ύ σΧύΓη&Κι ΓοΡΧηη πΓ η

ύοπξ&ΧάσΓμΚηάοπΚ

&ι Κι ΚπΧμ&νοποΛυσ Κά σΓμυσμ Γ Κά σΓμυσξύοποΛΓη&υσ

XOCRDMADuFVHi XHpCκ NWDVDMρύ AYr CGρTMAGRTκ

ηr DNEOpMρ

mΈHti (δ4ïO« a1ψδΣ/δ

χΦnΣΣΓ3mΓ(«41ïΦ3

aΈïξ HζΈΦf ΣδΙmδΣOδHΦΈΦnHΦf ΣΈδΣ/δ

l NHΦ3.; ..;

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Εμπιστοσύνη & Εξαπάτηση στα Πολυπρακτορικά Συστήματα

Διπλωματική Εργασία

Ρούσσης Δημήτριος

Επιβλέπουσα: Δασκαλοπούλου Ασπασία

Βόλος 2020

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Trust & Deception in Multi-Agent Systems

Diploma Thesis

Roussis Dimitrios

Supervisor: Daskalopoulou Aspasia

Volos 2020

Some make the world believe that they believe what they do not believe;

others, in greater number, make themselves believe it.

- Michel de Montaigne

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω την κυρία Δασκαλοπούλου πρώτον, γιατί μου κίνησε το ενδιαφέρον και με έφερε κοντά σε ένα αντικείμενο με απίστευτες διαστάσεις που δεν είχα φανταστεί πως μπορεί να με ταξιδέψει, δεύτερον, γιατί τα μαθήματα της ήταν από τα πιο διδακτικά και ταυτόχρονα από τα πιο ευχάριστα του προγράμματος και τρίτον, για την εμπιστοσύνη που μου έδειξε και την υπομονή που έκανε τα χρόνια που πέρασαν από όταν αποφασίσαμε το θέμα της εργασίας μέχρι την τελική της παράδοση.

Την εργασία αυτή, την αφιερώνω, στην οικογένεια μου με ένα τεράστιο ευχαριστώ για όλα. Δεν θα μπορούσα όμως ιδιαίτερα να μην ευχαριστήσω τρία άτομα του στενού μου περιβάλλοντος. Το πρώτο για την σωστή παρέμβαση που μου άλλαξε τη ζωή, το δεύτερο γιατί κάθε μέρα, καθόλη τη διάρκεια των σπουδών μου, ήταν πάντα εκεί και το τρίτο γιατί έκανε την υπέρβαση και μου έδειξε πως τίποτα δεν μπορεί να σταθεί εμπόδιο στους στόχους και στα όνειρα μου.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου Κατερίνα και Γιώργο για την βοήθεια τους όσον αφορά στην επιμέλεια του κειμένου.

ξύΧξΡξσΚΫΥΚπυηΚΫύΧνΓμ ΥΚι ΩμΚηΫΥΧοσ&οποΛΤ ηΐμ ΓΫύσΧξι &Γμυσ
ΥΓμ Γυι &υσ

úr TΈΗ«1ΓΤΈ/ψΥtiΣΓ4tiΥΣfΥΤΈΤï- Υ4Φf YN(ΦfΈΤ1/ΈΥΤf(δ4ïΟ- Υ.ïΟδïti(φ4tiΥw
.ΓΗ- Ytiΐ1Γ4φN4ïΤΈδ1ΦnΣδïΈΗti(δ4ïΟ«Τ1ψδΣ/δwΩδΜ-3ΩδïΐΗδΤΗΤΟ41ΦΥïΟφδ1dT/δΩδï
ΈΓψδ/ΦïΩΟ-.ïΟΤ3ΈΦfδΥδΈ4ndΜΤΟδΥ«Ω41ΦΕΦï«ΜΤΟδΥΣ4δΈΗδ/Σïδδf4«3Ω4Γ3ΤΤ1ψδΣ/δ3w
δΈΦ4ΤΗΤ/δΈΦΟΗΤïΣ4ïΟφΈΕ1ΦΝΥΈ1ΦΣtiΈïΟ«3Ω(ΦfΤ1ψδΣ/δ3wΤΥΈΕ1ΦΣξφΗΗΤïΟφΜΤ(Φ1z«3
.ïΟδï-(δ4δ.ïδΥΦΤ4ïΟ«3Ωï.ïΦΟ4ΓΣ/δ3wΈ1ΦΣtiΈïΟΝ4Γ4δ3ΟδïΈ1ΦΣtiΈïΟ-Υ.T.Φ(ζΥtiΥ
41/4tiΥw.TYΈΤ1ïζdTïζ1ψδέTïΣzΦ1ζ3Ω41/4tiΥψïδΩ4δΩΦΕΦ/δΩδΈδï4T/4δïΩφ.Tïδ4tiΥ
.Γ(ïΦf1ψ-Υé.ïΟδïΦndtiΥΟδïΩTΥΤ/Υδï΄Ε1ΦΝΥ(Τ1ïΟ«3Ω«ΦΗïΟ«3ΩδΥ4ïψ1δz«3wΩΦΕΓψζ3ΩΤ
ΈΦfΩd1ΓΣï(ΦΕΦï«ΜΤΟδΥΈΤ1ïΦ1/ΚΦΥ4δïΩΣ4ï3Ωξïξ΄ΗïΦψ1δzïΟζ3ΩδΥδzΦ1ζ3ΟδïΩ(ΝΥΦΥΟδï
ΈΗΓ1ΦnΥΩ4Φf3ΟδΥΝΥΤ3Ω4Γ3ΩΤΈïΣ4Γ(ΦΥïΟ«3ΩΈδ1φΜΤΣΓ3Ε,δΩΣΓ(Τ/δΩΝΕΦfΩζdtï
d1ΓΣï(ΦΕΦï«ΣΤïΩï.ζΤ3wΩΟΤ/(ΤΥΦwΩδ1dT/δΩ«έΟδïΩΕΓψζ3ΩφΗΗtiΥΩΣfψψ1δzζtiΥwΩδΥδzζ1ΦΥ4δïΩ
Tf.ïφΟ1ï4δΩΣ4ΦΩΟΤ/(ΤΥΦΩ(Τ4ΓΥΟδ4φΗΗΓΗΓΈδ1δΈΦ(Έ«ΟδïΩΓΩΣdT4ïΟ«ΩδΥδzΦ1φ
ΈΤ1ïΗδ(ξφΥΤ4δïΩΣ4ΦΩ4(«(δΩ4tiΥΩξïξ΄ΗïΦψ1δzïΟ-ΥΩδΥδzΦ1-Υ(ΤΩΈΗ«1ΓΩΈΤ1ïψ1δz«Ε
iΥδΗδ(ξφΥtiΩΈΗ«1ti3wΩδ4Φ(ïΟφΩΟδïΩΕ1ΦΣtiΈïΟφwΩΝΗΤ3Ω4ï3ΩΥΦ(ïΟζ3ΩΟδïΩïΦ1ΟΓ4ïΟζ3ΩΣfΥζΈΤïΤ3
ΈΦfΩnΥδ4δïΩΥδΩΕ1ΦΟnc ΦfΥΩΣ4ΓΥΩΈΤ1/Ε4tiΣΓΩΟδ4φΩ4ΓΥΩΦΕΦ/δΩδΈΦ.TïdΜΤ/wΩïδd1ΦΥïΟφwΩΝ4ïΩ
ΓΩΤ1ψδΣ/δΩδf4«»Ω4(«(δΩ4ΓΩ3ΩΤΥΩ(ΦfΩδΥ«ΟΤïΩïΝ4ïΩΤ/Υδï΄Ε1ΦΝΥΗΦψΦΟΗΦΕ«3ί

k&[redacted]

ΓθΦ

kΥ

h(

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία έχει ως θέμα την εμπιστοσύνη και την εξαπάτηση στα πολυπρακτορικά συστήματα. Αρχικά γίνεται μια εισαγωγή στην έννοια της Τεχνητής Νοημοσύνης και στην στάση της Ευρωπαϊκής Ένωσης απέναντι σε αυτή (Κεφάλαιο 1). Στα τρία επόμενα κεφάλαια, αναλύονται: οι πράκτορες, τα πολυπρακτορικά συστήματα καθώς και η αλληλεπίδραση μεταξύ των πρακτόρων σε αυτά (Κεφάλαιο 2), η εμπιστοσύνη, η διαχείριση της και η εμπιστοσύνη σε ατομικό και συστημικό επίπεδο με ένα παράδειγμα για το καθένα (Κεφάλαιο 3) καθώς και η εξαπάτηση, η ανίχνευση της και η αυταπάτη (Κεφάλαιο 4). Η εργασία ολοκληρώνεται με τον επίλογος στον οποίο γίνεται μια σύντομη αναφορά στα προηγούμενα κεφάλαια.

ABSTRACT

The purpose of this thesis is to analyze the trust and deception in multi-agent systems. To begin with, an introduction to Artificial Intelligence is made and the views of the European Union when it comes to AI are described (Chapter 1). In the next three chapters the following are presented: agents, multi-agent systems and their interactions in them (Chapter 2), trust, trust management, as well as individual and system level trust with an example for each level (Chapter 3), deception, deception detection and self-deception (Chapter 4). The thesis ends with the conclusion where a brief reference to the rest of the chapters is made.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η επιστήμη της Τεχνητής Νοημοσύνης είναι πολύ νέα. Οι πρώτες εργασίες σχετικά με το πεδίο αυτό ξεκίνησαν κατά τη διάρκεια του Β' Παγκόσμιου Πολέμου ενώ το όνομα "Τεχνητή Νοημοσύνη" εμφανίστηκε μόλις το 1956. Η επιστήμη αυτή έκανε την εμφάνιση της ως αποτέλεσμα μιας προσπάθειας χιλιάδων ετών, να αντιληφθούμε και να κατανοήσουμε τον τρόπο με τον οποίο σκεφτόμαστε. Πως δηλαδή μια "χούφτα" κυττάρων αντιλαμβάνεται, καταλαβαίνει, προβλέπει και εκμεταλλεύεται έναν κόσμο πολύ μεγαλύτερο σε μέγεθος και πολύ πολυπλοκότερο σε λειτουργία από αυτή. [1]

Το πεδίο της Τεχνητής Νοημοσύνης καταλαμβάνει τμήματα από όλες τις επιστήμες αλλά και τις τέχνες, αποτελώντας έτσι ένα από τα ευρύτερα επιστημονικά πεδία με πολλές δυνατότητες έρευνας, ανάπτυξης και εφαρμογής.

Τι είναι όμως Τεχνητή Νοημοσύνη;[1] Πως μπορεί να οριστεί μια έννοια όπως η νοημοσύνη η οποία για πολλούς έχει να κάνει με την "ψυχή" και για άλλους με τη χημεία; Πως μπορεί να οριστεί ως νοημοσύνη κάτι το οποίο προγραμματίζεται; Ιστορικά για τον όρο Τεχνητή Νοημοσύνη έχουν δοθεί πολλαπλοί ορισμοί οι οποίοι διαφοροποιούνται ανάλογα με την προσέγγιση στην οποία εστιάζουν. Ακολουθούν μερικά παραδείγματα τέτοιων ορισμών:

1. "Η συναρπαστική νέα προσπάθεια να κάνουμε τους υπολογιστές να σκέπτονται … μηχανές με νοημοσύνη με την πλήρη και κυριολεκτική έννοια" [3]
2. "Η μελέτη των νοητικών ικανοτήτων με τη χρήση υπολογιστικών μοντέλων" [4]
3. "Η τέχνη της δημιουργίας μηχανών που πραγματοποιούν λειτουργίες οι οποίες απαιτούν νοημοσύνη όταν πραγματοποιούνται από ανθρώπους." [5]
4. "Η τεχνητή νοημοσύνη ασχολείται με την ευφυή συμπεριφορά των τεχνουργημάτων" [6]

---

[1] Στο ερώτημα "Τι είναι Τεχνητή Νοημοσύνη;" οι ερευνητές του χώρου δίνουν πολλές διαφορετικές απαντήσεις, φαινόμενο που δεν απαντά σε άλλους επιστημονικούς χώρους, όπως η Φυσική, η Χημεία, η Ιατρική κ.ά. Ωστόσο, όλοι φαίνεται να συμφωνούν πως η Τεχνητή Νοημοσύνη είναι επιστήμη και όχι απλώς ένας κλάδος της τεχνολογίας λογισμικού. Κατά τον Winston [2], διευθυντή του εργαστηρίου ΤΝ του Πανεπιστημίου MIT, πρωταρχικός σκοπός της Τεχνητής Νοημοσύνης είναι "να κάνει τις μηχανές πιο έξυπνες". Σε αυτό συμφωνούν οι περισσότεροι από τους ερευνητές που αντιμετωπίζουν την Τεχνητή Νοημοσύνη ως αναζήτηση μεθόδων οι οποίες θα κάνουν τους ηλεκτρονικούς υπολογιστές πιο έξυπνους και, συνεπώς, πιο χρήσιμους από όσο είναι σήμερα.

Οι ορισμοί αυτοί επικεντρώνονται στη σκέψη και τη συμπεριφορά και αξιολογούν την επιτυχία με βάση τον άνθρωπό ή την ορθολογικότητα, όπου μια οντότητα ενεργεί ορθολογικά όταν ενεργεί με το βέλτιστο τρόπο βάσει όσων γνωρίζει.

Η Τεχνητή Νοημοσύνη αναπτύσσεται με ταχείς ρυθμούς και μπορεί να ωφελήσει ευρύ φάσμα τομέων. Ανάμεσα σε αυτούς η ιατροφαρμακευτική περίθαλψη (π.χ. επιτρέποντας πιο ακριβείς διαγνώσεις, διευκολύνοντας την καλύτερη πρόληψη των ασθενειών), η αύξηση της αποτελεσματικότητας της γεωργίας, ο μετριασμός της κλιματικής αλλαγής και η προσαρμογή σε αυτήν, η διαχείριση των χρηματοοικονομικών κινδύνων, η αύξηση της αποτελεσματικότητας των συστημάτων παραγωγής μέσω της προληπτικής συντήρησης, ο εντοπισμός περιπτώσεων απάτης και απειλών για την κυβερνοασφάλεια προσφέροντας βοήθεια στις αρχές επιβολής του νόμου για να καταπολεμούν αποτελεσματικότερα την εγκληματικότητα, η ενίσχυση της ασφάλειας των πολιτών και η συμβολή με πολλούς άλλους τρόπους που μόλις τώρα αρχίζουμε σταδιακά να φανταζόμαστε. Ταυτόχρονα όμως η τεχνητή νοημοσύνη συνεπάγεται ορισμένους δυνητικούς κινδύνους, όπως οι αδιαφανείς διαδικασίες λήψης αποφάσεων, οι έμφυλες ή άλλες διακρίσεις, η εισβολή στην προσωπική ζωή ή η χρήση της τεχνολογίας αυτής για εγκληματικούς σκοπούς. Τέλος, η Τεχνητή Νοημοσύνη συνεπάγεται επίσης νέες προκλήσεις για το μέλλον της εργασίας και εγείρει νομικά ζητήματα.

Η αξιόπιστη Τεχνητή Νοημοσύνη θα πρέπει να τηρεί όλες τις νομοθετικές και κανονιστικές διατάξεις, καθώς και μια σειρά βασικών απαιτήσεων [7]:

- Ανθρώπινη παρέμβαση και εποπτεία: τα συστήματα Τεχνητής Νοημοσύνης θα πρέπει να διευκολύνουν την προαγωγή ισότιμων κοινωνιών, υποστηρίζοντας την ανθρώπινη παρέμβαση και τα θεμελιώδη δικαιώματα και να μη μειώνουν, περιορίζουν ή παραπλανούν την ανθρώπινη αυτονομία.
- Στιβαρότητα και ασφάλεια: για την επίτευξη αξιόπιστης Τεχνητής Νοημοσύνης απαιτούνται αλγόριθμοι επαρκώς ασφαλείς, αξιόπιστοι και στιβαροί ώστε να μπορούν να αντιμετωπίζουν σφάλματα ή ασυνέπειες σε όλες τις φάσεις του κύκλου ζωής των συστημάτων.
- Ιδιωτικότητα και διακυβέρνηση των δεδομένων: οι πολίτες θα πρέπει να έχουν απόλυτο έλεγχο των προσωπικών τους δεδομένων, ενώ τα δεδομένα που τους αφορούν δεν θα χρησιμοποιούνται για να τους βλάψουν ή για να εισάγουν διακρίσεις εις βάρος τους.

- Διαφάνεια: θα πρέπει να εξασφαλίζεται η ιχνηλάτηση των συστημάτων Τεχνητής Νοημοσύνης.
- Διαφορετικότητα, απαγόρευση των διακρίσεων και δικαιοσύνη: τα συστήματα Τεχνητής Νοημοσύνης θα πρέπει να λαμβάνουν υπόψη όλο το φάσμα των ανθρώπινων ικανοτήτων, δεξιοτήτων και αναγκών και να διασφαλίζουν την προσβασιμότητα.
- Κοινωνική και περιβαλλοντική ευημερία: τα συστήματα Τεχνητής Νοημοσύνης θα πρέπει να χρησιμοποιούνται για την ενίσχυση των θετικών κοινωνικών αλλαγών και την ενίσχυση της βιωσιμότητας και της οικολογικής ευθύνης.
- Λογοδοσία: θα πρέπει να δημιουργηθούν μηχανισμοί μέσω των οποίων θα διασφαλίζεται η υπευθυνότητα και η λογοδοσία για τα συστήματα Τεχνητής Νοημοσύνης και τα αποτελέσματα τους.

Σε ένα περιβάλλον έντονου παγκόσμιου ανταγωνισμού, η Ευρώπη επιθυμεί να βρίσκεται στην πρώτη γραμμή των εξελίξεων[2] στον τομέα της Τεχνητής Νοημοσύνης. Η Ευρωπαϊκή Επιτροπή, για να τονώσει την ανταγωνιστικότητα της, ενισχύει τη συνεργασία στον τομέα της Τεχνητής Νοημοσύνης σε ολόκληρη την Ευρωπαϊκή Ένωση. Για την διαχείριση των ευκαιριών και των προκλήσεων της Τεχνητής Νοημοσύνης, η Ευρωπαϊκή Ένωση ενεργεί ως ενιαίο σύνολο και καθορίζει με τον δικό της τρόπο, με γνώμονα τις ευρωπαϊκές αξίες, την προώθηση της ανάπτυξης και της χρήσης της Τεχνητής Νοημοσύνης.[3]

Διαφαίνεται η ανάγκη για μια παγκόσμια συναίνεση και σύγκλιση σχετικά με τα ηθικά ζητήματα, που ανακύπτουν στον τομέα των τεχνολογιών της Τεχνητής Νοημοσύνης. Σε αυτήν την προσπάθεια σημαντικό και καίριο ρόλο θα πρέπει να αναλάβουν οι διεθνείς οργανισμοί, ικανοί για την προώθηση της κοινής συνεργασίας και συνεννόησης μεταξύ

---

[2] Πολλά πρόσφατα επιτεύγματα στον τομέα της Τεχνητής Νοημοσύνης προέρχονται από Ευρωπαϊκά εργαστήρια. Περίπου το ένα τέταρτο του συνόλου των βιομηχανικών και επαγγελματικών υπηρεσιών ρομποτικής παράγονται από Ευρωπαϊκές εταιρίες. [8]

[3] Στις 19 Φεβρουαρίου 2020, η Ευρωπαϊκή Επιτροπή παρουσίασε τη Λευκή Βίβλο [9] της για την Τεχνητή Νοημοσύνη. Το νέο πλαίσιο προσδιορίζει τους κανόνες που θα επιτρέψουν στην Ευρώπη να αναπτύξει ψηφιακές τεχνολογίες αιχμής και θα ενισχύσει τις ικανότητες της στον τομέα της Κυβερνοασφάλειας. Με αυτό τον τρόπο η Ευρωπαϊκή Ένωση ευελπιστεί ότι οι ψηφιακές εφαρμογές θα είναι επωφελείς στους πολίτες και τις επιχειρήσεις. Εκτιμά ότι θα βελτιωθούν οι δημόσιες υπηρεσίες, οι παροχές υγείας και περίθαλψης, η ενεργειακή απόδοση, θα εξοικονομηθούν πόροι και θα προωθηθούν τα ευφυή συστήματα μεταφορών.

των κρατών. Ωστόσο, προς επίτευξη του παραπάνω στόχου, δεν θα πρέπει να θυσιαστεί η πολιτιστική και ηθική πολυφωνία[4].

[4] Ο αντιπρόεδρος της Επιτροπής και επίτροπος Ψηφιακής Ενιαίας Αγοράς Άντρους Άνσιπ δήλωσε τα εξής: "Η δεοντολογική διάσταση της Τεχνητής Νοημοσύνης δεν είναι πολυτέλεια ή κάτι πρόσθετο. Μόνο με την εμπιστοσύνη μπορεί η κοινωνία μας να επωφεληθεί πλήρως από την τεχνολογία. Η δεοντολογική Τεχνητή Νοημοσύνη αποτελεί πρόταση επωφελή για όλους, η οποία μπορεί να αποτελέσει ανταγωνιστικό πλεονέκτημα για την Ευρώπη, επειδή θα της επιτρέψει να αποκτήσει ηγετική θέση για μια ανθρωποκεντρική Τεχνητή Νοημοσύνη την οποία θα μπορούν να εμπιστεύονται οι πολίτες." [8]

ΚΕΦΑΛΑΙΟ 2

ΠΡΑΚΤΟΡΕΣ & ΠΟΛΥΠΡΑΚΤΟΡΙΚΑ ΣΥΣΤΗΜΑΤΑ

2.1 Πράκτορας

Η εμφάνιση των πρακτόρων και ειδικότερα των ευφυών πρακτόρων στον κλάδο της Τεχνητής Νοημοσύνης και γενικότερα στην περιοχή της Επιστήμης των Υπολογιστών υπόσχεται ριζικές αλλαγές στην επικοινωνία μεταξύ χρήστη και λογισμικού στο σημερινό διασυνδεδεμένο και δικτυωμένο ψηφιακό κόσμο. Πρόκειται δηλαδή για σύγχρονα συστήματα Τεχνητής Νοημοσύνης, στα οποία δυνητικά μπορούν να χρησιμοποιηθούν επιλεκτικά και σε συνδυασμό μέθοδοι αναπαράστασης γνώσης και επίλυσης προβλημάτων με τεχνολογίες. Ήδη, γίνεται αισθητή η παρουσία τους σε πληθώρα εφαρμογών, όπως είναι η αναζήτηση και το φιλτράρισμα των πληροφοριών στο διαδίκτυο, η παροχή έξυπνων υπηρεσιών βοήθειας και εξυπηρέτησης πελατών, καθώς και ο έλεγχος σωστής λειτουργίας μεγάλων εργοστασιακών μονάδων. [11]

Ένας επιπλέον ορισμός που ανταποκρίνεται στα χαρακτηριστικά τους είναι το μοντέλο Hayes-Roth [12], που δίνει έμφαση στη συλλογιστική. “Οι ευφυείς πράκτορες συνεχώς εκτελούν τρεις λειτουργίες: αντιλαμβάνονται τις δυναμικές συνθήκες του περιβάλλοντος, δρουν στο περιβάλλον, ώστε να το αλλάξουν, και συλλογίζονται, ώστε να ερμηνεύσουν αυτά που αντιλαμβάνονται, να λύσουν προβλήματα, να εξάγαγουν συμπεράσματα, για να καθορίσουν τη δράση τους.”

Βασικό κοινό χαρακτηριστικό των πρακτόρων είναι η αυτονομία τους. Ως αυτόνομες οντότητες ελέγχουν τις καταστάσεις και τις συμπεριφορές τους, χωρίς να κατευθύνονται εξωγενώς βήμα προς βήμα. Άλλα κοινά χαρακτηριστικά[5] που διαθέτουν είναι η κοινωνικότητα και η δυνατότητα συνεργασίας με άλλους πράκτορες:

---

[5] Το 1995 προτάθηκε από τους Wooldridge & Jennings [13] μια λίστα γνωρισμάτων που πρέπει να έχει ένα υπολογιστικό σύστημα για να θεωρείται ευφυής πράκτορας. Τα χαρακτηριστικά αυτά είναι:

- Αυτονομία: Ένας πράκτορας βρίσκεται σε μια κατάσταση και πρέπει να είναι σε θέση να παίρνει αποφάσεις χωρίς την άμεση παρέμβαση μιας τρίτης οντότητας.
- Αντιδραστικότητα: Ένας πράκτορας βρίσκεται σε ένα περιβάλλον και πρέπει να μπορεί να το αντιλαμβάνεται και να ανταποκρίνεται σε αυτό έγκαιρα.
- Ενεργητικότητα: Ένας πράκτορας πρέπει να κάνει ενέργειες, ώστε να ικανοποιήσει το σκοπό για τον οποίο δημιουργήθηκε.
- Κοινωνική Ικανότητα: Ένας πράκτορας πρέπει να μπορεί να επικοινωνήσει με άλλους πράκτορες - ή και ανθρώπους.

- Αυτονομία: Οι πράκτορες ενεργούν αυτόνομα χωρίς άμεση παρέμβαση από χρήστες ή άλλους πράκτορες, με πλήρη έλεγχο των πράξεων τους (αυτοέλεγχο) και της εσωτερικής τους κατάστασης.
- Κοινωνικότητα: Οι πράκτορες αλληλεπιδρούν με τους χρήστες για την επίτευξη των στόχων τους και με άλλους πράκτορες μέσω μιας κοινά κατανοητής γλώσσας. Έτσι επιτυγχάνεται επικοινωνία μεταξύ των πρακτόρων για την ολοκλήρωση των ανεξάρτητων στόχων του καθενός ξεχωριστά και ενός κοινού στόχου με συνεργασία μεταξύ τους.
- Ορθολογικότητα: Αφορά την υπόθεση ότι ένας πράκτορας θα κάνει πάντα το σωστό, δηλαδή θα δρα καταλλήλως για την εκπλήρωση των στόχων του και όχι με τρόπο ο οποίος αποτρέπει την επίτευξη τους.

Οι ευφυείς πράκτορες διαθέτουν επιπλέον χαρακτηριστικά που αφορούν το βαθμό νοημοσύνης που διαθέτουν, όπως:

- Αντιδραστικότητα: Αφορά τον τρόπο με τον οποίο αντιλαμβάνονται το περιβάλλον και ανταποκρίνονται σε τυχόν αλλαγές του εντός συγκεκριμένων χρονικών πλαισίων .
- Προνοητικότητα: Οι πράκτορες δεν ανταποκρίνονται απλώς στις αλλαγές του περιβάλλοντός τους, αλλά είναι ικανοί να συμπεριφερθούν κατάλληλα σε αυτές τις αλλαγές θέτοντας επιμέρους στόχους (δηλαδή, αναλαμβάνουν πρωτοβουλία).
- Γνώση: Συγκεντρωμένη γνώση σχετική με τον τρόπο με τον οποίο λειτουργεί το περιβάλλον στο οποίο δρα έκαστος πράκτορας, η οποία έχει κατάλληλα αναπαρατεθεί, για να υποστηρίζει τη λήψη αποφάσεων.
- Πεποιθήσεις: Αποτελούν την άποψη του πράκτορα για το περιβάλλον του μια δεδομένη χρονική στιγμή, η οποία άποψη ενδέχεται να είναι εσφαλμένη.
- Επιθυμίες: Αφορούν την κρίση του πράκτορα για τις μελλοντικές καταστάσεις του περιβάλλοντός του, όπως, για παράδειγμα, αν μια μελλοντική κατάσταση είναι επιθυμητή ή όχι. Παρά ταύτα, δεν εξετάζεται αν μία επιθυμία του πράκτορα είναι εφικτή ή συγκρούεται με κάποια άλλη.
- Προθέσεις: Οι προθέσεις είναι υποσύνολο των στόχων, τους οποίους ο πράκτορας προσπαθεί να επιτύχει τη δεδομένη χρονική στιγμή. Δεδομένου ότι δεν είναι δυνατή η

ταυτόχρονη επίτευξη όλων των στόχων, επιλέγεται ένα υποσύνολο τους, βάσει ορισμένων κριτηρίων ιεράρχησης.

- Υποχρεώσεις: Αφορούν την υποχρέωση του πράκτορα να υπακούει σε ένα σύνολο κανόνων και να δρα σε ένα γενικότερο πλαίσιο, ώστε να επιτύχει το σκοπό για τον οποίο σχεδιάστηκε.
- Προσαρμοστικότητα: Ο πράκτορας προσαρμόζεται στο περιβάλλον του (ικανότητα μάθησης).

Με άλλα λόγια ένας πράκτορας είναι μια ανεξάρτητη οντότητα που μπορεί να αποφασίσει τι να κάνει λαμβάνοντας υπόψιν του το περιβάλλον στο οποίο βρίσκεται και επικοινωνώντας με άλλες οντότητες, με σκοπό να ικανοποιήσει το σχεδιαστικό του στόχο.

Εκτός των χαρακτηριστικών που προαναφέρθηκαν υπάρχουν και επιπλέον χαρακτηριστικά που απαντούν σε συγκεκριμένες κατηγορίες πρακτόρων προσφέροντας μια "ανθρωπόμορφη αρχιτεκτονική" λογισμικού, που επιτρέπει:

- Κινητικότητα: Είναι η ικανότητα ενός πράκτορα να μετακινείται ελεύθερα σε ένα φυσικό χώρο ή σε ένα δίκτυο.
- Συνεργασία μεταξύ πρακτόρων που βασίζεται σε
  - Φιλαλήθεια: Οι πράκτορες δε δίνουν εσκεμμένα λάθος πληροφορίες.
  - Αγαθή προαίρεση: Ο κάθε πράκτορας προσπαθεί να επιτύχει τους δικούς του στόχους, οι οποίοι βρίσκονται σε αρμονία με τους στόχους των υπολοίπων πρακτόρων του συστήματος.

Τα χαρακτηριστικά των πρακτόρων μπορούν να ομαδοποιηθούν σε τρεις άξονες:

- χαρακτηριστικά που αφορούν γενικά τη συνεργασία των πρακτόρων με τις γύρω τους οντότητες
- χαρακτηριστικά που αφορούν την κινητικότητα τους
- χαρακτηριστικά που αφορούν το βαθμό νοημοσύνης τους

Σύμφωνα με την παραπάνω κατηγοριοποίηση, οι ευφυείς πράκτορες χωρίζονται αρχικά σε δυο βασικές κατηγορίες, τους βιολογικούς και τους τεχνητούς. Οι βιολογικοί πράκτορες χρησιμοποιούν τις αισθήσεις τους, για να αντιληφθούν το κόσμο γύρω, τις

γνώσεις τους, για να συνάγουν συμπεράσματα γι' αυτόν, και τα μέρη του σώματος τους, για να εκτελέσουν τις ενέργειες που προκύπτουν από τη συλλογιστική τους. Οι τεχνητοί πράκτορες λειτουργούν με παρόμοιο τρόπο και χωρίζονται σε δύο υποκατηγορίες, τους ρομποτικούς και τους υπολογιστικούς πράκτορες. Οι ρομποτικοί πράκτορες έχουν ως αισθητήρες και μηχανισμούς δράσης μηχανικά ή ηλεκτρονικά μέρη και δρουν στον πραγματικό κόσμο. Στους υπολογιστικούς πράκτορες ανήκουν οι πράκτορες λογισμικού, προγράμματα που λειτουργούν συνεχώς και αυτόνομα στο πλαίσιο ενός υπολογιστικού συστήματος. Οι πράκτορες λογισμικού μπορούν να χωριστούν, ανάλογα με το σκοπό τους, σε πράκτορες επιμέρους εργασιών, πράκτορες ψυχαγωγίας. ιούς και υπολογιστικούς πράκτορες[6] [11]. Μια διαφορετική κατηγοριοποίηση των ευφυών πρακτόρων λογισμικού - που ενσωματώνει τις βασικές αρχές στις οποίες στηρίζονται όλα τα ευφυή συστήματα - είναι σε απλούς αντανακλαστικούς πράκτορες[7], αντανακλαστικούς πράκτορες βασισμένους σε μοντέλα[8], πράκτορες βασισμένους στο στόχο[9] και πράκτορες βασισμένους στη χρησιμότητα[10]. [1]

Με τη σειρά τους, τα περιβάλλοντα μέσα στα οποία δρουν οι πράκτορες, σύμφωνα με τους Russel και Nerving [1], χωρίζονται ανάλογα με τις ιδιότητες τους στις εξής κατηγορίες:

- Πλήρως ή Μερικώς Παρατηρήσιμα: Ένα περιβάλλον είναι πλήρως παρατηρήσιμο όταν ένας πράκτορας μπορεί να αποκτήσει πλήρεις, ακριβείς και ενημερωμένες πληροφορίες για την κατάσταση του.
- Αιτιοκρατικά ή Στοχαστικά: Ένα περιβάλλον είναι αιτιοκρατικό όταν οποιαδήποτε ενέργεια σε αυτό έχει ένα μοναδικό εγγυημένο αποτέλεσμα, με άλλα λόγια δεν

---

[6] Στους υπολογιστικούς πράκτορες ανήκουν οι πράκτορες τεχνητής ζωής που, επίσης, βασίζονται στο λογισμικό, αλλά έχουν δικά τους ιδιαίτερα χαρακτηριστικά, με βασικότερο ότι λειτουργούν σε εικονικά περιβάλλοντα.

[7] Το πιο απλό είδος πράκτορα, επιλέγεις τις ενέργειες τους με βάση την τρέχουσα κατάσταση, αγνοώντας το ιστορικό αντιλήψεων

[8] Είναι οι πράκτορες που χρησιμοποιούν ένα μοντέλο, σύμφωνα με το οποίο διατηρούν μια εσωτερική κατάσταση βασισμένη στο ιστορικό των αντιλήψεων για το κομμάτι του κόσμου που δεν μπορούν να παρατηρήσουν.

[9] Οι πράκτορες που εκτός από την τρέχουσα κατάσταση έχουν και πληροφορίες για τον στόχο τους, με άλλα λόγια, την επιθυμητή κατάσταση του περιβάλλοντος.

[10] Είναι οι πράκτορες που έχουν τη δυνατότητα να προτιμούν μια κατάσταση του περιβάλλοντος σε σχέση με μια άλλη ανάλογα με τη χρησιμότητα της. Κάτι τέτοιο είναι δυνατό μέσω μιας συνάρτησης χρησιμότητας, που απεικονίζει μια κατάσταση του περιβάλλοντος σε ένα πραγματικό αριθμό ανάλογα με το βαθμό ικανοποίησης.

υπάρχει αβεβαιότητα για την κατάσταση που θα προκύψει μετά την εκτέλεση μιας ενέργειας.

- Στατικό ή Δυναμικό: Ένα περιβάλλον είναι στατικό όταν μπορεί να υποτεθεί ότι παραμένει αμετάβλητο εκτός αν κάποιος πράκτορας εκτελέσει μια ενέργεια. Σε ένα δυναμικό περιβάλλον υπάρχουν και άλλες διεργασίες σε λειτουργία και άρα αλλάζει με τρόπους που δεν μπορεί να ελέγξει ο πράκτορας.
- Διακριτό ή Συνεχές: Ένα περιβάλλον είναι διακριτό όταν υπάρχει σταθερός και πεπερασμένος αριθμός ενεργειών και αντιλήψεων μέσα σε αυτό.

2.2 Πολυπρακτορικά Συστήματα

Η αυξανόμενη ανάγκη για νέα αυτόνομα υπολογιστικά συστήματα τα οποία θα είναι συνεχώς διασυνδεδεμένα, σε συνεργασία με τις τάσεις της Τεχνητής Νοημοσύνης δημιουργούν την έννοια της Κατανεμημένης Τεχνητής Νοημοσύνης.

Πολλές εφαρμογές λογισμικού είναι ανοιχτά κατανεμημένα συστήματα των οποίων τα συστατικά είναι αποκεντρωμένα, αλλάζουν συνεχώς και εξαπλώνονται σε όλο το δίκτυο. Για παράδειγμα, δίκτυα peer-to-peer, web services, σημασιολογικός ιστός, κοινωνικό δίκτυο, συστήματα σύστασης στο ηλεκτρονικό επιχειρείν αποτελούν τέτοια συστήματα. Αυτά τα συστήματα μπορούν να μοντελοποιηθούν ως ανοιχτά κατανεμημένα πολυπρακτορικά συστήματα στα οποία συχνά αλληλεπιδρούν μεταξύ τους αυτόνομοι πράκτορες σύμφωνα με ορισμένους μηχανισμούς επικοινωνίας και πρωτόκολλα. [15]

Τα πολυπρακτορικά συστήματα ή αλλιώς τα συστήματα πολλαπλών πρακτόρων, ανήκουν στην Κατανεμημένη Τεχνητή Νοημοσύνη και είναι στην ουσία κοινωνίες πρακτόρων, τα μέλη των οποίων επικοινωνούν και αλληλεπιδρούν μεταξύ τους μέσω δικτύου. Οι πράκτορες αυτοί πρέπει να έχουν τα χαρακτηριστικά που αναφέρθηκαν στην προηγούμενη ενότητα καθώς στα πλαίσια ενός πολυπρακτορικού συστήματος θα πρέπει να διαπραγματευτούν και συμφωνήσουν μεταξύ τους αφού στη γενικότερη περίπτωση καθένας από αυτούς θα εξυπηρετεί τα δικά του διαφορετικά συμφέροντα.

Ένα πολυπρακτορικό σύστημα αποτελείται από ένα σύνολο από πράκτορες που δρουν μαζί, για να επιλύσουν ένα πρόβλημα. Ένα τέτοιο σύστημα στοχεύει στη διασύνδεση και λειτουργία ήδη υπαρχόντων συστημάτων, καθώς και στην επίλυση προβλημάτων που πρώτον, είναι πέρα των δυνατοτήτων και της γνώσης ενός μόνο πράκτορα και δεύτερον, είναι από τη φύση τους κατανεμημένα. Τα πολυπρακτορικά συστήματα αποτελούν

βασικό τομέα της Κατανεμημένης Τεχνητής Νοημοσύνης από πλευράς χαλαρής θεώρησης των πρακτόρων, όπου η σχετική γνώση είναι κατανεμημένη σε διακριτές πηγές, όπως για παράδειγμα η υπάρχουσα εμπειρία στα επιμέρους γραφεία ενός οργανισμού. Παραδείγματα πολυπρακτορικών συστημάτων μπορεί κάποιος να συναντήσει σε διαδικτυακές εφαρμογές παροχής πληροφοριών σε σύνθετα διαδραστικά περιβάλλοντα.

Επομένως, η αλληλεπίδραση των πρακτόρων πραγματοποιείται σε ένα πολυπρακτορικό περιβάλλον. Οι πράκτορες σε ένα τέτοιο περιβάλλον ενδέχεται να έχουν αντικρουόμενους στόχους, όπως για παράδειγμα πράκτορες κρατήσεων θέσεων σε θέατρα, όταν έχουν ως στόχο την κράτηση των τελευταίων περιορισμένων θέσεων μιας θεατρικής παράστασης, ή μπορεί να έχουν αντίστοιχα κοινούς στόχους, όπως για παράδειγμα πράκτορες που συνεργάζονται, για να υποδείξουν στο χρήστη συμφέροντα αεροπορικά εισιτήρια για έναν προορισμό και ταυτόχρονα να του προτείνουν ξενοδοχεία για να μείνει και εταιρείες ενοικίασης αυτοκινήτου. Το πιο χαρακτηριστικό παράδειγμα πολυπρακτορικού περιβάλλοντος είναι το Διαδίκτυο[11]. Πρόκειται για ένα ανοικτό κατανεμημένο περιβάλλον συνεχώς εξελισσόμενο με αλλαγές στο μέγεθος του και στα δομικά του στοιχεία, μεταξύ των οποίων είναι και οι πράκτορες.

Στα χαρακτηριστικά των πολυπρακτορικών συστημάτων περιλαμβάνονται και τα ακόλουθα [11]:

- Κανένας πράκτορας δεν έχει πλήρη πληροφορία γιατί τα δεδομένα είναι κατανεμημένα
- Δεν υπάρχει κεντρικός έλεγχος στο σύστημα
- Οι υπολογισμοί γίνονται με ασύγχρονο τρόπο

Είναι επόμενο, τα χαρακτηριστικά ενός πολυπρακτορικού συστήματος να επιδρούν παράλληλα στον σχεδιασμό και τη λειτουργία των πρακτόρων που το αποτελούν.

Από τα παραπάνω συμπεραίνεται ότι είναι δυο τα βασικά χαρακτηριστικά των πολυπρακτορικών συστημάτων που τα διαχωρίζουν από τα συνηθισμένα συστήματα, η αυτονομία των πρακτόρων (η οποία όπως αναφέρθηκε αποτελεί και κύριο χαρακτηριστικό τους) και το ότι οι αλληλεπιδράσεις μέσα σε αυτά είναι μεταξύ ιδιοτελών οντοτήτων.

[11] Το Διαδίκτυο, με βάση τις κατηγορίες των περιβαλλόντων που δόθηκαν νωρίτερα, είναι ένα περιβάλλον δυναμικό και μερικώς παρατηρήσιμο. [14]

2.3 Αλληλεπίδραση μεταξύ των πρακτόρων

Η αρμονική συνύπαρξη των μελών ενός πολυπρακτορικού συστήματος θέτει τα θεμέλια για την αποδοτική και σταθερή λειτουργία του. Το πρόβλημα του πως αποφασίζουν οι πράκτορες με ποιόν και πότε θα αλληλεπιδράσουν έχει γίνει το κύριο θέμα έρευνας τα τελευταία χρόνια. Αυτό σημαίνει ότι πρέπει να διαπραγματευτούν με βαθμό αβεβαιότητας τη λήψη αποφάσεων κατά τη διάρκεια της αλληλεπίδρασης τους.

Από τα παραπάνω γίνεται εμφανές το ότι οι αλληλεπιδράσεις μεταξύ των πρακτόρων βρίσκονται στην καρδιά των πολυπρακτορικών συστημάτων με αποτέλεσμα να έχουν αναπτυχθεί διάφορα μοντέλα αλληλεπιδράσεων όπως αυτό του συντονισμού[12], της συνεργασίας[13] και των διαπραγματεύσεων[14]. Η διαπραγμάτευση στηρίζει τις προσπάθειες τόσο για συνεργασία όσο και για συντονισμό μεταξύ τόσο τεχνητών αλλά και ανθρώπινων πρακτόρων και είναι απαραίτητη τόσο όταν οι πράκτορες έχουν τα δικά τους συμφέροντα όσο και όταν είναι συνεργατικοί.

Όπως σημειώθηκε παραπάνω, τα πολυπρακτορικά συστήματα χρησιμοποιούνται κυρίως ως μέρος μεγάλων ανοιχτών κατανεμημένων συστημάτων. Τότε είναι που εμφανίζονται σημαντικές προκλήσεις που πρέπει να αντιμετωπιστούν. Πρώτον, όπως αναφέρθηκε ο κάθε πράκτορας έχει τα δικά του (και ενδεχομένως, διαφορετικά σε σχέση με τους υπόλοιπους πράκτορες στο σύστημα) συμφέροντα με αποτέλεσμα το πιο πιθανό σχέδιο ενεργειών να είναι αυτό που θα μεγιστοποιεί το δικό του συμφέρον. Δεύτερον, δεδομένου ότι το σύστημα είναι ανοιχτό, οι πράκτορες έχουν τη δυνατότητα να αποχωρούν και να εισέρχονται σε αυτό με διαφορετική ταυτότητα αποφεύγοντας έτσι πιθανές κυρώσεις για τις πράξεις τους όσο ήταν μέρος του συστήματος. Επιπλέον, ένα τέτοιο σύστημα θα περιέχει πράκτορες με διαφορετικά χαρακτηριστικά και ρόλους, οι

---

[12] Στα περισσότερα πολυπρακτορικά συστήματα, τα μέλη της κοινότητας έχουν εξειδίκευση στην επίλυση προβλημάτων που ενώ σχετίζεται είναι διακριτή μεταξύ τους, και η οποία συχνά πρέπει να συντονίζεται κατά την επίλυση προβλημάτων. Τέτοιες αλληλεπιδράσεις απαιτούνται λόγω των εξαρτήσεων μεταξύ των πράξεων των πρακτόρων, της ανάγκης αντιμετώπισης των γενικών περιορισμών και επειδή κανένας πράκτορας δεν έχει επαρκή ικανότητα, πόρους ή πληροφορίες για να λύσει ολόκληρο το πρόβλημα μόνος του. [16]

[13] Η συνεργασία στα πολυπρακτορικά συστήματα είναι κρίσιμη σε μια σειρά τομέων, συμπεριλαμβανομένων ομάδων διαστημικών σκαφών, μη επανδρωμένων αεροπορικών οχημάτων και πράκτορες για την υποστήριξη ανθρώπων [17]

[14] Ο πιο σημαντικός μηχανισμός για την διαχείριση εξαρτήσεων μεταξύ πρακτόρων κατά το run time είναι η διαπραγμάτευση. Η διαδικασία κατά την οποία μία ομάδα πρακτόρων κάνουν μια συμφωνία αποδεκτή από όλους για ένα ζήτημα. [18]

οποίοι ενδέχεται να παρέχουν τις ίδιες υπηρεσίες με διαφορετικά επίπεδα αποτελεσματικότητας. [19]

Απαραίτητη προϋπόθεση για τη λειτουργία ενός πολυπρακτορικού συστήματος, λόγω της πληθώρας των πρακτόρων και των αλληλεπιδράσεων μεταξύ τους, είναι η ύπαρξη πρωτοκόλλων επικοινωνίας και αλληλεπίδρασης. Χρειάζεται επιπλέον προσοχή στο ποια πρωτόκολλα οι σχεδιαστές τους θα υλοποιήσουν, καθώς το κάθε ένα από αυτά διέπεται από τους δικούς τους κανόνες που μπορεί να επιφέρουν διαφορετικά αποτελέσματα στις αλληλεπιδράσεις. Τα πρωτόκολλα επικοινωνίας ορίζουν μηχανισμούς μέσω των οποίων οι πράκτορες μεταδίδουν μεμονωμένα μηνύματα, ενώ τα πρωτόκολλα αλληλεπίδρασης κατευθύνουν την ανταλλαγή μίας σειράς μηνυμάτων μεταξύ των πρακτόρων, με άλλα λόγια πραγματοποιούν διάλογο. [20]

Όπως αναφέρθηκε, τα πρωτόκολλα επικοινωνίας καθορίζουν την επικοινωνία μεταξύ των πρακτόρων σε πρωτογενές επίπεδο. Οι πράκτορες, ιδιαίτερα αυτοί με εσωτερική κατάσταση, είναι οντότητες με ξεχωριστούς στόχους, επιθυμίες και πεποιθήσεις και απαιτούν πρωτόκολλα αλληλεπίδρασης που θα τους επιτρέψουν να λειτουργήσουν ως κοινωνία πρακτόρων. Στο πλαίσιο μιας τέτοιας κοινωνίας απαιτείται συντονισμός των πρακτόρων, να εμποδίζεται η άσχετη με το στόχο δραστηριότητα μέσω της αποφυγής καταστάσεων αδιεξόδου και ενεργούς αναμονής.

Τα πρωτόκολλα αλληλεπίδρασης προσδιορίζουν την τυποποίηση των μηνυμάτων που μπορούν να ανταλλάξουν οι πράκτορες μεταξύ τους. Ένα πρωτόκολλο αλληλεπίδρασης θα μπορούσε να ορίσει την ακόλουθη τυποποίηση των μηνυμάτων που ανταλλάσσουν δύο πράκτορες:

- πρόταση ενέργειας
- αποδοχή προτεινόμενης ενέργειας
- απόρριψη προτεινόμενης ενέργειας
- απόσυρση προτεινόμενης ενέργειας
- διαφωνία σε μία προτεινόμενη ενέργεια
- αντιπρόταση σε μία προτεινόμενη ενέργεια

Βασιζόμενοι σε αυτού του τύπου τα μηνύματα, μία τυπική επικοινωνία μεταξύ ενός πράκτορα x και ενός πράκτορα y, η οποία είναι παράλληλα ένα στιγμιότυπο του πρωτοκόλλου αλληλεπίδρασης για την διαπραγμάτευση, είναι η ακόλουθη:

1. Ο x προτείνει μια ενέργεια στον y
2. Ο y αναλύει την πρόταση και είτε
    a. Στέλνει μήνυμα αποδοχής στον x
    b. Στέλνει μήνυμα αντιπρότασης στον x
    c. Στέλνει μήνυμα διαφωνίας στον x
    d. Στέλνει μήνυμα απόρριψης στον x

Η επικοινωνία μεταξύ των πρακτόρων μπορεί να είναι σύγχρονη ή ασύγχρονη. Κατά τη σύγχρονη επικοινωνία, ο πράκτορας που θέτει μία ερώτηση είναι απαραίτητο να διακόψει τη λειτουργία του, μέχρι να πάρει μία απάντηση, ενώ, κατά την ασύγχρονη, η απάντηση μπορεί να έλθει οποιαδήποτε στιγμή μετά το χρόνο υποβολής της ερώτησης, χωρίς η λειτουργία του πράκτορα να διακόπτεται. Επιπρόσθετα, ένα άλλο θέμα που αφορά την επικοινωνία των πρακτόρων σε ένα σύστημα είναι ο βαθμός επικοινωνίας. Αυτός ορίζεται ως ο αριθμός των αποστολέων και των παραληπτών κατά την ανταλλαγή πληροφορίας. Ο βαθμός επικοινωνίας μπορεί να είναι, 1 προς 1, 1 προς N και N προς N. [11]

ΚΕΦΑΛΑΙΟ 3

ΕΜΠΙΣΤΟΣΥΝΗ & ΔΙΑΧΕΙΡΙΣΗ ΕΜΠΙΣΤΟΣΥΝΗΣ

3.1 Εμπιστοσύνη

Η εμπιστοσύνη αποτελεί αναπόσπαστο μέρος σε πολλές μορφές της ανθρώπινης αλληλεπίδρασης, επιτρέποντας στους ανθρώπους να ενεργούν υπό αβεβαιότητα και με κίνδυνο αρνητικών συνεπειών. Για παράδειγμα, η ανταλλαγή χρημάτων για μια υπηρεσία, η παροχή πρόσβασης στην ιδιοκτησία σας και η επιλογή μεταξύ αντικρουόμενων πηγών πληροφοριών, υποκρύπτουν κάποια μορφή εμπιστοσύνης. Στην επιστήμη των υπολογιστών, η εμπιστοσύνη είναι ένας ευρέως χρησιμοποιούμενος όρος του οποίου ο ορισμός διαφέρει μεταξύ των ερευνητών και των περιοχών εφαρμογής. [21]

Η εμπιστοσύνη αποτελεί μια σημαντική πτυχή τόσο της καθημερινής ζωής όσο και πολλών υπολογιστικών εφαρμογών για παρόμοιους λόγους. Στην πραγματικότητα, η εμπιστοσύνη[15] είναι ένα “άλμα πίστης” το οποίο είναι απαραίτητο να γίνει όποτε χρειαστεί να βασιστούμε σε τρίτους πράκτορες ή πληροφορίες. Αποφασίζουμε αν θα κάνουμε αυτό το άλμα πίστης βασιζόμενοι στην αξιολόγηση της αξιοπιστίας του πράκτορα ή των πληροφοριών. Σε γενικές γραμμές, όταν εμπιστευόμαστε ελλοχεύει ο κίνδυνος να στηριχθούμε σε αβέβαιες και πιθανώς απρόβλεπτες ενέργειες ή πληροφορίες. Μπορούμε να μειώσουμε έναν τέτοιο κίνδυνο και ένας τρόπος να επιτύχουμε αυτό το αποτέλεσμα είναι να μοιραστούμε τις εκτιμήσεις της εμπιστοσύνης και της αξιοπιστίας, μαζί με την προέλευσή τους, με τους υπόλοιπους, για να επιτρέψουμε την επαναχρησιμοποίησή τους και να αυξήσουμε την πιθανότητα σωστής εμπιστοσύνης χάρη στη διαθεσιμότητα αυτών των πληροφοριών. Επομένως, μια οντολογία για την αξιολόγηση της εμπιστοσύνης, ιδίως στη χρήση δεδομένων Web, μπορεί να υποδείξει τα βασικά στοιχεία που είναι απαραίτητα για τον καθορισμό της αξίας της εμπιστοσύνης. [22]

Η εμπιστοσύνη είναι ένα ευρέως διερευνημένο θέμα σε μια ποικιλία τομέων της επιστήμης των υπολογιστών. Συναντάται εκτενώς η χρήση της στον τομέα των ευφυών πρακτόρων. Τα περισσότερα παραδοσιακά στοιχεία για την αξιολόγηση της

[15] Στη σχέση εμπιστοσύνης εμπλέκονται δυο ρόλοι, ο πομπός/πηγή της πληροφορίας που είναι ο πράκτορας (ανθρώπινος ή τεχνητός) που “ζητά” και την εμπιστοσύνη και ο δέκτης που είναι ο πράκτορας που εμπιστεύεται.

εμπιστοσύνης στον φυσικό κόσμο, δεν είναι πια διαθέσιμα στα πολυπρακτορικά συστήματα. Κάθε φορά που μέσω των πρακτόρων οι άνθρωποι-χρήστες πρέπει να αλληλεπιδράσουν με εταίρους για τους οποίους δεν γνωρίζουν τίποτα, πρέπει παράλληλα να αντιμετωπίσουν την πρόκληση της λήψης αποφάσεων που ενέχουν κίνδυνο. Έτσι, η επιτυχία ενός πράκτορα μπορεί να εξαρτάται από την ικανότητα του να επιλέξει αξιόπιστους εταίρους. Προκύπτει, επομένως, ένα κρίσιμο ζήτημα, αναφορικά με το πως μπορεί ένας πράκτορας να εμπιστευτεί έναν άγνωστο εταίρο σε ένα ανοικτό και ως εκ τούτου άγνωστο και ενδεχομένως επικίνδυνο περιβάλλον. [23]

Σε αυτό το πλαίσιο, δεδομένου ότι οι πράκτορες, όπως και οι χρήστες τους, είναι πιθανόν να είναι ανέντιμοι, έχουν προταθεί διάφορα μοντέλα και μετρικές εμπιστοσύνης και φήμης. Αυτές οι προτάσεις εστιάζουν ουσιαστικά στην εκτίμηση του βαθμού της εμπιστοσύνης που μπορεί να επενδυθεί σε έναν συγκεκριμένο πράκτορα. [19] Ωστόσο, δεν υπάρχει ακόμη ένας ενιαίος και γενικά αποδεκτός ορισμός της εμπιστοσύνης στην ερευνητική κοινότητα.

Όπως για την Τεχνητή Νοημοσύνη, έτσι και για την εμπιστοσύνη έχουν δοθεί αρκετοί ορισμοί καθένας από τους οποίους δίνει έμφαση σε διαφορετικά σημεία της. Η εμπιστοσύνη έχει οριστεί με πολλούς διαφορετικούς τρόπους από τους ερευνητές από διάφορες οπτικές γωνίες. [15] Έχει αποτελέσει ένα ενεργό ερευνητικό θέμα σε διάφορους τομείς της επιστήμης των υπολογιστών, όπως η ασφάλεια και ο έλεγχος πρόσβασης στα δίκτυα υπολογιστών, η αξιοπιστία σε κατανεμημένα συστήματα, η θεωρία παιγνίων και τα πολυπρακτορικά συστήματα και οι πολιτικές για τη λήψη αποφάσεων σε περιβάλλον αβεβαιότητας. Από την πλευρά των υπολογιστών η εμπιστοσύνη ορίζεται ως η ποσοτικοποιημένη πεποίθηση από αυτόν που εμπιστεύεται σε σχέση με την ικανότητα, ειλικρίνεια, ασφάλεια και αξιοπιστία αυτού που εμπιστεύεται εντός ενός συγκεκριμένου πλαισίου.

Ο Gambetta [24] όρισε την εμπιστοσύνη ως την υποκειμενική πιθανότητα με την οποία ένας πράκτορας x αναμένει ότι ένας άλλος πράκτορας ή ομάδα πρακτόρων y θα εκτελέσει μια συγκεκριμένη ενέργεια από την οποία εξαρτάται η ευημερία του x. Με άλλα λόγια, ο x εμπιστεύεται τον y, όταν η πιθανότητα ο y να εκτελέσει μια πράξη ευεργετική ή το λιγότερο μη επιβλαβή για τον x είναι αρκετά μεγάλη.

Οι Grandison & Sloman [25] την όρισαν ως την ισχυρή πεποίθηση στην ικανότητα μιας οντότητας ότι μπορεί να ενεργήσει αξιόπιστα, με ασφάλεια και με ένα τρόπο που μπορεί να εξαρτηθεί από αυτήν, σε μια συγκεκριμένη περίπτωση.

Από τους παραπάνω ορισμούς είναι φανερό ότι η εμπιστοσύνη είναι χαρακτηριστικό σκεπτόμενων οντοτήτων - και άρα των πρακτόρων - καθώς και ότι η ίδια και η διαχείριση της είναι απαραίτητη σε περιβάλλοντα όπου δεν είναι εγγυημένη η ασφάλεια και η αξιοπιστία των συναλλαγών, καθώς και οι "καλές" προθέσεις των πρακτόρων.

3.2 Διαχείριση Επιστοσύνης

Οι Blaze, Feigenbaum & Lacy [26] όρισαν τη διαχείριση της εμπιστοσύνης ως μια ενοποιημένη προσέγγιση για τον προσδιορισμό και την ερμηνεία πολιτικών ασφαλείας, διαπιστευτηρίων και σχέσεων που επιτρέπουν την απ' ευθείας εξουσιοδότηση κρίσιμων ως προς την ασφάλεια ενεργειών.

Οι Josang, Kesey & Dimitrakos [27] την όρισαν ως τη δημιουργία συστημάτων και μεθόδων που επιτρέπουν στους εμπλεκόμενους να αξιολογούν και να παίρνουν αποφάσεις σχετικά με την αξιοπιστία πιθανών συναλλαγών που περιλαμβάνουν ρίσκο. Επιπλέον, επιτρέπουν σε αυτούς αλλά και στους διαχειριστές των συστημάτων αυτών να αυξάνουν και να παρουσιάζουν σωστά την αξιοπιστία τόσο των ιδίων όσο και των συστημάτων.

Όταν δύο πράκτορες αλληλεπιδρούν μεταξύ τους, ένας από αυτούς αναλαμβάνει το ρόλο του αιτούντος την υπηρεσία και ο άλλος ενεργεί ως πάροχος υπηρεσιών. Είναι σημαντικό να επισημάνουμε ότι μια σχέση εμπιστοσύνης μπορεί να περιλαμβάνει πολλαπλές διαστάσεις, ανάλογα με τη συγκεκριμένη προοπτική υπό την οποία αντιμετωπίζεται η αλληλεπίδραση μεταξύ αυτών των δύο παραγόντων. Μερικές συνήθεις διαστάσεις εμπιστοσύνης είναι οι εξής:

- Επάρκεια: Ένας αρμόδιος πράκτορας είναι σε θέση να διαμορφώνει σωστά και αποτελεσματικά τη διαμόρφωση των απαιτούμενων εργασιών.
- Ειλικρίνεια: Ένας ειλικρινής πράκτορας δείχνει μια αληθινή συμπεριφορά και δεν είναι παραπλανητικός.
- Ασφάλεια: Ένας ασφαλής αντιπρόσωπος διαχειρίζεται εμπιστευτικά τα προσωπικά δεδομένα και δεν επιτρέπει μη εξουσιοδοτημένη πρόσβαση σε αυτά.

• Αξιοπιστία: Ένας αξιόπιστος πράκτορας παρέχει αξιόπιστες υπηρεσίες.

Με άλλα λόγια, η ευθύνη μετράει τον βαθμό εμπιστοσύνης που μπορεί να δοθεί στις υπηρεσίες που παρέχονται από τον πράκτορα, συμπεριλαμβανομένης της αποτελεσματικότητας. Για καθεμία από τις πτυχές που εξετάστηκαν παραπάνω, η εμπιστοσύνη αυξάνει το φάσμα των αλληλεπιδράσεων μεταξύ δύο παραγόντων. [28]

Ανάλογα με που υπολογίζεται η εμπιστοσύνη/αξιοπιστία κάθε πράκτορα σε ένα πολυπρακτορικό σύστημα αυτή χωρίζεται σε "εμπιστοσύνη σε ατομικό επίπεδο" και σε "εμπιστοσύνη σε επίπεδο συστήματος", όπως αναλύουν οι Ramchurn, Huynh & Jennings. [19] Είναι σημαντικό να σημειωθεί πως οι παραπάνω προσεγγίσεις δεν αποκλείουν η μια την άλλη αλλά μπορούν να λειτουργήσουν συνδυαστικά, ώστε να καλύψουν η μια τυχόν αδυναμίες της άλλης και τελικά την διαχείριση της επιστοσύνης σε ένα πολυπρακτορικό σύστημα.

Η διαχείριση της εμπιστοσύνης βρίσκει εφαρμογή σε πολλές περιοχές στις οποίες κύριο ρόλο έχουν οι online συναλλαγές. Μερικές από αυτές είναι το ηλεκτρονικό εμπόριο, οι εφαρμογές διαμοιρασμού πόρων σε Peer-toPeer δίκτυα, οι φορητές συσκευές και η ασφάλεια κατανεμμημένων συστημάτων. [29]

3.3 Μοντέλα & Πρωτόκολλα Εμπιστοσύνης

Σε ατομικό επίπεδο συναντάμε μοντέλα εμπιστοσύνης, ενώ σε επίπεδο συστήματος πρωτόκολλα εμπιστοσύνης. Με τον όρο "μοντέλα εμπιστοσύνης" αναφερόμαστε στο πως αποτυπώνεται η δυνατότητα των πρακτόρων να μπορούν να αξιολογήσουν την αξιοπιστία και την ειλικρίνεια των ομοτίμων τους, ενώ αντίστοιχα με τον όρο "πρωτόκολλα εμπιστοσύνης" αναφερόμαστε στους μηχανισμούς ή/και τους κανόνες που διέπουν τις αλληλεπιδράσεις μεταξύ των πρακτόρων διαβεβαιώνοντας τους ότι θα έχουν κέρδος αν πραγματικά το αξίζουν με άλλα λόγια ότι κάποιος κακόβουλος πράκτορας δεν μπορεί να επηρεάσει τις ανταμοιβές των υπολοίπων πρακτόρων μέσα στο σύστημα.

Σε ατομικό επίπεδο τα μοντέλα εμπιστοσύνης χωρίζονται σε αυτά που βασίζονται στη μάθηση και τη φήμη[16] λαμβάνοντας υπόψη έμμεσα ή άμεσα το αποτέλεσμα

[16] Η φήμη είναι ένα μέτρο της εμπιστοσύνης που αντιλαμβάνεται ολόκληρη η κοινότητα σε σχέση με αυτόν που κρίνεται. Η φήμη αναλαμβάνει πολύ σημαντικό ρόλο όταν ένας πράκτορας δεν έχει επαρκή γνώση του αντιπάλου του.

παλαιότερων αλληλεπιδράσεων και σε αυτά που την κρίνουν συμπεριλαμβάνοντας κοινωνικο-γνωστικά[17] και πιο υποκειμενικά κριτήρια.

Τα μοντέλα που βασίζονται στη μάθηση υποθέτουν και άρα λειτουργούν πιο αποτελεσματικά όταν οι πράκτορες αλληλεπιδρούν πολλές φορές και όχι μόνο μια. Επιπλέον, στα μοντέλα αυτά, θεωρούμε ότι οι πράκτορες έχουν πιθανώς δόλιο σκοπό και άρα ενδέχεται να μην ικανοποιήσουν την εκάστοτε συμφωνία και αυτό γιατί το ενδεχόμενο να εξαπατήσουν τον άλλο πράκτορα εμπεριέχει καλύτερα για αυτούς αποτελέσματα. Έτσι, η εμπιστοσύνη είναι μια μετρική που αλλάζει και εξελίσσεται με βάση τη γνώση από προηγούμενες αλληλεπιδράσεις. Απαραίτητες προϋποθέσεις για το συνεχή υπολογισμό της εμπιστοσύνης είναι η δυνατότητα απεικόνισης της σε μετρήσιμη ποσότητα καθώς και η δυνατότητα αποθήκευσης του αποτελέσματος των προηγούμενων αλληλεπιδράσεων. Τα παραπάνω θα πρέπει κάθε πράκτορας να τα διατηρεί για καθέναν από τους πράκτορες που αλληλεπιδρά.

Τα μοντέλα που βασίζονται στη φήμη υπολογίζουν την εμπιστοσύνη με βάση τη γνώμη που έχει η κοινωνία πρακτόρων για τον εκάστοτε πράκτορα, με άλλα λόγια με βάση τη φήμη που αυτός έχει μέσα σε αυτή. Η πληροφορία για τη φερεγγυότητα ενός πράκτορα μέσα στην κοινωνία μπορεί να προέρχεται από ήδη υπάρχουσες σχέσεις μεταξύ μελών αυτής, από τη συσσώρευση πολλαπλών εκτιμήσεων για αυτόν από τα μέλη της κοινωνίας καθώς και από μετρικές που είναι αληθείς και πραγματικές. Τα μοντέλα αυτά δεν λαμβάνουν υπόψη τους ότι οι πράκτορες είναι εγωιστές και δεν θα μοιραστούν οποιαδήποτε πληροφορία αν δεν υπάρχει κάποιο όφελος από αυτό.

Τα μοντέλα που βασίζονται σε κοινωνικο-γνωστικούς παράγοντες, όπως αναφέρθηκε και νωρίτερα, υπολογίζουν την αξιοπιστία ενός πράκτορα λαμβάνοντας υπόψη και υποκειμενικά κριτήρια (περιβάλλον, χαρακτηρισικά αντιπάλου κτλ.), καθώς επιτρέπουν μια πιο περιεκτική ανάλυση των χαρακτηριστικών του άλλου πράκτορα. [30] Τέτοια κριτήρια μπορεί να είναι τα εργαλεία και οι ικανότητες του αντιπάλου οι οποίες μπορούν να κριθούν υποκειμενικά για το αν ο αντίπαλος μπορεί πράγματι να εκτελέσει/ ολοκληρώσει τη συμφωνηθείσα ενέργεια. Ειδικότερα, οι Castelfranchi & Falcone [31] επισημαίνουν την σημασία της γνωστικης άποψης της εμπιστοσύνης σε αντίθεση με την ποσοτική. Έτσι όταν ένας πρακτρορας x επιθυμεί να αναθέσει μια εργασία σε έναν πράκτορα y, αξιολογεί την εμπιστοσύνη που μπορεί να έχει σε αυτόν σκεπτόμενος και τις

[17] socio-cognitive based models

πεποιθήσεις που έχει (ο x) για τα κίνητρα του (y). Υποστηρίζουν ότι για την αξιολόγηση της εμπιστοσύνης του x προς τον y είναι ουσιώδεις οι:

- Πίστη στην επάρκεια: Η θετική αξιολόγηση του y από τον x ότι ο y μπορεί να ολοκληρώσει την εργασία που θα του αναθέσει, καθώς σε διαφορετική περίπτωση δεν έχει νόημα η αλληλεπίδραση.
- Πίστη στη θέληση: Ο x πιστεύει οτι ο y έχει αποφασίσει και σκοπεύει να ολοκληρώσει την εργασία που του ανατέθηκε. Αν ο x δεν πιστεύει στη θέληση του y, τότε ο y ενδέχεται να ψεύδεται για το αν θα ολοκληρώσει την εργασία και άρα η επιστοσύνη του x προς τον y μειώνεται.
- Πίστη στην επιμονή: Ο x πιστεύει οτι ο y είναι σταθερός στην πρόθεση του να κάνει αυτό που προτείνει. Στην περίπτωση που κάτι τέτοιο δεν ισχύει η αλληλεπίδραση εμπεριέχει ρίσκο και έτσι η επιστοσύνη του x προς τον y μειώνεται.
- Πίστη στα κίνητρα: Ο x πιστεύει οτι ο y έχει κίνητρα να βοηθήσει τον x και ότι τα κίνητρα αυτά θα είναι ισχυρότερα από κίνητρα που ενδέχεται να υπάρχουν για το αντίθετο. Τα κίνητρα αυτά ειναι ίδια με μακροπρόθεσμα οφέλη που θα έχει ο y αν ο x πετύχει τους στόχους του. Έτσι αν ο x πιστεύει ότι ο y έχει τα κίνητρα αυτά τότε τείνει να εμπιστεύεται τον y.

Σε επίπεδο συστήματος, τα πρωτόκολλα εμπιστοσύνης χωρίζονται περαιτέρω σε αυτά που ενθαρρύνουν αληθείς αλληλεπιδράσεις, αυτά που χτίζουν ένα σύστημα φήμης που καλλιεργεί αξιόπιστες συμπεριφορές και σε αυτά που αναπτύσσουν/χρησιμοποιούν μηχανισμούς ασφαλείας που εξασφαλίζουν ότι νέοι πράκτορες στο σύστημα είναι άξιοι εμπιστοσύνης.

Τα πρωτόκολλα που ενθαρρύνουν αληθείς αλληλεπιδράσεις, με άλλα λόγια την ανταλλαγή αληθών πληροφοριών μεταξύ των πρακτόρων που αλληλεπιδρούν, το πετυχαίνουν μέσω μηχανισμών και κανόνων που διέπουν τις αλληλεπιδράσεις και αποτρέπουν τους πράκτορες από το να ψεύδονται ή να υποθέτουν κατά τη διάρκεια τους καθώς πρέπει να ακολουθούν συγκεκριμένα βήματα ώστε να τις ολοκληρώσουν.

Σε αντίθεση με τα μοντέλα εμπιστοσύνης που για να λειτουργήσουν δεν λαμβάνουν υπόψη τους ότι οι πράκτορες δεν θα μοιράσουν πληροφορίες χωρίς όφελος, τα πρωτόκολλα που δημιουργούν ένα σύστημα φήμης, τη μοντελοποιούν σε επίπεδο συστήματος και οι αντίστοιχοι μηχανισμοί λειτουργούν είτε με συγκεντρωτικές είτε με

κατανεμημένες οντότητες. Έτσι τα αποτελέσματα των αλληλεπιδράσεων και η αξιολόγηση των συμμετεχόντων αποθηκεύονται στις οντότητες αυτές και δημοσιεύονται σε ολόκληρη την κοινωνία των πρακτόρων, ώστε να έχουν όλοι πρόσβαση στις μετρικές αυτές. Με τα πρωτόκολλα αυτά το σύστημα διαχειρίζεται τη συγκεντρωτική φήμη του κάθε πράκτορα, σε αντίθεση με τα αντίστοιχα μοντέλα στα οποία κάθε πράκτορας έχει την ευθύνη αυτή. Το αποτέλεσμα είναι το σύστημα να αποτρέπει κακόβουλες συμπεριφορές, όπως και ψευδή πληροφορίες στις αλληλεπιδράσεις μέσα σε αυτό.

Τα πρωτόκολλα που χρησιμοποιούν μηχανισμούς ασφαλείας βασίζονται στο ότι κάθε νέος πράκτορας στο σύστημα μπορεί να ταυτοποιηθεί. Το 2002 προτάθηκαν, από τους Poslad et al. [32], ένα πλήθος απαιτήσεων σχετικά με την ασφάλεια, που είναι απραίτητες ώστε οι πράκτορες να εμπιστευτουν ο ένας τον άλλον, όπως και τα μηνύματα[18] που ανταλλάσουν:

- Ταυτότητα: Η δυνατότητα ταυτοποίησης κάθε οντότητας μέσα στο σύστημα.
- Δικαιώματα Πρόσβασης: Η δυνατότητα να δίνονται σε κάθε πράκτορα στο σύστημα δικαιώματα ανάλογα με την ταυτότητα του.
- Ακεραιότητα Περιεχομένου: Η δυνατότητα να αναγνωρίζεται αν ένα μήνυμα έχει διαφοροποιηθεί από όταν στάλθηκε από τον αποστολέα.
- Απόρρητο Περιεχομένου: Η δυνατότητα να προσπελαύνουν τα μηνύματα που ανταλλάσσονται μόνο οι οντότητες για τις οποίες προορίζονται.

Η διασφάλιση τόσο της ακεραιότητας όσο και του απορρήτου του περιεχομένου μπορεί να επιτευχθεί είτε με σύγχρονους είτε με ασύγχρονους τρόπους. Ο παραδοσιακός τρόπος διασφάλισης της ασύγχρονης επικοινωνίας είναι να παρέχεται ένας ασφαλής “φάκελος” για καθένα από τα μηνύματα που στέλνονται, ενώ ένας τρόπος διασφάλισης της σύγχρονης επικοινωνίας συμπεριλαμβάνει την έννοια μιας ροής μηνυμάτων.

3.4 Παραδείγματα Μοντέλων & Πρωτοκόλλων

Από τα πρώτα μοντέλα εμπιστοσύνης (για τον υπολογισμό αυτής εντός του πράκτορα) είναι αυτό που προτάθηκε από το Marsh το 1994. Ο Marsh μοντελοποίσε την εμπιστοσύνη χωρίζοντας τη σε μικρότερους παράγοντες. Πρώτα όρισε το πόσο ο x εμπιστεύεται στον y ως $T_x(y)$ με τιμές στο διάστημα [-1, 1), όπου -1 είναι η παντελής

[18] Η επικοινωνία στα πολυπρακτορικά συστήματα βασίζεται κατα κύριο λόγο στην ανταλλαγή μηνυμάτων.

έλλειψη εμπιστοσύνης, 0 είναι η μη εμπιστοσύνη όταν ο x έχει καθόλου ή λιγες πληροφορίες για τον y ή είναι ουδέτερος, ενώ το 1 - η τυφλή εμπιστοσύνη - δεν συμπεριλαμβάνεται στις τιμές καθώς εννοεί ότι ο x δεν κρίνει καθόλου την εμπιστοσύνη του προς τον y, με άλλα λόγια δεν σκέφτεται, που είναι αντιθετο με το οτι η εμπιστοσύνη είναι χαρακτηριστικό σκεπτόμενων οντοτήτων. Στη συνέχεια εισήγαγε τη σημασία μιας κατάστασης α στον πράκτορα x ως $I_x(a)$ με τιμές στο διάστημα (-1, 1) και τη χρησιμότητα της κατάστασης στο x ως $U_x(a)$ και αυτή με τιμές στο διάστημα (-1, 1). Συνολικά όρισε την εμπιστοσύνη του x στον y σε μια κατάσταση α ως:

$$T_x(y, a) = T_x(y)U_x(a)I_x(a)$$

Ο Marsh θεωρεί ως τιμές για τα $I_x(a)$ και $U_x(a)$ προκύπτουν από το σύστημα/περιβάλλον ενώ η εμπιστοσύνη του x προς τον y, $T_x(y)$, προκύπτει ως μέσος όρος των παλαιότερων τιμών της εμπιστοσύνης του x στον y στην κατάσταση α, $T_x(y, a)$. Όταν υιοθετείται το μοντέλο του Marsh, θα πρέπει να αποφασίζεται αν ο πράκτορας θα συμπεριλαμβάνει σε αυτό το μέσο όρο όλες τις παλαιότερες αλληλεπιδράσεις με τον y ή μόνο αυτές που είναι σχετικές με την κατάσταση α καθώς η διαφοροποίηση αυτή θα έχει δραστικές επιπτώσεις στο αποτέλεσμα. [33]

Το πιο διαδεδομένο πρωτόκολλο στα πολυπρακτορικά συστήματα είναι αυτό των δημοπρασιών. Το πρωτόκολλο των δημοπρασιών χωρίζεται περαιτέρω ανάλογα με τους μηχανισμούς και τους κανόνες που τις διέπουν. Οι μονόπλευρες δημοπρασίες διακρίνονται σε τέσσερα βασικά πρωτόκολλα/μηχανισμούς:

- Αγγλική Δημοπρασία: Ο κάθε πλειοδότης μπορεί να αυξήσει την προσφορά του μέχρι κανένας άλλος πλειοδότης να μην επιθυμεί να αυξήσει τη δική του προσφορά. Ο πλειοδότης αυτός κερδίζει τη δημοπρασία.
- Ολλανδική Δημοπρασία: Ξεκινά με μια πολύ μεγάλη αρχική τιμή η οποία μειώνεται σταδιακά μέχρι ένας από τους πλειοδότες να προσφέρει τιμή κερδίζοντας τη δημοπρασία.
- Κλειστή Δημοπρασία Πρώτης Τιμής: Οι πράκτορες/χρήστες κάνουν την προσφορά τους χωρίς να ξέρουν τις προσφορές των υπολοίπων. Η μεγαλύτερη προσφορά κερδίζει τη δημοπρασία.

- Δημοπρασία Vickrey: Οι πράκτορες/χρήστες κάνουν την προσφορά τους χωρίς να ξέρουν τις προσφορές των υπολοίπων, όμως ο νικητής της δημοπρασίας - αυτός με την μεγαλύτερη προσφορά - πληρώνει την δεύτερη μεγαλύτερη προσφορά και όχι τη δική του.

Επομένως οι Αγγλικές και Ολλανδικές δημοπρασίες επιβάλλουν την ειλικρίνεια στον δημοπράτη καθώς ο πραγματικός νικητής και η προσφορά του δεν μπορούν να αλλοιωθούν αφού όλες οι προσφορές είναι δημόσιες. Επίσης οι τρεις πρώτοι μηχανισμοί δεν εγγυώνται ότι οι συμμετέχοντες θα φανερώσουν το πως αυτοί αξιολογούν πραγματικά το αντικείμενο της δημοπρασίας. Αντίθετα, η δημοπρασία Vickrey επιβάλλει την ειλικρίνεια στους πλειοδότες καθώς αν η προσφορά τους είναι μεγαλύτερη από την αξία του αντικειμένου, ο πράκτορας θα πληρώσει περισσότερο από αυτήν, ενώ στην περίπτωση που η προσφορά είναι μικρότερη ενδέχεται να χάσει τη δημοπρασία. Το μεγαλύτερο μειονέκτημα της είναι ότι ο δημοπράτης μπορεί να ψεύδεται αφού οι προσφορές, στο σύνολο τους, είναι γνωστές μόνο σε αυτόν. Έτσι, ο δημοπράτης μπορεί να εξαπατήσει τους συμμετέχοντες τόσο για το νικητή όσο και για το ύψος της δεύτερης προσφοράς. [19]

Για τον υπολογισμό της αξιοπιστίας ενός πράκτορα στις διαδικτυακές δημοπρασίες[19], χρησιμοποιείται είτε το σωρευτικό μοντέλο είτε η μεθοδολογία αναλογίας. Το σωρευτικό μοντέλο χρησιμοποιεί το άθροισμα των ανατροφοδοτήσεων των άλλων χρηστών που μπορούν να είναι είτε 0, 1 ή -1. Το μοντέλο αναλογίας αναφέρεται στην αναλογία των θετικών ανατροφοδοτήσεων προς το συνολικό αριθμό των ανατροφοδοτήσεων. Και τα δυο αυτά μοντέλα αγνοούν την πιθανότητα ύπαρξης κακόβουλων χρηστών και δεν λαμβάνουν υπόψην τους την αξιοπιστία του ατόμου που αξιολογεί τον άλλο πελάτη. Το σωρευτικό μοντέλο χρησιμοποιεί το άθροισμα των φημών που υπολογίζονται με την πάροδο του χρόνου και το ίδιο ισχύει και στην περίπτωση του μοντέλου αναλογίας με εξαίρεση το κομμάτι του αθροίσματος.

Στις διαδικτυακές δημοπρασίες, σημαντικό ρόλο έχουν η φήμη και η αξιοπιστία των συμμετεχόντων. Οι συμμετέχοντες δεν γνωρίζουν στοιχεία για τον πωλητή και άρα δεν έχουν άλλη επιλογή από το να εμπιστευτούν τους μηχανισμούς ανατροφοδότησης που παρέχονται. Έτσι, ο χρήστης με τις καλύτερες κριτικές (είτε με το μεγαλύτερο άθροισμα,

[19] Μια διαδικτυακή δημοπρασία ορίζεται ως μία εικονική αγορά που φιλοξενείται στο διαδίκτυο και ταιριάζει αγοραστές και πωλητές στον κόσμο χωρίς να λαμβάνει υπόψη τους φυσικούς περιορισμούς των παραδοσιακών δημοπρασιών όπως η γεωγραφική θέση, η φυσική παρουσία, ο χρόνος και ο χώρος.

κάνοντας χρήση του σωρευτικού μοντέλου είτε με την καλύτερη αναλογία ανατροφοδοτήσεων) θα επιλέγεται πιο συχνά από κάποιον με λιγότερες θετικές ή πολλές αρνητικές κριτικές. Οι τιμές αυτές είναι δημόσιες και αποθηκεύονται σε ένα κεντρικό σημείο ώστε να είναι προσβάσιμες από όλους τους συμμετέχοντες. Επεκτείνοντας τα παραπάνω, μπορεί να γίνει εισαγωγή βαρών[20] με σκοπό να δωθεί σημασία στη διαφορετική αντικειμενική αξία της εκάστοτε συναλλαγής. Έτσι, η φήμη και η αξιοπιστία των χρηστών μπορεί να υπολογιστεί εκ νέου. [34]

Όπως αναφέρθηκε νωρίτερα, τα μοντέλα και τα πρωτόκολλα εμπιστοσύνης μπορούν να λειτουργήσουν συνδυαστικά. Για παράδειγμα η εμπιστοσύνη που έχει ένας πράκτορας x προς ένα δημοπράτη σε μια δημοπρασία Vickrey, μπορεί να προκύπτει σαν αποτέλεσμα της δημόσιας πληροφορίας γι' αυτόν, στην οποία έχει πρόσβαση, και της "προσωπικής" του γνώμης (εμπιστοσύνη προς το δημοπράτη από παλιότερες αλληλεπιδράσεις, χρησιμότητα και σημασία της δημοπρασίας για τον ίδιο) από το μοντέλο του Marsh.

[20] Τιμών που ανήκουν στο κλειστό διάστημα [0,1]

ΚΕΦΑΛΑΙΟ 4

ΕΞΑΠΑΤΗΣΗ & ΑΥΤΑΠΑΤΗ

4.1 Εξαπάτηση

Στη φύση, η εξαπάτηση συνήθως χρησιμοποιείται είτε από αρπακτικά που μιμούνται ένα άλλο είδος που θεωρείται ακίνδυνο από το θήραμα τους είτε από το θήραμα που μιμείται ένα είδος, το οποίο δεν θα κυνηγούσαν οι θηρευτές τους. Η μίμηση ωφελεί τους θηρευτές επιτρέποντας τους να φτάσουν πολύ πιο κοντά στο θήραμα τους χωρίς να το ανησυχούν, ενώ το θήραμα επωφελείται με το να ξεφύγει από την προσοχή των θηρευτών του. Και οι δύο αυτές ικανότητες είναι σαφώς πλεονεκτικές και έχουν επιλεγεί μέσω της διαδικασίας της εξέλιξης. [49]

Οι άνθρωποι εξαπατούν ο ένας τον άλλον από την εμφάνιση του είδους τους. Η συμφωνία, η συνεργασία και η εμπιστοσύνη θα ήταν απλή αν δεν υπήρχε ποτέ η εξαπάτηση στις επικοινωνιακές αλληλεπιδράσεις τόσο μεταξύ των ανθρώπων όσο και μεταξύ των μηχανών. Οι μηχανές, εξαπατούν η μια την άλλη ή στην πραγματικότητα εξαπατούν τους ανθρώπους; Εάν και όταν το κάνουν, πως μπορεί αυτό να το εντοπιστεί; Οι μελλοντικές επιπτώσεις μιας μηχανής που θα έχει την ικανότητα να κατανοεί άλλα μυαλά (ανθρώπινα ή τεχνητά) και που έχει επίσης τους λόγους και τις προθέσεις να εξαπατήσει τους άλλους είναι σκοτεινές από ηθική σκοπιά. Ο εντοπισμός και η κατανόηση της ανέντιμης και ανήθικης συμπεριφοράς τέτοιων μηχανών είναι ζωτικής σημασίας για την τρέχουσα έρευνα στην Τεχνητή Νοημοσύνη. [48]

Για την έννοια της εξαπάτησης έχουν δοθεί πολλοί ορισμοί. Δύο από τους πιο κοινά αποδεκτούς είναι από τους Whaley και Burgoon. [49] Σύμφωνα με τον Whaley, η εξαπάτηση είναι “πληροφορία που έχει σχεδιαστεί για να χειραγωγεί τη συμπεριφορά των άλλων κάνοντας τους να αποδεχτούν μια ψευδή ή παραμορφωμένη παρουσίαση του περιβάλλοντος (φυσικού, κοινωνικού ή πολιτικού)”. Ο Burgoon ορίζει την εξαπάτηση ως “μια εσκεμμένη πράξη που διαπράττεται από έναν αποστολέα για να πείσει ένα παραλήπτη για το αντίθετο σε σχέση με αυτό που πιστεύει ο αποστολέας, ώστε ο δέκτης να βρίσκεται σε μειονεκτική θέση”. Είναι με άλλα λόγια, η πράξη της μετάδοσης ψευδών πληροφοριών στον αντίπαλο, με την πρόθεση ότι αυτές οι πληροφορίες θα δώσουν στον αντίπαλο μια ψευδή πεποίθηση. Η λανθασμένη πεποίθηση αποσκοπεί στο να προκαλέσει στο θύμα μια εξειδικευμένη παρανόηση που θα ωφελεί τον απατεώνα με

εξειδικευμένο τρόπο. Οι μελλοντικές αποφάσεις του θύματος θα βασίζονται σε ψευδείς πληροφορίες, επιτρέποντας σε σε αυτόν που εξαπατά να χρησιμοποιήσει την εξαπάτηση για να αλλάξει τη συμπεριφορά του αντιπάλου του σε πλεονέκτημα. Η διαδικασία σχεδιασμού εξαπάτησης περιλαμβάνει έναν αριθμό ανθρώπινων παραγόντων και είναι καθοδηγούμενη από πρόθεση, όπου οι προθέσεις είναι συνήθως κρυμμένες ή δεν παρατηρούνται εύκολα.

Οι Whaley και Burgoon χώρισαν την εξαπάτηση σε προσομοιωτική και διεγερτική. Η προσομοιωτική εξαπάτηση έχει σκοπό να δημιουργήσει ψευδείς πεποιθήσεις και χωρίζεται περαιτέρω σε μιμητική, επινοητική και δελεαστική. Η μιμητική εξαπάτηση είναι μια προσπάθεια προσομοίωσης άλλων πραγμάτων, η επινοητική δημιουργεί μια νέα πραγματικότητα διαφορετική από την τρέχουσα, ενώ η δελεαστική προσπαθεί να τραβήξει το θύμα της εξαπάτησης μακρυά από την αλήθεια. Αντίθετα με την προσομοιωτική, η διεγερτική εξαπάτηση επικεντρώνεται στην απόκρυψη της αλήθειας και διαιρείται με τη σειρά της σε εξαπάτηση συγκάλυψης, επανασυσκευασίας και εκθάμβωσης. Η πρώτη κρύβει την αλήθεια στο παρασκήνιο ή την κάνει αόρατη, η δεύτερη παρουσιάζει την αλήθεια δίνοντας της μια νέα και εντελώς διαφορετική εμφάνιση και η τρίτη προσπαθεί να αποκρύψει την αλήθεια προκαλώντας αβεβαιότητα, όταν η συγκάληψη ή επανασυσκευασία είναι αδύνατες.

Οι στρατηγικές εξαπάτησης, όταν ο κύκλος συμπεριφοράς ενός ατόμου/πράκτορα θεωρείται βρόχος OODA[21], μπορούν να θεωρηθούν ως προσπάθειες αλλαγής των πεποιθήσεων του αντιπάλου παρέχοντας παραπλανητικές πληροφορίες, τις οποίες το θύμα θα συλλέξει κατά τη διάρκεια του σταδίου της Παρατήρησης. Αυτές οι πληροφορίες αποθηκεύονται σαν μια πεποίθηση την οποία συμβουλεύεται ο πράκτορας κατά το βήμα Προσανατολισμού, όταν το θύμα εξαπάτησης προσπαθεί να ενημερώσει το μοντέλο που διατηρεί για τον κόσμο. Σε αυτό το σημείο, εάν η στρατηγική εξαπάτησης πέτυχε, τότε η αντιληπτή κατάσταση του θύματος για τον κόσμο θα περιλαμβάνει την πεποίθηση που του παρείχε ο εξαπατών. Το θύμα συνεχίζει με τα στάδια της Απόφασης και της Δράσης με βάση το διαστρευλομένο μοντέλο του κόσμου, το οποίο θα το αφήσει ευάλωτο στην εκμετάλλευση από τον απατεώνα. [49]

---

[21] Το μοντέλο βρόχου Παρατήρησης - Προσανατολισμού - Απόφασης - Δράσης (OODA) [48] είναι μια μέθοδος μοντελοποίησης του βρόχου συμβάντος ενός ατόμου. Αυτός ο κύκλος τεσσάρων βημάτων διαμορφώνει τη συλλογή πληροφοριών, τη λήψη αποφάσεων και τις ενέργειες ενός ατόμου με την πάροδο του χρόνου, με προηγούμενη συμπεριφορά να παρέχει ανατροφοδότηση για την τρέχουσα ανάλυση και απόφαση.

4.2 Ανίχνευση Εξαπάτησης

Για την έρευνα στην Τεχνητή Νοημοσύνη είναι ζωτικής σημασίας το να κατανοήσουμε την ανέντιμη και ανήθικη συμπεριφορά των μηχανών. Για την ανίχνευση της εξαπάτησης από μια μηχανή, αναμφισβήτητα απαιτείται ένα μοντέλο για το πως οι μηχανές μπορούν να εξαπατήσουν και πως μπορεί να εντοπιστεί μια τέτοια εξαπάτηση. [48]

Προκειμένου να εντοπιστεί εξαπάτηση, οποιοδήποτε αντίπαλο μοντέλο πρέπει να έχει τη δυνατότητα να συλλάβει την πρόθεση του αντιπάλου. Υπάρχουν πολλές προσεγγίσεις για την ανίχνευση της εξαπάτησης. Για τη μοντελοποίηση αντιπάλων ο Yuan πρότεινε ένα πλαίσιο που συνδυάζει έναν μηχανισμό ανίχνευσης εξαπάτησης με ένα μοντέλο συμπερασμάτων εχθρικής πρόθεσης. Ο σκοπός είναι να εντοπιστεί οποιαδήποτε παραπλανητική ενέργεια βάσει της πρόθεσης του αντιπάλου. [49]

Ο εντοπισμός της εξαπάτησης είναι δυνατός παρατηρώντας τα μοτίβα που παρουσιάζονται από τα διάφορα ειδής της. Οι Johnson και Grazioli [49] προτείνουν τη χρήση τεσσάρων διαδικασιών για την ανίχνευση της εξαπάτησης. Πρώτον, θα πρέπει να γίνει προσπάθεια παρατήρησης του περιβάλλοντος και εντοπισμού αναντιστοιχιών μέσα σε αυτό. Σε περίπτωση που βρεθούν τέτοιες αναντιστοιχίες, περαιτέρω διαδικασίες ανίχνευσης θα ενεργοποιηθούν. Δεύτερον, γίνεται προσπάθεια παραγωγής υπόθεσης εξαπάτησης σχετικά με ύποπτους χειρισμούς στο περιβάλλον σε μια διαδικασία που ονομάζεται “Παραγωγή Υπόθεσης”. Τρίτον, προκειμένου να αξιολογηθεί η παραπάνω υπόθεση, η αναπαράσταση του περιβάλλοντος που παρατηρείται επεξεργάζεται, έτσι ώστε να είναι συνεπής με τους υποθετικούς ύποπτους χειρισμούς. Τέλος, όλες οι αποδεκτές υποθέσεις συνδυάζονται για να παράγουν ένα τελικό αποτέλεσμα.

Οι Santos και Johnson [48] βασιζόμενοι στο μοντέλο των Johnson και Grazioli για την ανίχνευση της εξαπάτησης, πρότειναν μια μέθοδο σύγκρισης απόψεων μεταξύ εμπειρογνωμόνων που έχουν παρόμοιες γνώσεις. Αναφέρθηκαν σε ένα σύστημα όπου ένας άνθρωπος ή ένα πράκτορας (που έχει το ρόλο αυτού που αποφασίζει) παραθέτει μια λίστα με παρατηρήσεις σε μια ομάδα πρακτόρων. Αυτοί, με βάση τη λογική και τις γνώσεις τους επιστρέφουν σε αυτόν την άποψη τους για τις υπολοιπόμενες τυχαίες μεταβλητές του περιβάλλοντος. Στο σύστημα αυτό, οι πράκτορες έχουν κάποια κοινή γνώση και την αποθηκεύουν με ένα πιθανοτικό μοντέλο κάνοντας χρήση τυχαίων μεταβλητών. Ειδικότερα, για το πρώτο στάδιο από τα τέσσερα στάδια του μοντέλου των Johnson και Grazioli, το στάδιο της Ενεργοποίησης, παραθέτουν τεχνικές με τις οποίες

ένας πράκτορας μπορεί να προβλέψει τι απάντηση θα τους δώσει ένας πράκτορας, καθώς για την ανίχνευση της εξαπατάτησης θα πρέπει να έχει κάποια μορφη απάντησης ως αναμενόμενη δεδομένων κάποιων στοιχείων.

- Τεχνικές βασισμένες στη διαδικασία: Οι απαντήσεις των πρακτόρων μπορούν να ελεγχθούν. Έτσι, σε ένα σύστημα όπως αυτό που περιγράφηκε, οι τιμές των τυχαίων μεταβλητών θα πρέπει να ανήκουν στο διάστημα [0,1] και να αθροίζουν στο 1. Επίσης, θα πρέπει όλες οι αναμενόμενες μεταβλητές να περιλαμβάνονται στην απάντηση και όλες να είναι έγκυρες.
- Τεχνικές βάσει προτίμησης: Αν ο σχεδιαστής του συστήματος απαιτεί από όλους τους πράκτορες να προσφέρουν απάντηση για όλες τις υποθέσεις, η εξαπάτηση ανιχνεύεται με το αν όλοι έχουν απάντηση, με όλες τις μεταβλητές να είναι έγκυρες. Αν όμως το εύρος του προβλήματος είναι πολύ μεγάλο και δεν έχουν όλοι οι πράκτορες την εξειδικευμένη γνώση να απαντήσουν για όλες τις υποθέσεις, τότε θα πρέπει να αποθηκεύεται το ιστορικό των απαντήσεων κάθε πράκτορα ώστε να γίνεται σύγκριση με κάθε νέα απάντηση. Ενδεχόμενη εξαπάτηση ανιχνεύεται στην περίπτωση που κάποια υπόθεση δεν συμπεριλαμβανόταν στο ιστορικό απαντήσεων του πράκτορα
- Τεχνικές βασισμένες στην πρόθεση: Οι πράκτορες από τη φύση τους έχουν γνώση για το περιβάλλον στο οποίο ανήκουν και οι απόψεις/απαντήσεις που δίνουν είναι βασισμένες πάνω σε αυτή τη γνώση. Έτσι οι απόψεις τους πρέπει να είναι συνεπείς και άρα προβλέψιμες. Με αυτή την υπόθεση, είναι δυνατή η επαλήθευση της άποψης κάθε πράκτορα κάνοντας χρήση των απόψεων των υπολοίπων. Η επαλήθευση γίνεται μετρώντας την απόσταση μιας άποψης με τις υπόλοιπες, με χρήση της συσχέτισης Pearson[22] και της πρόβλεψης GroupLens[23]. Αν η απόσταση είναι αρκτετά μεγάλη, με άλλα λόγια η άποψη ενός πράκτορα είναι πολύ διαφορετική από τους άλλους, ενδέχεται να έχει ανιχνευθεί εξαπάτηση. Στην πράξη, γίνεται η υπόθεση ότι το σφάλμα μεταξύ της προβλεπόμενης και της ακριβούς γνώμης έχει κανονική κατανομή. Εάν η

[22] Συσχέτιση Pearson είναι μια στατιστική τιμή που μετρά τη γραμμική συσχέτιση μεταξύ δυο μεταβλητών x και y. Παίρνει τιμές που ανήκουν στο διάστημα [-1, 1] όπου -1 είναι αρνητική γραμμική συσχέτιση, 0 είναι καμία γραμμική συσχέτιση και 1 απόλυτη γραμμική συχέτιση. [49]

[23] GroupLens είναι ένα σύστημα για συλλογικό φιλτράρισμα των netnews, που βοηθά τους χρήστες να βρουν άρθρα που θα τους ενδιαφέρουν μέσα στην τεράστια ροή των διαθέσιμων άρθρων. [48] Στη συγκεριμένη περίπτωση αντί να προβλέπει άρθρα που μπορεί να ενδιαφέρουν χρήστες θα προβλέπει την άποψη ενός πράκτορα.

διαφορά μεταξύ της ακριβούς γνώμης και της προβλεπόμενης γνώμης είναι πάνω από τρεις φορές την τυπική απόκλιση, τότε λέμε ότι έχει εντοπιστεί πιθανή εξαπάτηση.

4.3 Αυταπάτη

Μερικοί άνθρωποι, ειδικά οι φιλόσοφοι, αντιλαμβάνονται την αυταπάτη σαν μια αντίφαση σχετικά με το πώς μπορεί ο εαυτός να εξαπατηθεί, αφού αυτό απαιτεί να γνωρίζει τι δεν ξέρει. Η αντίφαση επιλύεται εύκολα ορίζοντας τον εαυτό ως συνειδητό μυαλό, έτσι ώστε η αυταπάτη να συμβαίνει όταν ο εαυτός διατηρείται στο σκοτάδι[24], και όταν παραπλανά το συνειδητό μυαλό. Έτσι, το κλειδί για τον ορισμό της αυταπάτης είναι ότι οι αληθινές πληροφορίες αποκλείονται κατά προτίμηση από τη συνείδηση και, εάν διατηρούνται, διατηρούνται στο ασυνείδητο. [49]

Η ψυχολογική προκατάληψη είναι ένας αναπόφευκτος παράγοντας όταν οι άνθρωποι που αποφασίζουν έρχονται αντιμέτωποι με πολύπλοκες αποφάσεις σε ένα αβέβαιο περιβάλλον. Οι πεποιθήσεις μας δεν διαμορφώνονται απλώς και μόνο από τα διαθέσιμα στοιχεία, αλλά επηρεάζονται από τις επιθυμίες μας και τις προθέσεις μας. Η έρευνα πάνω στην ανθρώπινη συμπεριφορά έχει εντοπίσει μια πληθώρα λογικών και φαινομενικά παράλογων τάσεων για το πως οι άνθρωποι διαχειρίζονται τις πεποιθήσεις τους και λαμβάνουν αποφάσεις. Η έρευνα πάνω στα ανθρώπινα συναισθήματα έχει αναλύσει διαφορετικούς τρόπους με τους οποίους οι άνθρωποι τείνουν να απορρίψουν τις αγχωτικές πεποιθήσεις και να διατηρήσουν αυτές με τις οποίες νιώθουν άνετα, όπως η άρνηση και η ευσεβής σκέψη. Η έρευνα πάνω στη γνωστική ασυμφωνία έχει αποδείξει ότι οι άνθρωποι συχνά προσπαθούν να επιτύχουν συνοχή μεταξύ των πεποιθήσεων και των συμπεριφορών τους και επικεντρώνονται στον τρόπο που αλλάζουμε τις πεποιθήσεις μας για να επιλύσουμε ασυνέπειες μεταξύ μιας επιθυμητής εικόνας που έχουμε οι ίδιοι για τον εαυτό μας και τη συμπεριφορά μας. Ομοίως, έχει βρεθεί μια τάση για αυτό που ονομάζεται παρακινημένη παρέμβαση, με άλλα λόγια, η τάση για εξαγωγή συμπερασμάτων και πεποιθήσεων, με βάση τη συνέπεια των κινήτρων κάποιου και όχι μόνο με βάση τις πράξεις του. Η υπολογιστική μοντελοποίηση αυτών των μηχανισμών διατήρησης των ανθρώπινων πεποιθήσεων έχουν καταστεί σημαντικοί για ένα ευρύ φάσμα εφαρμογών. [48]

[24] όταν ο μεγαλύτερος οργανισμός διατηρεί τις αληθινές πληροφορίες από τη συνείδηση.

Η ψυχολογική βιβλιογραφία σχετικά με την αυταπάτη συνήθως αναφέρεται στην πράξη της αυταπάτης ως τις εσωτερικές διαδικασίες προκατάληψης που εμπλέκονται στην υιοθέτηση μιας επιθυμητής πεποίθησης απέναντι σε πιθανώς αντιφατικά αποδεικτικά στοιχεία. Είναι δηλαδή η πίστη στο P όταν το σύνολο των διαθέσιμων ικανών στοιχεία δείχνουν ¬P. Εστιάζοντας σε αυτές τις διαδικασίες προκατάληψης, συχνά ο ορισμός και η προδιαγραφή της επιθυμητής κατάστασης πεποίθησης παραμένει πολύ αφηρημένη. Εναλλακτικά, ένα περιβάλλον αβεβαιότητας είναι εκείνο στο οποίο οι αντικειμενικές πιθανότητες είναι εξ ολοκλήρου ή μερικώς άγνωστες απαιτώντας τη διατύπωση των υποκειμενικών εκτιμήσεων πιθανότητας που αντανακλούν τις πεποιθήσεις του λήπτη αποφάσεων. [49]

Η αυταπάτη είναι ένας ειδικός τύπος εξαπάτησης, όπου ο απατεώνας (αυτός που εξαπατά) και το θύμα είναι το ίδιο άτομο. Ενώ κάτι τέτοιο μπορεί να φαίνεται αντιπαραγωγικό στην καλύτερη περίπτωση, υπάρχουν κάποιες εξηγήσεις για το πως μπορεί η αυταπάτη να είναι ευεργετική. Ο Trivers προτείνει ότι η αυταπάτη μπορεί να είναι επωφελής εάν χρησιμοποιείται για την υποστήριξη του ψέματος. Η υπόθεση είναι ότι η αυταπάτη χρησιμοποιείται για να πείσει αυτόν που εξαπατά ότι πιστεύει το δικό του ψέμα. Όταν ένα άτομο επικοινωνεί ένα ψέμα στο θύμα για το οποίο το προορίζει, το θύμα θα λάβει επίσης ένα δευτερεύον μήνυμα μέσω της γλώσσας του σώματος του απατεώνα που δείχνει την αλήθεια του μηνύματος. Εάν χρησιμοποιείται αυταπάτη, ο εξαπατών πιστεύει ότι στο ψέμα και η γλώσσα του σώματός του θα δείξει ότι είναι αλήθεια. Ωστόσο, εάν είναι η αυταπάτη δεν χρησιμοποιείται, τότε η γλώσσα του σώματος του παραπλανητή θα δείξει ότι ψεύδονται και το θύμα μπορεί να καταλάβει την πρόθεση για εξαπάτηση. Αφού εξαπατηθεί το θύμα, ο απατεώνας επαναφέρει τη σωστή πίστη στη μνήμη του. Ο απατεώνας μπορεί στη συνέχεια να επωφεληθεί από την εσφαλμένη αντίληψη που έδωσε στον αντίπαλό του. Αυτή η μέθοδος χρησιμοποιεί την αυταπάτη ως μηχανισμό υποστήριξης της εξαπάτησης. [49] Αναλογικά, στην περίπτωση ενός πράκτορα, όταν αυτός θέλει να εξαπατήσει τον αντίπαλο του για την αλήθεια μιας δήλωσης P, ενώ γνωρίζει ότι το ¬P ισχύει στο περιβάλλον, θα πρέπει πρώτα να θεωρήσει και να πιστέψει το ¬(¬P), με άλλα λόγια το P. Αφού ο αντίπαλος έχει πειστεί για την αλήθεια του P, ο ίδιος θα θεωρήσει και θα πιστέψει ¬P, ώστε να επαναφέρει τον εαυτό του στην αρχική κατάσταση.

Ο Ramachandran υποστηρίζει ότι η αυταπάτη δεν μπορεί να βοηθήσει τους απατεώνες με αυτόν τον τρόπο, η παραπλανητική πίστη κρύβει τον στόχο της εξαπάτησης από τον εξαπατόντα. Αυτό καθιστά αδύνατο το να επωφεληθούν αργότερα από την εξαπάτηση. Η θεωρία του Ramachandran για τους σκοπούς της αυταπάτης είναι ότι η αυταπάτη χρησιμοποιείται ως μηχανισμός άμυνας. Το άτομο χρησιμοποιεί την αυταπάτη για να δημιουργήσει μια συνεκτική δομή πεποιθήσεων για τον εαυτό του, που θα επιβάλει σταθερότητα στη συμπεριφορά τους. Τα άτομα μπορούν επομένως να κρύψουν πληροφορίες από τον εαυτό τους που διαφωνεί με τις βασικές πεποιθήσεις τους. Αυτή η θεωρία συμφωνεί επίσης με την πεποίθηση των ψυχολόγων και ψυχίατρων ότι η αυταπάτη χρησιμοποιείται για την προστασία του χρήστη από επιβλαβείς αναμνήσεις, καταστέλλοντας τες. [49]

ΚΕΦΑΛΑΙΟ 5

ΕΠΙΛΟΓΟΣ

Πολλοί συγγραφείς διακρίνουν την ιστορία της Τεχνητής Νοημοσύνης σε τέσσερις περιόδους:

- Προϊστορική: η Τεχνητή Νοημοσύνη ουσιαστικά προαναγγέλλεται σε διηγήματα επιστημονικής φαντασίας.
- Κλασική (μέχρι τα μέσα της δεκαετίας του 1960): αναπτύχθηκαν συστήματα που έπαιζαν παιχνίδια και έλυναν γρίφους.[25]
- Ρομαντική (μέχρι τα μέσα της δεκαετίας του 1970): οι προσπάθειες επικεντρώνονται στην ανάπτυξη συστημάτων που κατανοούν ιστορίες και διάλογους σε φυσική γλώσσα.
- Μοντέρνα (μέχρι τα τέλη της δεκαετίας του 1980): χαρακτηρίζεται από την ανάπτυξη συστημάτων που βασίζονται στη γνώση και την εμπορική εκμετάλλευση των αποτελεσμάτων της έρευνας γύρω από την Τεχνητή Νοημοσύνη.

Αυτήν την εποχή βιώνουμε τη μετα-μοντέρνα περίοδο στην οποία η Τεχνητή Νοημοσύνη καλείται να παίξει ένα σημαντικό ρόλο σε ένα νέο πληροφοριακό περιβάλλον, του οποίου κύρια χαρακτηριστικά είναι η εξάπλωση του διαδικτύου και η διείσδυση των υπολογιστικών συστημάτων σε κάθε είδους συσκευές ευρείας και καθημερινής χρήσης. [49]

Η επιστήμη των υπολογιστών έχει μετακινηθεί από το παράδειγμα ενός απομωνομένου μηχανήματος στο παράδειγμα ενός δικτύου συστημάτων και κατανεμημένων υπολογιστών. Η Τεχνητή Νοημοσύνη κινείται από το παράδειγμα μιας απομονωμένης και μη τοποθετημένης νοημοσύνης στο παράδειγμα της τοποθετημένης, κοινωνικής και συλλογικής νοημοσύνης. Ειδικότερα αυτό που κυριαρχεί σήμερα είναι μια πιο πρακτοροκεντική άποψη[26]. [48]

---

[25] Ο Alan Turing, ο οποίος θεωρείται ο πατέρας της Τεχνητής Νοημοσύνης, εμπνεύστηκε το 1950 ένα τεστ (Turing test), για την αναγνώριση ευφυών μηχανών. Το Turing test βασίζεται σε μία σειρά από ερωτήσεις που υποβάλλει κάποιος σε έναν άνθρωπο και μία μηχανή, χωρίς να ξέρει εκ των προτέρων ποιος είναι ποιος. Αν στο τέλος δεν καταφέρει να ξεχωρίσει τον άνθρωπο από τη μηχανή, τότε η μηχανή περνάει το τεστ και θεωρείται ευφυής.

[26] Εν ολίγοις, ένας πράκτορας είναι ένα ενθυλακωμένο σύστημα υπολογιστή που είναι τοποθετημένο σε κάποιο περιβάλλον, που είναι ικανός για ευέλικτη και αυτόνομη δράση μέσα σε αυτό προκειμένου να να ανταποκριθεί στους στόχους σχεδιασμού του. [49]

Τα πολυπρακτορικά συστήματα αντιπροσωπεύουν ένα νέο τρόπο σύλληψης και υλοποίησης του κατανεμημένου λογισμικού. Τα πολυπρακτορικά συστήματα αποτελούνται από ετερογενείς και αυτοτελείς πράκτορες που αλληλεπιδρούν, διαπραγματεύονται και συντονίζονται[27] μεταξύ τους, σε ανοιχτά περιβάλλοντα, μόνο για να επιτύχουν τους δικούς τους στόχους ή εργασίες[28]. Έτσι, οι πράκτορες δεν είναι απαραίτητα καλοπροαίρετοι, συνεργάσιμοι ή ειλικρινείς. Για να επιτύχουν τους δικούς τους στόχους παρά την παρέμβαση και τον ανταγωνισμό των άλλων πρακτόρων θα κινηθούν στρατηγικά και θα εξαπατήσουν. [1] Επιπλέον, η εξαπάτηση δεν πρέπει απαραίτητα να σχετίζεται με λογικούς ή σκεπτόμενους πράκτορες. Ωστόσο, για να υπάρχει εξαπάτηση δεν χρειάζεται οι πράκτορες να είναι λογικοί. Το ίδιο ισχύει και για τους τεχνητούς πράκτορες. Για παράδειγμα, ακόμα και απλοί αντιδραστικοί πράκτορες (που δεν υπολογίζουν καμία σκοπιμότητα) μπορούν να μάθουν ή να αναπτύξουν απατηλές συμπεριφορές, καθώς αυτό μπορεί να είναι επωφελές σε ανταγωνιστικές καταστάσεις. Αναπόσπαστο κομμάτι των αλληλεπιδράσεων είναι η εμπιστοσύνη ανάμεσα στους πράκτορες καθώς και η διαχείριση αυτής. [48]

Η εμπιστοσύνη μεταξύ των πρακτόρων θεωρείται ως ένα από τα πιο σημαντικά θεμέλια βάσει των οποίων οι πράκτορες αποφασίζουν να αλληλεπιδρούν μεταξύ τους. Έτσι, το πρόβλημα του πως αποφασίζουν οι πράκτορες να αλληλεπιδρούν μεταξύ τους, μπορεί να περιοριστεί σε ένα: στο πώς οι πράκτορες υπολογίζουν την εμπιστοσύνη τους προς τους συνεργάτες τους. Όσο περισσότερο εμπιστεύεται ένας πράκτορας έναν συνεργάτη, τόσο μεγαλύτερη η πιθανότητα να αποφασίζει να αλληλεπιδράσει με αυτόν τον συνεργάτη. Ωστόσο, η ικανότητα δυσπιστίας, όταν δικαιολογείται, μας επιτρέπει να αποφύγουμε βλάβες όταν αντιμετωπίζουμε αναξιόπιστα συστήματα ή ανέντιμους ανθρώπους και οργανισμούς. Ομοίως, η εμπιστοσύνη, όταν είναι αδικαιολόγητη, οδηγεί σε έκθεση σε κινδύνους. Η εμπιστοσύνη είναι σαν μια πυξίδα για να μας καθοδηγεί με ασφάλεια σε έναν κόσμο αβεβαιότητας, κινδύνου και ηθικών κινδύνων.

Πώς μπορούμε να διαχειριστούμε την εμπιστοσύνη; Σε γενικές γραμμές, υπάρχουν δύο βασικές προσεγγίσεις για την εμπιστοσύνη στα πολυπρακτορικά συστήματα. Πρώτον, για να επιτρέψουμε στους πράκτορες να εμπιστεύονται ο ένας τον άλλον, υπάρχει ανάγκη να τους δοθεί η δυνατότητα να συλλογιστούν την αμοιβαία φύση, την αξιοπιστία ή την

[27] Ανταγωνίζονται ή συνεργάζονται.

[28] Που εκχωρούνται από τους χρήστες τους.

ειλικρίνεια των ομολόγων τους. Αυτή η ικανότητα αποτυπώνεται μέσω μοντέλων εμπιστοσύνης. Ο τελικός στόχος είναι να επιτρέπουν στους πράκτορες να υπολογίζουν το ποσό της εμπιστοσύνης που μπορούν να έχουν στους συνεργάτες τους. Ένας υψηλός βαθμός εμπιστοσύνης σε έναν πράκτορα θα σήμαινε ότι είναι πιθανό να επιλεγεί ως συνεργάτης αλληλεπίδρασης και (πιθανώς) μια αμοιβαία στρατηγική που χρησιμοποιείται για αυτόν σε πολλαπλές αλληλεπιδράσεις. Αντίθετα, ένας χαμηλός βαθμός εμπιστοσύνης θα είχε ως αποτέλεσμα να μην επιλεγεί (εάν υπάρχουν άλλοι, πιο αξιόπιστοι, συνεργάτες αλληλεπίδρασης) ή μια μη αμοιβαία στρατηγική που υιοθετείται εναντίον του σε πολλαπλές αλληλεπιδράσεις (εάν δεν υπάρχει καλύτερη εναλλακτική λύση). Με αυτόν τον τρόπο, τα μοντέλα εμπιστοσύνης στοχεύουν να καθοδηγήσουν τη λήψη αποφάσεων ενός πράκτορα για να αποφασίσει πως, πότε και με ποιον να αλληλεπιδράσει.

Για να γίνει αυτό, τα μοντέλα εμπιστοσύνης απαιτούν αρχικά από τους πράκτορες να συγκεντρώνουν κάποιες γνώσεις σχετικά με τα χαρακτηριστικά των ομότιμων τους. Αυτό μπορεί να επιτευχθεί με πολλούς διαφορετικούς τρόπους, όπως μέσω συμπερασμάτων που προκύπτουν από τα αποτελέσματα πολλαπλών άμεσων αλληλεπιδράσεων με αυτούς τους συνεργάτες ή μέσω έμμεσων πληροφοριών που παρέχονται από άλλους. Η περίπτωση άμεσης αλληλεπίδρασης μας οδηγεί να εξετάσουμε μεθόδους με τις οποίες οι πράκτορες μπορούν να μάθουν ή να αναπτύξουν καλύτερες στρατηγικές για να αντιμετωπίσουν ειλικρινείς και ανέντιμους πράκτορες, έτσι ώστε οι αποδόσεις να μεγιστοποιούνται μακροπρόθεσμα. Η περίπτωση έμμεσης αλληλεπίδρασης απαιτεί από τους πράκτορες να είναι σε θέση να αναπτύξουν μεθόδους, για να αποκτήσουν αξιοπιστία και να αιτιολογήσουν τις πληροφορίες που συλλέγονται από άλλους παράγοντες. [19]

Οι άνθρωποι χρησιμοποιούσαν πάντα την εξαπάτηση στην κοινωνική τους αλληλεπίδραση. Θα συνεχίσουν δε να εξαπατούν ο ένας τον άλλον, ενώ αλληλεπιδρούν μέσω υπολογιστών και δικτύων. Ο υπολογιστής, ως μέσο, θα παρέχει απλώς νέες ευκαιρίες και τρόπους εξαπάτησης, και ίσως μερικές φορές, τρόπους πρόληψης της εξαπάτησης ή υπεράσπισης από αυτή[29]. [48] Η ικανότητα των πρακτόρων να εξαπατούν αποτελεί σοβαρή μελλοντική απειλή για τη σχέση μεταξύ ανθρώπων και τεχνητής

[29] Για παράδειγμα, μέσω μιας απλής και γρήγορης πρόσβασης σε πιστοποίηση - όπως συμβαίνει με πιστωτικές κάρτες - ή με οδηγίες από κάποια αξιόπιστη αρχή.

νοημοσύνης. Αυτό είναι ιδιαίτερα απειλητικό στη σχέση εμπιστοσύνης μεταξύ ανθρώπων και τεχνητών παραγόντων. [48]

Η αυταπάτη μπορεί να χρησιμοποιηθεί είτε σαν βοηθητικό εργαλείο για την εξαπάτηση ενός τρίτου είτε σαν μηχανισμός άμυνας για τον ίδιο τον πράκτορα. Έτσι, σε ένα σύστημα όπου οι πράκτορες πιθανώς να αυταπατώνται, πως πρέπει να υπολογίζεται η εμπιστοσύνη για τις αλληλεπιδράσεις με ένα τέτοιο πράκτορα; Στην περίπτωση, για παράδειγμα, που ένας πράκτορας x επικοινώνησε μια πληροφορία σε ένα πράκτορα y και αυτή η πληροφορία είναι πραγματικά ψευδής, αλλά αληθής για τον x λόγω αυταπάτης, πως θα πρέπει ο x να κριθεί από τον y (σε ατομικό επίπεδο) ή/και από το περιβάλλον (σε επίπεδο συστήματος); Αν η αυταπάτη του x είχε σκοπό την εξαπάτηση, θα πρέπει αυτός να τιμωρηθεί. Τι γίνεται όμως στην περίπτωση που ήταν αποτέλεσμα άλλων παραγόντων, πχ. μηχανισμός άμυνας; Σε τέτοιου είδους συστήματα, τα μοντέλα και τα πρωτόκολλα της εμπιστοσύνης που θα χρησιμοποιηθούν, θα πρέπει να ενισχυθούν, έτσι ώστε να λαμβάνουν υπόψη τους και τον παράγοντα της “αγνής” αυταπάτης και να τιμωρούν την κακόβουλη χρήση της.[30]

[30] Ο Jones [48] περιγράφει μια παρόμοια κατάσταση, από την καθημερινή ζωή, αναφερόμενος στο παράδειγμα μιας γυναίκας και της συμπεριφοράς της.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. S. Russell and P. Norvig, *Artificial intelligence: A modern approach: United States edition*. Upper Saddle River, NJ: Pearson, 1994.
2. Winston, *Artificial Intelligence*, 3rd ed. Upper Saddle River, NJ: Pearson, 1992.
3. J. Haugeland, *Artificial intelligence: The very idea*. Cambridge, Mass.: MIT Press, 1985.
4. E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*. London, England: Addison Wesley, 1985.
5. R. Kurzweil, *The age of intelligent machines*. Cambridge, Mass.: MIT Press, 1990.
6. N. J. Nilsson, *Artificial intelligence: A new synthesis*. San Francisco, CA: Morgan Kaufmann, 1998.
7. "Οικοδόμηση εμπιστοσύνης στην ανθρωποκεντρική τεχνητή νοημοσύνη," *Europa.eu*, 2019. [Online]. Available: https://ec.europa.eu/transparency/regdoc/rep/1/2019/EL/COM-2019-168-F1-EL-MAIN-PART-1.PDF. [Accessed: 14-Sep-2020].
8. "Τεχνητή νοημοσύνη: Η Επιτροπή παρουσιάζει μια ευρωπαϊκή προσέγγιση για την τόνωση των επενδύσεων και τον καθορισμό δεοντολογικών κατευθυντήριων γραμμών," *Europa.eu*, 2018. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/el/IP_18_3362. [Accessed: 14-Sep-2020].
9. "Τεχνητή νοημοσύνη - Η ευρωπαϊκή προσέγγιση της αριστείας και της εμπιστοσύνης," *Europa.eu*, 2020. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1/language-el/format-PDF. [Accessed: 14-Sep-2020].
10. "European group on ethics in science and new technologies," *Europa.eu*, 2018. [Online]. Available: http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf. [Accessed: 14-Sep-2020].
11. Κ. Γεωργούλη, *Τεχνητή Νοημοσύνη, Μια Εισαγωγική Προσέγγιση*. Αθήνα: Σύνδεσμος Ελληνικων Ακαδημαϊκων Βιβλιοθηκών Εθνικό Μετσόβιο Πολυτεχνείο, 2015.
12. B. Hayes-Roth, "An architecture for adaptive intelligent systems," *Artif. Intell.*, vol. 72, no. 1–2, pp. 329–365, 1995.
13. M. Wooldridge and N. R. Jennings, "Intelligent agents: theory and practice," *Knowl. Eng. Rev.*, vol. 10, no. 2, pp. 115–152, 1995.
14. M. Wooldridge, *Εισαγωγή Στα Πολυπρακτορικά Συστήματα*. Αθήνα: Εκδόσεις Κλειδάριθμος, 2008.
15. M. H. Nguyen and D. Q. Tran, "A Combination Trust Model for Multi-Agent Systems," *International Journal of Innovative Computing, Information and Control*, vol. 9, no. 15, pp. 2405–2420, 2012.
16. N. R. Jennings, "Commitments and conventions: The foundation of coordination in multi-agent systems," *Knowl. Eng. Rev.*, vol. 8, no. 3, pp. 223–250, 1993.

17. D. V. Pynadath and M. Tambe, “Multiagent teamwork: Analyzing the optimality and complexity of key theories and models,” in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 2 - AAMAS ’02*, 2002.
18. N.R. Jennings, P. Faratin, A.R. Lomuscio, S. Parsons, M.J. Wooldridge & C. Sierra, “Automated Negotiation: Prospects, Methods and Challenges,” *Group Decision and Negotiation*, vol. 10, no. 2, pp. 199–215, 2001.
19. S. D. Ramchurn, D. Huynh, and N. R. Jennings, “Trust in multi-agent systems,” *Knowl. Eng. Rev.*, vol. 19, no. 1, pp. 1–25, 2004.
20. M. N. Huhns and L. M. Stephens, *Multiagent systems: A modern approach to distributed artificial intelligence*. Cambridge, Mass.: MIT Press, 1999.
21. D. Artz and Y. Gil, “A survey of trust in computer science and the Semantic Web,” *Web Semant.*, vol. 5, no. 2, pp. 58–71, 2007.
22. D. Ceolin, A. Nottamkandath, W. Fokkink, and V. Maccatrozzo, “Towards the Definition of an Ontology for Trust in (Web) Data,” 2014.
23. Κ. Κράβαρη, “Διαλειτουργικότητα και Αξιοπιστία της Αλληλεπίδρασης Ευφυών Πρακτόρων στον Σημασιολογικό Ιστό,” Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη, 2015.
24. D. Gambetta, “Can We Trust Trust?,” in *Trust: Making and Breaking Cooperative Relations*, Department of Sociology, University of Oxford, 1990, pp. 213–237.
25. T. Grandison and M. Sloman, “A Survey of Trust in Internet Applications,” *IEEE Communications Surveys & Tutorials*, vol. 3, no. 4, pp. 2–16, 2000.
26. M. Blaze, J. Feigenbaum, and J. Lacy, “Decentralized Trust Management,” in *IEEE Symposium on Security and Privacy*, 1996.
27. A. Jøsang, C. Keser, and T. Dimitrakos, “Can We Manage Trust?,” in *Trust Management, Third International Conference*, 2005.
28. D. Rosaci, G. M. L. Sarnè, and S. Garruzzo, “Integrating trust measures in multiagent systems,” *International Journal of Intelligent Systems*, vol. 27, no. 1, pp. 1–15, 2012.
29. Μ. Μπαλταγιάννης, “Αναπαράσταση και Διαχείριση Εμπιστοσύνης σε Ανοιχτά Πολυπρακτορικά Συστήματα,” Πανεπιστήμιο Θεσσαλίας, Θεσσαλία, 2006.
30. P. Dasgupta, “Trust as a commodity,” in *Trust: Making and Breaking Cooperative Relations*, 1998, pp. 49–72.
31. C. Castelfranchi and R. Falcone, “Principles of trust for MAS: cognitive anatomy, social importance, and quantification,” in *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)*, 1998.
32. S. Poslad, P. Charlton, and M. Calisti, “Specifying standard security mechanisms in multi-agent systems,” in *The 2002 international conference on Trust, reputation, and security: theories and practice*, 2002, pp. 227–237.
33. H. A. Abbass, G. Leu, and K. Merrick, “A review of theoretical and practical challenges of trusted autonomy in big data,” IEEE Access, vol. 4, pp. 2808–2830, 2016.

34. E. Sathiyamoorthy, N. C. S. N. Iyenger, and V. Ramachandran, “Agent based trust management model based on weight value model for online auctions,” International Journal of Network Security & Its Applications, vol. 1, no. 3, 2010.
35. L. Brumley, C. Kopp, and K. Korb, “Misperception, Self-Deception and Information Warfare,” 2006.
36. S. Sarkadi, A. R. Panisson, R. H. Bordini, and P. McBurney, “Modelling deception using theory of mind in multi-agent systems,” 2019.
37. E. Santos Jr., D. Li, and X. Yuan, “On deception detection in multi-agent systems and deception intent,” in *Modeling and Simulation for Military Operations III*, 2008.
38. A. Luft, “The OODA Loop and the half-beat,” *Thestrategybridge.org*, 17-Mar-2020. [Online]. Available: https://thestrategybridge.org/the-bridge/2020/3/17/the-ooda-loop-and-the-half-beat. [Accessed: 14-Sep-2020].
39. P. E. Johnson, S. Grazioli, K. Jamal, and R. Glen Berryman, “Detecting deception: adversarial problem solving in a low base-rate world,” *Cognitive Science A Multidisciplinary Journal*, vol. 25, no. 3, pp. 355–392, 2001.
40. E. Santos and G. Johnson, “Towards Detecting Deception in Intelligent Systems,” in *SPIE - The International Society for Optical Engineering*, 2004.
41. “Συσχέτιση,” *Upatras.gr*. [Online]. Available: https://thalis.math.upatras.gr/~adk/lectures/ida/lab6/slides6.pdf. [Accessed: 14-Sep-2020].
42. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An open architecture for collaborative filtering of netnews,” in *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW ’94*, 1994.
43. R. L. Trivers, *Deceit and self-deception: Fooling yourself the better to fool others*. Harlow, England: Penguin Books, 2014.
44. J. Y. Ito, D. V. Pynadath, and S. C. Marsella, “Self-deceptive decision making: normative and descriptive insights,” in *8th International Joint Conference on Autonomous Agents and Multiagent Systems*, 2009.
45. J. Y. Ito, D. V. Pynadath, and S. C. Marsella, “Modeling self-deception within a decision-theoretic framework,” in *Autonomous Agents and Multi-Agent Systems*, 2008, vol. 20, pp. 322–333.
46. A. J. I. Jones, “On the Logic of Self-deception,” *South American Journal of Logic*, 2016.
47. Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, and Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*. Εκδόσεις Γκιούρδα, 2006.
48. C. Castelfranchi, “Artificial liars: Why computers will (necessarily) deceive us and each other,” *Ethics and Information Technology*, vol. 2, no. 2, pp. 113–119, 2000.
49. M. Wooldridge, “Agent-based software engineering,” *IEE Proc. Softw. Eng*, vol. 144, no. 1, pp. 26–37, 1997.