Ατομική Διπλωματική Εργασία

ΜΕΛΕΤΗ ΣΧΕΣΗΣ ΑΥΤΟΕΛΕΓΧΟΥ ΚΑΙ ΣΥΝΕΙΔΗΤΟΤΗΤΑΣ

Άννα Γεωργίου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2015

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μελέτη σχέσης αυτοελέγχου και συνειδητότητας

Άννα Γεωργίου

Επιβλέπων Καθηγητής
Δρ. Χρίστος Χριστοδούλου

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2015

Ευχαριστίες

Θα ήθελα να εκφράσω ιδιαίτερες ευχαριστίες προς τον επιβλέποντα καθηγητή μου Δρ. Χρίστο Χριστοδούλου για τη συνεχή στήριξη και καθοδήγησή του όλους αυτούς τους μήνες. Επίσης θα ήθελα να ευχαριστήσω τον διδακτορικό φοιτητή Βασίλη Βασιλειάδη για την πολύτιμη βοήθεια που μου προσέφερε καθ' όλη τη διάρκεια της εργασίας μου. Τέλος, ένα μεγάλο ευχαριστώ στην οικογένειά μου που είναι πάντα δίπλα μου εμψυχώνοντας με και δίνοντάς μου δύναμη για το μέλλον.

Περίληψη

Στόχος αυτής της διπλωματικής εργασίας είναι η μελέτη της έννοιας του αυτοέλεγχου (self-control) σε συνδυασμό με την έννοια της συνειδητότητας (consciousness). Στην παρούσα διπλωματική εργασία μοντελοποιούμε το άνω και το κάτω μέρος του εγκεφάλου με ένα βασικό υπολογιστικό μοντέλο υπό μορφή πινάκων (Q-tables). Οι πίνακες αυτοί αντιπροσωπεύουν δύο πράκτορες (ένα για κάθε μέρος του εγκεφάλου) οι οποίοι παίζουν το παιγνίδι του επαναλαμβανόμενου διλήμματος του φυλακισμένου (Iterated Prisoner's Dilemma) με τελικό σκοπό να συμβιβαστούν, αναπτύσσοντας έτσι την έννοια του αυτοελέγχου. Οι πράκτορες εκπαιδεύονται με τη μέθοδο της ενισχυτικής μάθησης και πιο συγκεκριμένα με την τεχνική SARSA (State-Action-Reward-State-Action), η οποία βασίζεται στη μέθοδο χρονικών διαφορών (Temporal Difference). Η επιλογή επόμενης κίνησης από τον κάθε πράκτορα γίνεται με βάση μία πιθανότητα epsilon (ε), ακολουθώντας την πολιτική ε-greedy. Η συνειδητότητα από την άλλη μεριά μοντελοποιείται μέσω του πίνακα αμοιβών του παιχνιδιού. Από αυτόν τον πίνακα οι πράκτορες θα λαμβάνουν αμοιβή μετά από κάθε κίνησή τους. Ο πίνακας αποτελείται από τις τιμές S,P,R,T και η συνειδητότητα θα αναπαρίσταται από διαφορές ανάμεσα σε ζεύγη τιμών, εκ των οποίων η μία αντιστοιχεί σε αμοιβή συνεργασίας και η άλλη αμοιβή μη συνεργασίας του παίκτη. Όσο μεγαλύτερη διαφορά υπάρχει στις τιμές (S-T, S-P και R-T) του πίνακα, τόσο μεγαλύτερη η συνειδητότητα. Για να εξακριβώσουμε τη σχέση αυτοελέγχου και συνειδητότητας, χρησιμοποιούμε την τεχνική αναρρίχηση λόφων (hill climbing) και διαφοροποιούμε τις τιμές του πίνακα αμοιβών κάνοντας χρήση της γκαουσιανής κατανομής (Gaussian Distribution). Εφαρμόζοντάς τους νέους πίνακες στο υπολογιστικό μοντέλο με τα δύο μέρη του εγκεφάλου, παρατηρήσαμε πως αυτές οι αλλαγές επηρεάζουν τη συνεργασία ανάμεσα στους πράκτορες και κατ' επέκταση την ανάπτυξη του αυτοελέγχου. Διεξάχθηκαν διάφορα πειράματα, με διάφορες τιμές αμοιβών χρησιμοποιώντας τρεις διαφορετικές μεθόδους. Με τυχαίους πίνακες αμοιβών, κρατώντας σταθερό P και R στον πίνακα αμοιβών και τροποποίηση ολόκληρου του πίνακα αμοιβών. Τα αποτελέσματα που παίρνουμε δείχνουν τελικά πως αυτοέλεγχος και συνειδητότητα συνδέονται μέσω μιας αντιστρόφως ανάλογης σχέσης και αποδεικνύουν ότι εν όσον η συνειδητότητα μειώνεται, ο αυτοέλεγχος αυξάνεται. Άρα με λίγα λόγια άτομα με ισχυρό αυτοέλεγχο, δε χρειάζονται κατά ανάγκη μεγάλη συνειδητότητα για να

αποφασίσουν ποια είναι η σωστή συμπεριφορά που πρέπει να ακολουθήσουν σε κάθε περίσταση.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Στόχος Διπλωματικής Εργασίας

Αυτή η διπλωματική εργασία έγινε με απώτερο σκοπό να ερευνήσουμε τη σχέση ανάμεσα στον αυτοέλεγχο και την συνειδητότητα του ανθρώπου. Ο αυτοέλεγχος είναι μια πολύ σημαντική «αρετή» για τον καθένα από εμάς αφού καθορίζει πολλές φορές την συμπεριφορά και τις επιλογές μας. Γι' αυτό το λόγο η έννοια του αυτοελέγχου θα μπορούσε να οριστεί ως η επιλογή μιας μεγάλης καθυστερημένης αμοιβής, έναντι μιας μικρότερης αμοιβής επί του παρόντος (Christodoulou et al., 2010), (Rachlin, 2000). Όλοι μας καθημερινά ενεργούμε με κινήσεις οι οποίες είναι αποτελέσματα έλλειψης ισχυρού αυτοελέγχου αφού ενώ ξέρουμε ποιο είναι το σωστό για εμάς, δεν ενεργούμε κατά αυτόν τον τρόπο. Ουσιαστικά πριν από κάθε κίνησή μας υπάρχει μια σύγκρουση ανάμεσα σε αυτές τις δυο επιλογές που έχουμε, το να κάνουμε το σωστό (cognition) λαμβάνοντας αργότερα μια μεγάλη αμοιβή ή να ενεργήσουμε με βάση την επιθυμία/ένστικτο λαμβάνοντας τώρα μια αμοιβή μικρότερης τιμής, όποια κι αν είναι αυτή. Η σύγκρουση αυτή της γνωστικής λειτουργίας (cognition) με το κίνητρο (motivation) (Rachlin, 1995) συνδέεται με την σύγκρουση που συμβαίνει ανάμεσα στα δυο μέρη του εγκεφάλου άνω και κάτω (Bjork et al., 2004) πριν από την απόφαση για κάποια ενέργεια.

Τα μέρη του εγκεφάλου θα μοντελοποιηθούν με ένα υπολογιστικό βασικό μοντέλο υπό μορφή πινάκων (Q-tables) βασισμένο σε παλιότερο μοντέλο αυτοελέγχου (Christodoulou et al., 2010). Το μοντέλο θα αποτελείται από δύο πίνακες (ένα για κάθε μέρος του

εγκεφάλου) οι οποίοι θα παίζουν το παιγνίδι του επαναλαμβανόμενου διλήμματος του φυλακισμένου (Iterated Prisoner's Dilemma - IPD) (Rappoport & Chammah, 1965). Το IPD αποτελεί ουσιαστικά ένα παιχνίδι γενικού αθροίσματος το οποίο επαναλαμβάνεται με σκοπό την εκπαίδευση των πρακτόρων. Οι δυο πράκτορες λειτουργούν ανταγωνιστικά αφού καθένας από αυτούς αντιλαμβάνεται το περιβάλλον και ενεργεί με διαφορετικό τρόπο σε κάθε επανάληψη. Και οι δύο πράκτορες όμως σε κάθε επανάληψη θα έχουν τις επιλογές C και D για την επόμενη τους κίνηση, δηλαδή είτε να συνεργαστούν (C), είτε όχι (D). Οι πίνακες/πράκτορες θα εκπαιδεύονται με ενισχυτική μάθηση (Sutton & Barto, 1998) και για την εκπαίδευση τους θα χρησιμοποιηθεί μία από τις δύο τεχνικές, Q-Learning ή SARSA (State-Action-Reward-State-Action), οι οποίες βασίζονται στη μέθοδο χρονικών διαφορών (Sutton, 1988) (Temporal Difference). Μέσα από την εκπαίδευση τους οι πράκτορες θα μαθαίνουν να συνεργάζονται και οι δυο, καταλήγοντας έτσι στο συμβιβασμό.

Από θέμα σχεδίασης οι δύο πράκτορες θα είναι ίδιοι με μόνη διαφορά τον παράγοντα έκπτωσης (γ) στις εξισώσεις μάθησης. Από τη μια πλευρά για το κάτω μέρος του εγκεφάλου που προτιμά την γρήγορη μικρή αμοιβή (SS-Smaller Sooner) θα χρησιμοποιούμε μικρό παράγοντα έκπτωσης κοντά στο 0 έτσι ώστε η αξία των μελλοντικών κινήσεων να επιδέχεται μεγάλη έκπτωση και ο πράκτορας να συμπεριφέρεται με βάση το ένστικτο. Από την άλλη, για το άνω μέρος του εγκεφάλου το οποίο επιδιώκει την μεγάλη καθυστερημένη αμοιβή (LL-Larger Later) (Rachlin, 1995) θα χρησιμοποιείται μεγάλος (κοντά στο 1) παράγοντας έκπτωσης για να δίνεται στον πράκτορα μια πιο λογική συμπεριφορά.

Οι πίνακες θα είναι δισδιάστατοι με μέγεθος ίσο με τις πιθανές καταστάσεις στις οποίες μπορεί να βρεθούν οι παίκτες ανάλογα με τις κινήσεις τους, επί τις κινήσεις που μπορεί να εκτελέσει ο παίκτης/πράκτορας την επόμενη στιγμή. Οι κινήσεις στη δική μας περίπτωση θα είναι μόνο δύο, C όταν ο πράκτορας συνεργάζεται, D όταν ο πράκτορας δεν συνεργάζεται και οι πιθανές καταστάσεις τέσσερις (CC, CD, DC, DD). Στα κελιά του πίνακα θα περιέχονται οι τιμές Q(s,a) που θα λαμβάνει ο πράκτορας ως απάντηση από το περιβάλλον ανάλογα με το ποια ήταν η κατάσταση (s) πριν και ποια κίνηση (a) επέλεξε ο πράκτορας να κάνει. Για τη διαδικασία επιλογής νέας κίνησης (a) από τους πράκτορες θα χρησιμοποιηθεί ο αλγόριθμος ε-greedy (Tokic, 2010), ο οποίος ανάλογα με την τιμή

της πιθανότητας ε, θα καθορίζει κατά πόσο ο πράκτορας θα εκτελεί κινήσεις για τις οποίες ήδη ξέρει τις αμοιβές (τεχνική εκμετάλλευσης-exploitation) ή αν θα επιχειρεί να εξερευνάει νέες κινήσεις (τεχνική εξερεύνησης-exploration). Σε κάθε περίπτωση οι τιμές στα κελιά του πίνακα θα αλλάζουν ανάλογα με την ανταμοιβή ή τιμωρία που θα λάβει ο παίκτης μετά την εκτέλεση μιας κίνησης.

Εφόσον οι δύο πράκτορες μάθουν να συμβιβάζονται θα μπορούμε να εντάξουμε την έννοια της συνειδητότητας στην έρευνά μας. Η συνειδητότητα θα αναπαρίσταται μέσω των τιμών του πίνακα αμοιβών τις οποίες θα λαμβάνει ο πράκτορας ως ανταμοιβή ή τιμωρία στο παιγνίδι IPD. Ο πίνακας αυτός θα περιέχει την μεταβλητή T η οποία θα αντιστοιχεί στο ενισχυτικό σήμα που λαμβάνει ο πράκτορας που δεν συνεργάστηκε την στιγμή που ο άλλος πράκτορας συνεργάστηκε και πήρε S. Τη μεταβλητή R η οποία είναι η αμοιβή που λαμβάνουν και οι δυο πράκτορες εάν συνεργαστούν και την P η οποία είναι η τιμωρία που λαμβάνουν και οι δυο εάν δεν συνεργαστούν. Ξέροντας αυτά μπορούμε να μοντελοποιήσουμε τη συνειδητότητα μέσω της διαφοράς αυτών των τιμών. Πιο συγκεκριμένα θα αναπαρίσταται από διαφορές ανάμεσα σε ζεύγη τιμών, εκ των οποίων η μία αντιστοιχεί σε αμοιβή συνεργασίας και η άλλη αμοιβή μη συνεργασίας του παίκτη. Με αυτό τον τρόπο και λόγω της διαφοράς των τιμών ο παίκτης θα έρχεται αντιμέτωπος με δίλημμα και θα παρουσιάζεται εσωτερική σύγκρουση σε αυτόν. Επίσης, όσο πιο πολύ διαφέρουν οι τιμές S-T, S-P, R-T τόσο πιο μεγάλη θα είναι η εσωτερική σύγκρουση στον εγκέφαλο ενός ατόμου ο οποίος θα προσπαθεί να λειτουργήσει λογικά ή με το ένστικτο. Όσο πιο μεγάλη αυτή η εσωτερική σύγκρουση, τόσο μεγαλύτερος είναι και ο βαθμός συνειδητότητας του ατόμου.

Για να μελετήσουμε διάφορες περιπτώσεις, με μεγάλη αλλά και μικρή συνειδητότητα, θα κάνουμε διάφορα πειράματα με διαφορετικές τιμές (S, P, R, T) και θα χρησιμοποιήσουμε τον αλγόριθμο αναρρίχησης λόφων (hill climbing) σε τρεις διαφορετικές μεθόδους τροποποίησης του πίνακα αμοιβών. Στην πρώτη μέθοδο το παιγνίδι θα παίζεται σε κάθε γύρο με νέο τυχαίο πίνακα αμοιβών, στη δεύτερη μέθοδο θα μεταβάλλονται μόνο οι τιμές S και T του πίνακα και στην τρίτη μέθοδο θα μεταβάλλονται και οι τέσσερις τιμές του πίνακα μας. Τα πειράματα θα μας δώσουν μια ένδειξη ποια είναι τελικά η σχέση της συνειδητότητας με τον αυτοέλεγχο και πώς οι τιμές του πίνακα αμοιβών επηρεάζουν το δρόμο των πρακτόρων προς το να συμβιβαστούν και να καταλήξουν στον αυτοέλεγχο.

Στα επόμενα κεφάλαια θα δούμε γιατί επιλέχθηκαν αυτές οι τεχνικές, πως υλοποιήθηκαν και στο τέλος ποια αποτελέσματα επέφερε η έρευνα. Τέλος αναφέρονται τα συμπεράσματα της μελέτης και πως μπορεί να εξελιχθεί.

1.2 Προηγούμενη δουλειά – έρευνα

Η έννοια του αυτοελέγχου έχει μοντελοποιηθεί ήδη τα προηγούμενα χρόνια υπό μορφή δυο νευρωνικών δικτύων (Banfield, 2006), (Cleanthous, 2010) τα οποία αναπαριστούσαν τα δύο ξεχωριστά μέρη του εγκεφάλου. Χρησιμοποιήθηκε ενισχυτική μάθηση για την εκπαίδευσή τους, ενώ για την αλλαγή των βαρών του δικτύου χρησιμοποιήθηκαν διαφορετικές προσεγγίσεις σε κάθε νευρωνικό δίκτυο. Το κάτω μέρος του εγκεφάλου εκτελούσε αλλαγές στα βάρη με βάση τον κανόνα της επιλεκτικής εξερεύνησης (selective bootstrap) και από την άλλη, το άνω μέρος του εγκεφάλου χρησιμοποιούσε τη μέθοδο χρονικών διαφορών (temporal difference). Τα δίκτυα έπαιζαν και αυτά το παιγνίδι του Επαναλαμβανόμενου δίλημματος του Φυλακισμένου (Iterated Prisoner's Dilemma-IPD) και εκπαιδεύονταν μέσα από τις επαναλήψεις, όπως και στην δική μας περίπτωση οι δύο πίνακες.

Η απόφασή μας να αναπαραστήσουμε τα δυο κομμάτια του εγκεφάλου με Q-tables και όχι νευρωνικά δίκτυα πάρθηκε μετά από συζήτηση για τα δεδομένα που θα είχε να χειριστεί το πρόγραμμα και την πολυπλοκότητά του. Λόγω του ότι δεν θα έχουμε μεγάλο πλήθος καταστάσεων και κινήσεων, δεν θα έχουμε ούτε μεγάλο όγκο δεδομένων άρα ο στόχος μας μπορεί να επιτευχθεί και με πίνακες. Έτσι έχουμε ένα πρόγραμμα με λιγότερη πολυπλοκότητα, πιο γρήγορο και πιο εύκολο στο να κατανοηθεί.

Έρευνα σε αυτό το πεδίο έγινε και το 2010 μέσα από τη διδακτορική διατριβή του Αριστόδημου Κλεάνθους (Cleanthous, 2010) στην οποία πραγματοποιήθηκε μια εξερεύνηση της δομής της εσωτερικής σύγκρουσης που δημιουργείται ανάμεσα στους πράκτορες που παίζουν το παιχνίδι του επαναλαμβανόμενου διλήμματος του φυλακισμένου (IPD). Παράλληλα εξετάσθηκε το πως αυτή επηρεάζει τον αυτοέλεγχο του ατόμου μέσω διαφοροποίησης του πίνακα αμοιβών, ανάλογα με το πόσο μεγάλη είναι η σύγκρουση (conflict).

Κεφάλαιο 2

Γνωσιολογικό Υπόβαθρο

2.1 Αυτοέλεγχος

Ο εγκέφαλος θεωρείται από πολλούς ένα από τα ομορφότερα όργανα του ανθρώπινου σώματος. Ακόμη και σήμερα όμως, λειτουργίες του ανθρώπινου εγκεφάλου παραμένουν δύσκολες στο να εξηγηθούν έως και ανεξήγητες, πράγμα που τον καθιστά εξαιρετικά ενδιαφέρον θέμα για έρευνα και μελέτη.

Σε αυτό το υποκεφάλαιο θα δούμε την πτυχή του ανθρώπινου εγκεφάλου που ονομάζεται αυτοέλεγχος. Ο αυτοέλεγχος είναι μια συμπεριφορά η οποία έχει δεχθεί πολλούς ορισμούς τα τελευταία χρόνια, με τους περισσότερους να ορίζουν αυτή την έννοια ως τον έλεγχο των συναισθημάτων, των επιθυμιών αλλά και των πράξεων του εαυτού μας. Γι' αυτό και κατά τον Αριστοτέλη της αρχαίας Ελλάδας, ο άνθρωπος που μπορεί και ξεπερνά τις επιθυμίες του είναι γενναιότερος και από τον άνθρωπο που νικά του εχθρούς του. Μια λιγότερο θεωρητική εξήγηση είναι αυτή του Rachlin (Rachlin, 1995) η οποία αναφέρει τον αυτοέλεγχο ως την επιλογή μιας μεγάλης μακροπρόθεσμης ανταμοιβής (LL) έναντι μιας μικρότερης ανταμοιβής επί του παρόντος (SS). Ωστόσο εκτός από τον επιστημονικό

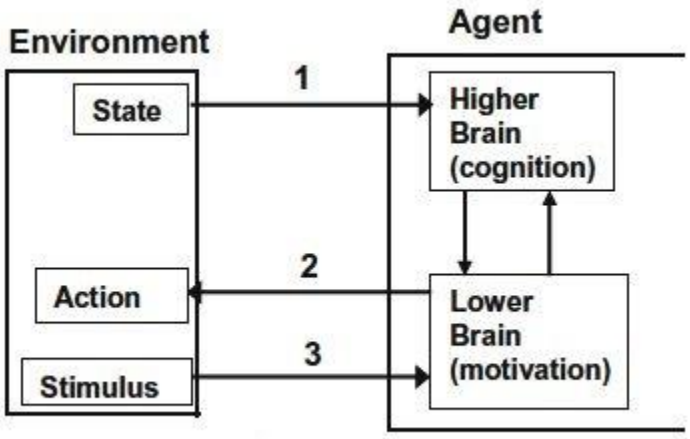
τομέα, η ανάγκη για να εξηγήσουμε την έννοια του αυτοελέγχου εμφανίζεται και στον θρησκευτικό τομέα. Ο Απόστολος Παύλος στην αποστολή του προς του Γαλάτες (5:22-23) περιγράφει τον αυτοέλεγχο ως έναν από τους καρπούς του πνεύματος, ενώ ο Απόστολος Πέτρος αναφέρει ότι η ανάπτυξη αυτοελέγχου είναι απαραίτητη για την σωτηρία του χριστιανού (Προς Πέτρο 1:5-8).

Ωστόσο, η συμπεριφορά του αυτοελέγχου χαρακτηρίζεται από πολλούς ως προβληματική και δύσκολη καθώς για κάποιους ανθρώπους είναι εύκολο να σκέφτονται λογικά και να πράττουν το σωστό ενώ για κάποιους άλλους όχι. Πολλές φορές ενώ γνωρίζουμε ποιο είναι το σωστό και ποια κίνηση θα ήταν πιο λογική, ενεργούμε με τον αντίθετο τρόπο λόγω απουσίας του κινήτρου που θα μας ωθούσε στην ορθή κατεύθυνση. Αυτή η αντιθετική συμπεριφορά ανάμεσα στη γνωστική λειτουργία και το κίνητρο οφείλεται στην έλλειψη αυτοελέγχου και έχει συνδεθεί με το άνω και κάτω μέρος του εγκεφάλου (Bjork et al., 2004), (Christodoulou et al., 2010).Τα δύο αυτά μέρη του εγκεφάλου συγκρούονται πριν από κάθε μας ενέργεια, αφού το πάνω μέρος αντιπροσωπεύει την λογική-σωστή συμπεριφορά ενώ το κάτω μέρος αντιπροσωπεύει τα πιο πρωτόγονα αισθήματα του αυθορμητισμού και του ενστίκτου. Άρα έχοντας υπόψη τον ορισμό που έδωσε ο Rachlin το 1995, ένας άνθρωπος ο οποίος λειτουργεί ενστικτωδώς δεν έχει αυτοέλεγχο και προτιμά να λάβει την βραχυπρόθεσμη, μικρότερη σε αξία, ανταμοιβή (SS) ενώ ο άνθρωπος που βλέπει μακροπρόθεσμα επιδιώκοντας την ανταμοιβή μεγαλύτερης αξίας (LL) έχει αυτοέλεγχο.

Για να δούμε πιο πρακτικά τη συμπεριφορά του αυτοελέγχου και να την κατανοήσουμε καλύτερα ας δούμε ένα παράδειγμα από τη καθημερινή μας ζωή. Έστω ένας φοιτητής πανεπιστημίου για τον οποίον έχουμε 2 περιπτώσεις ανταμοιβών. Έστω η μακροπρόθεσμη μεγάλης αξίας (LL) ανταμοιβή η οποία είναι να πάρει καλούς βαθμούς στο τέλος του εξαμήνου, ενώ η βραχυπρόθεσμη μικρής αξίας (SS) ανταμοιβή είναι να πάει για ποτό το βράδυ. Στην αρχή του εξαμήνου ο στόχος κάθε φοιτητή είναι οι καλοί βαθμοί στα μαθήματα, αλλά όταν δεχθούν πρόσκληση για το βράδυ η αξία της SS αμοιβής αρχίζει να μεγαλώνει έναντι των καλών βαθμών LL. Ένας φοιτητής ο οποίος διακατέχεται από αυτοέλεγχο όμως θα επιλέξει να μην πάει στο μπαρ για ποτό το βράδυ και να μείνει στο σπίτι να διαβάσει, έτσι ώστε να καταφέρει να φτάσει στην LL

ανταμοιβή. Με λίγα λόγια ο αυτοέλεγχος δίδει στο φοιτητή τη δύναμη να αντισταθεί στον πειρασμό, που στο παράδειγμα μας πιο πάνω είναι η βραδινή έξοδος (SS).

Στο σχήμα που ακολουθεί παρουσιάζεται ένα μοντέλο που περιγράφει την διαδικασία της σύγκρουσης των δύο μερών του εγκεφάλου (Rachlin, 2000). Ο άνθρωπος βρίσκεται σε μία κατάσταση (state) στην οποία καλείται να πάρει μια απόφαση για την επόμενη του ενέργεια. Λόγω αυτού, εισέρχονται αρχικά στον εγκέφαλο του ανθρώπου πληροφορίες και ερεθίσματα τα οποία καταλήγουν πρώτα στο πάνω μέρος του εγκεφάλου (βέλος 1). Το μέρος αυτό, ονομάζεται και γνωστικό σύστημα αφού είναι υπεύθυνο για τις γνώσεις και τη λογική συμπεριφορά. Οι νέες πληροφορίες θα αναμιχθούν με τα αισθήματα και την μνήμη από το κάτω μέρος του εγκεφάλου ή αλλιώς λιμπικό σύστημα. Η ανάμειξη θα «κατέβει» και πάλι στο ενδότερα του εγκεφάλου (κάτω μέρος) και θα επιστρέψει στο περιβάλλον την συμπεριφορά/ενέργεια (βέλος 2). Ανάλογα με το ποια συμπεριφορά επιλέχθηκε τελικά, ο άνθρωπος θα λάβει την σχετική ανταμοιβή και το ανάλογο ερέθισμα προς το λιμπικό σύστημα (βέλος 3).



Σχήμα 2.1: *Σχεδιάγραμμα μοντέλου, με τα δύο μέρη του εγκεφάλου (άνω και κάτω) να αλληλοεπιδρούν με το περιβάλλον.*

2.2 Δίλημμα του Φυλακισμένου (ΔΦ)

Για την μοντελοποίηση της σύγκρουσης που περιγράψαμε πιο πάνω αλλά και για την μοντελοποίηση των διαπροσωπικών συγκρούσεων χρησιμοποιήθηκε ένα παιγνίδι γενικού αθροίσματος το οποίο ονομάζεται Δίλημμα του Φυλακισμένου (Prisoner’s

Dilemma PD). Το παιγνίδι αυτό είναι ουσιαστικά παιγνίδι γενικού αθροίσματος κατά το οποίο αναλύονται αποφάσεις σε καταστάσεις στρατηγικής αλληλεξάρτησης. Επίσης το Δίλημμα του Φυλακισμένου έχει χρησιμοποιηθεί ευρέως για την ανάλυση καταστάσεων κοινωνικών διλημμάτων αλλά και στη μοντελοποίηση της συνεργασίας μεταξύ ανθρώπων.

Το παιγνίδι πήρε αυτή την ονομασία αφού η αρχική ιδέα το ήθελε να «παίζεται» από δύο ανθρώπους οι οποίοι είναι κατηγορούμενοι για κάποιο έγκλημα. Η αστυνομία στην προσπάθειά της να ανακαλύψει ποιος από τους δύο είναι ο πραγματικός ένοχος, βάζει τους δύο υπόπτους σε δύο διαφορετικά κελιά έτσι ώστε να μην έχουν επικοινωνία μεταξύ τους και εξετάζει έναν κρατούμενο τη φορά. Στο κάθε κρατούμενο γίνεται η εξής ίδια πρόταση από τον εισαγγελέα:

- Αν ο κρατούμενος ο οποίος εξετάζεται κατηγορήσει τον άλλο κρατούμενο, χωρίς να γνωρίζει ότι ο άλλος κρατούμενος αποφάσισε να μην μιλήσει, τότε αφήνεται ελεύθερος γιατί συνεργάστηκε με την αστυνομία. Ενώ ο άλλος κρατούμενος ο οποίος έμεινε σιωπηλός φυλακίζεται για μεγάλο χρονικό διάστημα.
- Αν κανένας από τους δύο κρατούμενους δε μιλήσει, τότε φυλακίζονται και οι δύο αλλά με μειωμένη ποινή φυλάκισης ο καθένας. Δηλαδή φυλακίζονται για λιγότερα χρόνια από ότι στην προηγούμενη περίπτωση.
- Αν και οι δύο κρατούμενοι κατηγορήσουν ο ένας τον άλλον, τότε φυλακίζονται με μια μέτρια ποινή φυλάκισης ο καθένας.

Άρα καθένας από τους κρατούμενους έχει δύο επιλογές. Ή να μιλήσει με την αστυνομία και να κατηγορήσει τον άλλο κρατούμενο (Defects), ή να μην μιλήσει καθόλου (Cooperates). Το γεγονός όμως ότι δεν γνωρίζουν τι θα επιλέξει ο άλλος κρατούμενος, καθιστά αβέβαιο το αν θα καταφέρουν τελικά το στόχο τους, δηλαδή να αφεθούν ελεύθεροι. Και οι δύο βλέπουν ουσιαστικά το προσωπικό τους όφελος παρόλο που εάν ήθελαν το συλλογικό καλό, το οποίο θα τους διασφάλιζε σίγουρα το ελάχιστο χρονικό διάστημα φυλάκισης, θα επέλεγαν να παραμείνουν σιωπηλοί και οι δύο.

Αν καθορίσουμε τις ποινές που είναι πιθανό να λάβουν οι δύο κρατούμενοι καταλήγουμε στην πιο κάτω ανισότητα η οποία αποτελεί τον βασικό κανόνα για το παιχνίδι του Διλήμματος του Φυλακισμένου:

T > R > P > S (κανόνας 1)

- S (Sucker) είναι η μεγάλη ποινή φυλάκιση που παίρνει ο κρατούμενος ο οποίος παρέμεινε σιωπηλός ενώ ο άλλος κρατούμενος τον κατηγόρησε.
- P (Punishment) είναι η μέτριας τιμής ποινή φυλάκισης που παίρνουν και οι δύο κρατούμενοι όταν κατηγορήσουν ο ένας τον άλλον.
- R (Reward) είναι η μικρότερη ποινή φυλάκισης που παίρνουν και οι δύο κρατούμενοι όταν παραμείνουν σιωπηλοί (συνεργαστούν).
- T (Temptation) αντιστοιχεί στην απελευθέρωση του κρατούμενου που συνεργάστηκε με την αστυνομία την στιγμή που ο άλλος κρατούμενος παρέμεινε σιωπηλός.

Με τις πιο πάνω τιμές δημιουργείται ο πίνακας αμοιβών (Σχήμα 2.2) που θα χρησιμοποιείται στο παιγνίδι του ΔΦ, ο οποίος είναι δισδιάστατος λόγω της παρουσίας δύο παικτών και στην προκειμένη περίπτωση δύο κρατουμένων.

	Cooperate	**Defect**
Cooperate	R , R	S , T
Defect	T , S	P , P

Σχήμα 2.2: *Πρότυπος πίνακας αμοιβών για το παιγνίδι του ΔΦ. T (Temptation), P (Punishment), R (Reward), S (Sucker)*

2.3 Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου (ΕΔΦ)

Στην παρούσα διπλωματική εργασία δε θα χρησιμοποιήσουμε το παιγνίδι ενός γύρου ΔΦ, αλλά το Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου, στο οποίο δύο παίκτες παίζουν επανειλημμένα το παιγνίδι ΔΦ. Έτσι τους δίνεται η ευκαιρία να επιλέξουν ενέργεια περισσότερες φορές και όχι απλά μία φορά. Όταν το ΔΦ παίζεται μόνο μία φορά, οι παίκτες έχοντας κατά νου το προσωπικό τους όφελος, τείνουν προς το να μην συνεργαστούν (cooperate) και να πάρουν το ρίσκο να κατηγορήσουν ο ένας τον άλλον. Όταν όμως το παιγνίδι παίζεται πολλές φορές η επιλογή του συμβιβασμού, δηλαδή να σιωπήσουν και οι δύο, φαντάζει πιο δελεαστική όσο περνάνε οι γύροι. Λόγω αυτής της επαναληπτικής συμπεριφοράς το παιγνίδι απέκτησε μεγαλύτερη πολυπλοκότητα και του προστέθηκε άλλο ένα κανόνας,

2R > T + S (κανόνας 2)

έτσι ώστε να διασφαλιστεί ότι κανένας παίκτης δε θα βρίσκεται σε καλύτερη θέση έναντι του άλλου.

Τώρα που εξηγήσαμε και το Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου, ας δούμε πως το παράδειγμα αυτοελέγχου που θέσαμε πριν με το φοιτητή και το μπαρ μπορεί να εφαρμοστεί στο παιγνίδι. Με αυτό τον τρόπο θα καταφέρουμε να εξηγήσουμε τη μοντελοποίηση του παιγνιδιού μέσα από πραγματικό παράδειγμα. Έστω λοιπόν πάλι ο φοιτητής στην αρχή του εξαμήνου. Ο μακροπρόθεσμος στόχος του, ο οποίος θα του δώσε τη μεγαλύτερη αμοιβή (LL) είναι οι καλοί βαθμοί στο τέλος του εξαμήνου. Όταν δέχεται μια πρόσκληση για να πάει σε μπαρ με φίλους του το βράδυ (SS), γεννιέται ένα δίλημμα για το τι πρέπει να κάνει. Το γεγονός ότι την επόμενη κιόλας ημέρα θα πρέπει να παραδώσει μια εργασία σε μάθημα του πανεπιστημίου, εντείνει περισσότερο το δίλημμα του φοιτητή. Δημιουργείται έτσι μια εσωτερική σύγκρουση στον φοιτητή, ανάμεσα στο λογικό κομμάτι του εαυτού του το οποίο του λέει να μείνει στο σπίτι για να διαβάσει (LL) και στο συναισθηματικό κομμάτι του το οποίο τον ωθεί στο να βγει με τους φίλους του να διασκεδάσει (SS). Αυτά τα δύο κομμάτια μπορούν να αναπαρασταθούν από δύο πράκτορες που παίζουν το ΕΔΦ. Ο κάθε πράκτορας μπορεί ή να συνεργαστεί (cooperate) ή όχι (defect), επιδιώκοντας το καλύτερο προσωπικό του όφελος μόνο. Άρα ο φοιτητής θα καταλήξει σε μία από τις 4 καταστάσεις ανάλογα με το πώς θα ενεργήσει ο καθένας από τους δύο πράκτορες:

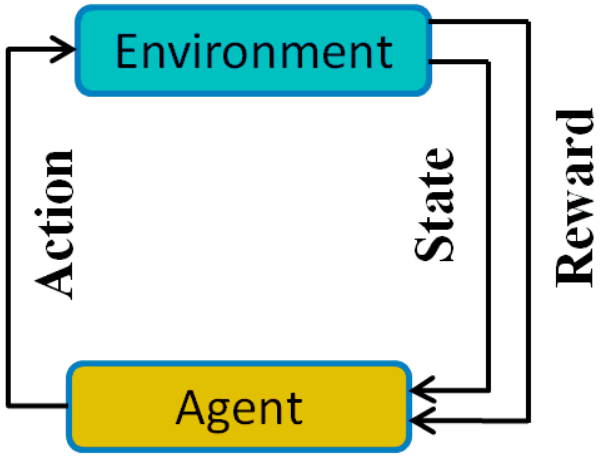
- **Κατάσταση 1 (CC) :** Και οι δύο πράκτορες συνεργάζονται. Ο φοιτητής πάει στο μπαρ για ένα γρήγορο ποτό και επιστρέφει νωρίς στο σπίτι για διάβασμα.
- **Κατάσταση 2 (CD) :** Ο πράκτορας της λογικής συνεργάζεται, ενώ ο πράκτορας των συναισθημάτων επιμένει. Ο φοιτητής πάει στο μπαρ και περνάει καλά.
- **Κατάσταση 3 (DC) :** Ο πράκτορας της λογικής επιμένει, ενώ ο πράκτορας των συναισθημάτων συνεργάζεται. Ο φοιτητής μένει στο σπίτι για να διαβάσει.
- **Κατάσταση 4 (DD) :** Κανείς από τους δύο δεν συνεργάζεται, και οι δύο πράκτορες επιμένουν. Ο φοιτητής δεν κάνει τίποτα από τις δύο επιλογές του διλήμματος. Δηλαδή είτε μένει στο σπίτι αλλά δεν μπορεί να συγκεντρωθεί στο διάβασμα, είτε πάει στο μπαρ αλλά δεν περνάει καλά επειδή νιώθει τύψεις που δεν διαβάζει.

Το πιο πάνω παράδειγμα μπορεί να γίνει πιο γενικό έτσι ώστε να μπορούμε να μιλούμε για μοντελοποίηση του αυτοελέγχου. Το δίλημμα και η εσωτερική σύγκρουση του φοιτητή θα αντιστοιχίζεται με την εσωτερική σύγκρουση που συμβαίνει στον εγκέφαλο κάθε ανθρώπου πριν από την απόφαση για κάποια ενέργεια. Άρα το παιγνίδι του Επαναλαμβανόμενου Διλήμματος του Φυλακισμένου θα παίζεται με πράκτορες το πάνω και το κάτω μέρος του εγκεφάλου που περιγράψαμε σε προηγούμενο υποκεφάλαιο. Στόχος του παιχνιδιού θα είναι ο άνθρωπος να πάρει την μακροπρόθεσμη μεγάλη ανταμοιβή που στοχεύει, καταλήγοντας όπως είπαμε σε αυτοέλεγχο. Ο αυτοέλεγχος όμως προϋποθέτει τη συνεργασία και των δύο πρακτόρων για το μέγιστο συλλογικό καλό. Άρα το παιγνίδι του ΔΦ θα επαναλαμβάνεται μέχρι να καταλήξουμε στην κατάσταση CC η οποία θα δώσει και τη μέγιστη ανταμοιβή. Για να καταλήξουν στην κατάσταση της αμοιβαίας συνεργασίας όμως οι πράκτορες θα πρέπει με κάποιο τρόπο να εκπαιδεύονται και να μαθαίνουν μέσα από τις επαναλήψεις του παιγνιδιού.

2.4 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση αποτελεί τη μία από τις τρεις κατηγορίες μάθησης που υπάρχουν (Επιβλεπόμενη, Μη-Επιβλεπόμενη και Ενισχυτική) για την εκπαίδευση συστημάτων με πράκτορες. Είναι όμως η πιο «φυσική» καθώς δεν υπάρχει δάσκαλος που να επιβλέπει τον πράκτορα αλλά στη θέση του έχουμε το περιβάλλον το οποίο αλληλοεπιδρά με τον πράκτορα. Αυτή η μέθοδος μάθησης μπορεί να χαρακτηριστεί και trial and error μάθηση, αφού με την απουσία δασκάλου ο πράκτορας μαθαίνει από τις συνέπειες των πράξεων του και επιλέγει τις κινήσεις του μαθαίνοντας από τις προηγούμενες επιλογές του αλλά και από καινούργιες επιλογές. Όταν ο πράκτορας βασίζεται σε παλιότερες κινήσεις του τότε λέμε ότι ακολουθεί τεχνική εκμετάλλευσης, ενώ όταν ρισκάρει να ανακαλύψει ποιες θα είναι οι συνέπειες από νέες κινήσεις λέμε ότι ακολουθεί τεχνική εξερεύνησης. Οι συνέπειες που έχει ο πράκτορας μετά από κάθε κίνηση, δεν είναι τίποτα άλλο από ένα ενισχυτικό αριθμητικό σήμα το οποίο μεταφράζεται σαν αμοιβή ή σαν τιμωρία ανάλογα με την τιμή του. Ο στόχος του πράκτορα ο οποίος μαθαίνει είναι πάντα να μεγιστοποιήσει τις αμοιβές του. Ένας ορισμός για την ενισχυτική μάθηση δόθηκε από τον Barto ως εξής: «Common-sense idea that if an action is followed by a satisfactory state of affairs, then the tendency to produce that action is strengthened i.e. Reinforced».

Η ενισχυτική μάθηση εφαρμόζεται συνήθως στο πλαίσιο διαδικασίας απόφασης Markov (πήρε το όνομα από τον Andrey Markov, 1856-1922) και τα απαραίτητα στοιχεία που πρέπει να υπάρχουν για να εφαρμοστεί η ενισχυτική μάθηση, είναι τρία: το περιβάλλον, ο πράκτορας και η κίνηση/ενέργεια. Το περιβάλλον παρέχει είσοδο (state) στον πράκτορα και δέχεται την έξοδο (action) του πράκτορα. Ανάλογα με την κίνηση που έκανε ο πράκτορας, το περιβάλλον του επιστρέφει ένα ενισχυτικό σήμα/ανταμοιβή (reward) το οποίο δεν λέει στον πράκτορα ποια κίνηση να κάνει, απλά αξιολογεί την παρούσα κίνηση. Άρα ο πράκτορας πρέπει να είναι σε θέση να δει σε ποια κατάσταση βρίσκεται το περιβάλλον και να κάνει μια ενέργεια είτε με εξερεύνηση, είτε με εκμετάλλευση γνώσης. Στο σχήμα 2.3 βλέπουμε την αλληλεπίδραση πράκτορα-περιβάλλοντος που περιεγράφηκε πιο πάνω.



Σχήμα 2.3: *Σχεδιάγραμμα με συστατικά τον πράκτορα και το περιβάλλον. Με την ενισχυτική μάθηση τα δυο συστατικά αλληλοεπιδρούν λαμβάνοντας και στέλνοντας πληροφορίες.*

Στη διπλωματική εργασία αυτή οι δύο πράκτορες θα εκπαιδεύονται με το είδος ενισχυτικής μάθησης το οποίο ονομάζεται Μέθοδος Χρονικών Διαφορών (Sutton, 1988). Η Μέθοδος Χρονικών Διαφορών είναι μια μέθοδος πρόβλεψης που χρησιμοποιείται στο πλαίσιο της Ενισχυτικής Μάθησης με σκοπό συνήθως την πρόβλεψη της αναμενόμενης αμοιβής για κάθε κατάσταση. Βασικός κανόνας αυτής της μεθόδου είναι ο εξής:

NewEstimate = OldEstimate + StepSize [Target – OldEstimate]

Και όπως βλέπουμε από την εξίσωση η πρόβλεψη για την νέα αμοιβή βασίζεται στην προηγούμενη πρόβλεψη. Στα επόμενα υποκεφάλαια θα δούμε τις δύο τεχνικές με τις οποίες μπορεί να εφαρμοστεί η Μέθοδος Χρονικών Διαφορών, τον αλγόριθμο Q-Learning και τον αλγόριθμο SARSA.

2.4.1 Αλγόριθμος Q-Learning

Κύριο χαρακτηριστικό του αλγορίθμου Q-Learning (Watkins et al., 1992) (Watkins et al., 1989) είναι ότι ονομάζεται και αλγόριθμος εκτός πολιτικής αφού η πολιτική εκτίμησης μπορεί να είναι διαφορετική από την πολιτική με την οποία θα συμπεριφέρεται τελικά ο πράκτορας. Αρχικά χρησιμοποιεί μια οποιαδήποτε πολιτική για να υπολογίσει τη βέλτιστη τιμή Q* και όταν βρει τη βέλτιστη πολιτική ακολουθεί μόνο αυτή. Η εξίσωση που εκτελείται για τις νέες τιμές των πρακτόρων είναι η εξής:

$$\mathbf{Q(s,a) \leftarrow (1-\alpha)*Q(s,a) + \alpha*[\, R(s) + \gamma*\max_{a'} Q(s',a')]}$$

Όπου **Q(s,a)** είναι η εκτίμηση της εκπτωτικής συσσωρευμένης αμοιβής που θα λάβει ο πράκτορας ξεκινώντας από την κατάσταση **s** και κάνοντας την κίνηση **a**. Επίσης $\mathbf{\max_{a'} Q(s',a')}$ είναι η βέλτιστη πολιτική που πρέπει να ακολουθήσει ο πράκτορας ενώ βρίσκεται στην κατάσταση **s**. Αρχικά όμως αυτή η βέλτιστη τιμή Q δεν υπάρχει, αλλά δημιουργείται με τις επαναλήψεις του προγράμματος και καταλήγει εν τέλει στο βέλτιστο ή στο όσο πιο βέλτιστο γίνεται. Το **γ** που έχουμε στην εξίσωση είναι ο εκπτωτικός παράγοντας ο οποίος ανάλογα με την τιμή του δίνει λιγότερη ή περισσότερη αξία στις μελλοντικές αμοιβές, ενώ το **α** είναι ο ρυθμός μάθησης του πράκτορα, δηλαδή πόσο γρήγορα ή αργά μαθαίνει. Λόγω του περιορισμένου πλήθους καταστάσεων και κινήσεων αλλά και ότι το περιβάλλον δεν αλλάζει, ο αλγόριθμος Q-Learning είναι σχεδόν σίγουρο ότι στο τέλος της εκτέλεσης θα συγκλίνει στις βέλτιστες τιμές.

2.4.2 Αλγόριθμος SARSA

Ο αλγόριθμος SARSA (Sutton, 1996) είναι ακρώνυμο από τις λέξεις State-Action-Reward-State-Action και η κύρια διαφορά του από τον αλγόριθμο Q-Learning είναι ότι είναι αλγόριθμος εντός πολιτικής. Εκτός αυτού μια άλλη διαφορά είναι ότι για την εκτίμηση της αμοιβής μιας πράξης λαμβάνεται υπόψη και η επόμενη κίνηση του πράκτορα και όχι η καλύτερη δυνατή κίνηση όπως γίνεται στον Q-Learning, εξ ου και το

όνομα του (State-Action-Reward-State-Action). Αυτό φαίνεται και από την εξίσωση υπολογισμού των τιμών Q:

$$\mathbf{Q(s,a) \leftarrow (1-\alpha)*Q(s,a) + \alpha*[R(s) + \gamma*Q(s',a')]}$$

Όπου και πάλι **Q(s,a)** είναι η εκτίμηση της εκπτωτικής συσσωρευμένης αμοιβής που θα λάβει ο πράκτορας ξεκινώντας από την κατάσταση **s** και κάνοντας την κίνηση **a**, Το **γ** που έχουμε στην εξίσωση είναι ο εκπτωτικός παράγοντας ο οποίος ανάλογα με την τιμή του δίνει λιγότερη ή περισσότερη αξία στις μελλοντικές αμοιβές, ενώ το **α** είναι ο ρυθμός μάθησης του πράκτορα, δηλαδή πόσο γρήγορα ή αργά μαθαίνει. Εκπτωτικός παράγοντας κοντά στο 0 (0.1) κάνει τον πράκτορα να επιδιώκει τις πλησιέστερες αμοιβές (SS) ενώ παράγοντας κοντά στο 1 (0.9), κάνει τον πράκτορα να ενδιαφέρεται μόνο για μακροπρόθεσμες μεγάλες αμοιβές (LL). Ουσιαστικά στον αλγόριθμο SARSA ακολουθείται μια τυχαία πολιτική και ο πράκτορας προσπαθεί να την κάνει καλύτερη μαθαίνοντας μέσα από τις κινήσεις εξερεύνησης που κάνει.

Όπως και ο αλγόριθμος Q-Learning έτσι και ο SARSA εν τέλει συγκλίνει στη βέλτιστη λύση αλλά σε μικρότερο χρονικό διάστημα, πράγμα που του δίνει ένα πλεονέκτημα στη χρήση του. Για να εξακριβωθεί αν στον εγκέφαλο έχουμε ενισχυτική μάθηση και τι είδους μάθηση, έχουν γίνει ήδη κάποιες έρευνες όπως αυτή των (Morris et al., 2006). Το 2006 λοιπόν, προσπάθησαν να εκπαιδεύσουν πιθήκους μέσω ενός πειράματος όπου παρουσιάζονταν στους πίθηκους κάποιες υποδείξεις οι οποίες και προέβλεπαν την ανταμοιβή με διαφορετικές πιθανότητες ανάλογα με την επιλογή που έκανε ο πίθηκος. Μετά από μερικές επαναλήψεις του πειράματος οι πίθηκοι είχαν την δυνατότητα να διαλέξουν ανάμεσα σε δύο ενδείξεις που τους παρουσιάζονταν ταυτόχρονα. Σε αυτούς τους γύρους του πειράματος οι ερευνητές κατέγραψαν ότι η ένδειξη που επέλεγε ο πίθηκος προκαλούσε εκτόξευση ντοπαμινεργικών ουσιών στον εγκέφαλό τους. Το ποσοστό της ντοπαμίνης που καταγραφόταν, παρατηρήθηκε ότι ταίριαζε με το ποσοστό σφάλματος της επόμενης επιλογής του πίθηκου. Αυτό μπορεί εύκολα να συνδεθεί με τον τρόπο που εκτελείται ο αλγόριθμος SARSA. Στην παρούσα διπλωματική εργασία αφού θέλουμε να παρατηρήσουμε τη σχέση αυτοελέγχου και συνειδητότητας θα χρησιμοποιήσουμε τον αλγόριθμο SARSA για το κομμάτι της συνειδητότητας, αφού είναι πιο κοντά στο πως λειτουργεί ο εγκέφαλος.

2.5 Πολιτική ε-greedy

Όπως είπαμε πιο πάνω οι πράκτορες θα εκπαιδεύονται με τον αλγόριθμο ενισχυτικής μάθησης SARSA. Το δίλημμα που γεννάται τώρα, είναι ποια τακτική θα ακολουθούν οι πράκτορες στην απόφασή τους για την επόμενη τους κίνηση. Θα ακολουθούν τακτική εξερεύνησης ή εκμετάλλευσης γνώσης; Το μόνο σίγουρο είναι ότι αρχικά και οι δύο θα πρέπει να εξερευνήσουν και να βρεθούν σε νέες κινήσεις έτσι ώστε να αποκτήσουν μια πρώτη εικόνα του περιβάλλοντος και των αμοιβών. Εφόσον οι πράκτορες βρουν κάποιες κινήσεις που τους δίνουν μεγάλες ανταμοιβές, θα μπορούν να ακολουθούν αυτές τις κινήσεις έτσι ώστε να μεγιστοποιήσουν τις αμοιβές τους. Μια καλή πολιτική που θα χρησιμοποιήσουμε μαζί με την ενισχυτική μάθηση είναι η πολιτική ε-greedy. Η πολιτική αυτή δεν είναι ούτε άπληστη, ούτε μη άπληστη, αλλά κάτι ενδιάμεσο λόγω του *έψιλον*.

Η πολιτική ε-greedy εξισορροπεί τις τακτικές εξερεύνησης και εκμετάλλευσης βάσει του *έψιλον (ε)* το οποίο αποτελεί την πιθανότητα με την οποία ο πράκτορας θα εξερευνά. Αν για παράδειγμα το $\varepsilon=0.1$ τότε ο πράκτορας θα εξερευνά νέες κινήσεις με μια πιθανότητα 0.1 ενώ θα εκμεταλλεύεται την ήδη υπάρχουσα γνώση με πιθανότητα 0.9. Δηλαδή οι πράκτορες μας θα επιλέγουν μια κίνηση η οποία πιστεύουν ότι έχει την καλύτερη μακροπρόθεσμη αμοιβή βάσει πιθανότητας $1-\varepsilon$, ενώ θα επιλέγουν τυχαία μια νέα κίνηση με πιθανότητα ε. Άρα το έψιλον παίρνει τιμές από $0 < \varepsilon < 1$.

2.6 Συνειδητότητα

Το δεύτερο κομμάτι αυτής της διπλωματικής, εάν θέσουμε ως το πρώτο να είναι το κομμάτι του αυτοελέγχου, είναι η συνειδητότητα. Αν και η έννοια της συνειδητότητας αποτελεί μία από τις έννοιες που σχετίζονται με τον εγκέφαλο και είναι δύσκολο να εξηγηθεί, θα προσπαθήσουμε να την εξηγήσουμε, να την μοντελοποιήσουμε αλλά και να την συνδέσουμε με τον αυτοέλεγχο. Μερικοί ορισμοί για την έννοια της συνειδητότητας την αναφέρουν ως το «να έχει κανείς αντίληψη, σκέψεις, συναισθήματα και ευαισθητοποίηση» (Sutherland, 1989). Ένας καθηγητής νευροψυχολογίας την ίδια χρονική περίοδο χαρακτήρισε την «συνειδητότητα ως το πιο προφανές αλλά και το πιο μυστήριο πράγμα στον εγκέφαλό μας» (Dennett, 1991). Με τον ίδιο περίπου τρόπο περιγράφεται και στο λεξικό "Random House Dictionary of the English Language", όπου η συνειδητότητα αναφέρεται ως η αντίληψη της ύπαρξής μας, των αισθήσεων, σκέψεων

αλλά και ό,τι μας περιβάλλει. Επίσης στο ίδιο λεξικό δίνεται και ένας δεύτερος ορισμός ο οποίος χαρακτηρίζει την συνειδητότητα ως τις σκέψεις και τα συναισθήματα συλλογικά ενός ατόμου ή μιας ομάδας ατόμων. Πρώτος από όλους για τη συνειδητότητα μίλησε ο Ρενέ Ντεκάρτ (Καρτέσιος) (Ρενέ, 1629) και την χαρακτήρισε ως την ενσυνείδητη κατάσταση η οποία είναι μια άμεση, ολοφάνερη και θεμελιώδης εμπειρία ζωής. Αυτό αποτέλεσε και για τον Ρενέ Ντεκάρτ τη σταθερή αφετηρία στο έργο του για την «ανάπλαση» της πραγματικότητας στο οποίο και ανέλυε περαιτέρω την συνειδητότητα.

Πρακτικά τώρα, έρευνες (Morsella et al., 2009a), (Morsella et al., 2009b), έχουν γίνει ήδη μέσω πειραμάτων σε διάφορους ανθρώπους με σκοπό να παρατηρήσουν την συνειδητότητα τους σε διαφορετικές περιπτώσεις. Χρησιμοποιήθηκαν πειράματα που περιλάμβαναν ασκήσεις τύπου Stroop task (Stroop, 1935), και Flanker task (Eriksen & Schultz, 1979) και μέσω αυτών μετρήθηκε το ποσοστό της συνειδητότητας που χρειάζονταν έτσι ώστε να ανταποκριθούν σωστά στα τεστ. Όση μεγαλύτερη σύγκρουση και αντίθεση υπήρχε στις εργασίες που παρουσιάζονταν στους ανθρώπους, τόση μεγαλύτερη συνειδητότητα χρειαζόταν έτσι ώστε να εκτελεστεί η σωστή ενέργεια ή να απαντηθεί ορθά και γρήγορα η ερώτηση χωρίς λάθος. Για παράδειγμα σε μια εργασία όπου οι ερωτώμενοι έπρεπε να απαντήσουν τι χρώμα είναι η λέξη που τους παρουσίαζαν, περισσότερα λάθη αλλά και μεγαλύτερος χρόνος ανταπόκρισης παρατηρήθηκε στις περιπτώσεις όπου η λέξη και το χρώμα δεν ήταν τα ίδια (π.χ. η λέξη RED παρουσιαζόταν με μπλε χρώμα). Σε τέτοιες περιπτώσεις παρατηρήθηκε ότι χρειαζόταν μεγαλύτερη συνειδητότητα από τους ανθρώπους έτσι ώστε να απαντήσουν σωστά, δηλαδή όσο μεγαλύτερη ήταν η αντίθεση/conflict στην άσκηση, τόσο περισσότερη η συνειδητότητα. Σε αυτή την παρατήρηση βασιζόμαστε και εμείς για την μοντελοποίηση της συνειδητότητας στην εργασία μας.

Επίσης μέσα από τη διδακτορική διατριβή του Αριστόδημου Κλεάνθους (Cleanthous, 2010) πραγματοποιείται μια εξερεύνηση της δομής της εσωτερικής σύγκρουσης που δημιουργείται ανάμεσα στα μέρη του εγκεφάλου και πως αυτή επηρεάζει τον αυτοέλεγχο τους. Αυτό επιτυγχάνεται με την διαφοροποίηση του πίνακα με τις τιμές που λαμβάνουν οι πράκτορες ως αμοιβή/τιμωρία ανάλογα με το πόσο μεγάλη είναι η σύγκρουση (conflict). Εφαρμόζοντάς τους νέους πίνακες στο υπολογιστικό μοντέλο με τα δύο μέρη του εγκεφάλου, μας δίνεται η δυνατότητα να παρατηρήσουμε πως αυτές οι αλλαγές

επηρεάζουν τη συνεργασία ανάμεσα στους πράκτορες και κατ' επέκταση την ανάπτυξη του αυτοελέγχου. Συνδέοντας τις δύο αυτές παρατηρήσεις έχουμε τη συνειδητότητα στην εργασία αυτή να αναπαρίσταται μέσα από το μέγεθος της σύγκρουσης (conflict) μεταξύ άνω και κάτω μέρους του εγκεφάλου. Αυτό θα φαίνεται από τον πίνακα αμοιβών που θα παίρνει ο πράκτορας σε κάθε του κίνηση.

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο ο πίνακας αυτός αποτελείται από τις τιμές T (Temptation), P (Punishment), R (Reward), S (Sucker). Όσο πιο πολύ διαφέρουν οι τιμές S-T, S-P, R-T τόσο πιο μεγάλη θα είναι η εσωτερική σύγκρουση στον εγκέφαλο ενός ατόμου ο οποίος θα προσπαθεί να λειτουργήσει λογικά ή με το ένστικτο. Στην πρώτη περίπτωση της διαφοράς S-T, έχουμε την εσωτερική σύγκρουση αφού S είναι η αμοιβή που λαμβάνει ο πράκτορας ο οποίος δεν συνεργάζεται, ενώ T είναι η αμοιβή που λαμβάνει ο πράκτορας που συνεργάζεται. Άρα η διαφορά ανάμεσα σε αυτές τις τιμές αποτελεί σύγκρουση για τον κάθε πράκτορα, όσο αφορά το ποια κίνηση να διαλέξει. Το ίδιο ισχύει και για τις περιπτώσεις S-P και R-T, αφού με την ίδια λογική η μία τιμή είναι αμοιβή που δίνεται για συνεργασία του πράκτορα και η άλλη για μη συνεργασία του. Όσο πιο μεγάλη αυτή η εσωτερική σύγκρουση, τόσο μεγαλύτερος είναι και ο βαθμός συνειδητότητας του ατόμου. Μέσα από το πρόγραμμα θα έχουμε διάφορες τιμές στον πίνακα αμοιβών για να παρατηρήσουμε διαφορετικούς βαθμούς συνειδητότητας και πως επηρεάζεται ο αυτοέλεγχος.

Στο σχήμα πιο κάτω (Σχήμα 2.4) παρουσιάζεται ο τρόπος με τον οποίο ο Αριστόδημος Κλεάνθους τροποποιούσε την εσωτερική σύγκρουση στη διδακτορική διατριβή του (Cleanthous, 2010). Όπως εξηγήσαμε, αλλάζοντας τη διαφορά ανάμεσα στα ζεύγη τιμών S-T, S-P, R-T, άλλαζε και η εσωτερική σύγκρουση. Εδώ βλέπουμε τρεις διαφορετικές εσωτερικές συγκρούσεις όπως αυτές παρουσιάζονται στη διατριβή. Αυξάνοντας τη διαφορά μεταξύ των τιμών T-S, μεγαλώνει και η εσωτερική σύγκρουση και στη δική μας περίπτωση θα μεγαλώνει παράλληλα και η συνειδητότητα, αφού όπως έδειξαν οι έρευνες εσωτερική σύγκρουση και συνειδητότητα συνδέονται.

(i)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 1.5
	Defect	1.5, -1.3	-1.2, -1.2

Μικρή εσωτερική σύγκρουση:
T-S = 1.5-(-1.3) = 2.8
P-S = -1.2-(-1.3) = 0.1
T-R = 1.5-1.4 = 0.1

(ii)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 7
	Defect	1.5, -5	-1.2, -1.2

Μέτρια εσωτερική σύγκρουση:
T-S = 1.5-(-5) = 6.5
P-S = -1.2-(-5) = 3.8
T-R = 1.5-1.4 = 0.1

(iii)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 14
	Defect	1.5, -12	-1.2, -1.2

Μεγάλη εσωτερική σύγκρουση:
T-S = 1.5-(-12) = 13.5
P-S = -1.2-(-12) = 10.8
T-R = 1.5-1.4 = 0.1

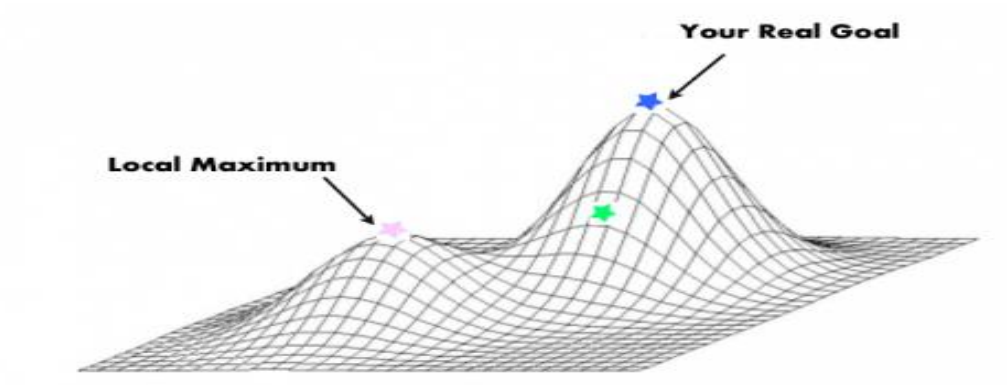
Σχήμα 2.4: Τρεις διαφορετικές εσωτερικές συγκρούσεις όπως παρουσιάζονται και τροποποιούνται στη διδακτορική διατριβή του Αριστόδημου Κ. (Cleanthous, 2010). Αλλάζοντας τη διαφορά ανάμεσα στα ζεύγη τιμών S-T, S-P, R-T, αλλάζει και η εσωτερική σύγκρουση.

2.7 Hill Climbing

Ο αλγόριθμος αναρρίχησης λόφων ή αλλιώς Hill Climbing θα χρησιμοποιηθεί με σκοπό την βελτιστοποίηση των τιμών στον πίνακα αμοιβών έτσι ώστε οι πράκτορες μας να συνεργάζονται και να λαμβάνουν τις μέγιστες αμοιβές αλλά και για να παρατηρηθούν διαφορές στη συνειδητότητα. Ας δούμε όμως και πως λειτουργεί αυτός ο αλγόριθμος. Αρχικά ο αλγόριθμος αναρρίχησης λόφων (Russell & Norvig, 2003) ανήκει στην κατηγορία αλγορίθμων τοπικής εξερεύνησης. Είναι ένας επαναληπτικός αλγόριθμος ο

οποίος παίρνει ως αρχική παράμετρο μια αυθαίρετη λύση στο πρόβλημα που θέλουμε να λύσουμε και μέσα από τις επαναλήψεις επιδιώκει να βρει μια καλύτερη λύση. Σε κάθε επανάληψη αυξάνεται ή μειώνεται ανάλογα μια παράμετρος/στοιχείο της λύσης και εάν αυτή η αλλαγή μας δώσει καλύτερη λύση, κρατάμε αυτή την λύση και συνεχίζουμε της αλλαγές πάνω σε αυτήν. Η ίδια αυτή διαδικασία επαναλαμβάνεται έως ότου ο αλγόριθμος Hill Climbing δεν βρίσκει νέα καλύτερη λύση. Η λογική του αλγορίθμου είναι απλή και χωρίς μεγάλη πολυπλοκότητα, γι' αυτό και αποτελεί ένα από τους δημοφιλέστερους αλγορίθμους ανάμεσα στους αλγορίθμους βελτιστοποίησης. Επίσης ευρεία χρήση βρίσκει και στο πεδίο της τεχνητής νοημοσύνης, σε προβλήματα στα οποία πρέπει να φτάσουμε από μια αρχική κατάσταση σε μια τελική κατάσταση/στόχο.

Εάν θέλουμε τώρα να εξηγήσουμε τον αλγόριθμο αναρρίχησης λόφων πιο μαθηματικά τότε μπορούμε να τον χαρακτηρίσουμε ως αλγόριθμο ο οποίος προσπαθεί να μεγιστοποιήσει μια συνάρτηση $f(x)$. Η παράμετρος x της συνάρτησης είναι διάνυσμα από τιμές, οι οποίες θα αλλάζουν σε κάθε επανάληψη με σκοπό μια καλύτερη τιμή από τη συνάρτηση $f(x)$. Κάθε αλλαγή στο x η οποία μας βελτιστοποιεί την συνάρτηση, αποθηκεύεται και οι επαναλήψεις συνεχίζουν με αυτό το x, μέχρι να φτάσουν στη μέγιστη τιμή. Στη συνάρτηση υπάρχουν και τοπικά ελάχιστα (χαμηλότερες κορυφές) αλλά σκοπός είναι να φτάσουμε (αναρριχηθούμε) στο ολικό μέγιστο (πιο ψηλή κορυφή). Η γραφική της συνάρτησης μοιάζει συνήθως όπως αυτή του σχήματος 2.4.2, με διάφορες κορυφές, γι' αυτό και ο αλγόριθμος ονομάζεται αναρρίχηση λόφων.



Σχήμα 2.4.2: *Τρισδιάστατη γραφική αναπαράσταση μιας συνάρτησης, όπως την βλέπει ο αλγόριθμος hill climbing.*

Στη δική μας περίπτωση η συνάρτηση $f(\mathrm{x})$ θα δέχεται ως διάνυσμα x, τον πίνακα με τις αμοιβές και τρέχοντας τον αλγόριθμο hill climbing θα τροποποιούμε τις τιμές του πίνακα προς το βέλτιστο. Τελικός σκοπός της βελτιστοποίησης είναι οι πράκτορες να καταλήγουνε σε συνεργασία-αυτοέλεγχο και η σύγκρουση (διαφορά) των τιμών να έχει μειωθεί ανάλογα. Ουσιαστικά αυτό που θέλουμε να δείξουμε είναι πως ξεκινώντας με μεγάλο ποσοστό σύγκρουσης στις τιμές (μεγάλη συνειδητότητα) οι πράκτορες αδυνατούν να συνεργαστούν ενώ με τις τροποποιήσεις η σύγκρουση μειώνεται και οι πράκτορες συνεργάζονται και αναπτύσσουν αυτοέλεγχο.

Κεφάλαιο 3

Περιγραφή Συστήματος

3.1 Εισαγωγή

Το σύστημα για τη διπλωματική αυτή γράφτηκε στη γλώσσα προγραμματισμού Java και σχεδιάστηκε με βάση τον αντικειμενοστραφή προγραμματισμό. Ο κώδικας του συστήματος δεν είναι πολύ μεγάλος, παρόλα αυτά είναι αποτελεσματικός και με μικρή πολυπλοκότητα. Πιο κάτω περιγράφονται οι δύο κλάσεις (**Player** και **MainProgram**) του συστήματος, η λειτουργία τους και τι περιέχει η κάθε μια. Τέλος θα δούμε πως συνδέονται και πως το πρόγραμμα δίνει ως έξοδο τα αποτελέσματα που θέλουμε.

3.2 Πράκτορες

Το άνω και κάτω μέρος του εγκεφάλου τα οποία συγκρούονται όπως αναφέραμε και νωρίτερα, θα αναπαρασταθούν από δύο πράκτορες, ένα για το κάθε μέρος. Έτσι δημιουργήσαμε το αντικείμενο **Player**, ένας πράκτορας ο οποίος θα συμμετέχει στο Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου. Ο κάθε πράκτορας έχει ως χαρακτηριστικά (μεταβλητές) το ρυθμό μάθησης (*learning_rate*), το έψιλον (*epsilon*), το ρυθμό έκπτωσης (*discount*), τον δισδιάστατο πίνακα τιμών Q (*Q_table),* τον πίνακα αμοιβών του (*payoff_matrix*), την τωρινή (*current_action*) και την επόμενη κίνηση (*current_action*) του. Στην αρχή του προγράμματος όταν δημιουργούνται οι πράκτορες, δημιουργούνται δύο αντικείμενα τύπου **Player** και στον constructor τους δημιουργούνται οι δύο πίνακες *Q_table* και *payoff_matrix*. Ο πρώτος είναι δισδιάστατος, μεγέθους 4x2

αφού έχουμε 4 καταστάσεις (CC, CD, DC, DD) και 2 κινήσεις (C, D) τις οποίες μπορεί να εκτελέσει ο πράκτορες. Η κλάση **Player** περιέχει επίσης τις απαραίτητες συναρτήσεις για ορισμό τιμών στις μεταβλητές και στους πίνακες, αλλά και συναρτήσεις οι οποίες επιστρέφουν τις τιμές αυτές. Ακόμη, περιέχει τη συνάρτηση επιλογής επόμενης κίνησης (*choose_action*) η οποία εφαρμόζει την πολιτική ε-greedy που είπαμε στο προηγούμενο κεφάλαιο. Τελευταία και πιο σημαντική συνάρτηση που υπάρχει στην κλάση Player είναι η *update_Q*, η οποία ενημερώνει τον πίνακα τιμών Q και ουσιαστικά τον πράκτορα κατά τη διάρκεια της ενισχυτικής μάθησης. Σε αυτή τη συνάρτηση έχουμε την εξίσωση για τον αλγόριθμο που χρησιμοποιούμε, είτε αυτός είναι Q-Learning, είτε είναι SARSA.

3.3 Βασικό Μοντέλο

Η δεύτερη κλάση του συστήματός μας όπου και κτίζεται το βασικό μοντέλο με τους πράκτορες και παίζεται το ΕΔΦ ονομάζεται **MainProgram**. Στο αρχείο αυτό το παιγνίδι παίζεται 100 φορές (*runs*). Κάθε *run* αποτελείται από 500 *executions* και κάθε *execution* αποτελείται από 50 *trials* των 1000 *episodes*. Στην αρχή κάθε *run* έχουμε αρχικοποίηση του πίνακα αμοιβών *SPRT*, ενός μονοδιάστατου πίνακα 4 θέσεων με τις τιμές S, P, R, T. Σε κάθε *trial* αρχικοποιούμε τους πράκτορες μέσω της συνάρτησης *setPlayers* η οποία παίρνει ως παράμετρο τον πίνακα αμοιβών *SPRT_old* εάν είναι το πρώτο *execution* του *run*, αλλιώς καλείται με παράμετρο τον τροποποιημένο πίνακα αμοιβών *SPRT_new*. Οι πράκτορες παίρνουν πάντα το ίδιο παράγοντα έκπτωσης έτσι ώστε να διαφοροποιούνται και να αναπαριστούν τα δύο μέρη του εγκεφάλου. Το κάτω μέρος παίρνει παράγοντα έκπτωσης 0.1 και το πάνω μέρος 0.9.

Αφού θέσουμε στους πράκτορες τον πίνακα αμοιβών εκτελούμε τον αλγόριθμο ενισχυτικής μάθησης για 1000 (*episodes*) φορές σε κάθε *trial* έτσι ώστε να εκπαιδευτούν οι πράκτορες μας. Κατά τη διάρκεια της μάθησης αποθηκεύονται στον πίνακα *states_results* τα ποσοστά για τις τέσσερις καταστάσεις του παιγνιδιού.

Μετά το τέλος των επεισοδίων κάθε *trial* υπολογίζουμε τη συνειδητότητα, βρίσκοντας τις διαφορές των τιμών στον πίνακα αμοιβών. Υπολογίζεται η διαφορά ανάμεσα στις τιμές S-T, S-P και T-R και αποθηκεύονται στον πίνακα *STdifference*, για να εκτυπωθούν στο τέλος του προγράμματος σε αρχείο. Στο πρώτο *execution* κάθε *run* απλά

αποθηκεύουμε το ποσοστό συμβιβασμού του τελευταίου επεισοδίου, μετά την εκπαίδευσή των πρακτόρων, έτσι ώστε να μπορούμε να το συγκρίνουμε με τα ποσοστά των επόμενων εκτελέσεων (*execution*). Εάν δεν είμαστε στο πρώτο *execution* συγκρίνουμε το ποσοστό συμβιβασμού αυτού του *execution* με του προηγούμενου. Εάν ο συμβιβασμός έχει αυξηθεί, τότε κρατάμε τις νέες τιμές του πίνακα *SPRT_new* στον πίνακα *SPRT_old* και το νέο ποσοστό συμβιβασμού *CC_rate_new* στην μεταβλητή *CC_rate_old*. Εκτός αυτών αποθηκεύουμε και τα ποσοστά των τεσσάρων καταστάσεων και για τα 1000 επεισόδια από τον πίνακα *states_results* στον πίνακα *best_rates*. Ο πίνακας *best_rates* θα κρατά τα ποσοστά για τα οποία έχουμε καλύτερο συμβιβασμό στο τελευταίο επεισόδιο. Αν το νέο ποσοστό συμβιβασμού *CC_rate_new* όμως δεν έχει αυξηθεί σε σχέση με το προηγούμενο, συνεχίζουμε το πρόγραμμα με τις προηγούμενες τιμές του πίνακα αμοιβών *SPRT_old*.

Επίσης σε κάθε εκτέλεση καλείται η συνάρτηση *HCNewPayoffs* η οποία εκτελεί τον αλγόριθμο Hill Climbing. Με το hill climbing όπως εξηγήσαμε στο Κεφάλαιο 2 τροποποιούμε τις τιμές του πίνακα αμοιβών. Οι αμοιβές περιορίζονται από το πρόγραμμά μας στο εύρος τιμών -50 με 50. Μέσα στη συνάρτηση αυτή καλείται μια άλλη συνάρτηση, η *Gaussian*. Αυτή η συνάρτηση παράγει αριθμούς τους οποίους προσθέτουμε στις τιμές του πίνακα αμοιβών. Οι «γκαουσιανοί» αριθμοί μπορεί να είναι θετικοί ή αρνητικοί. Σε κάθε περίπτωση φροντίζουμε έτσι ώστε το άθροισμα να κρατά την τιμή της αμοιβής μέσα στο εύρος τιμών αλλά και να επαληθεύει τις δύο ανισότητες/κανόνες του Επαναλαμβανόμενου Διλήμματος του Φυλακισμένου.

Ανάλογα με το ποια μέθοδο χρησιμοποιούμε για την τροποποίηση του πίνακα (όπως αναφέραμε σε προηγούμενο κεφάλαιο δοκιμάσαμε 3 μεθόδους) έχουμε και την αντίστοιχη εκτέλεση της συνάρτησης *HCNewPayoffs*. Εάν αλλάζουμε και τις 4 τιμές θα καλείται 4 φορές, εάν θα αλλάζουμε μόνο τις τιμές S και T θα καλείται 2 φορές. Η συνάρτηση παίρνει ως παράμετρο τον παλιό πίνακα αμοιβών *SPRT_old* και επιστρέφει μια νέα αμοιβή που αποθηκεύεται στην κατάλληλη θέση του νέου πίνακα *SPRT_new*.

Ακόμη σε κάθε run προσθέτουμε τις τιμές του πίνακα *best_rates* με αυτές του πίνακα *avgstates_afterHill* και τις αποθηκεύουμε στην αντίστοιχη θέση του δεύτερου πίνακα. Στο τέλος του προγράμματος, βρίσκουμε το μέσο όρο για όλα τα runs διαιρώντας κάθε

θέση του πίνακα *avgstates_afterHill* δια το πλήθος των runs και εκτυπώνουμε τα ποσοστά αυτά στο αρχείο avgstates_afterHill.txt. Σε αρχείο (conflicts.txt) εκτυπώνονται επίσης και οι διαφορές των τιμών του πίνακα αμοιβών, έτσι ώστε να μπορούμε να μελετήσουμε την συνειδητότητα. Στο αρχείο conflicts.txt εκτυπώνονται 6 στήλες. Στην πρώτη στήλη παρουσιάζονται οι μεταβολές της διαφοράς των τιμών S-T για όλα τα runs, στην τρίτη στήλη οι μεταβολές της διαφοράς των τιμών S-P και στην πέμπτη στήλη οι μεταβολές της διαφοράς των τιμών R-T. Ενδιάμεσα αυτών των στηλών, δηλαδή στη δεύτερη, τέταρτη και έκτη στήλη, εκτυπώνονται τα ποσοστά του συμβιβασμού για διευκόλυνση της δημιουργίας γραφικών παραστάσεων ανάμεσα σε τιμές αμοιβών (συνειδητότητα) και ποσοστών συμβιβασμού (αυτοέλεγχος). Αφού σκοπός μας είναι να δούμε τη σχέση συνειδητότητας και αυτοελέγχου, πρέπει να φτιάξουμε γραφικές παραστάσεις που να δείχνουν πως επηρεάζει η μια έννοια την άλλη. Στο επόμενο κεφάλαιο θα παρουσιάσουμε και θα συζητήσουμε για γραφικές παραστάσεις σαν και αυτές αλλά και γραφικές παραστάσεις για τις τέσσερις καταστάσεις του παιγνιδιού.

Κεφάλαιο 4

Αποτελέσματα και Συζήτηση

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δούμε και θα συζητήσουμε τα αποτελέσματα της εργασίας. Αρχικά θα δούμε αποτελέσματα αυτοελέγχου παίζοντας το επαναλαμβανόμενο δίλημμα του φυλακισμένου στο βασικό μοντέλο και δίνοντας στους πράκτορες διάφορους ρυθμούς μάθησης αλλά και έψιλον για την πολιτική ε-greedy. Στη συνέχεια περνώντας στο κομμάτι της συνειδητότητας θα δούμε τα αποτελέσματα από τις τρεις προσπάθειες που έγιναν για μοντελοποίησή της μέσω τροποποίησης του πίνακα αμοιβών του πράκτορα. Πρώτα θα παρουσιάσουμε τα αποτελέσματα που πήραμε τρέχοντας το βασικό μοντέλο και το παιχνίδι να παίζεται κάθε φορά με διαφορετικούς τυχαίους πίνακες αμοιβών στην αρχή κάθε γύρου. Δεύτερον θα παρουσιάσουμε τα αποτελέσματα εκτελώντας το πρόγραμμα και τροποποιώντας στον πίνακα αμοιβών μόνο τις τιμές S και T που αντιπροσωπεύουν την εσωτερική σύγκρουση στον πράκτορα και κρατώντας σταθερές τις τιμές P και R. Τέλος θα δούμε τα αποτελέσματα που είχαμε ως έξοδο από

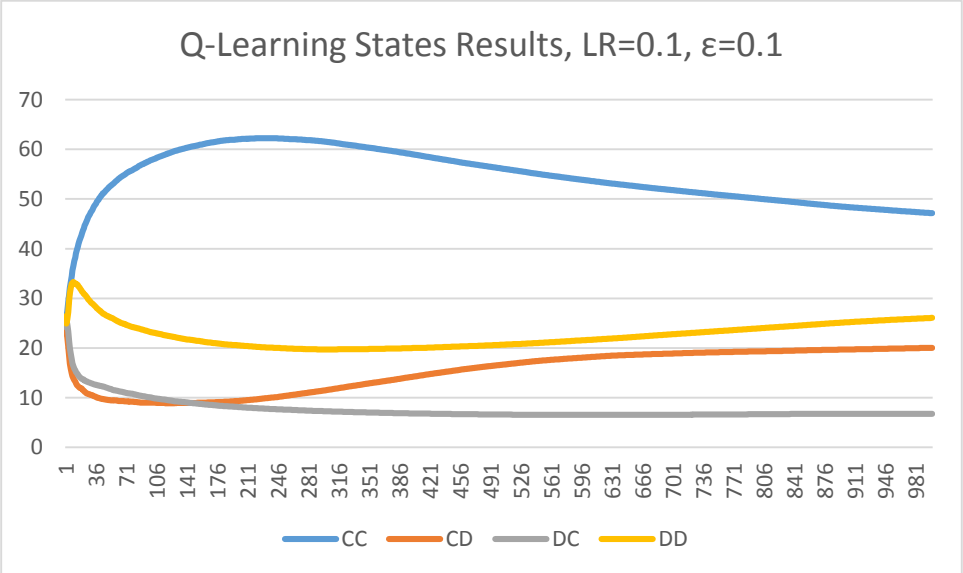
το πρόγραμμα εκτελώντας το με ένα πίνακα αμοιβών στον οποίο τροποποιούσαμε όλες του τις τιμές.

4.2 Αποτελέσματα Αυτοέλεγχου

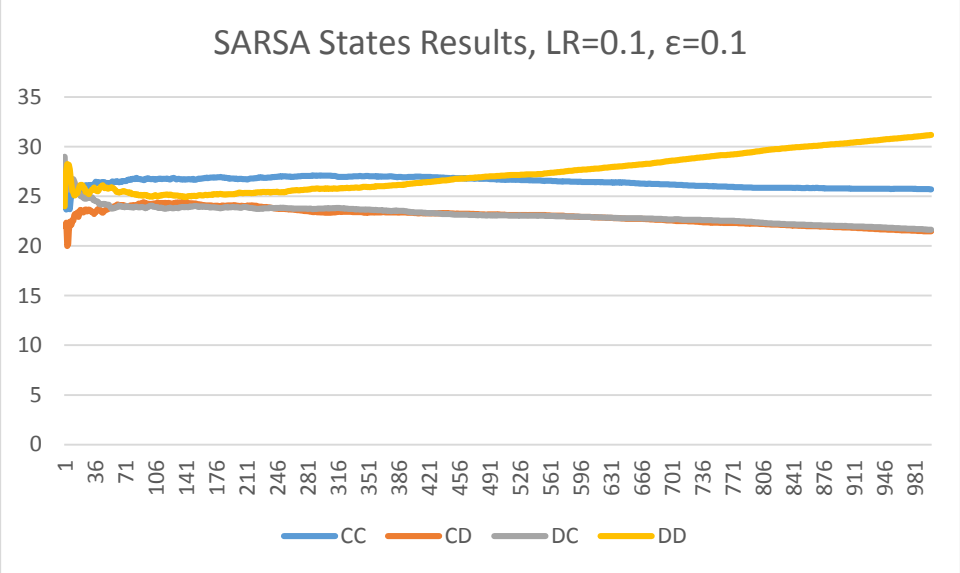
Στα πρώτα στάδια της διπλωματικής εργασίας, αρχικός στόχος ήταν να δημιουργήσουμε το βασικό μοντέλο με τους δύο πράκτορες οι οποίοι θα παίζουν το επαναλαμβανόμενο ΔΦ και να δούμε τα ποσοστά συνεργασίας-αυτοελέγχου. Χρησιμοποιήθηκαν οι αλγόριθμοι Q-Learning και SARSA με διάφορες τιμές για το ρυθμό μάθησης. Οι γραφικές παρουσιάζουν το μέσο όρο σε ποσοστό (τοις εκατό %) που οι πράκτορες επέλεγαν την κάθε κατάσταση κατά τη διάρκεια των 1000 επεισοδίων. Το πρόγραμμα εκτελείται 50 φορές από 1000 επεισόδια κάθε φορά.

Στις πιο κάτω γραφικές παραστάσεις βλέπουμε τα αρχικά αποτελέσματα για τις 4 καταστάσεις του προβλήματος (CC, CD, DC, DD) χρησιμοποιώντας και στους δύο αλγορίθμους ρυθμό μάθησης (LR = Learning Rate) ίσο με 0.1 και έψιλον στην πολιτική πάλι ίσο με 0.1. Ο πίνακας αμοιβών ο οποίος δεν μεταβάλλεται σε αυτό το σημείο έχει τις τιμές **4, -3, 5, -2** για τις μεταβλητές (R, S, T, P).

4.2.1 Βασικό Μοντέλο – Δίλημμα Φυλακισμένου



Σχήμα 4.1: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης:Q-Learning. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1. Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*



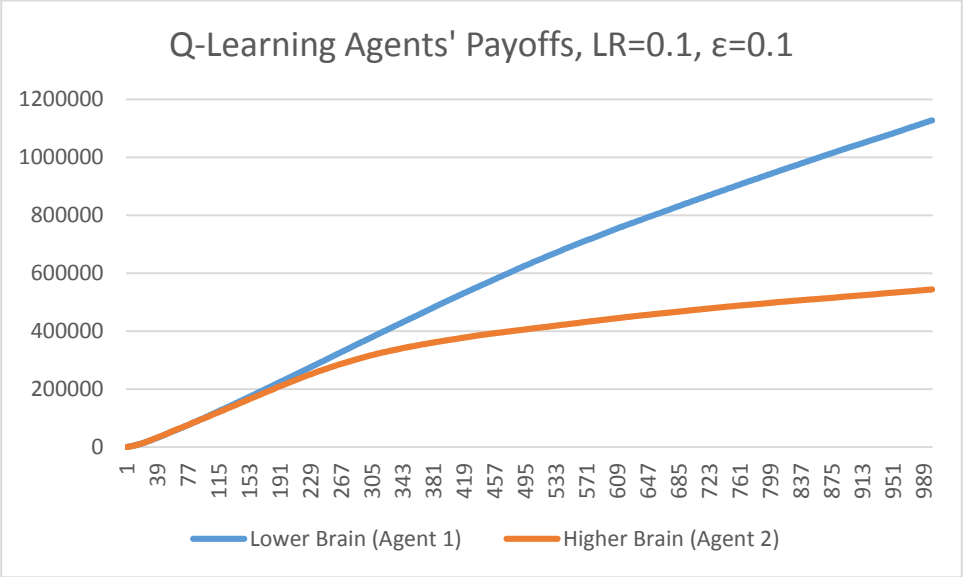
Σχήμα 4.2: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*

Οι δύο γραφικές διαφέρουν κατά πολύ στα αποτελέσματα που παρουσιάζουν, αφού στην πρώτη περίπτωση (Σχήμα 4.1) οι πράκτορες καταφέρνουν να μάθουν μέσω του αλγόριθμου ενισχυτικής μάθησης Q-Learning και στο τέλος της εκτέλεσης καταλήγουν να συνεργάζονται με το ποσοστό του συμβιβασμού CC να είναι το υψηλότερο (47%). Άρα αφού τα δύο μέρη του εγκεφάλου (πράκτορες) συμβιβάζονται τις περισσότερες φορές, τότε έχουμε αυτοέλεγχο στον άνθρωπο. Παρόλο όμως που το ποσοστό του συμβιβασμού υπερισχύει σε όλα τα επεισόδια έναντι των άλλων καταστάσεων, παρουσιάζει μείωση ενώ είχε φτάσει και σε ποσοστό 62%.

Στο Σχήμα 4.2 βλέπουμε τα αποτελέσματα του αλγορίθμου SARSA τα οποία όπως φαίνεται δεν είναι τόσο καλά όσο του άλλου αλγορίθμου και δεν πλησιάζουν στο στόχο μας που είναι οι πράκτορες να συμβιβαστούν (CC). Βλέπουμε ότι ναι μεν το ποσοστό συμβιβασμού είναι αρχικά το πιο ψηλό αλλά εν τέλει υπερισχύει η κατάσταση DD για τα υπόλοιπα επεισόδια. Στην κατάσταση DD κανείς από τους δύο πράκτορες δεν

συνεργάζεται και η ανάπτυξη αυτοελέγχου που θέλουμε δεν επιτυγχάνεται. Όμως, οι διαφορές στα ποσοστά των 4 καταστάσεων δεν είναι πολύ μεγάλες, κυμαίνονται από 20%-32%, άρα σίγουρα υπάρχουν περιθώρια βελτίωσης.

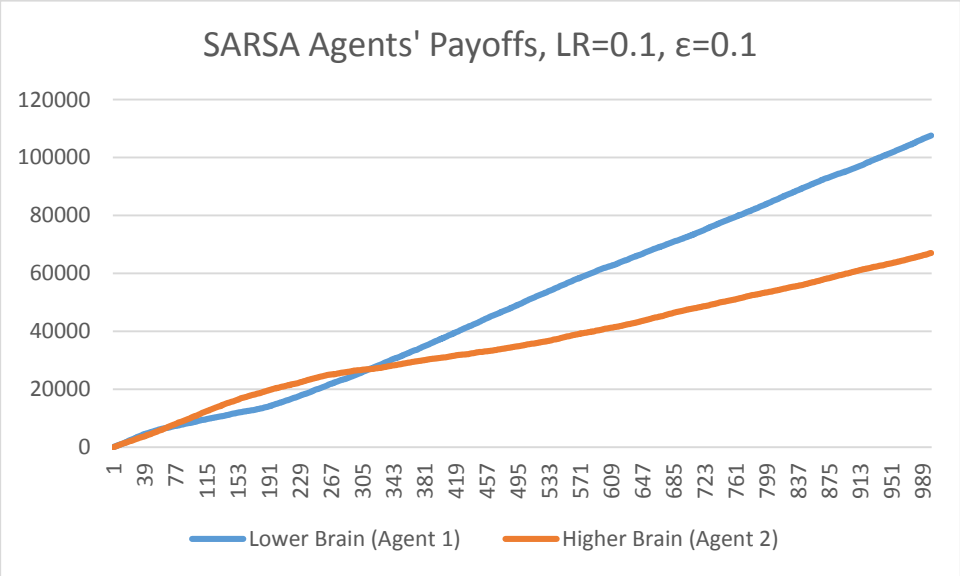
Για να επιβεβαιώσουμε ότι οι πράκτορες μαθαίνουν μπορούμε να παρατηρήσουμε και τις αμοιβές που λαμβάνουν από την αρχή του παιχνιδιού μέχρι τη λήξη του. Εάν οι αμοιβές μεγιστοποιούνται τότε οι πράκτορες μαθαίνουν ότι για να πάρουν τη βέλτιστη αμοιβή πρέπει να συνεργαστούν. Αρχικά ας δούμε τη γραφική παράσταση (Σχήμα 4.3) με τις αμοιβές που λαμβάνουν οι πράκτορες για το προηγούμενο παράδειγμα εκτέλεσης (S=-3, P=-2, R=4, T=5, LR=0.1, ε=0.1).



Σχήμα 4.3: *Γραφική παράσταση συνολικών αμοιβών των πρακτόρων κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: Q-Learning. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1. Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*

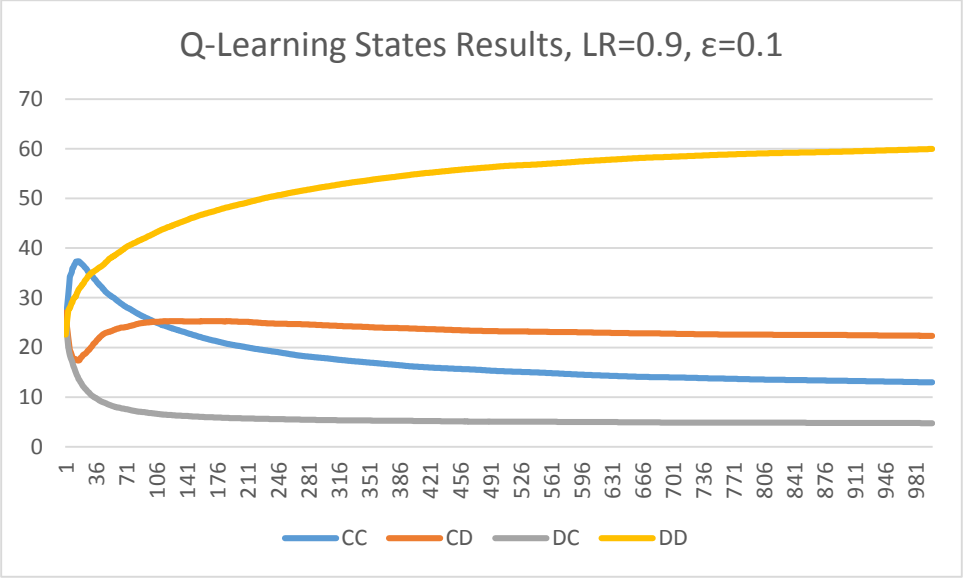
Με την πάροδο των επεισοδίων οι συνολικές αμοιβές αυξάνονται και για τους δύο πράκτορες. Οι αμοιβές για τον πράκτορα 1 φτάνουν μέχρι και πιο πάνω από 1000000 ενώ του πράκτορα 2 λίγο πιο κάτω από 600000. Αυτό επιβεβαιώνει την προηγούμενη γραφική παράσταση για το Q-Learning η οποία μας έδειχνε ότι η κατάσταση του συμβιβασμού έχει το μεγαλύτερο ποσοστό, άρα οι πράκτορες μας αναπτύσσουν αυτοέλεγχο μέσω της συνεργασίας.

Από την άλλη μεριά όταν τρέξαμε το πρόγραμμα αλλά με αλγόριθμο SARSA είδαμε ότι αρχικά οι πράκτορες συνεργάζονται ενώ στη συνέχεια όχι. Ως εκ τούτου αναμένουμε και στη γραφική των αμοιβών μειωμένες τιμές. Στο σχήμα 4.4 πιο κάτω φαίνεται ότι σε σύγκριση με το Σχήμα 4.3 οι αμοιβές και των δύο πρακτόρων είναι αισθητά μειωμένες, ακριβώς λόγω του γεγονότος ότι δεν συνεργάζονταν καθ' όλη τη διάρκεια του παιγνιδιού, όπως είδαμε στο Σχήμα 4.2.

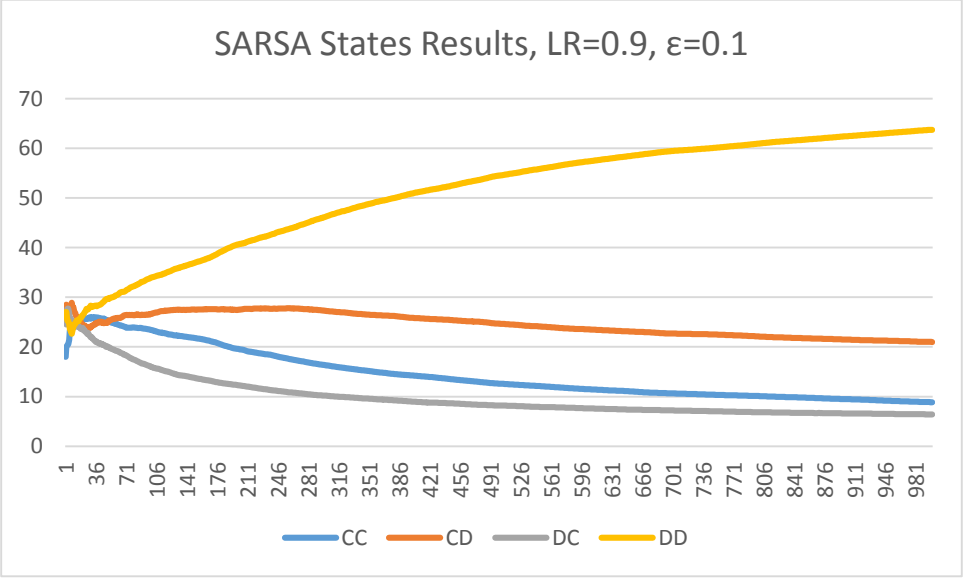


Σχήμα 4.4: *Γραφική παράσταση συνολικών αμοιβών των πρακτόρων κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1. Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*

Αυξάνοντας τον ρυθμό μάθησης και στους δύο αλγορίθμους, τα αποτελέσματα είναι εντελώς αντίθετα με αυτό που θέλουμε να επιτύχουμε, αφού όπως φαίνεται στις γραφικές παραστάσεις πιο κάτω, Σχήμα 4.5 και Σχήμα 4.6, Q-Learning και SARSA αντίστοιχα υπερισχύει η κατάσταση DD και οι πράκτορες δεν συνεργάζονται.



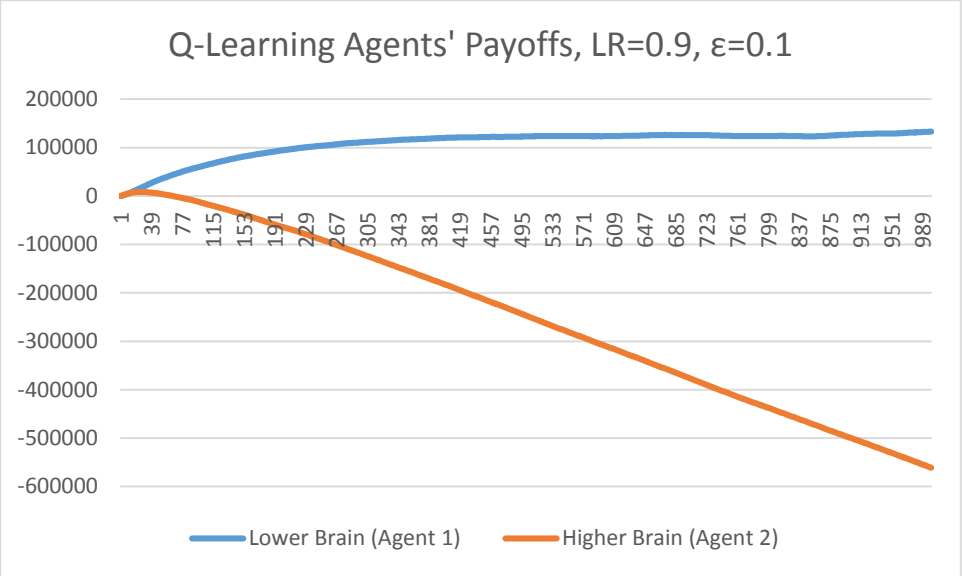
Σχήμα 4.5: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: Q-Learning. Ρυθμός μάθησης: 0.9. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*



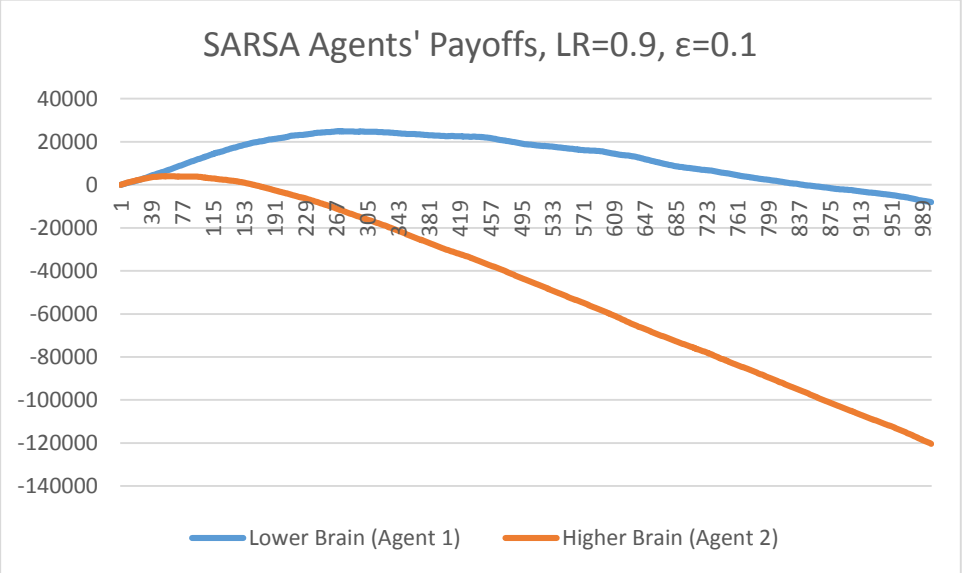
Σχήμα 4.6: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.9. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*

Τα πιο πάνω επιβεβαιώνονται και από τις γραφικές των αμοιβών που παίρνουν οι πράκτορες, Σχήμα 4.7 και Σχήμα 4.8 για Q-Learning και SARSA αντίστοιχα. Λόγω του μεγάλου ρυθμού μάθησης οι πράκτορες υπερκεπαιδεύονται και έχουμε ανεπιθύμητα

αποτελέσματα. Η κατάσταση DD έχει το πιο μεγάλο ποσοστό και κανένας από τους δύο πράκτορες δηλαδή δεν συνεργάζεται άρα δεν λαμβάνουν μεγάλες αμοιβές. Στον πίνακα αμοιβών υπάρχουν επίσης και αρνητικές αμοιβές-τιμωρίες και έτσι οι συνολικές αμοιβές των πρακτόρων είναι είτε πολύ μικρές (πράκτορας 1) είτε αρνητικές (πράκτορας 2).



Σχήμα 4.7: *Γραφική παράσταση συνολικών αμοιβών των πρακτόρων κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: Q-Learning. Ρυθμός μάθησης: 0.9. Έψιλον: 0.1. Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*



Σχήμα 4.8: *Γραφική παράσταση συνολικών αμοιβών των πρακτόρων κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.9. Έψιλον: 0.1. Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5*

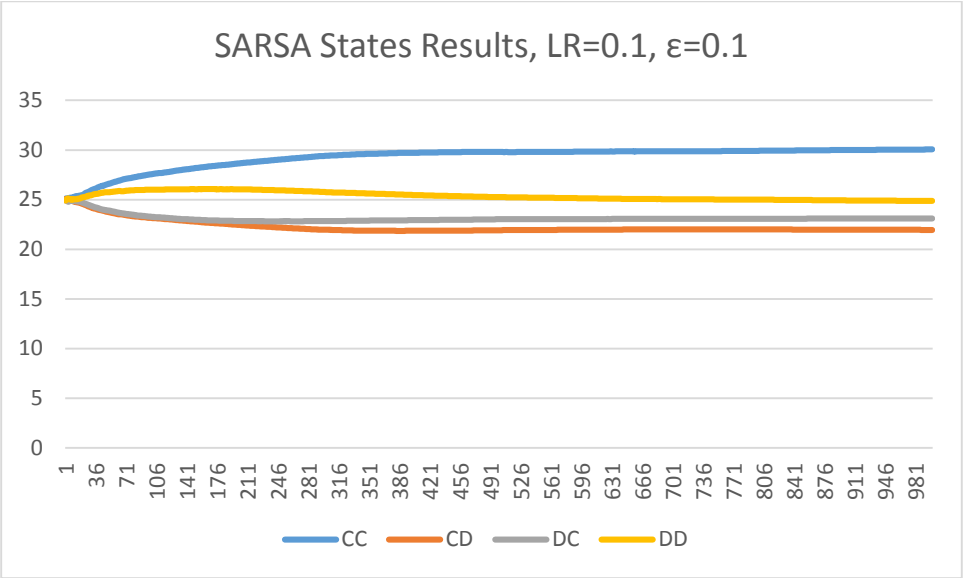
Διαφορές έχουν και μεταξύ τους οι αλγόριθμοι αφού στον αλγόριθμο SARSA (Σχήμα 4.8) έχουμε μεγάλες αρνητικές τιμές στις αμοιβές του πράκτορα 2, συμπεραίνοντας έτσι ότι το μεγαλύτερο μέρος της διάρκειας του παιχνιδιού ο πράκτορας αυτός δεν συνεργαζόταν.

Για να καταφέρουμε να εκπαιδεύσουμε τους πράκτορες μας και να αναπτύσσουν αυτοέλεγχο τα δύο μέρη του εγκεφάλου θα χρησιμοποιήσουμε τους τρεις τρόπους τους που αναφέραμε στην εισαγωγή και τους οποίους θα δούμε στα επόμενα υποκεφάλαια. Η ιδέα για τους τρεις αυτούς τρόπους προήλθε μέσα από την προσπάθειά μας να μοντελοποιήσουμε τη συνειδητότητα και παράλληλα να μεγιστοποιήσουμε το ποσοστό αυτοελέγχου. Αφού καταλήξουμε σε αυτοέλεγχο θα εξετάσουμε και τη σύνδεσή του με τη συνειδητότητα. Εκτός αυτού, να σημειώσουμε ότι στα επόμενα υποκεφάλαια θα ασχοληθούμε μόνο με τα αποτελέσματα του αλγόριθμου SARSA σε συνδυασμό με το hill climbing αφού όπως είδαμε στο υποκεφάλαιο 2.4.2 είναι πιο κοντά στον τρόπο λειτουργίας του εγκεφάλου.

4.2.2 Εκτέλεση μοντέλου με τυχαίους πίνακες αμοιβών

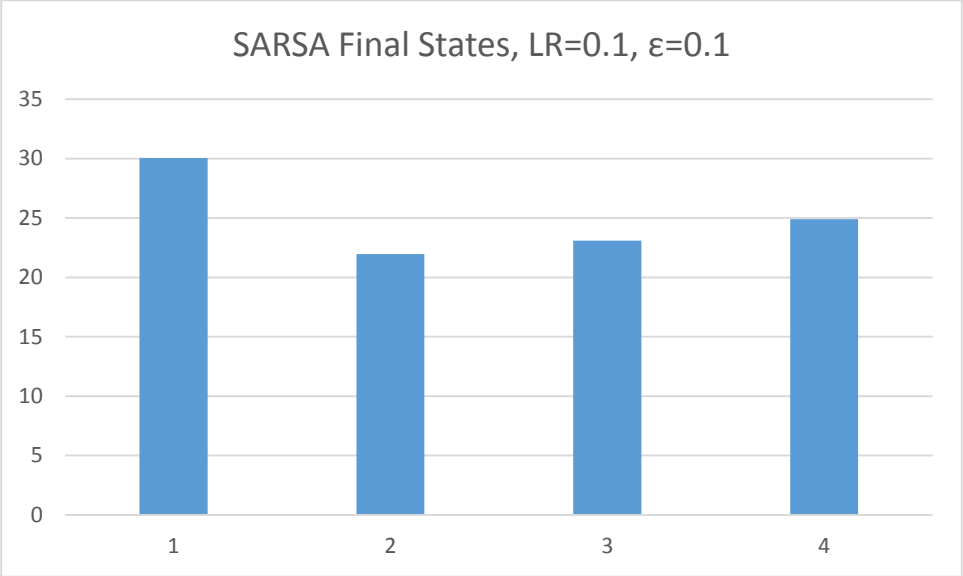
Ως πρώτη προσπάθεια θα προσθέσουμε εντολές στο πρόγραμμα έτσι ώστε να ξεκινά με ένα αρχικό πίνακα αμοιβών που του δίνουμε εμείς και θα εκτελείται η ίδια διαδικασία όπως πριν για 50 εκτελέσεις. Για κάθε μια εκτέλεση όμως το πρόγραμμα δε θα ξεκινά ξανά με τον ίδιο πίνακα που ορίσαμε στην αρχή (όπως στη δεύτερη και τρίτη μέθοδο) αλλά θα δημιουργείται νέος τυχαίος πίνακας αμοιβών για τους δύο πράκτορες και ο αλγόριθμος θα χρησιμοποιεί αυτό τον νέο πίνακα. Τιμές για τον αρχικό πίνακα θα έχουμε και πάλι τις **S=-3, P=-2, R=4, T=5**. Στον τερματισμό του προγράμματος θα παίρνουμε τον μέσο όρο και των 50 γύρων για τα ποσοστά των 4 καταστάσεων του παιχνιδιού.

Με ίδιο ρυθμό μάθησης και έψιλον όπως πριν (και τα δύο 0.1) έχουμε την εξής βελτίωση στην γραφική μας παράσταση:



Σχήμα 4.9α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα:* Τυχαίοι αρχικοί πίνακες αμοιβών.

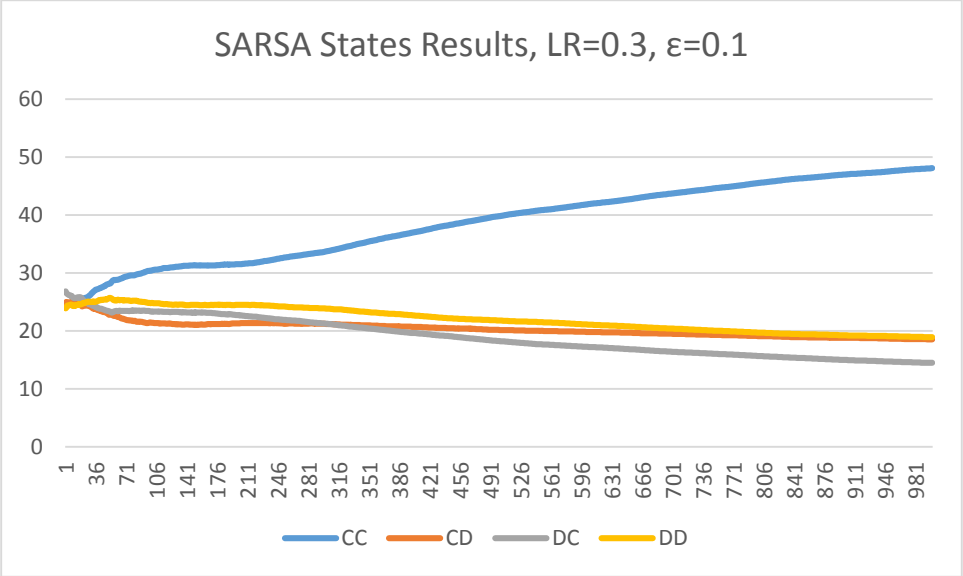
Σε αντίθεση με το Σχήμα 4.2 στο οποίο είχαμε πάλι τον αλγόριθμο SARSA με τις ίδιες παραμέτρους, στο Σχήμα 4.9α η κατάσταση της συνεργασίας CC υπερισχύει πιο έντονα από τις άλλες καταστάσεις. Παρόλο που τα ποσοστά των καταστάσεων είναι και πάλι κοντά το ένα στο άλλο και κυμαίνονται από 22% μέχρι και 30+%, όπως φαίνεται και στο Σχήμα 4.9β, μπορούμε να πούμε με επιφύλαξη ότι οι πράκτορες μαθαίνουν και αναπτύσσουν σε κάποιο βαθμό αυτοέλεγχο.



Σχήμα 4.9β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα:* Τυχαίοι αρχικοί πίνακες αμοιβών.

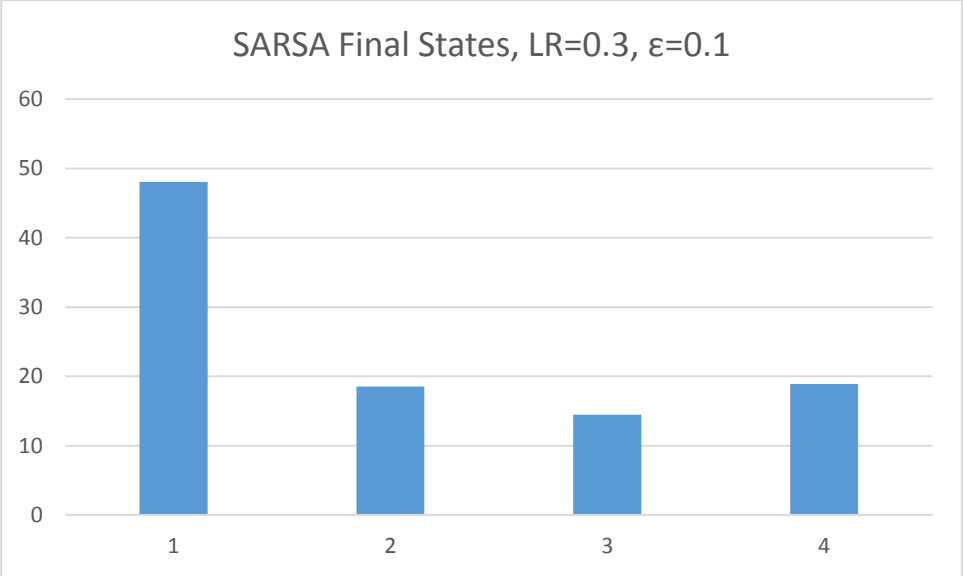
Στο πιο πάνω σχήμα φαίνονται τα ποσοστά του τελευταίου επεισοδίου για τις τέσσερις καταστάσεις του παιγνιδιού, κατά μέσο όρο. Όπου 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Όπως ήδη έχουμε αναφέρει, τα ποσοστά είναι κοντά μεταξύ τους, οπότε η κατάσταση συμβιβασμού (στήλη 1) δεν υπερισχύει ξεκάθαρα έναντι των υπολοίπων.

Εάν αυξήσουμε το ρυθμό μάθησης και στους δύο πράκτορες σε 0.3 παρατηρούμε ότι το «τοπίο» ξεκαθαρίζει περισσότερο και το ποσοστό της συνεργασίας αυξάνεται ακόμα περισσότερο. Άρα οι πράκτορες καταλήγουν πιο εύκολα σε αυτοέλεγχο και όπως βλέπουμε στο Σχήμα 4.10 και το ποσοστό συνεργασίας αγγίζει σχεδόν το 50% :



Σχήμα 4.10α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα:* Τυχαίοι αρχικοί πίνακες αμοιβών.

Η νέα διαφορά των καταστάσεων εμφανίζεται στο Σχήμα 4.10β πιο καθαρά, όπου παρατηρούμε ότι όντως το ποσοστό της κατάστασης του συμβιβασμού (στήλη 1), έχει όντως αυξηθεί, ενώ τα ποσοστά των υπόλοιπων καταστάσεων μειώθηκαν. Άρα επιβεβαιώνεται για άλλη μια φορά ότι οι πράκτορές μας αναπτύσσουν αυτοέλεγχο μέσω αυτής της τεχνικής.

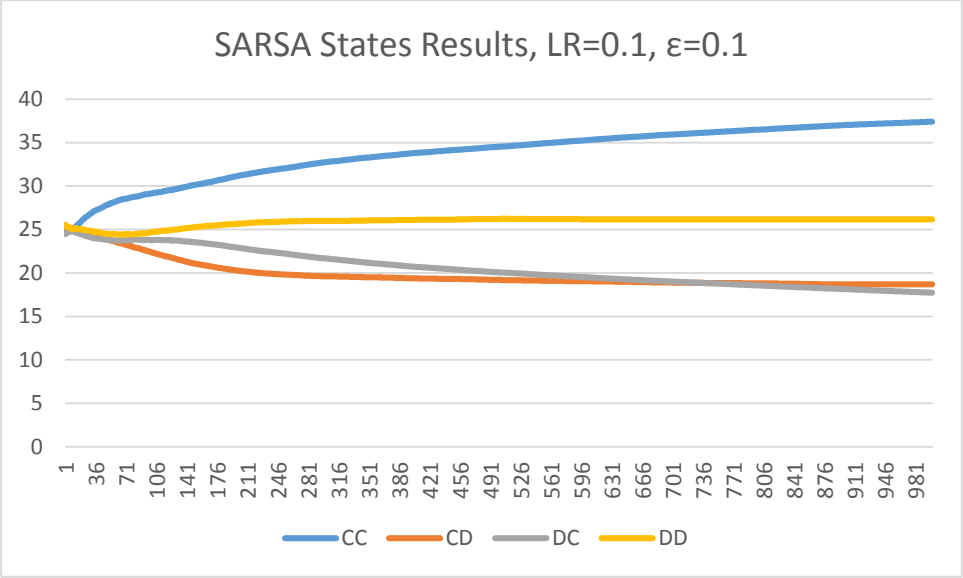


Σχήμα 4.10β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα:* Τυχαίοι αρχικοί πίνακες αμοιβών.

4.2.3 Μέθοδος με σταθερό P και R στον πίνακα αμοιβών

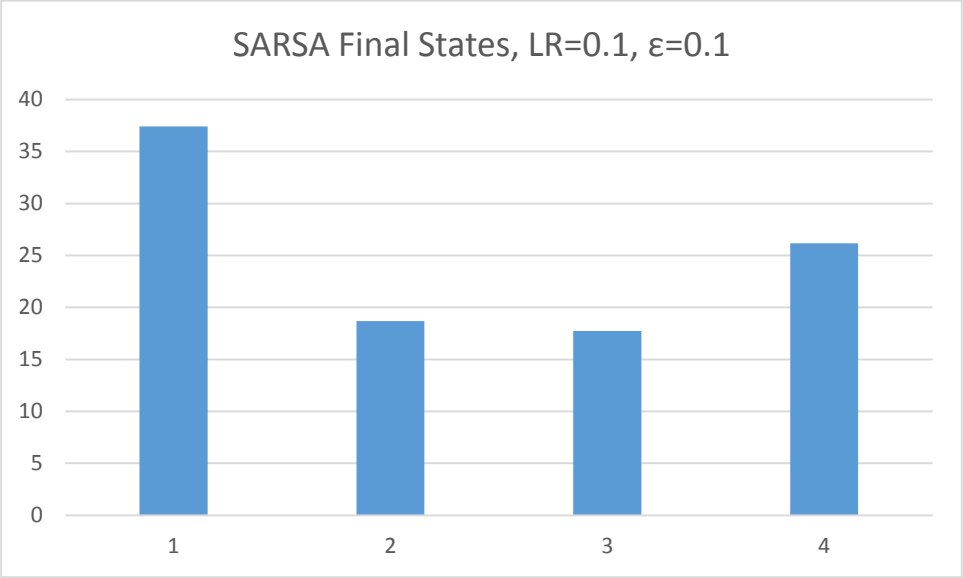
Ξέροντας ότι μέσα από τις τιμές του πίνακα αμοιβών μπορούμε να αναπαραστήσουμε τη συνειδητότητα, ένας άλλος τρόπος για να προσπαθήσουμε να αυξάνουμε το ποσοστό του αυτοελέγχου είναι να κρατούμε σταθερές τις τιμές P και R. Όπως είδαμε η διαφορά ανάμεσα στις υπόλοιπες δύο τιμές (S και T) του πίνακα αντιπροσωπεύει την εσωτερική σύγκρουση ανάμεσα στα δύο μέρη του εγκεφάλου. Πιο κάτω θα δούμε κάποια από τα αποτελέσματα που πήραμε, κρατώντας αυτές τις τιμές (P και R) σταθερές όλη τη διάρκεια του παιχνιδιού και μεταβάλλοντας μόνο τις άλλες δύο. Ο αρχικός πίνακας θα είναι ίδιος με τα προηγούμενα υποκεφάλαια και στην αρχή κάθε γύρου το πρόγραμμα θα ξεκινά ξανά με τον αρχικό πίνακα που ορίσαμε, ενώ στο τέλος θα έχουμε αποτελέσματα με τον μέσο όρο των γύρων.

Πρώτα τρέχουμε το νέο μας πρόγραμμα με έψιλον=0.1 και ρυθμό μάθησης=0.1 όπως προηγουμένως και παρατηρούμε ότι στη γραφική παράσταση (Σχήμα 4.11α) των καταστάσεων τα αποτελέσματα είναι καλύτερα από την αντίστοιχη περίπτωση του υποκεφαλαίου 4.2.2.



Σχήμα 4.11α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5, Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

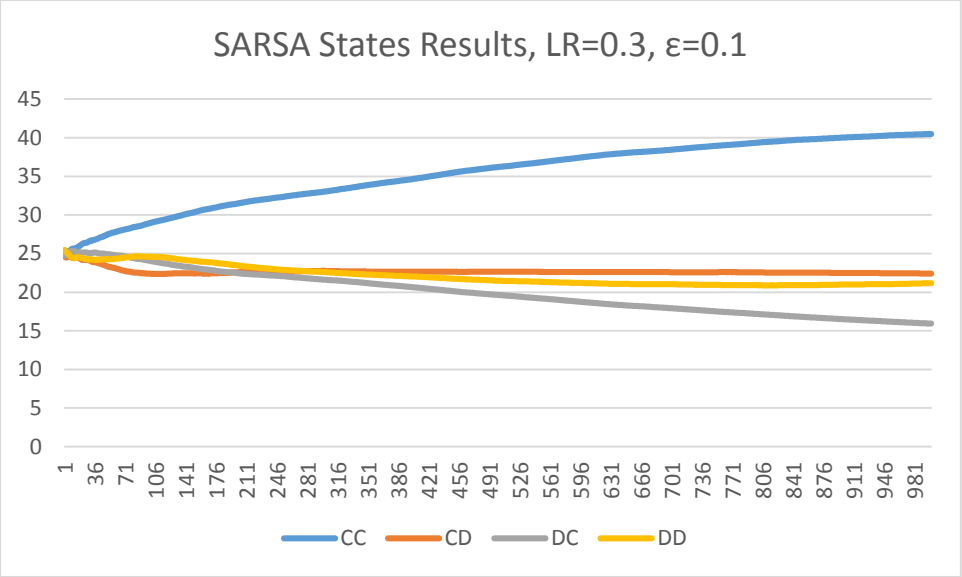
Οι πράκτορές μας συνεργάζονται και καταλήγουν σε αυτοέλεγχο, με το ποσοστό της κατάστασης συνεργασίας να είναι εμφανές το μεγαλύτερο (37%) όπως βλέπουμε και στο Σχήμα 4.11β στην επόμενη σελίδα.



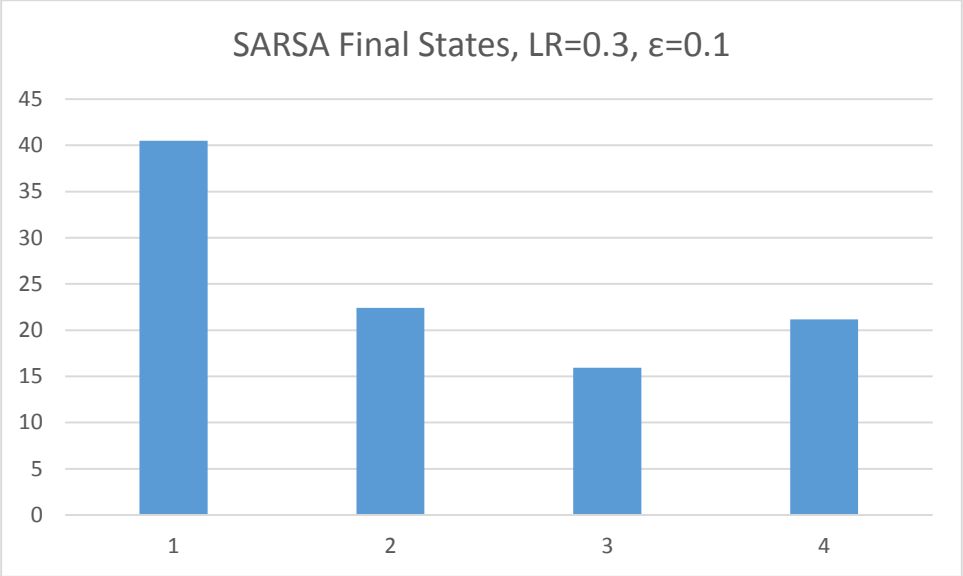
Σχήμα 4.11β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας*

αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.

Για να έχουμε μέτρο σύγκρισης με τα προηγούμενα εκτελούμε το πρόγραμμα και με ρυθμό μάθησης=0.3 και παραθέτουμε τα αποτελέσματα στο Σχήμα 4.12α πιο κάτω. Με μεγαλύτερο ρυθμό (0.3) το ποσοστό συνεργασίας αυξάνεται κι άλλο και έχουμε μεγαλύτερο αυτοέλεγχο από την προηγούμενη εκτέλεση της μεθόδου με ρυθμό μάθησης 0.1, αλλά μικρότερο από την αντίστοιχη περίπτωση στο υποκεφάλαιο 4.2.2 με τους τυχαίους πίνακες.



Σχήμα 4.12α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*



Σχήμα 4.12β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

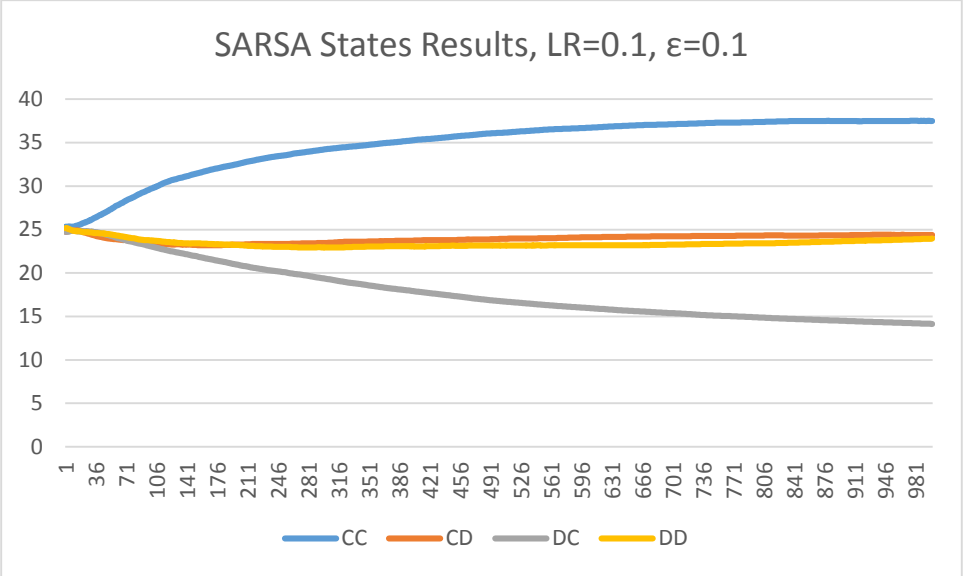
Στο πιο πάνω σχήμα (Σχήμα 4.12β) φαίνονται και πάλι τα ποσοστά του τελευταίου επεισοδίου για τις τέσσερις καταστάσεις του παιγνιδιού, κατά μέσο όρο. Όπου 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Όλες οι καταστάσεις έχουν μειωθεί και η κατάσταση του συμβιβασμού αυξήθηκε και έχει το πιο μεγάλο ποσοστό.

4.2.4 Μέθοδος τροποποίησης ολόκληρου του πίνακα αμοιβών

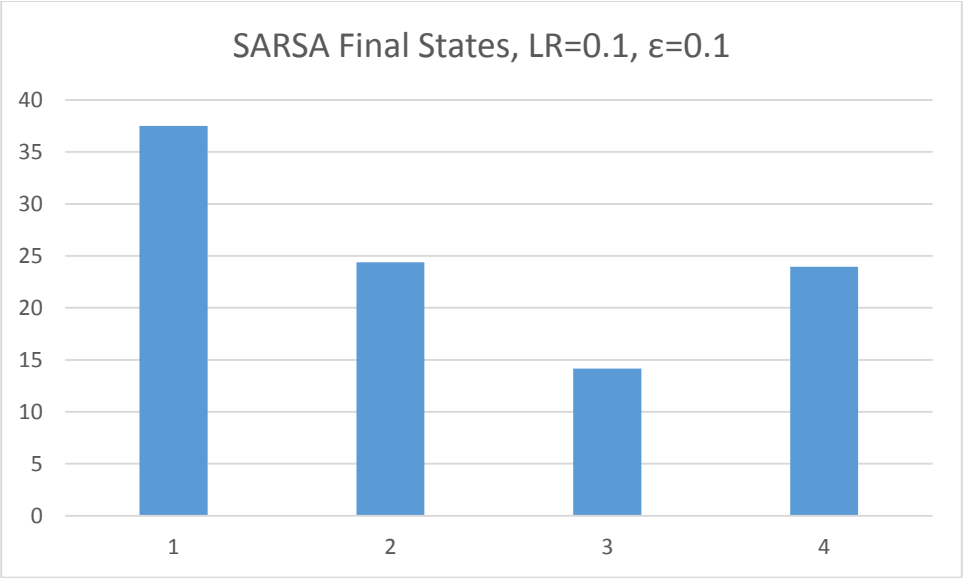
Στην τρίτη και τελευταία μέθοδο που χρησιμοποιήσαμε τροποποιούσαμε τον πίνακα με hill climbing όπως και στις προηγούμενες μεθόδους αλλά με τη διαφορά ότι αλλάζαμε όλες τις τιμές του πίνακα. Σε κάθε εκτέλεση το πρόγραμμα ξεκινά από τον ίδιο πίνακα αμοιβών και όχι με νέους τυχαίους-όπως στο υποκεφάλαιο 4.2.1 (μέθοδος πρώτη) και στο τέλος θα πάρουμε μέσο όρο των καταστάσεων για όλους τους γύρους.

Αρχικά εκτελέσαμε το πρόγραμμα με ρυθμό μάθησης = 0.1 και έψιλον = 0.1 και τα αποτελέσματα που παρουσιάζονται στο Σχήμα 4.13α είναι σχεδόν παρόμοια με αυτά της

αντίστοιχης περίπτωσης στο υποκεφάλαιο 4.2.3 (Δεύτερη Μέθοδος –Σχήμα 4.11α). Αν επικεντρωθούμε στο ποσοστό της κατάστασης CC, τότε έχουμε ίδια αποτελέσματα αυτοελέγχου με μόνη διαφορά ότι το ποσοστό στην παρούσα μέθοδο αυξάνεται πιο νωρίς (χρειάζεται λιγότερα επεισόδια για να αυξηθεί).



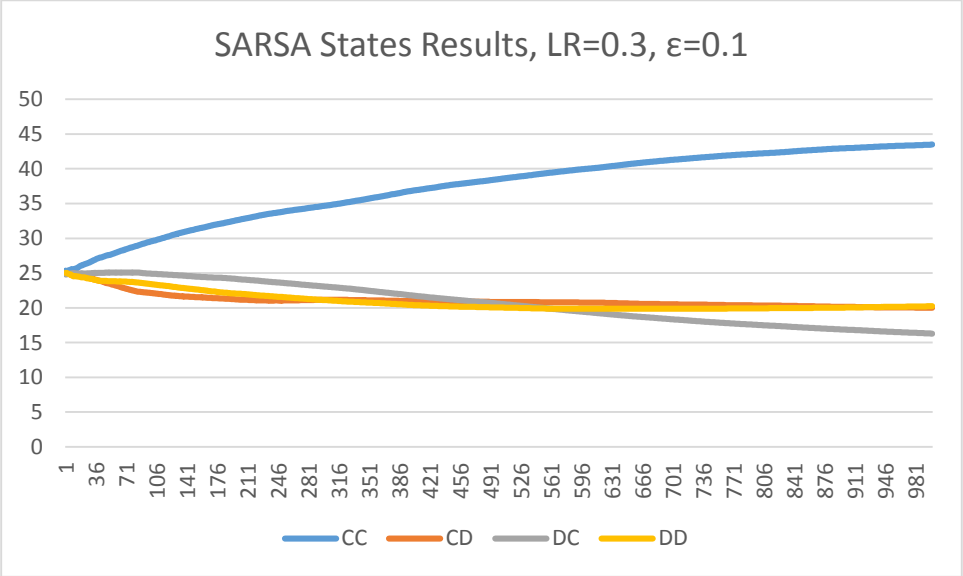
Σχήμα 4.13α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*



Σχήμα 4.13β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.1. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

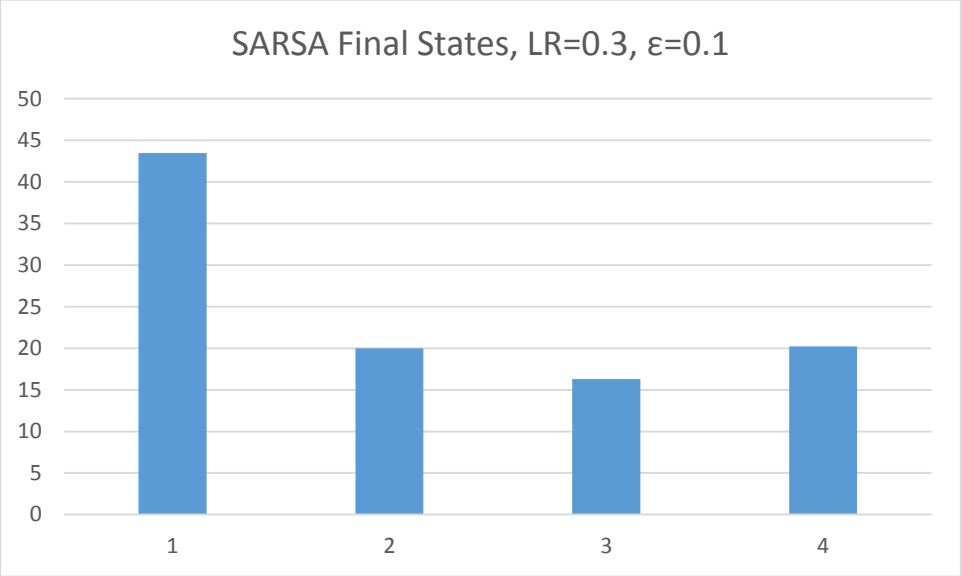
Το Σχήμα 4.13β επιβεβαιώνει τις παρατηρήσεις που καταγράψαμε πιο πάνω αφού παρουσιάζει τα ποσοστά των καταστάσεων στο τελευταίο επεισόδιο με το ποσοστό συμβιβασμου-αυτοελέγχου να κυμαίνεται και πάλι στο 37%. Ωστόσο παρατηρούνται αλλαγές στις καταστάσεις CD (στήλη 2) και DC (στήλη 3) αφού η CD αυξήθηκε και η DC παρουσίασε μείωση.

Στη συνέχεια, αυξάνοντας και πάλι τον ρυθμό μάθησης στα 0.3, ώστε να συγκρίνουμε με τα πιο πάνω, γίνεται αυτό που αναμέναμε και το ποσοστό του αυτοελέγχου αυξάνεται κι άλλο. Στο Σχήμα 4.14α και στο Σχήμα 4.14β παρατηρούμε το ποσοστό της κατάστασης CC να είναι το μεγαλύτερο από τις 4 καταστάσεις και να φτάνει σε τιμές σχεδόν 45%. Σε σύγκριση με την αντίστοιχη περίπτωση της προηγούμενης μεθόδου (Σχήμα 4.12α) βλέπουμε ότι το ποσοστό είναι μεγαλύτερο.



Σχήμα 4.14α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3.*

Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.



Σχήμα 4.14β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών: S=-3, P=-2, R=4, T=5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

4.3 Αποτελέσματα Συνειδητότητας

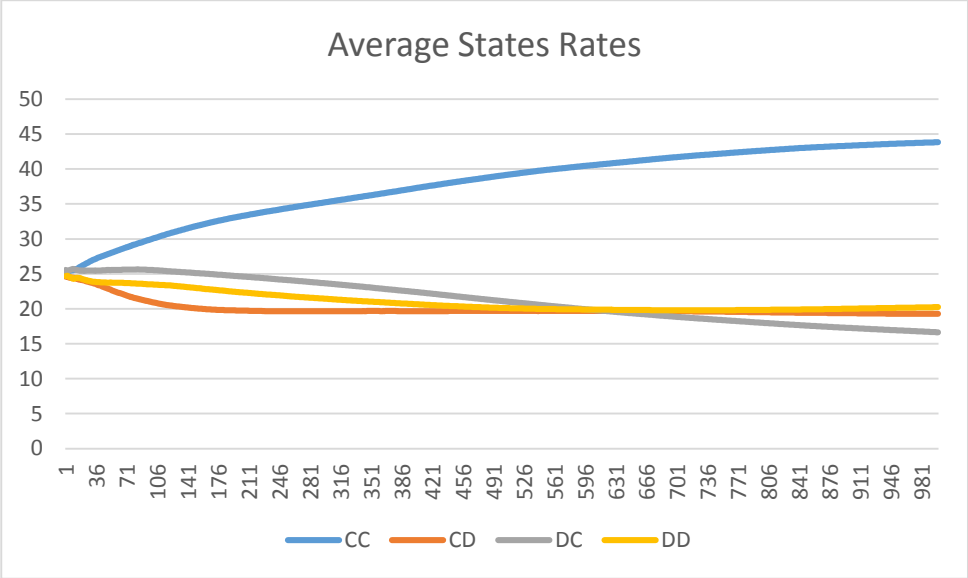
Σε αυτό το υποκεφάλαιο στόχος είναι να μελετήσουμε τα αποτελέσματα που πήραμε από τα πειράματα που έγιναν για τη έννοια τη συνειδητότητας, αλλά και να παρατηρήσουμε πως συνδέεται με τον αυτοέλεγχο. Γι' αυτό το λόγο τα αποτελέσματα τα οποία θα παρουσιαστούν είναι βασισμένα στη δεύτερη και τρίτη μέθοδο που περιεγράφηκαν πιο πάνω (Μέθοδος με σταθερό P και R στον πίνακα αμοιβών & Μέθοδος τροποποίησης ολόκληρου του πίνακα αμοιβών). Επιλέχθηκαν αυτές οι δύο μέθοδοι καθώς μπορούμε άμεσα να συσχετίσουμε και να παρατηρήσουμε τη συνειδητότητα μέσα από τις τιμές του πίνακα αμοιβών οι οποίες δε θα αλλάζουν με τυχαίο τρόπο. Εκτός αυτού τα αποτελέσματα της πρώτης μεθόδου συνολικά ήταν λιγότερο καλά συγκριτικά με τις άλλες δύο μεθόδους.

Πριν προχωρήσουμε με τα αποτελέσματα, να θυμηθούμε πως η συνειδητότητα ορίζεται μέσα από τις τιμές του πίνακα αμοιβών. Ο πίνακας αμοιβών αποτελείται από τις 4 τιμές S-Sucker, P-Punishment, R-Reward, T-Temptation. S ως αμοιβή λαμβάνει ο πράκτορας ο οποίος συνεργάστηκε ενώ ο άλλος πράκτορας όχι, λαμβάνοντας T ως αμοιβή. R είναι η αμοιβή που λαμβάνουν και οι δύο εάν συνεργαστούν και P η αμοιβή όταν κανείς από τους δύο δεν συνεργαστεί. Όπως αναφέραμε σε προηγούμενα κεφάλαια η συνειδητότητα αντιστοιχεί στην εσωτερική σύγκρουση που έχουμε ή με απλά λόγια στο δίλημμα που αντιμετωπίζουμε. Η εσωτερική σύγκρουση των πρακτόρων αφορά το αν θα συνεργαστούν ή όχι. Άρα λαμβάνοντας υπόψη τα πιο πάνω μπορούμε να πούμε ότι η εσωτερική σύγκρουση/συνειδητότητα ενισχύεται όταν έχουμε μεγάλη διαφορά ανάμεσα στα εξής ζεύγη τιμών: T-S, T-R, P-S.

Πιο κάτω λοιπόν, παρουσιάζονται αποτελέσματα με διάφορους αρχικούς πίνακες αμοιβών, με μεγάλη αλλά και μικρή αρχική σύγκρουση, έτσι ώστε να δούμε πως επηρεάζεται το ποσοστό του αυτοελέγχου. Οι πράκτορες παραμένουν ως έχουν με ίδιο παράγοντα έκπτωσης (0.1-κάτω και 0.9-άνω), εκπαιδεύονται με **SARSA** και θα έχουν και οι δύο ρυθμό μάθησης **0.3** αφού όπως είδαμε στα πειράματα για τον αυτοέλεγχο αποδίδει καλύτερα. Απώτερος σκοπός είναι βέβαια να αποδειχθεί η αντίστροφη σχέση μεταξύ συνειδητότητας και αυτοελέγχου.

4.3.1 Μέθοδος με σταθερό P και R στον πίνακα αμοιβών

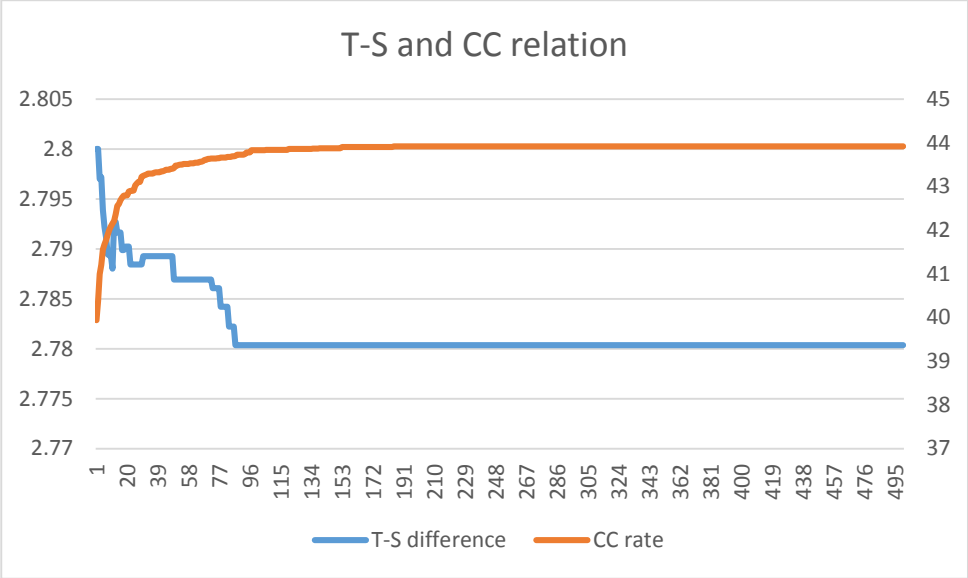
Οι τιμές P και R παραμένουν σταθερές όλη τη διάρκεια του παιχνιδιού και το μόνο που αλλάζει είναι οι άλλες τιμές S και T οι οποίες σχετίζονται άμεσα με τη συνειδητότητα. Στην πρώτη εκτέλεση έχουμε μικρή διαφορά ανάμεσα στις τιμές, άρα μικρό ποσοστό συνειδητότητας. Ο πίνακας αμοιβών (S, P, R, T) έχει ως εξής : **-1.3, -1.2, 1.4, 1.5**



Σχήμα 4.15α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών (S, P, R, T): -1.3, -1.2, 1.4, 1.5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

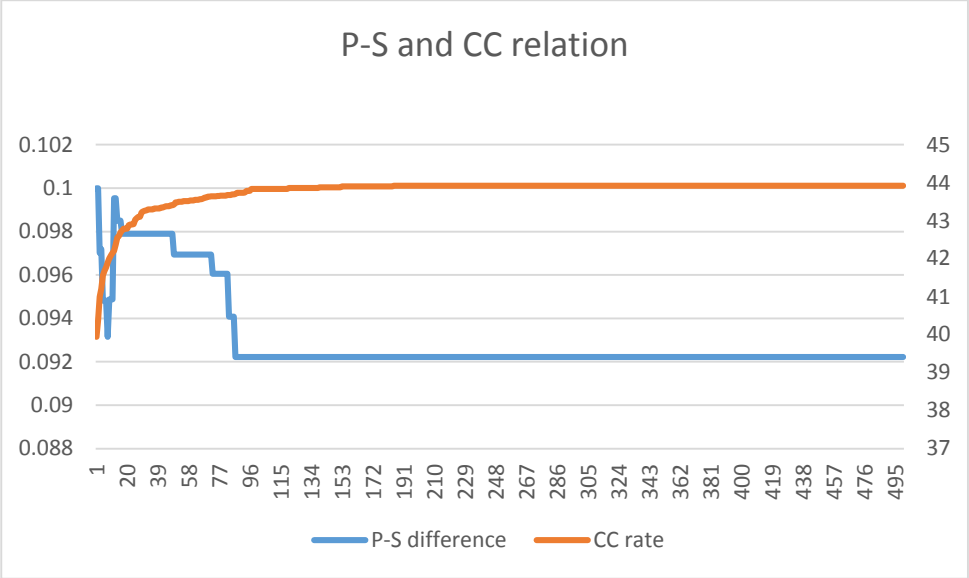
Στο Σχήμα 4.15α παρουσιάζεται η κλασσική γραφική παράσταση για τις τέσσερις καταστάσεις του παιχνιδιού, όπου και πάλι φαίνεται η κατάσταση του συμβιβασμού (μπλε γραμμή) να αυξάνεται. Ωστόσο το ποσοστό του συμβιβασμού είναι σχετικά μικρό (λιγότερο από 45%), άρα δεν έχουμε μεγάλη ανάπτυξη αυτοελέγχου. Ας δούμε όμως πως επηρεάστηκε το ποσοστό του συμβιβασμού (αυτοελέγχου) μέσα από τις εναλλαγές τις συνειδητότητας.

Πιο κάτω παρατίθενται τρεις γραφικές παραστάσεις για τα τρία ζεύγη τιμών που ορίζουν την συνειδητότητα σε σχέση με το ποσοστό αυτοελέγχου. Στις γραφικές υπάρχουν δύο κάθετοι ψ-άξονες, ένας για τις τιμές του αυτοελέγχου και ένας για τις τιμές της διαφοράς των τιμών αμοιβής.



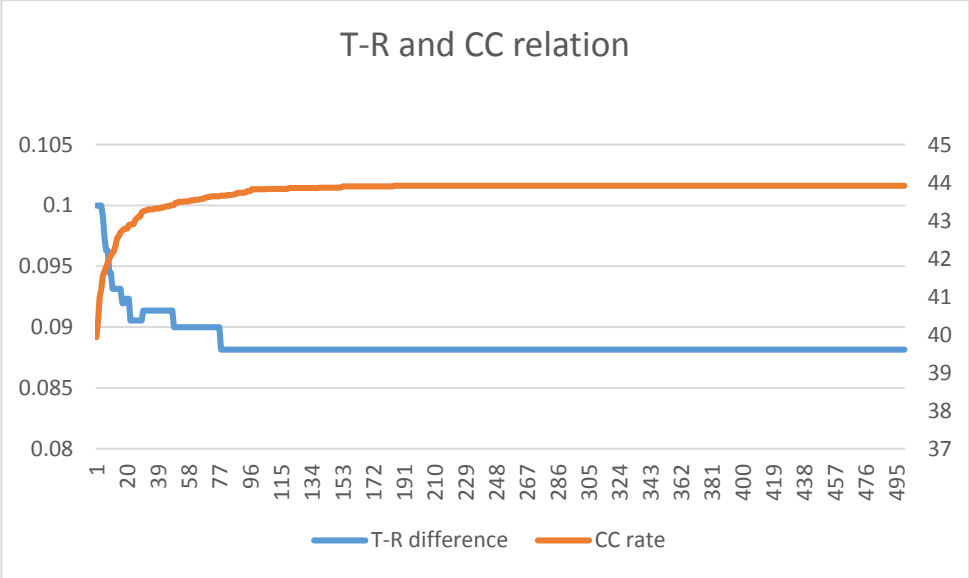
Σχήμα 4.15β: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών (S, P, R, T): -1.3, -1.2, 1.4, 1.5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

Στο Σχήμα 4.15β έχουμε την συνειδητότητα να αντιπροσωπεύεται από τη διαφορά μεταξύ των τιμών T και S (μπλε γραμμή) ενώ ο αυτοέλεγχος να αντιπροσωπεύεται από το ποσοστό συμβιβασμού (πορτοκαλί γραμμή). Παρατηρούμε ότι κατά τη διάρκεια των γύρων του παιχνιδιού η διαφορά μεταξύ των τιμών της συνειδητότητας μειώνεται ενώ παράλληλα το ποσοστό συμβιβασμού αυξάνεται. Οι αλλαγές ωστόσο δεν είναι σημαντικά μεγάλες, οπότε είναι δύσκολο εκ πρώτης όψεως να βγάλουμε συμπεράσματα.



Σχήμα 4.15γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών Ρ-Σ (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών (S, P, R, T): -1.3, -1.2, 1.4, 1.5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

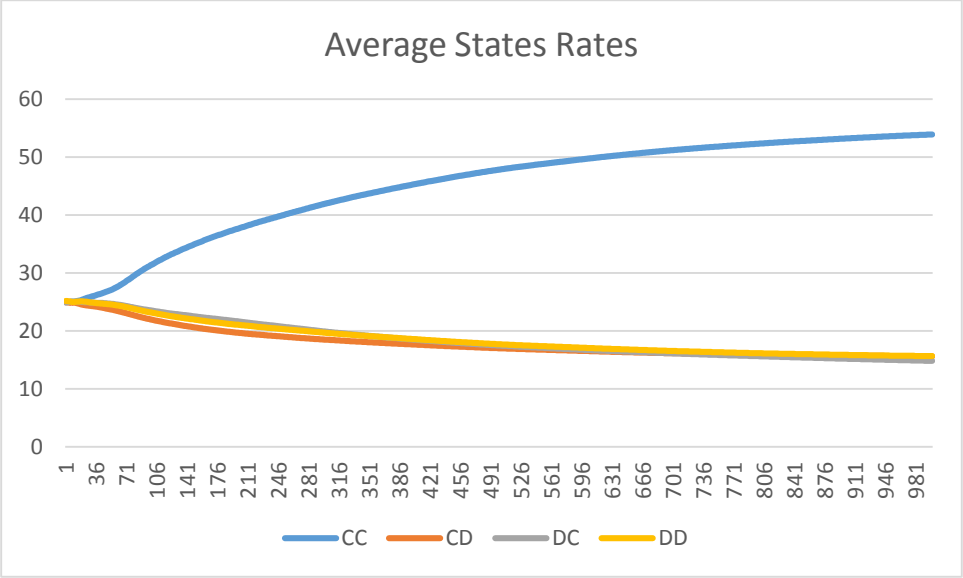
Το Σχήμα 4.15γ είναι αντίστοιχο του Σχήματος 4.15β με τη μόνη διαφορά ότι τώρα η συνειδητότητα αντιπροσωπεύεται από τις τιμές S και P. Με την πάροδο των επεισοδίων η διαφορά μεταξύ των τιμών S και P μειώνεται, όπως βλέπουμε στην μπλε γραμμή, ενώ την ίδια ώρα το ποσοστό συμβιβασμού παρουσιάζει αύξηση. Και πάλι οι αυξομειώσεις δεν είναι έντονες, αποτρέποντάς μας από το να καταλήξουμε σε κάποιο συμπέρασμα για τη συνειδητότητα και τον αυτοέλεγχο.



Σχήμα 4.15δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών (S, P, R, T): -1.3, -1.2, 1.4, 1.5. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

Τελευταίο σχήμα για τον πίνακα αμοιβών με μικρή σύγκρουση είναι το πιο πάνω (Σχήμα 4.15δ), στο οποίο παρουσιάζεται η διαφορά των τιμών R και T εκ μέρους της συνειδητότητας και το ποσοστό συνεργασίας εκ μέρους του αυτοελέγχου. Τα αποτελέσματα δεν είναι και τόσο καλά πάλι λόγω της μικρής σύγκρουσης που υπάρχει στις τιμές του πίνακα, αλλά έστω και με μικρές μεταβολές φαίνεται ότι η διαφορά T-R μειώνεται, άρα οι τιμές έρχονται πιο κοντά.

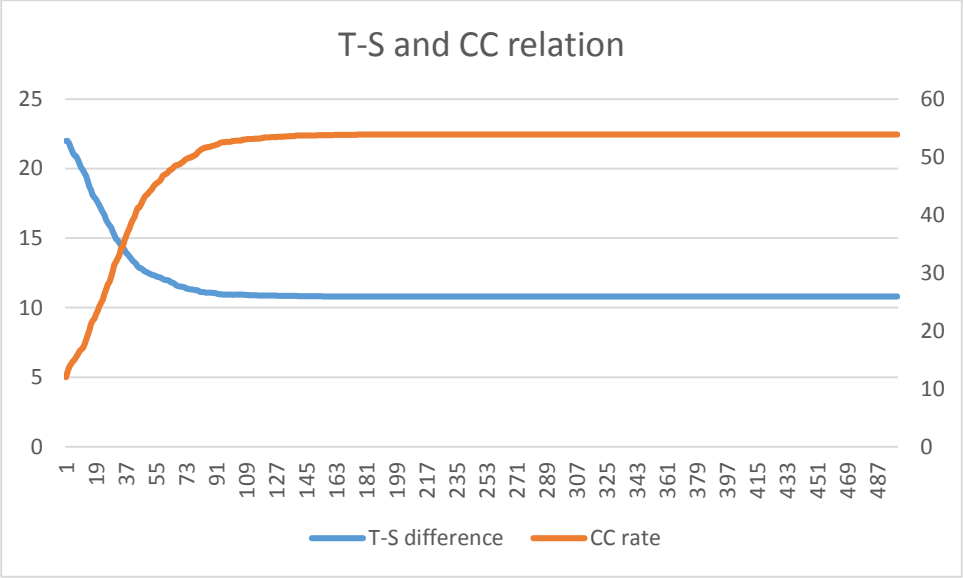
Ας εξετάσουμε μια μεγαλύτερη εσωτερική σύγκρουση για να δούμε ποιες είναι οι διαφορές. Το δεύτερο πείραμα που θα δούμε εκτελέστηκε με τους πράκτορές μας να παίρνουν αμοιβές από τον πίνακα με της εξής τιμές: **S=-15, P=-9, R=1.4, T=7**. Σε αυτή την περίπτωση έχουμε μεγαλύτερη αρχική σύγκρουση από πριν καθώς αυξήσαμε τις τιμές των αμοιβών. Άρα έχουμε μεγαλύτερο επίπεδο συνειδητότητας αρχικά αφού πλέον η διαφορά ανάμεσα στις τιμές T-S, T-R και P-S έχει αυξηθεί.



Σχήμα 4.16α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

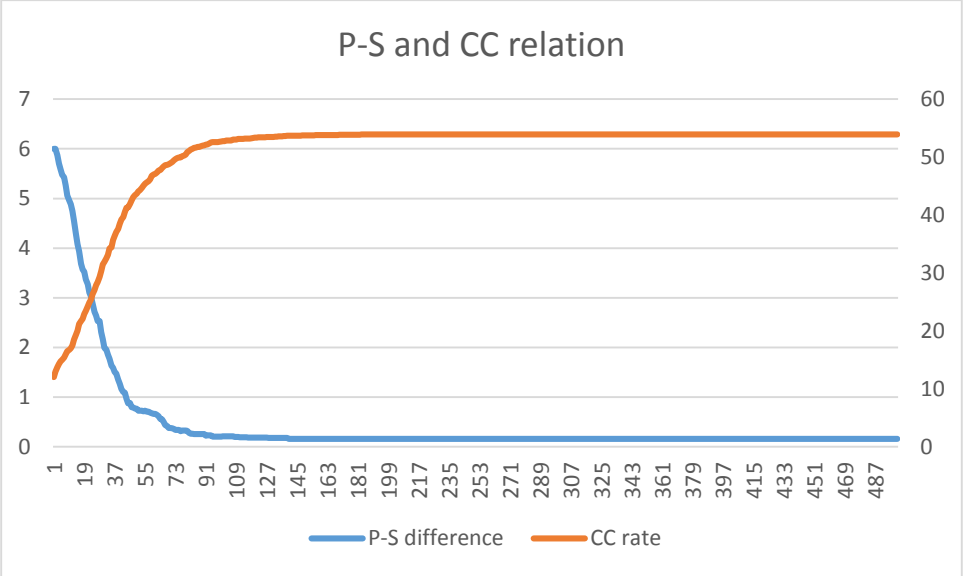
Στο Σχήμα 4.16α παρατηρούμε ότι η μεγαλύτερη αρχική σύγκρουση που δώσαμε στον πίνακα, μας έδωσε μεγαλύτερο ποσοστό αυτοελέγχου (ποσοστό κατάστασης CC-μπλε γραμμή) και η τιμή του ποσοστού υπερβαίνει το 50%.

Στις γραφικές παραστάσεις πιο κάτω θα δούμε και πάλι πως μεταβλήθηκε η συνειδητότητα κρατώντας σταθερές τις τιμές P και R σε αυτή την περίπτωση.



Σχήμα 4.16β*: Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών Τ-Σ (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

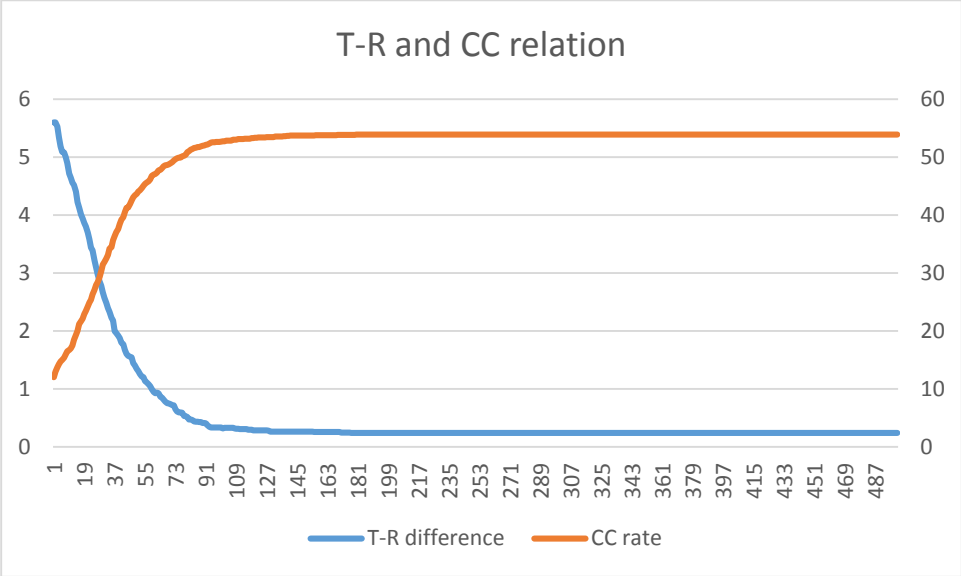
Αρχικά στην γραφική αυτή έχουμε τη συνειδητότητα (διαφορά S-T) και τον αυτοέλεγχο (ποσοστό συνεργασίας) να ακολουθούν αντίθετη πορεία, όπως αναμέναμε. Η συνειδητότητα μειώνεται καθώς ο αυτοέλεγχος αυξάνεται. Το ίδιο συμβαίνει και στην επόμενη γραφική-Σχήμα 4.16γ.



Σχήμα 4.16γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών P-Σ (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

Η διαφορά συνειδητότητας και αυτοελέγχου είναι ακόμα πιο έντονη στο Σχήμα 4.16γ αφού η συνειδητότητα μειώνεται κατά πολύ λόγω της μείωσης της διαφοράς ανάμεσα στις αμοιβές S και P. Το τοπίο αρχίζει να ξεκαθαρίζει και κατευθυνόμαστε προς το αποτέλεσμα και κατάληξη που θέλουμε.

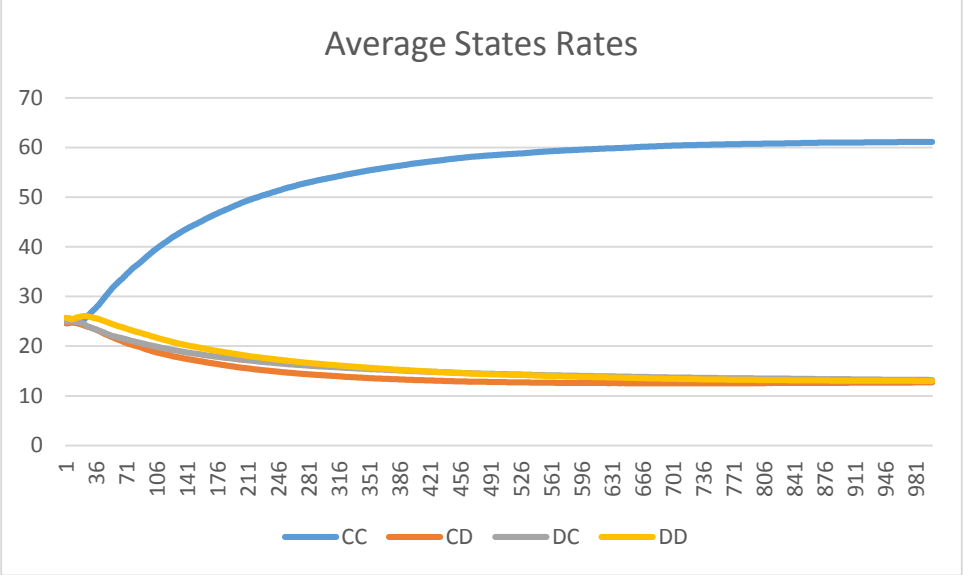
Συμπληρωματικά, μπορούμε να δούμε και την τρίτη γραφική παράσταση για το συγκεκριμένο πείραμα η οποία παρουσιάζει την συνειδητότητα μέσα από το τρίτο ζευγάρι αμοιβών (T-R) και τον αυτοέλεγχο να διασταυρώνονται. Δηλαδή και πάλι η διαφορά μεταξύ των αμοιβών μειώνεται, άρα μειώνεται η συνειδητότητα και ο αυτοέλεγχος ακολουθεί την ανοδική του πορεία μέχρι και λίγο πιο πάνω από το 50%. η συνειδητότητα μειώνεται αισθητά αφού η διαφορά μεταξύ των αμοιβών φτάνει σχεδόν το μηδέν. Όλα αυτά παρουσιάζονται πιο κάτω στη γραφική παράσταση Σχήμα 4.16δ.



Σχήμα 4.16δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3.*

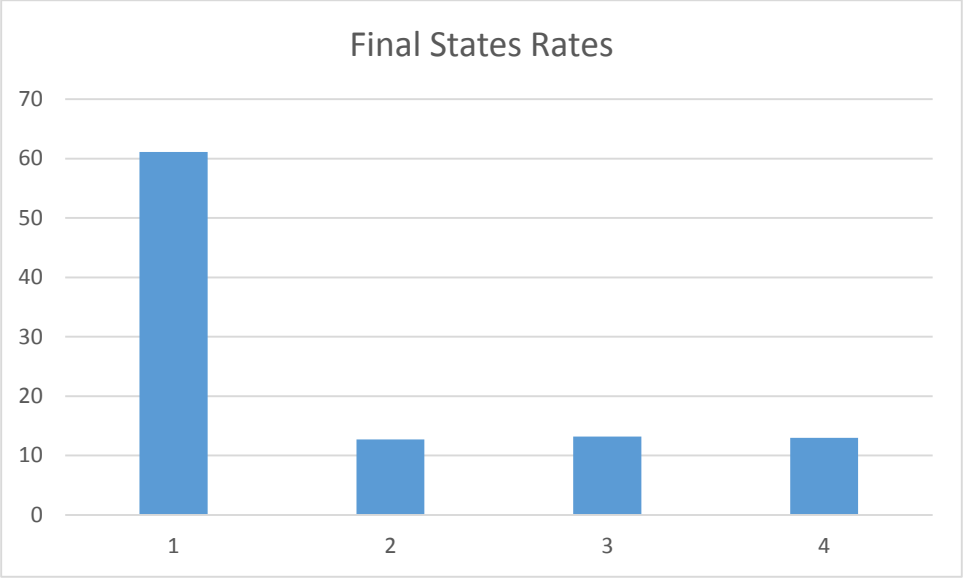
Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.

Στο τελευταίο πείραμα για τη μέθοδο με σταθερό P και R στον πίνακα αμοιβών, προσπαθήσαμε να δείξουμε τη σχέση συνειδητότητας και αυτοελέγχου αρχικοποιώντας τον πίνακα αμοιβών με πολύ μεγάλη σύγκρουση. Ο πίνακας που χρησιμοποιήθηκε ήταν ο εξής: **S=-55, P=-48, R=-2, T=10**.



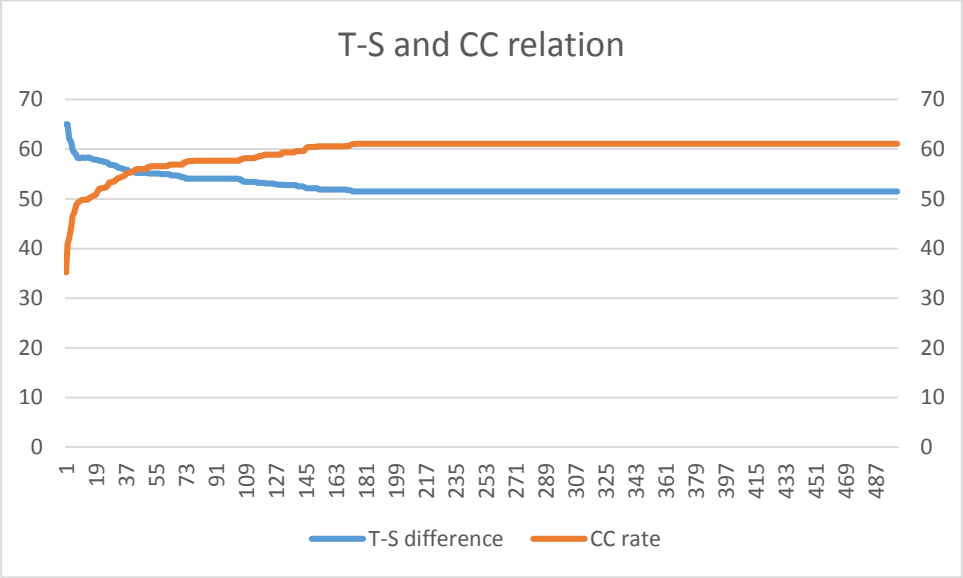
Σχήμα 4.17α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

Στο Σχήμα 4.17α έχουμε και πάλι την κλασσική γραφική παράσταση με τα ποσοστά για τις τέσσερις καταστάσεις του παιγνιδιού. Αυτό που είναι αξιοσημείωτο είναι ότι οι γραμμές είναι πιο ομαλές και η μπλε γραμμή η οποία αντιπροσωπεύει τον αυτοέλεγχο έχει αυξηθεί ακόμα περισσότερο από προηγουμένως. Το ποσοστό αυτό ξεπερνά και το 60% λόγω της μεγαλύτερης αρχικής σύγκρουσης που ορίσαμε στον πίνακα.



Σχήμα 4.17β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

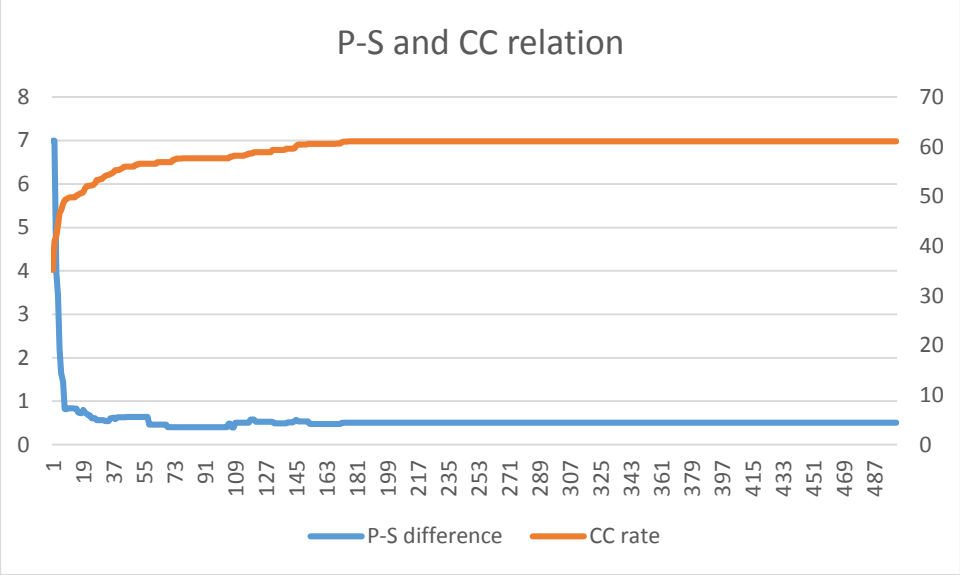
Παραθέτουμε και το Σχήμα 4.17β στο οποίο φαίνεται πιο έντονα η διαφορά ανάμεσα στα ποσοστά των καταστάσεων αλλά και η αύξηση του ποσοστού συμβιβασμού (στήλη 1) σε 60% έναντι των άλλων καταστάσεων οι οποίες κυμαίνονται λίγο πιο πάνω από το 10%.



Σχήμα 4.17γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού*

(πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.

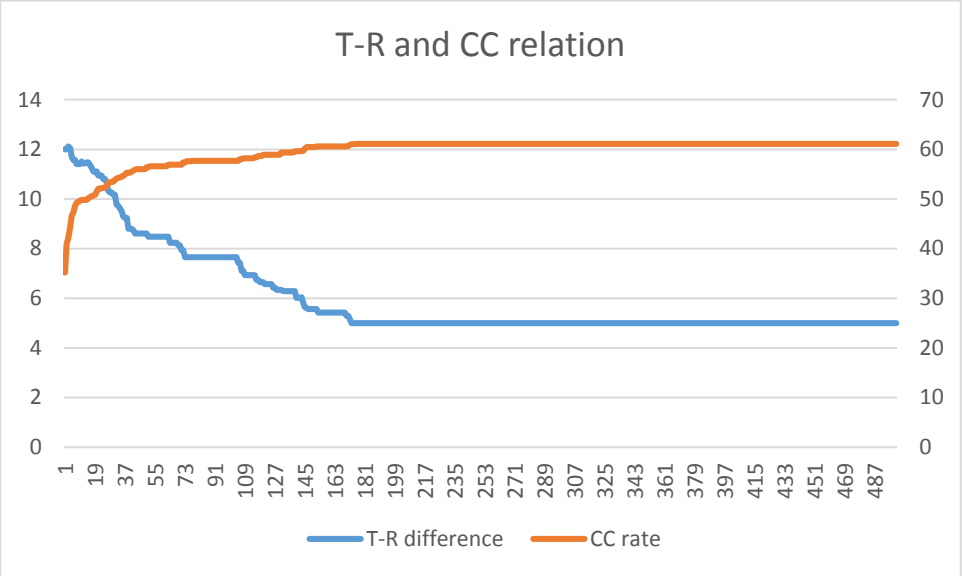
Πιο πάνω στο Σχήμα 4.17γ παρατηρείται η συνειδητότητα να μειώνεται μέχρι και 15 μονάδες, μετά τις εναλλαγές των τιμών S και T από το hill climbing. Επίσης την ίδια ώρα ο αυτοέλεγχος και το ποσοστό συμβιβασμού αυξάνονται όπως είδαμε και πριν μέχρι την τιμή 60%.



Σχήμα 4.17δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών P-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

Η συνειδητότητα στο Σχήμα 4.17δ μειώνεται και πάλι σε τεράστιο βαθμό και η αντίθεση που θέλουμε να δείξουμε φαίνεται καθαρά, αφού η διαφορά μεταξύ S και P μειώνεται άρα και η συνειδητότητα, ενώ το ποσοστό συνεργασίας αυξάνεται άρα και ο αυτοέλεγχος. Επίσης να τονίσουμε ότι η αρχική διαφορά μεταξύ S και P στο τέλος του παιχνιδιού μειώνεται τόσο που σχεδόν γίνεται μηδενική.

Από την άλλη, στη γραφική παράσταση που δίνουμε στο Σχήμα 4.17ε έχουμε μεγάλη αντίθεση στις τιμές συνειδητότητας και αυτοελέγχου και η αντίστροφη σχέση τους είναι επίσης έντονη. Όπως φαίνεται και στη γραφική η συνειδητότητα μειώνεται ραγδαία από τα αρχικά στάδια του παιχνιδιού και έτσι δεν έχουμε την ομαλή καμπύλη που είχαμε στα προηγούμενα παραδείγματα. Παρόλα αυτά είναι εύκολο να καταλήξουμε ότι η συνειδητότητα και ο αυτοέλεγχος τελικά συνδέονται με αντιστρόφως ανάλογη σχέση.



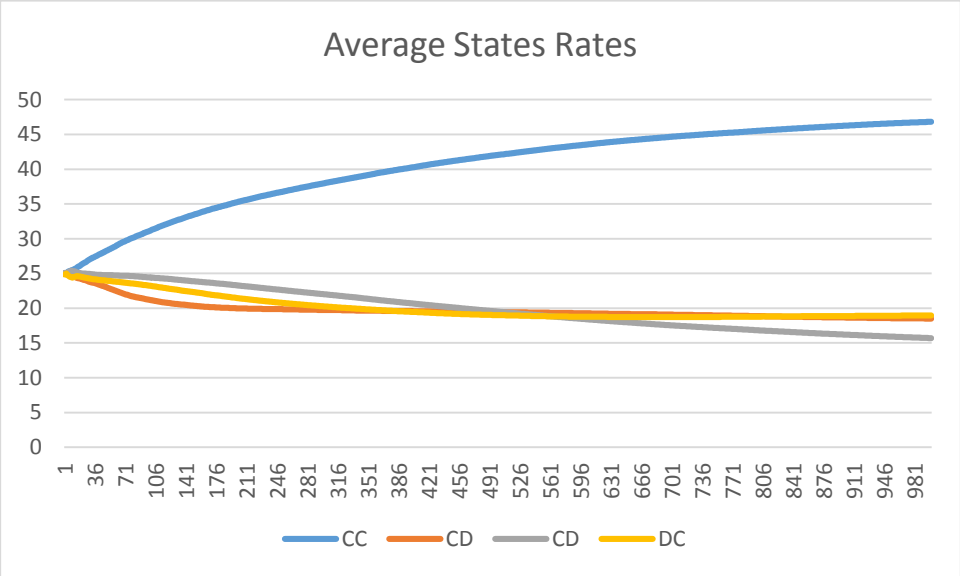
Σχήμα 4.17ε: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Σταθερό P και R στον πίνακα αμοιβών.*

4.3.2 Μέθοδος τροποποίησης ολόκληρου του πίνακα αμοιβών

Συνεχίζοντας με τα πειράματά μας θα μελετήσουμε τη συνειδητότητα εκτελώντας το παιχνίδι με τη μέθοδο τροποποίησης ολόκληρου του πίνακα αμοιβών. Με λίγα λόγια αυτή τη φορά το hill climbing θα μεταβάλλει και τις τέσσερις τιμές αμοιβών των πρακτόρων. Οι ρυθμοί μάθησης, παράγοντες έκπτωσης και τιμή του έψιλον παραμένουν όπως στο υποκεφάλαιο 4.3.1. Επίσης τα πειράματα τα οποία θα εξετάσουμε θα είναι με τους ίδιους πίνακες όπως πριν για να μπορούμε να συγκρίνουμε τα αποτελέσματα.

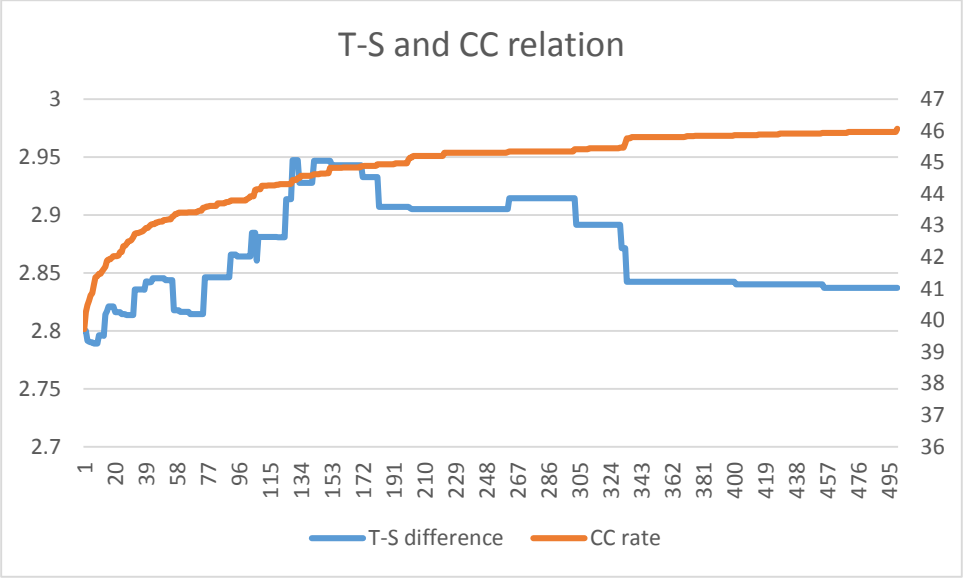
Αρχικά θα δούμε ένα πείραμα όπου ο πίνακας έχει μικρή σύγκρουση και ο πίνακας (S, P, R, T) έχει ως εξής : **-1.3, -1.2, 1.4, 1.5**

Στη γραφική παράσταση με τα ποσοστά των καταστάσεων (Σχήμα 4.18α) έχουμε και πάλι το ποσοστό του αυτοελέγχου (μπλε γραμμή) να υπερισχύει αλλά να μη φτάνει σε πολύ μεγάλες τιμές. Το ποσοστό που μας αφορά φτάνει μόλις λίγο πάνω από το 45.%

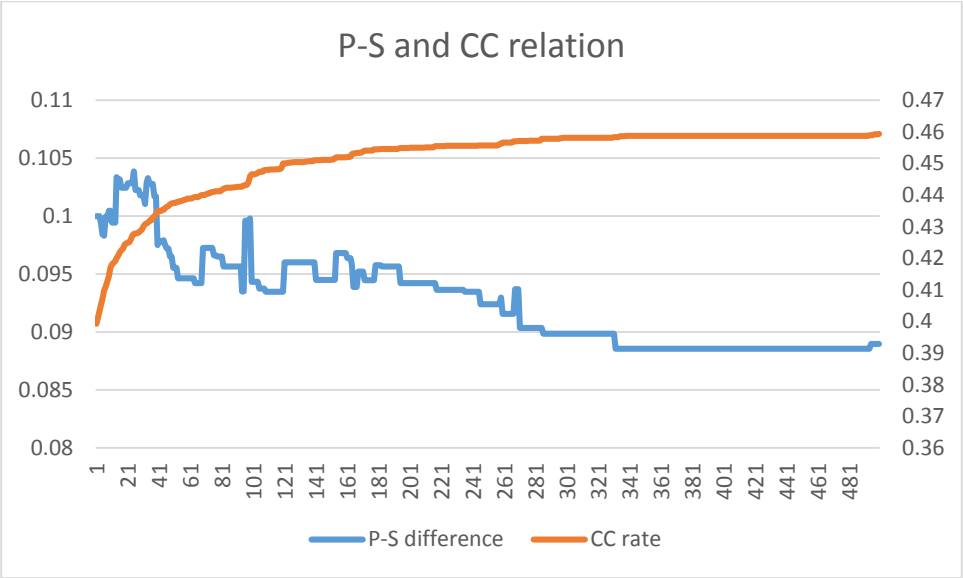


Σχήμα 4.18α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-1.3, P=-1.2, R=1.4, T=1.5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Στην δεύτερη γραφική παράσταση (Σχήμα 4.18β) βλέπουμε ότι η καμπύλη για τη διαφορά των τιμών S-T εκτός του ότι δεν είναι ομαλή, παρουσιάζει αύξηση στην αρχή και στη συνέχεια μείωση. Ωστόσο, οι αυξομειώσεις αυτές κινούνται σε μικρά πλαίσια των 15% και 10%, πράγμα που μας δίνει τη δυνατότητα να τις θεωρήσουμε αμελητέες και να τις αποδώσουμε στο γεγονός ότι οι τιμές του πίνακα είναι η μια πολύ κοντά στην άλλη. Η μικρή διαφορά μεταξύ των αρχικών τιμών δυσκολεύει τη δουλειά του hill climbing το οποίο προσπαθεί να τις φέρει όσο πιο κοντά γίνεται.



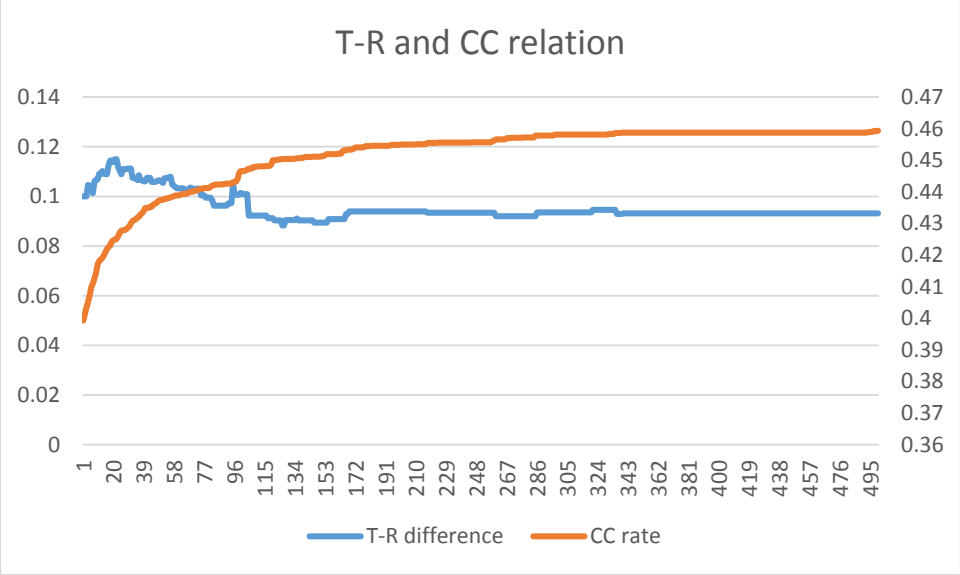
Σχήμα 4.18β: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-1.3, P=-1.2, R=1.4, T=1.5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*



Σχήμα 4.18γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών P-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3.*

Έψιλον: 0.1 Πίνακας αμοιβών S=-1.3, P=-1.2, R=1.4, T=1.5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.

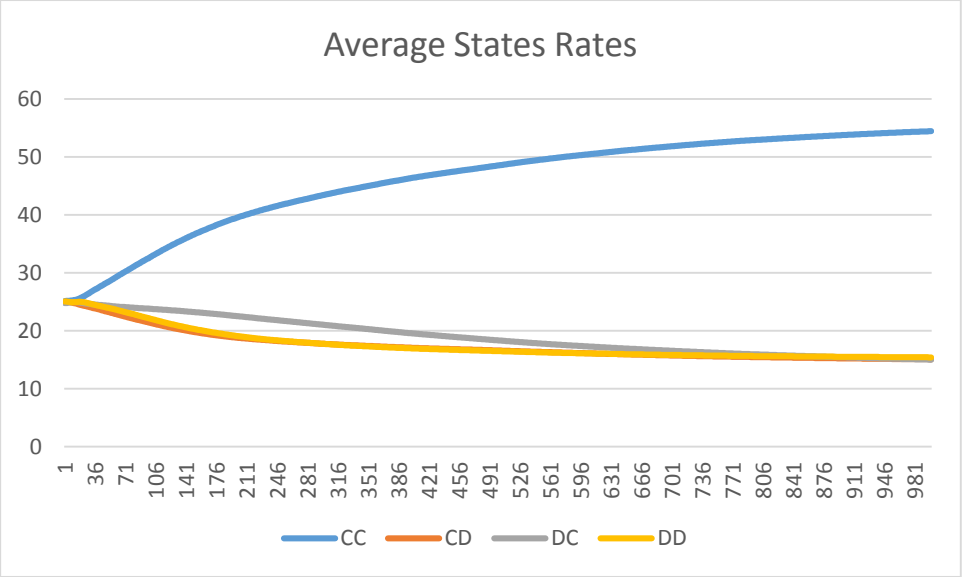
Στο Σχήμα 4.18γ έχουμε τη γραφική παράσταση για τη διαφορά των τιμών S και P που αντιπροσωπεύουν τη συνειδητότητα και το ποσοστό του αυτοελέγχου (πορτοκαλί γραμμή). Όπως και στο Σχήμα 4.18β η καμπύλη της συνειδητότητας δεν είναι ομαλή και παρουσιάζει αυξομειώσεις. Παρόλα αυτά εν τέλει ακολουθεί μια καθοδική πορεία σημειώνοντας μείωση.



Σχήμα 4.18δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-1.3, P=-1.2, R=1.4, T=1.5. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

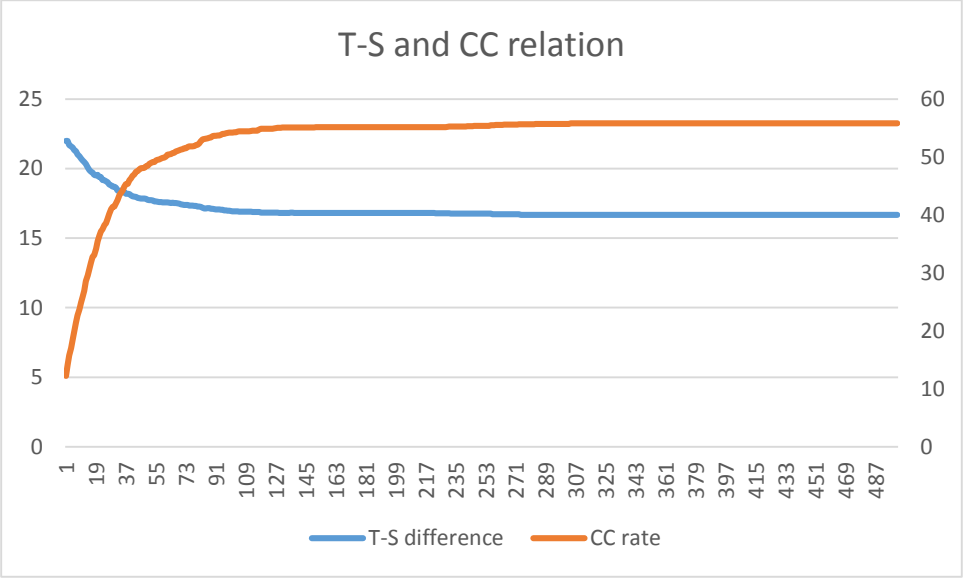
Η τελευταία γραφική παράσταση για αυτό το πείραμα (Σχήμα 4.18δ) παρουσιάζει τον αυτοέλεγχο και τη συνειδητότητα σε πιο ομαλές μορφές σε σχέση με τις δυο προηγούμενες γραφικές. Η αντίθεση μεταξύ συνειδητότητας (διαφορά R-T) και αυτοελέγχου (ποσοστό συμβιβασμού) δεν είναι πολύ μεγάλη, ωστόσο φαίνεται σε κάποιο βαθμό η αντιστρόφως ανάλογη σχέση των δύο εννοιών.

Συνεχίζοντας εξετάζουμε ένα δεύτερο πείραμα, αυξάνοντας τη σύγκρουση και μετατρέποντας τον πίνακα ως εξής: **S=-15, P=-9, R=1.4, T=7**. Αυξάνοντας τη διαφορά ανάμεσα στις τιμές, ξεκινούμε το παιγνίδι με μεγάλη συνειδητότητα. Λόγω αυτού αναμένουμε μετά την εκτέλεση του hill climbing να μειωθεί η διαφορά και η συνειδητότητα. Επίσης όπως φαίνεται και στη γραφική του Σχήματος 4.19α ο αυτοέλεγχος μαζί με το ποσοστό του συμβιβασμού αυξάνεται περισσότερο από το προηγούμενο πείραμα με μικρή σύγκρουση. Ο αυτοέλεγχος κυμαίνεται όπως και στην προηγούμενη μέθοδο λίγο πιο πάνω από 50% και λίγο περισσότερο.



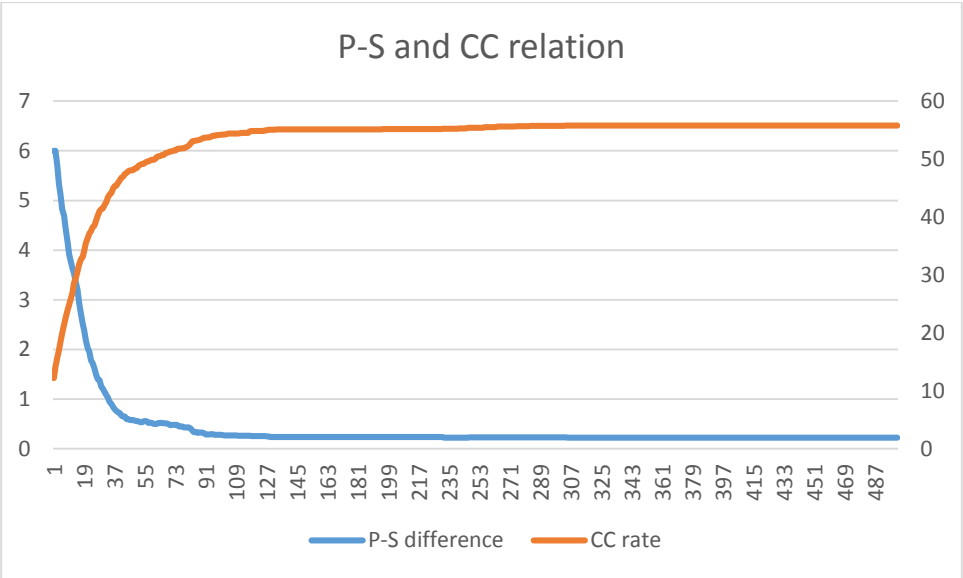
Σχήμα 4.19α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Με την αύξηση της διαφοράς ανάμεσα στις 4 τιμές του πίνακα αμοιβών και συνάμα αύξηση της σύγκρουσης οι γραφικές παραστάσεις για τη συνειδητότητα είναι πιο ομαλές. Πιο κάτω βλέπουμε στο Σχήμα 4.19β τη συνειδητότητα να αναπαρίσταται από τη διαφορά S και T και με την πάροδο των γύρων του παιχνιδιού να μειώνεται. Η μείωση είναι μεγαλύτερη από ότι στο προηγούμενο παράδειγμα και οι καμπύλες συνειδητότητας και αυτοελέγχου σχηματίζουν μια ωραία καθαρή αντίθεση.



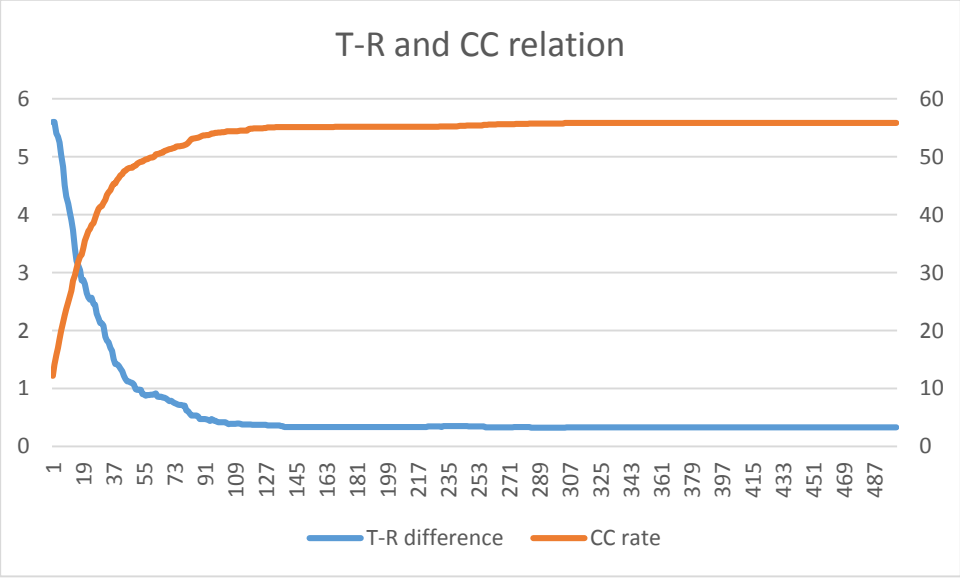
Σχήμα 4.19β: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Επίσης ξεκάθαρη αντίθεση παρουσιάζεται και στη γραφική παράσταση του Σχήματος 4.19γ αφού η διαφορά μεταξύ των τιμών S και P μειώνεται σε πολύ μεγάλο βαθμό (κοντά στο 0) την ίδια ώρα που το ποσοστό του αυτοελέγχου ακολουθεί ανοδική πορεία προς το 60%.



Σχήμα 4.19γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών P-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

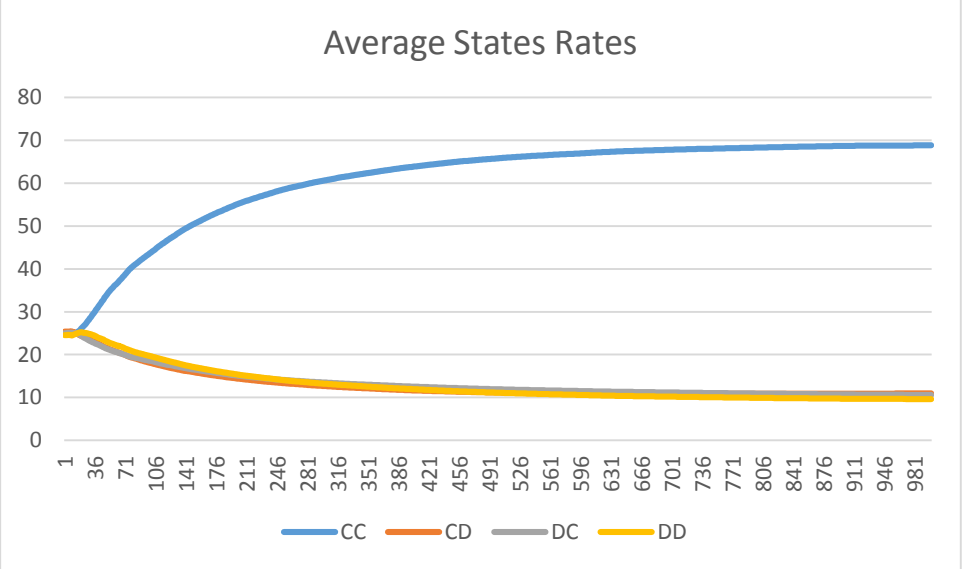
Εκτός των πιο πάνω την αντιστρόφως ανάλογη σχέση συνειδητότητας και αυτοελέγχου παρουσιάζει ξεκάθαρα η γραφική παράσταση πιο κάτω (Σχήμα 4.19δ). Στο Σχήμα αυτό έχουμε τη διαφορά ανάμεσα στις τιμές R και T του πίνακα από τη μια και το ποσοστό συνεργασίας από την άλλη. Για ακόμα μια φορά η μεγάλη αρχική σύγκρουση που ορίσαμε στον αρχικό πίνακα έχει ως αποτέλεσμα την διασταύρωση των δύο καμπυλών που βλέπουμε στη γραφική. Η καμπύλη του αυτοελέγχου αυξάνεται σταθερά (πορτοκαλί γραμμή) ενώ η καμπύλη της συνειδητότητας μειώνεται προς το μηδέν (μπλε γραμμή), επιβεβαιώνοντας για άλλη μια φορά την αντιστρόφως ανάλογη σχέση που θέλουμε να αποδείξουμε.



Σχήμα 4.19δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Τελευταίο πείραμα το οποίο θα παρουσιάσουμε είναι ένα πείραμα με πολύ μεγάλη εσωτερική σύγκρουση στους πράκτορες του παιχνιδιού, αντίστοιχη με αυτή που δώσαμε στο προηγούμενο υποκεφάλαιο με τη μέθοδο τροποποίησης δύο εκ των τεσσάρων τιμών του πίνακα αμοιβών.

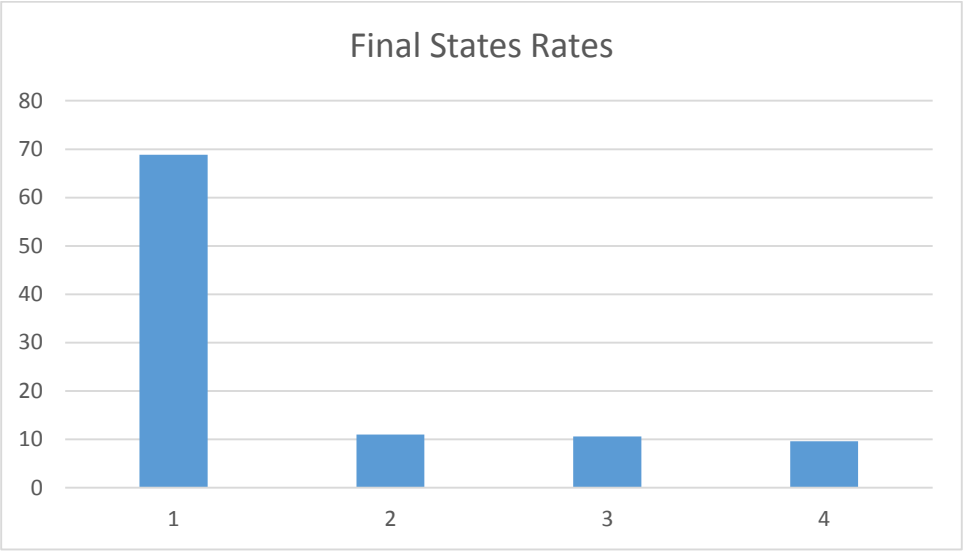
Οι τιμές του πίνακα έχουν αυξηθεί και είναι οι εξής: : **S=-55, P=-48, R=-2, T=10**. Όπως φαίνεται στη γραφική παράσταση πιο κάτω (Σχήμα 4.20α) το ποσοστό του αυτοελέγχου έχει αυξηθεί ακόμα περισσότερο, αφήνοντας σε πολύ μικρές τιμές τα υπόλοιπα ποσοστά. Ο αυτοέλεγχος φτάνει σχεδόν σε ποσοστό 70%.



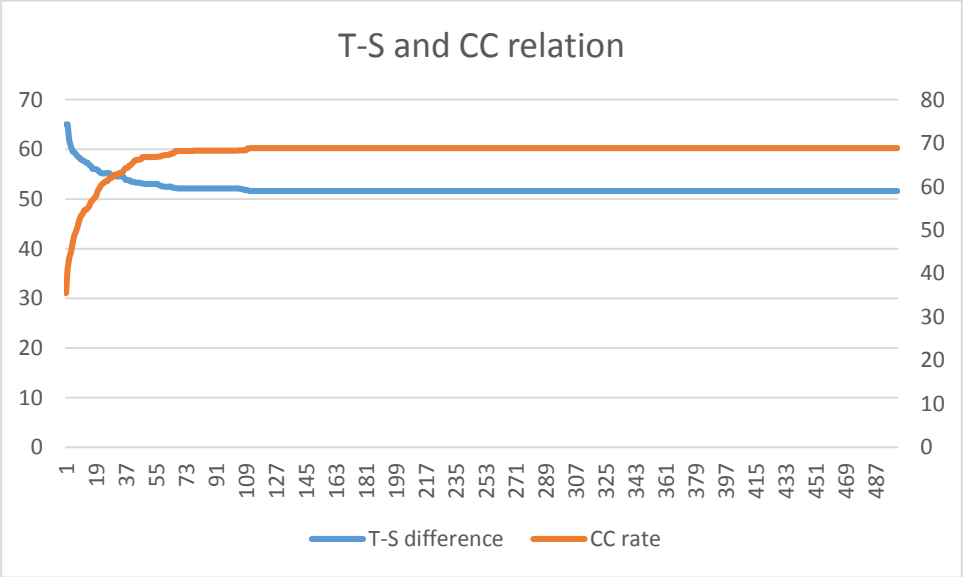
Σχήμα 4.20α: *Γραφική παράσταση ποσοστών των τεσσάρων καταστάσεων του παιγνιδιού κατά τη διάρκεια 1000 επεισοδίων. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-15, P=-9, R=1.4, T=7. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Την αυξημένη τιμή του ποσοστού της συνεργασίας και κατ' επέκταση και του αυτοελέγχου επιβεβαιώνει το Σχήμα 4.20β, όπου φαίνονται τα ποσοστά σε μορφή στηλών. Η πρώτη στήλη είναι το ποσοστό του αυτοελέγχου και όπως εύκολα μπορούμε να παρατηρήσουμε είναι μόλις λίγα εκατοστά κάτω από το 70%. Από την άλλη η καταστάσεις CD-δεύτερη στήλη, DC-τρίτη στήλη και DD-τέταρτη στήλη περιορίζονται κοντά στο 10%. Το ποσοστό του αυτοελέγχου εδώ αποτελεί το μεγαλύτερο από όλα τα

πειράματα που κάναμε μέχρι τώρα και αναμένουμε στις επόμενες γραφικές παραστάσεις να δούμε και ανάλογα καλά αποτελέσματα από πλευράς συνειδητότητας.



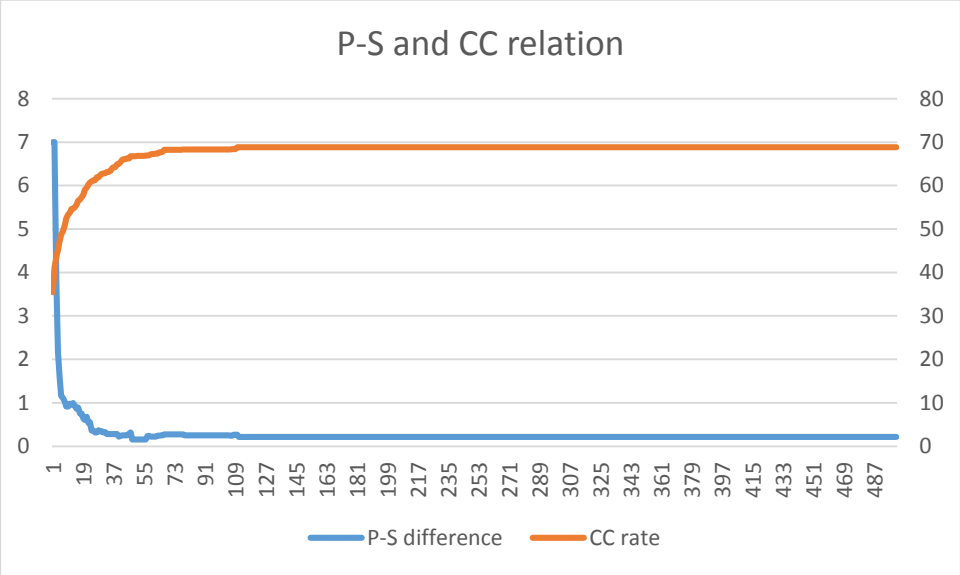
Σχήμα 4.20β: *Γραφική παράσταση με τα ποσοστά των 4 καταστάσεων του παιχνιδιού στο τελευταίο επεισόδιο. 1-κατάσταση CC, 2-κατάσταση CD, CD, 3-κατάσταση CD και 4-κατάσταση DD. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*



Σχήμα 4.20γ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού*

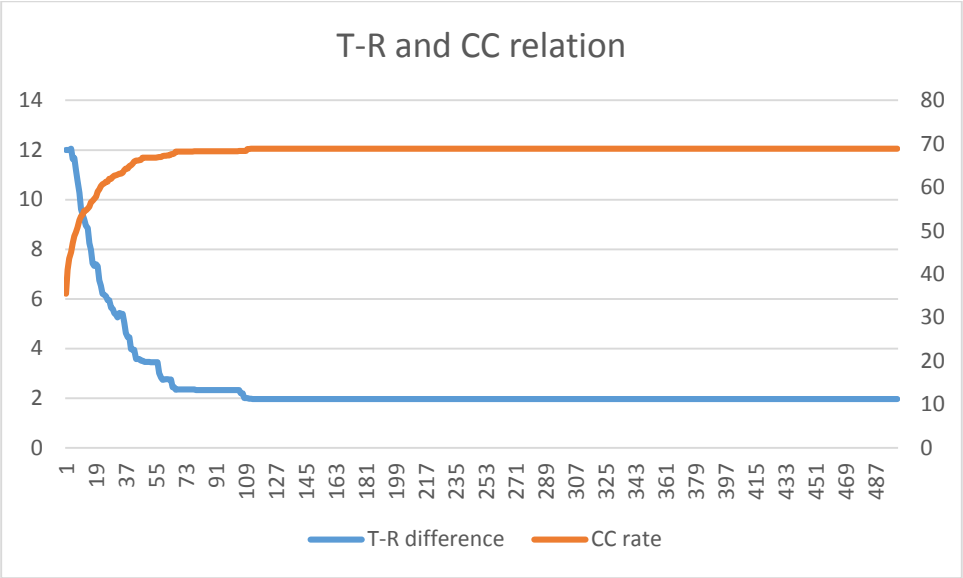
(πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.

Στο Σχήμα 4.20γ βλέπουμε πως μεταβλήθηκε η διαφορά των τιμών S και T μετά από το hill climbing. Όπως φαίνεται και από την μπλε γραμμή, η διαφορά αυτή μειώθηκε σχεδόν 15 μονάδες, ενώ το ποσοστό του συμβιβασμού διπλασιάστηκε. Άρα έχουμε από τη μια πλευρά τη συνειδητότητα να μειώνεται και από την άλλη τον αυτοέλεγχο να αυξάνεται.



Σχήμα 4.20δ: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών P-S (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Στη γραφική παράσταση πιο πάνω, η αντιθετική πορεία του αυτοελέγχου και της συνειδητότητας είναι πιο έντονη όπως μπορούμε να παρατηρήσουμε. Η μπλε γραμμή η οποία σχηματίζεται μέσα από τη διαφορά των τιμών S και P, αντιπροσωπεύει τη συνειδητότητα και μειώνεται ραγδαία. Η πορτοκαλί γραμμή όμως σχηματίζεται μέσα από το ποσοστό συμβιβασμού των δύο πρακτόρων και αντιπροσωπεύει τον αυτοέλεγχο ο οποίος αυξάνεται. Για άλλη μια φορά επιβεβαιώνεται η αντιστρόφως ανάλογη σχέση συνειδητότητας και αυτοελέγχου.



Σχήμα 4.20ε: *Γραφική παράσταση συνειδητότητας και αυτοελέγχου για τους 500 γύρους παιχνιδιού. Διαφορά αμοιβών T-R (μπλε γραμμή)=συνειδητότητα, ποσοστό συμβιβασμού (πορτοκαλί γραμμή)=αυτοέλεγχος. Αλγόριθμος μάθησης: SARSA. Ρυθμός μάθησης: 0.3. Έψιλον: 0.1 Πίνακας αμοιβών S=-55, P=-48, R=-2, T=10. Μέθοδος τροποποίησης πίνακα: Τροποποίηση ολόκληρου του πίνακα αμοιβών.*

Τελευταία γραφική παράσταση του κεφαλαίου, το Σχήμα 4.20ε, όπου και πάλι αυτοέλεγχος και συνειδητότητα φαίνονται να κατευθύνονται σε αντίθετες κατευθύνσεις και επιβεβαιώνουν την αντίστροφη σχέση που έχουν. Η αντιστρόφως ανάλογη σχέση τους αποκαλύπτεται από το γεγονός ότι η διαφορά R και T μειώνεται (άρα η συνειδητότητα μειώνεται) και λόγω αυτής της μείωσης το ποσοστό του συμβιβασμού αυξάνεται (άρα ο αυτοέλεγχος αυξάνεται).

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Εργασία

5.1 Σύνοψη και Συμπεράσματα

Μέσα από αυτή την διπλωματική μελετήσαμε την έννοια του αυτοελέγχου του ανθρώπου αλλά και την έννοια της συνειδητότητας του. Ο αυτοέλεγχος μας, όπως είδαμε στο Κεφάλαιο 1, εξαρτάται από τα δύο μέρη του εγκεφάλου μας τα οποία συμβάλλουν στις αποφάσεις μας. Το άνω μέρος και το κάτω μέρος του εγκεφάλου μας «συγκρούονται» κάθε φορά που πρέπει να πράξουμε κάτι, αφού το άνω αντιπροσωπεύει την ορθολογιστική σκέψη ενώ το κάτω τον αυθορμητισμό. Στο κεφάλαιο 3 είδαμε πως μοντελοποιήσαμε τα μέρη του εγκεφάλου με σκοπό να μελετήσουμε τον αυτοέλεγχο στη συνέχεια. Τα δύο μέρη αντιπροσωπεύονται στην εργασία μας από δύο πράκτορες οι οποίοι συμμετέχουν σε ένα παίγνιο ολικού αθροίσματος, το Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου. Αυτό που διαφοροποιεί τους πράκτορες και τους κάνει να συμπεριφέρονται με τρόπο όμοιο των μερών του εγκεφάλου είναι ο παράγοντας έκπτωσης. Για το κάτω μέρος τους εγκεφάλου θέσαμε παράγοντα έκπτωσης 0.1, ενώ για το πάνω μέρος παράγοντα έκπτωσης 0.9. Επίσης οι πράκτορες εκπαιδεύονται με Ενισχυτική Μάθηση και λαμβάνουν αμοιβές-τιμωρίες μετά από κάθε τους κίνηση. Πρώτιστος σκοπός μας ήταν να μάθουν οι πράκτορες να συνεργάζονται, άρα τα δύο μέρη του εγκεφάλου να συμβιβάζονται σε μια μέση λύση και το άτομο να πράττει με αυτοέλεγχο. Όπως είδαμε και στα αποτελέσματα του Κεφαλαίου 4 η προσομοίωση του αυτοελέγχου δούλευε καλά με τον αλγόριθμο ενισχυτικής μάθησης Q-Learning αλλά όχι τόσο καλά με τον αλγόριθμο SARSA, αφού με τον πρώτο αλγόριθμο οι πράκτορες εκπαιδεύονταν εύκολα και ανέπτυσσαν μεγάλο ποσοστό αυτοελέγχου ενώ με τον δεύτερο αλγόριθμο όχι τόσο. Μετά από την πρόσθεση του hill climbing στο πρόγραμμα

και τη χρήση των τριών μεθόδων που είδαμε στο προηγούμενο κεφάλαιο (Κεφάλαιο 4), οι πράκτορες ανέπτυσσαν αυτοέλεγχο και με τον αλγόριθμο SARSA.

Το hill climbing όπως εξηγήσαμε και νωρίτερα (Κεφάλαιο 2) χρησιμοποιήθηκε για τροποποίηση του πίνακα αμοιβών που λαμβάνουν οι πράκτορες, έτσι ώστε να βρούμε τις κατάλληλες τιμές οι οποίες δίνουν το μεγαλύτερο ποσοστό αυτοελέγχου. Από τις τρεις μεθόδους (Μέθοδος με τυχαίους πίνακες αμοιβών, Μέθοδος με σταθερό P και R στον πίνακα αμοιβών, Μέθοδος τροποποίησης ολόκληρου του πίνακα αμοιβών) παρατηρήθηκε ότι καλύτερα αποτελέσματα για αυτοέλεγχο έχει η $3^{η}$ μέθοδος η οποία μας έδωσε μέχρι και ποσοστό αυτοελέγχου 70%.

Στη συνέχεια ενσωματώσαμε την έννοια της συνειδητότητας στο πρόγραμμά μας, μέσα από τις τιμές του πίνακα αμοιβών. Για τους λόγους που αναφέραμε στα προηγούμενα κεφάλαια η συνειδητότητα αντιπροσωπεύεται μέσα από τις διαφορές των τιμών T-S, T-R και P-S. Όσο πιο μεγάλη η διαφορά ανάμεσα σε αυτές τις τιμές, τόσο πιο μεγάλη η συνειδητότητα που έχουμε. Ο αλγόριθμος hill climbing έπαιρνε ως είσοδο αυτές τις τιμές και τις τροποποιούσε δίνοντας στους πράκτορες διαφορετικά ποσοστά συνειδητότητας έτσι ώστε να έχουμε αυτοέλεγχο. Από αυτή τη διαδικασία παρατηρήσαμε πως οι μεταβολές του πίνακα επηρεάζουν την έννοια του αυτοελέγχου. Τα πειράματα του Κεφαλαίου 4 έδειξαν πως τελικά, αυτοέλεγχος και συνειδητότητα συνδέονται με αντιστρόφως ανάλογη σχέση και τα αποτελέσματά μας συνάδουν με την αρχική μας υπόθεση, αλλά και επιβεβαιώνουν το άρθρο “The Essence of Conscious Conflict” (Morsella et al., 2009b).

Όσο το hill climbing μείωνε τη συνειδητότητα, μειώνοντας τη διαφορά ανάμεσα στις τιμές του πίνακα αμοιβών, τόσο πιο πολύ ώθηση είχαν οι πράκτορες στο να συνεργαστούν. Γι’ αυτό στις γραφικές παραστάσεις του Κεφαλαίου 4 βλέπουμε δύο καμπύλες (αυτοέλεγχος-συνειδητότητα) να πορεύονται προς αντίθετες κατευθύνσεις. Η συνειδητότητα μειώνεται, κάνοντας έτσι τον αυτοέλεγχο να αυξάνεται. Επίσης κάτι άλλο που παρατηρήθηκε ήταν ότι όσο πιο μεγάλες ήταν οι διαφορές ανάμεσα στις τιμές τόσο πιο πολύ αυτοέλεγχο είχαμε. Δηλαδή εάν αρχίζαμε το πρόγραμμα με μεγάλη συνειδητότητα και μεγαλύτερη σύγκρουση, τότε το hill climbing κατάφερνε να μειώσει όσο πιο πολύ μπορούσε τη διαφορά έτσι ώστε να αναπτυχθεί μεγάλος αυτοέλεγχος.

Ολοκληρώνοντας, και αφού καταλήξαμε στο συμπέρασμα ότι όντως συνειδητότητα και αυτοέλεγχος έχουν μια αντιστρόφως ανάλογη σχέση θεωρητικά και πειραματικά να πούμε πως λειτουργούν στην πραγματικότητα. Ο αυτοέλεγχος πηγάζει από την ανάγκη του ανθρώπου να μην παρασύρεται από τις παράλογες επιθυμίες του και να πράττει με γνώμονα το σωστό. Ακόμη όταν δεν ενεργούμε όπως πρέπει, λέμε ότι δεν έχουμε αρκετό αυτοέλεγχο ώστε να βάλουμε όρια στον εαυτό μας, όταν για παράδειγμα καπνίζουμε πολύ ή πίνουμε πολύ αλκοόλ. Ξέρουμε ποιο είναι το σωστό για μας αλλά και πάλι δεν το κάνουμε. Γι' αυτό και ο αυτοέλεγχος θεωρείται από πολλούς προβληματική συμπεριφορά. Το γεγονός ότι γνωρίζουμε ποιο είναι το σωστό, υποδηλώνει την ύπαρξη συνειδητότητας. Αντιλαμβανόμαστε πώς θα έπρεπε να συμπεριφερθούμε αλλά συνειδητά πράττουμε το αντίθετο. Λόγω αυτού λέμε ότι μειωμένα ποσοστά αυτοελέγχου, συνδέονται με μεγάλα ποσοστά συνειδητότητας και το αντίστροφό. Δηλαδή, σε περιπτώσεις που ο άνθρωπος έχει ανεπτυγμένο αυτοέλεγχο, αποφασίζει πώς πρέπει να ενεργήσει, χωρίς να έχει απαραίτητα μεγάλη συνειδητότητα.

5.2 Μελλοντική Εργασία

Τελειώνοντας τη διπλωματική αυτή, ναι μεν αποδείξαμε ως ένα σημείο τη σύνδεση αυτοελέγχου και συνειδητότητας, το σίγουρο είναι όμως ότι πάντα θα υπάρχουν περιθώρια βελτίωσης και αλλαγών. Μια εισήγηση η οποία θα μπορούσε να εφαρμοστεί και να γίνει έρευνα στα νέα αποτελέσματα που θα αποφέρει είναι να χρησιμοποιηθεί άλλος αλγόριθμος στη θέση του hill climbing. Το hill climbing αποτελεί ένα από τους δημοφιλέστερους αλγορίθμους ευρετικής αναζήτησης, υστερεί όμως λόγω του ότι δεν είναι σίγουρο ότι πάντα θα καταλήγει σε ολικό μέγιστο. Ο αλγόριθμος που χρησιμοποιήσαμε, φέρει πάντα και τον κίνδυνο να παγιδευτούμε σε τοπικό μέγιστο και όχι ολικό μέγιστο όπως στοχεύουμε.

Έτσι εάν θέλουμε να κτίσουμε ένα πρόγραμμα το οποίο θα λειτουργεί χωρίς τον εγγενή κίνδυνο παγίδευσης που αναφέραμε, μπορούμε να αντικαταστήσουμε τον αλγόριθμο αναρρίχησης λόφων με γενετικούς αλγορίθμους. Οι γενετικοί αλγόριθμοι μοιάζουν κατά πολύ με την αναρρίχηση λόφων αφού ανήκει στον τομέα της βέλτιστης επίλυσης προβλημάτων όπως το hill climbing. Με την αντικατάσταση αυτή ο αλγόριθμος μας δε

θα παγιδεύεται εύκολα σε τοπικές βέλτιστες λύσεις και θα έχουμε πιο σίγουρα αποτελέσματα.

Αναφορές

Banfield, G. (2006). *Simulation of Self-Control through Precommitment Behaviour in an Evolutionary System.* Birkbeck: PhD Thesis, University of London.

Bjork, J. M.; Knutson, B.; Fong, G. W.; Daniel, M., Caggiano; Shannon, M., Bennett; Daniel, W., Hommer. (2004). Incentive-elicited brain activation in adolescents: similarities and differences from young adults. *Journal of Neuroscience, 24*(8), 1793-1802.

Christodoulou, C., Banfield, G. and Cleanthous, A. (2010). Self-control with spiking and non-spiking neural networks playing games. Journal of Physiology - Paris, 104, 108-117.

Cleanthous, A. (2010). In search of self-control through computational modelling of internal conflict. Cyprus: PhD Thesis, University of Cyprus.

Dennett, D. (1991). *Consciousness Explained.* Boston: Little, Brown & Co.

E, M., CC, B., & SC, K. (2010). Cognitive and neural components of the phenomenology of agency. *Neurocase: The Neural Basis of Cognition, 17*(3), 209-230.

Charles W. Eriksen, Derek W. Schultz (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception and Psychophysics, 25*(4), 249–263.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9*(8), 1057-1063.

Ezequiel Morsella, Lilian E. Wilson, Christopher C. Berger, Mikaela Honhongva, Adam Gazzaley, and John A. Bargh (2009a). Subjective aspects of cognitive control at different stages of processing. *Attention, Perception & Psychophysics, 71*(8), 1807-1824.

Morsella E, Gray JR, Krieger SC, Bargh JA. (2009b). The Essence of Conscious Conflict: Subjective Effects of Sustaining Incompatible Intentions. *Emotion, 9*(5), 717-728.

Rachlin H. (1995). Self-control: beyond commitment. *Behavioral and Brain Sciences, 18*(1), 139-140.

Rachlin H. (2000). *The Science of Self-Control.* Cambirdge, MA: Harvard University Press.

Rappoport, A. and Chammah, A.M. (1965). Prisoner's Dilemma: a Study in Conflict and Cooperation (Ann Arbor, MI: University of Michigan Press.

Stroop JR. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643-662.

Sutherland, N.S. (1989). *Macmillan Dictionary of Psychology.* London: Macmillan.

Sutton, R.S., (1988). Learning to predict by the method of temporal differences. Machine Learning 3, 9-44

Sutton, R.S., (1996). Generalization in reinforcement learning: Successful examples using sparse coding. *Advances in Neural Information Processing*, 1038-1044.

Sutton, R.S., Barto, A.G., (1998). Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Tokic, M. (2010). Adaptive ε-greedy Exploration in Reinforcement. *KI'10 Proceedings of the 33rd annual German conference on Advances in artificial intelligence* (203-210). Heidelberg : Springer-Verlag Berlin.

Watkins, C. J. C. H., Dayan, P. (1992). Q-learning. *Machine Learning 8*, 279-292.

Watkins; C. J. C. H. . (1989). Learning from delayed rewards. Ph. D. Dissertation. King's College, University of Cambridge, UK.

Ρενέ, Ν. (1629). Κανόνες για την καθοδήγηση του πνεύματος (Regulae ad directionem ingenii-ημιτελές). μετάφραση-σχόλια: Γ. Δαρδιώτης· εισαγωγή: Ν. Αυγελής, Θεσσαλονίκη: Εγνατία 1974 (επανέκδοση).

Παράρτημα Α

Κλάση Player.java

```
public class Player {

   private double learning_rate;
   private double epsilon;
   private double discount;
   private double[][] Q_table;
   private double[] payoff_matrix;
   private int current_action;
   private int next_action;

   public Player(){
      this.Q_table = new double[4][2];
      this.payoff_matrix = new double[4];
   }

   public void setDiscount(double value){
      this.discount = value;
   }

   public double getDiscount(){
      return this.discount;
   }

   public void setEpsilon(double value){
      this.epsilon = value;
   }

   public double getEpsilon(){
      return this.epsilon;
   }

   public void setLearning_rate(double value){
      this.learning_rate = value;
   }

   public double getLearning_rate(){
      return this.learning_rate;
   }

   public void setQ_table(int row, int column, double value){
      this.Q_table[row][column] = value;
   }
```

```
  public double getQ_table(int row, int column){
    return this.Q_table[row][column];
  }

  public void setPayoff_matrix(double value1, double value2, double value3, double value4){
    payoff_matrix[0] = value1;
    payoff_matrix[1] = value2;
    payoff_matrix[2] = value3;
    payoff_matrix[3] = value4;
  }

  public double getPayoff_matrix(int row){
    return this.payoff_matrix[row];
  }

  public void setCurrent_action(int action){
    this.current_action = action;
  }

  public int getCurrent_action(){
    return this.current_action ;
  }

  public void setNext_action(int action){
    this.next_action = action;
  }

  public int getNext_action(){
    return this.next_action ;
  }

  public int choose_action(int state){
    int action;
    double random = Math.random();

    if(random<epsilon){     //explore
      action = (int) (Math.random() * 2);    // values: 0(cooperate) or 1(defect)
    }

    else {    //exploit
      // find best known action
      if (this.getQ_table(state, 0) > this.getQ_table(state, 1))
        action = 0;
      else if(this.getQ_table(state, 0) < this.getQ_table(state, 1))
        action = 1;
      else
        action = (int) (Math.random() * 2);    // values: 0(cooperate) or 1(defect)
```

```
        }

        return  action;
    }

    public void update_Q(int current_state, int next_state){
        double Q;

        // Calculate Q and set value in player's Q table
        Q = ((1 - this.learning_rate) * this.getQ_table(current_state,
this.getCurrent_action())) + (this.learning_rate *
(this.getPayoff_matrix(current_state) + (this.getDiscount() *
this.getQ_table(next_state, this.getNext_action()))));

        this.setQ_table(current_state, this.getCurrent_action(), Q);
    }

}
```

Παράρτημα Β

Κλάση MainProgram.java

```
import java.io.*;
import java.util.Random;

/**
 * Created by annageorgiou on 20/01/15.
 */
public class MainProgram {

   public static Player lower;
   public static Player higher;

   public static double[][][] states_results;
   public static double[][][] payoff_results;
   public static double[][] avgstates_afterHill;

   public static int current_state, next_state; // Takes values 0-3

   public static int get_state(int action_lower, int action_higher) {
      int next_state;

      if (action_lower == 0 && action_higher == 0)
         next_state = 0;
      else if (action_lower == 0 && action_higher == 1)
         next_state = 2;
      else if (action_lower == 1 && action_higher == 0)
         next_state = 1;
      else
         next_state = 3;

      return next_state;
   }


   public static void update_states_results(int episode, int state, int trial) {
      states_results[episode][trial][0] = states_results[episode - 1][trial][0];
      states_results[episode][trial][1] = states_results[episode - 1][trial][1];
      states_results[episode][trial][2] = states_results[episode - 1][trial][2];
      states_results[episode][trial][3] = states_results[episode - 1][trial][3];
      states_results[episode][trial][state]++;
   }
```

```
   public static void update_payoff_results(int episode, double lower_payoff, double higher_payoff, int trial) {

      payoff_results[episode][trial][0] = payoff_results[episode - 1][trial][0];
      payoff_results[episode][trial][1] = payoff_results[episode - 1][trial][1];

      payoff_results[episode][trial][0] += lower_payoff;
      payoff_results[episode][trial][1] += higher_payoff;

   }


   public static void statistics(int numberOfTrials, int numberOfEpisodes) {

      int e, t, s, p;
      double sumAll_states;

      for (e = 0; e < numberOfEpisodes; e++) {
         for (t = 0; t < numberOfTrials; t++) {   // find sum of states visits from each trial of current episode and save in last column
            for (s = 0; s < 4; s++) {     // states
               states_results[e][numberOfTrials][s] += states_results[e][t][s];
            }

            for (p = 0; p < 2; p++) {     // payoffs
               payoff_results[e][numberOfTrials][p] += payoff_results[e][t][p];
            }
         }

         sumAll_states = 0;

         for (s = 0; s < 4; s++) {     // sum of state fields in last column (#trials+1)
            sumAll_states += states_results[e][numberOfTrials][s];
         }

         for (s = 0; s < 4; s++) {     // normalise
            states_results[e][numberOfTrials][s] /= sumAll_states;
         }
      }
   }

   static double gaussian(double mu, double sigma) {

      Random randomno = new Random();
      double p = 1.0, p1 = 0, p2;

      while (p >= 1.0) {
         p1 = 2 * randomno.nextDouble() - 1;
         p2 = 2 * randomno.nextDouble() - 1;
```

```
        p = p1 * p1 + p2 * p2;
    }

    return mu + sigma * p1 * Math.sqrt(-2.0 * Math.log(p) / p);
}

public static double HCNewPayoffs(double[] SPRT, int payoff_value) {

    double newpayoff = SPRT[payoff_value];
    double P, R, S, T;
    T = SPRT[3];
    R = SPRT[2];
    P = SPRT[1];
    S = SPRT[0];

double temp;
    double rand_prob;
    do {
        //generate a random value using gaussian distribution
        rand_prob = gaussian(0, 1);

        temp = SPRT[payoff_value] + rand_prob;
    }while (temp<-50 || temp>50);

    if (payoff_value == 0) {
        if ((T > R) && (R > P) && (P > temp) && (2 * R > T + temp)) {
            SPRT[payoff_value] = temp;
        }
    } else if (payoff_value == 1) {
        if ((T > R) && (R > temp) && (temp > S) && (2 * R > T + temp)) {
            SPRT[payoff_value] = temp;
        }
    } else if (payoff_value == 2) {
        if ((T > temp) && (temp > P) && (P > S) && (2 * temp > T + S)) {
            SPRT[payoff_value] = temp;
        }
    } else if (payoff_value == 3) {
        if ((temp > R) && (R > P) && (P > S) && (2 * R > temp + S)) {
            SPRT[payoff_value] = temp;
        }
    }

    T = SPRT[3];
    R = SPRT[2];
    P = SPRT[1];
    S = SPRT[0];

    //if the new value is within the rules : T>R>P>S and 2R>T+S keep and return
    if ((T > R) && (R > P) && (P > S) && (2 * R > T + S)) {
```

```
            newpayoff = SPRT[payoff_value];
        }
        return newpayoff;
    }

    public static void setPlayers(double[] SPRT_old) {

        lower = new Player();
        lower.setDiscount(0.1); //do not change
        lower.setLearning_rate(0.3);
        lower.setEpsilon(0.1);

lower.setPayoff_matrix(SPRT_old[2],SPRT_old[3],SPRT_old[0],SPRT_old[1]);

        higher = new Player();
        higher.setDiscount(0.9); //do not change
        higher.setLearning_rate(0.3);
        higher.setEpsilon(0.1);

higher.setPayoff_matrix(SPRT_old[2],SPRT_old[0],SPRT_old[3],SPRT_old[1]);

    }

    public static void SarsaLearning(int episodes_before, int t) {

        int i;

        int initial_state = (int) (Math.random() * 4);

        current_state = initial_state;

        states_results[0][t][current_state]++;

        boolean is_first = true;

        lower.setNext_action(lower.choose_action(current_state));
        higher.setNext_action(higher.choose_action(current_state));

        for (i = 1; i < episodes_before; i++) {
            lower.setCurrent_action(lower.getNext_action());
            higher.setCurrent_action(higher.getNext_action());

            next_state=get_state(lower.getCurrent_action(),
higher.getCurrent_action());

            lower.setNext_action(lower.choose_action(next_state));
            higher.setNext_action(higher.choose_action(next_state));
```

```
        if (!is_first) {
          lower.update_Q(current_state, next_state);
          higher.update_Q(current_state, next_state);
        }

        current_state = next_state;

        update_states_results(i, current_state, t);
        update_payoff_results(i, lower.getPayoff_matrix(current_state),
higher.getPayoff_matrix(current_state), t);

        is_first = false;
      }
    }

    public static double[] randomPayoffs() {

      double[] p_values = new double[4];
      double S = 0, P = 0, R = 0, T = 0;
      double lowLimit = -60.0;
      double highLimit = 60;
      Random randomno = new Random();

      while (((T <= R) || (R <= P) || (P <= S) && (2 * R <= T + S))) {

        p_values[3] = (highLimit - lowLimit) * randomno.nextDouble() +
lowLimit;
        highLimit = p_values[3];
        for (int i = 2; i >= 0; i--) {
          p_values[i] = (highLimit - lowLimit) * randomno.nextDouble() +
lowLimit;
        }
        T = p_values[3];
        R = p_values[2];
        P = p_values[1];
        S = p_values[0];
      }
      return p_values;
    }

    public static void main(String[] args) throws IOException {
      int episodes_before = 1000, episodes_after = 0;
      int trials = 50;
      int executions = 500;
      int runs=10;
      double[] SPRT_old2 = new double[4];
      double[] SPRT_old;
      double[] SPRT_new;
      double [][] bestRates = new double[episodes_before][4];
```

```
double [][] STdifference = new double[executions][4];
double CC_rate_new, CC_rate_old, prob=0.6;

states_results = new double[episodes_before][trials + 1][4];
payoff_results = new double[episodes_before][trials + 1][2];
avgstates_afterHill = new double[episodes_before][4];

PrintWriter writer2 = new PrintWriter("avgstates_afterHill.txt");
PrintWriter writer4 = new PrintWriter("conflicts.txt");

for (int r = 0; r < runs; r++) {
   SPRT_old = new double[]{-55, -48, -2, 10};


   SPRT_new = new double[]{-55, -48, -2, 10};


   CC_rate_new = CC_rate_old  = 0;

   states_results = new double[episodes_before][trials + 1][4];
   payoff_results = new double[episodes_before][trials + 1][2];

   //random new payoffs – remove if 1st method is used to change SPRT array
  /* if (r > 0) {
      SPRT_old = SPRT_new = randomPayoffs();
      CC_rate_new = CC_rate_old = CC_new = CC_old = 0;

      states_results = new double[episodes_before][trials + 1][4];
      payoff_results = new double[episodes_before][trials + 1][2];
   }*/

   for (int ex = 0; ex < executions; ) {

      for (int t = 0; t < trials; t++) {

         if (ex == 0) {
            setPlayers(SPRT_old);
         }
         else {
            setPlayers(SPRT_new);
         }

         SarsaLearning(episodes_before, t);

      }
```

```
            statistics(trials, episodes_before);
            STdifference[ex][0] += SPRT_old[3] - SPRT_old[0];
            STdifference[ex][1] += SPRT_old[1] - SPRT_old[0];
            STdifference[ex][2] += SPRT_old[3] - SPRT_old[2];

            if (ex == 0) {

                //last CC rate
                CC_rate_old = states_results[episodes_before - 1][trials][0];
            } else {
                //last execution CC rate
                CC_rate_new = states_results[episodes_before - 1][trials][0];

                if (CC_rate_new > CC_rate_old) {

                    SPRT_old[0] = SPRT_new[0];
                    SPRT_old[1] = SPRT_new[1];
                    SPRT_old[2] = SPRT_new[2];
                    SPRT_old[3] = SPRT_new[3];
                    CC_rate_old = CC_rate_new;

                    for (int i=0; i<episodes_before; i++){
                        System.arraycopy(states_results[i][trials], 0, bestRates[i], 0, 4);
                    }

                } else {
                    SPRT_new[0] = SPRT_old[0];
                    SPRT_new[1] = SPRT_old[1];
                    SPRT_new[2] = SPRT_old[2];
                    SPRT_new[3] = SPRT_old[3];
                }
            }

            STdifference[ex][3] += CC_rate_old;

            SPRT_old2[0] = SPRT_old[0];
            SPRT_old2[1] = SPRT_old[1];
            SPRT_old2[2] = SPRT_old[2];
            SPRT_old2[3] = SPRT_old[3];


            for (int j = 0; j < 4; j++) {
//remove comments if 2nd method is used to change the SPRT method
              /* if (j == 0 || j == 3) { */
                //change payoff values by a probability
                double rand_prob = Math.random();
                if (rand_prob < prob) {
                    SPRT_new[j] = HCNewPayoffs(SPRT_old2, j);
                }
```

```
            /* } */
               }
            ex++;
         }


         for (int i = 0; i < episodes_before; i++) {
            for (int j = 0; j < 4; j++) {
               avgstates_afterHill[i][j] += bestRates[i][j];
            }
         }

         System.out.println("FINAL PAYOFF VALUES: " + SPRT_old[0] + " " +
SPRT_old[1] + " " + SPRT_old[2] + " " + SPRT_old[3]);
      }

      for (int i = 0; i < episodes_before; i++) {
         for (int j = 0; j < 4; j++) {
            avgstates_afterHill[i][j] = avgstates_afterHill[i][j] / runs;
            writer2.print(avgstates_afterHill[i][j]+" ");
         }
         writer2.println();
      }

      writer4.println("S-T CC_rate S-P CC_rate R-T CC_rate");
      for (int i=0; i<executions; i++){
         for (int j=0; j<3; j++) {
            writer4.print(STdifference[i][j] / runs + " "+STdifference[i][3] / runs +
" ");
         }
         writer4.println();
      }

      writer2.close();
      writer4.close();
   }
}
```