Expert Demonstrations*. 2018. arXiv: 1801.10459 [cs.AI].

[78] Thomas B Sheridan. "Human–robot interaction: status and challenges". In: *Human factors* 58.4 (2016), pp. 525–532.

[79] Anthony R Lanfranco et al. "Robotic surgery: a current perspective". In: *Annals of surgery* 239.1 (2004), p. 14.

[80] Won Hyuk Chang and Yun-Hee Kim. "Robot-assisted therapy in stroke rehabilitation". In: *Journal of stroke* 15.3 (2013), p. 174.

[81] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. "Social robotics". In: *Springer handbook of robotics* (2016), pp. 1935–1972.

[82] Tony Belpaeme et al. "Social robots for education: A review". In: *Science robotics* 3.21 (2018), eaat5954.

[83] Shelly Sicat et al. "Playing the mirror game with a humanoid: Probing the social aspects of switching interaction roles". In: Aug. 2017, pp. 1078–1083. DOI: 10.1109/ROMAN.2017.8172437.

[84] Gregory Kahn et al. *Self-supervised Deep Reinforcement Learning with Generalized Computation Graphs for Robot Navigation*. 2018. arXiv: 1709.10489 [cs.LG].

[85] Lucia Liu et al. "Robot Navigation in Crowded Environments Using Deep Reinforcement Learning". In: Oct. 2020, pp. 5671–5677. DOI: 10.1109/IROS45743.2020.9341540.

[86] Stephen James and Edward Johns. *3D Simulation for Robot Arm Control with Deep Q-Learning*. 2016. arXiv: 1609.03759 [cs.RO].

[87] Fangyi Zhang et al. *Towards Vision-Based Deep Reinforcement Learning for Robotic Motion Control*. 2015. arXiv: 1511.03791 [cs.LG].

[88] Marwan Qaid Mohammed, Kwek Lee Chung, and Chua Shing Chyi. "Review of Deep Reinforcement Learning-Based Object Grasping: Techniques, Open Challenges, and Recommendations". In: *IEEE Access* 8 (2020), pp. 178450–178481. DOI: 10.1109/ACCESS.2020.3027923.

[89] Shirin Joshi, Sulabh Kumra, and Ferat Sahin. *Robotic Grasping using Deep Reinforcement Learning*. 2020. arXiv: 2007.04499 [cs.RO].

[90] Victoria Hodge, Richard Hawkins, and Rob Alexander. "Deep reinforcement learning for drone navigation using sensor data". In: *Neural Computing and Applications* 33 (Mar. 2021). DOI: 10.1007/s00521-020-05097-x.

[91] Wu Chunxue et al. "UAV Autonomous Target Search Based on Deep Reinforcement Learning in Complex Disaster Scene". In: *IEEE Access* PP (Aug. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2019.2933002.

[92] Tuomas Haarnoja et al. *Learning to Walk via Deep Reinforcement Learning*. 2019. arXiv: 1812.11103 [cs.LG].

[93] Ali Shafti et al. *Real-World Human-Robot Collaborative Reinforcement Learning*. 2020. arXiv: 2003.01156 [cs.RO].

[94] Fotios Lygerakis, Maria Dagioglou, and Vangelis Karkaletsis. "Accelerating Human-Agent Collaborative Reinforcement Learning". In: *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*. PETRA '21. Corfu, Greece: Association for Computing Machinery, 2021, pp. 90–92. ISBN: 9781450387927. DOI: 10.1145/3453892.3454004. URL: https://doi.org/10.1145/3453892.3454004.

[95] Andrei A. Rusu et al. *Progressive Neural Networks*. 2022. arXiv: 1606.04671 [cs.LG].

[96] André Barreto et al. *Successor Features for Transfer in Reinforcement Learning*. 2018. arXiv: 1606.05312 [cs.AI].

[97] Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. "Comparing Objective and Subjective Measures of Usability in a Human-Robot Dialogue System". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Ed. by Keh-Yih Su et al. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 879–887. URL: https://aclanthology.org/P09-1099.

[98] Annika Silvervarg and Arne Jönsson. "Subjective and Objective Evaluation of Conversational Agents in Learning Environments for Young Teenagers". In: (May 2012).

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

[100] Mateusz Paliga. "The Relationships of Human-Cobot Interaction Fluency with Job Performance and Job Satisfaction among Cobot Operatorsmdash;The Moderating Role of Workload". In: *International Journal of Environmental Research and Public Health* 20.6 (2023). ISSN: 1660-4601. URL: https://www.mdpi.com/1660-4601/20/6/5111.

[101] Wei Yang et al. "Human grasp classification for reactive human-to-robot handovers". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 11123–11130.

[102] Sadaf Hussain et al. "Trait Based Trustworthiness Assessment in Human-Agent Collaboration Using Multi-Layer Fuzzy Inference Approach". English. In: *IEEE Access* 9 (2021). Funding Information: This work was supported in part by the Korea Institute for Advanced Study (KIAS) under Grant CG076601, and in part by the Sejong University Faculty Research Fund. Publisher Copyright: © 2013 IEEE., pp. 73561–73574. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3079838.

[103] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. "How Do You Like Your Virtual Agent?: Human-Agent Interaction Experience through Nonverbal Features and Personality Traits". In: *Human Behavior Understanding*. Ed. by Hyun Soo Park et al. Cham: Springer International Publishing, 2014, pp. 1–15.

[104] Gerald Matthews et al. “Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems”. In: *Personality and Individual Differences* 169 (2021). Celebrating 40th anniversary of the journal in 2020, p. 109969. ISSN: 0191-8869. DOI: https://doi.org/10.1016/j.paid.2020.109969. URL: https://www.sciencedirect.com/science/article/pii/S0191886920301586.

[105] Goldberg. “An alternative "description of personality": The Big-Five factor structure.” In: 59 (1990). DOI: 10.1037/0022-3514.59.6.1216.

[106] Ioannis Tsaousis. “The traits personality questionnaire (TPQue): A Greek measure for the five factor model”. In: *Personality and Individual Differences* 26.2 (1998), pp. 271–283. ISSN: 0191-8869. DOI: https://doi.org/10.1016/S0191-8869(98)00131-7. URL: https://www.sciencedirect.com/science/article/pii/S0191886998001317.

[107] Shalom H. Schwartz. “An Overview of the Schwartz Theory of Basic Values”. In: *Online Readings in Psychology and Culture* 2.1 (2012). Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License. DOI: 10.9707/2307-0919.1116. URL: https://doi.org/10.9707/2307-0919.1116.

[108] Shalom H. Schwartz. “Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries”. In: ed. by Mark P. Zanna. Vol. 25. Advances in Experimental Social Psychology. Academic Press, 1992, pp. 1–65. DOI: https://doi.org/10.1016/S0065-2601(08)60281-6. URL: https://www.sciencedirect.com/science/article/pii/S0065260108602816.

[109] Astrid Schepman and Paul Rodway. “Initial validation of the general attitudes towards Artificial Intelligence Scale”. In: *Computers in Human Behavior Reports* 1 (2020), p. 100014. ISSN: 2451-9588. DOI: https://doi.org/10.1016/j.chbr.2020.100014. URL: https://www.sciencedirect.com/science/article/pii/S2451958820300142.

[110] Astrid Schepman and Paul Rodway. “The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust”. In: *International Journal of Human-Computer Interaction* 39 (June 2022), pp. 1–18. DOI: 10.1080/10447318.2022.2085400.

[111] Yufei Wang and Tianwei Ni. *Meta-SAC: Auto-tune the Entropy Temperature of Soft Actor-Critic via Metagradient*. 2020. arXiv: 2007.01932 [cs.LG].

[112] Yaosheng Xu et al. *Target Entropy Annealing for Discrete Soft Actor-Critic*. 2021. arXiv: 2112.02852 [cs.LG].