Πανεπιστήμιο Δυτικής Αττικής

Σχολή Μηχανικών

Τμήμα Βιομηχανικής Σχεδίασης και Παραγωγής

Διπλωματική εργασία

Αλληλεπίδραση Ανθρώπου-Ρομπότ και Εμπιστοσύνη

Κουστένη Ζωή
71445062

Επιβλέπων:

Γρηγόριος Νικολάου
Λέκτορας Εφαρμογών

Αιγάλεω - Αθήνα, Σεπτέμβριος, 2023

Εγκρίθηκε από την εξεταστική επιτροπή τον Οκτώβριο 2023.

GRIGORIOS NIKOLAOU Digitally signed by GRIGORIOS NIKOLAOU Date: 2023.11.15 08:54:37 +02'00'

Νικολάου Γρηγόριος
Λέκτορας Εφαρμογών

Soultana Vasileiadou Digitally signed by Soultana Vasileiadou Date: 2023.11.22 11:12:36 +02'00'

Βασιλειάδου Σουλτάνα
Επίκουρη Καθηγήτρια

Christos Drosos Digitally signed by Christos Drosos Date: 2023.11.21 12:07:55 +02'00'

Δρόσος Χρήστος
Ε.ΔΙ.Π

Κουστένη Ζωή
Τμήμα Βιομηχανικής Σχεδίασης και Παραγωγής
Πανεπιστήμιο Δυτικής Αττικής

Η κάτωθι υπογεγραμμένη Κουστένη Ζωή του Πολυχρόνη, με αριθμό μητρώου 71445062 φοιτήτρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Βιομηχανικής Σχεδίασης και Παραγωγής, δηλώνω υπεύθυνα ότι: «Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από εμένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Η Δηλούσα

Κουστένη Ζωή

Στην οικογένειά μου, στους φίλους μου και στην τεράστια υποστήριξή τους.

Ευχαριστίες

Ευχαριστώ από καρδιάς του φίλους και συνεργάτες της φοιτητικής μου ζωής που αποτέλεσαν την έμπνευση αυτής της εργασίας. Ταυτόχρονα είμαι ευγνώμων για τον επιβλέπων καθηγητή μου Γρηγόρη Νικολάου για την απέραντη εμπιστοσύνη και βοήθεια που μου προσέφερε στις σπουδές μου αλλά και στην διπλωματική μου.

Κυριακή 17 Σεπτεμβρίου 2023
Κουστένη Ζωή

Περίληψη

Η αλληλεπίδραση ανθρώπου-ρομπότ αποτελεί ένα ολοένα και πιο σημαντικό θέμα προς συζήτηση και έρευνα, με την εμπιστοσύνη και τα γενικά συναισθήματα που προκαλούνται στον χειριστή, να είναι στο προσκήνιο. Η παρούσα διπλωματική εργασία εξετάζει και αναλύει σε βάθος τα προηγούμενα σχετικά ερευνητικά άρθρα για να συγκεντρώσει τα σημαντικότερα στοιχεία γύρω από την εμπιστοσύνη του ανθρώπου στον ρομπότ-βοηθό του: Αρχικά αναφέρονται οι διαφορές μεταξύ της διαπροσωπικής εμπιστοσύνης με εκείνη του ανθρώπου-αυτοματισμού αλλά και τους παράγοντες επιρροής της, όπου και χωρίζονται σε τρεις κατηγορίες, την εμπιστοσύνη προδιάθεσης, περιστασιακή εμπιστοσύνη και επίκτητη εμπιστοσύνη. Έπειτα, μετά τα διάφορα μοντέλα και πλαίσια εμπιστοσύνης που κατάγραφηκαν ερευνητικά, υπογραμμίζονται οι διάφορες προκλήσεις και προοπτικές στην ανάπτυξη αξιόπιστων ρομπότ, με έμφαση στον σχεδιασμό τους, την απόκτηση και την διατήρηση της εμπιστοσύνης αλλά και στρατηγικές επιδιόρθωσής της. Τέλος, ακολουθεί μία λεπτομερής ενότητα αναφερόμενη στην αλληλεπίδραση ανθρώπου-κοινωνικού ρομπότ, την συναισθηματική προδιάθεση του ανθρώπου, παράγοντες που επηρεάζουν την αντίληψή τους για αυτά και τα ανθρωπομορφικά χαρακτηριστικά του ρομπότ που βοηθούν στην αξιοπιστία του.

Λέξεις Κλειδιά: Αλληλεπίδραση ανθρώπου-μηχανής, Εμπιστοσύνη, Κοινωνικά ρομπότ, Επιδιόρθωση εμπιστοσύνης, Ανάπτυξη εμπιστοσύνης

Abstract

In recent years, the relationship between humans and robots has become a prominent topic of discussion. One of the key factors in this relationship is trust, and how operators feel about their robot assistants. To gain a deeper understanding of human trust in robots, this thesis delves into relevant research articles in detail. The first section of the thesis explores the differences between interpersonal trust and trust in human automation, and categorizes the factors that influence trust into dispositional trust, situational confidence, and acquired trust. The second section focuses on various models and frameworks of trust, as well as the challenges and perspectives in developing trustworthy robots. This includes a detailed discussion on the design, acquisition, and maintenance of confidence, as well as strategies to repair trust. The third and final section of the thesis is dedicated to human-social robot interaction, human emotional predisposition, and the factors that influence perception of robots. Additionally, it examines the anthropomorphic characteristics that help establish trust between humans and robots. By analyzing all of these factors, the thesis aims to provide a comprehensive understanding of human trust in robots, which is crucial in further developing successful human-robot relationships.

Keywords: HRI, Trust, Social Robots, Building trust, Restoring trust, Human-Robot interaction

Πίνακας περιεχομένων

Πίνακας σχημάτων

Κεφάλαιο 1

Εισαγωγή

Καθώς η αυτονομία και οι δυνατότητες των ρομποτικών συστημάτων αυξάνονται, αναμένεται ότι θα παίζουν το ρόλο του συμπαίκτη και όχι του εργαλείου και ότι θα αλληλεπιδρούν με τους ανθρώπινους συνεργάτες με πιο ρεαλιστικό τρόπο, δημιουργώντας μια πιο ανθρώπινη σχέση. Δεδομένης της επίδρασης της εμπιστοσύνης που παρατηρείται στην αλληλεπίδραση ανθρώπου-ρομπότ (Human-Robot Interaction), η κατάλληλη εμπιστοσύνη στους ρομποτικούς συνεργάτες είναι ένας από τους κύριους παράγοντες που επηρεάζουν την απόδοση της αλληλεπίδρασης ανθρώπου-ρομπότ. Η απόδοση των ομάδων μπορεί να μειωθεί εάν οι άνθρωποι δεν εμπιστεύονται κατάλληλα τα ρομπότ, αχρηστεύοντάς τα ή κάνοντας κακή χρήση με βάση την περιορισμένη εμπειρία τους.

Μέσω της παρούσας διπλωματικής ο σκοπός αυτής είναι η ανάλυση και επεξήγηση της εμπιστοσύνης μεταξύ ανθρώπου-αυτοματισμού, μέσω μίας εκτενούς συστημικής έρευνας σε ποικίλα επιστημονικά άρθρα.

1.1 Ορισμοί

- Η *αλληλεπίδραση ανθρώπου-αυτοματισμού* (Human-Automation Interaction ή HAI) αναφέρεται στη μελέτη και το σχεδιασμό των αλληλεπιδράσεων μεταξύ ανθρώπων και αυτοματοποιημένων συστημάτων. Επικεντρώνεται στην κατανόηση και τη βελτιστοποίηση της συνεργασίας και της επικοινωνίας μεταξύ ανθρώπων και τεχνολογιών αυτοματισμού, όπως τα ρομπότ, η τεχνητή νοημοσύνη και τα αυτόνομα συστήματα. Η αλληλεπίδραση ανθρώπου-αυτοματισμού περιλαμβάνει την ανάπτυξη διεπαφών, αλγορίθμων και τεχνικών που επιτρέπουν στον άνθρωπο να αλληλεπιδρά αποτελεσματικά με τα αυτοματοποιημένα συστήματα και να τα ελέγχει, καθώς και την ικανότητα του αυτοματισμού να κατανοεί και να ανταποκρίνεται στην ανθρώπινη εισροή και ανατροφοδότηση. Ο στόχος είναι η δημιουργία απρόσκοπτης και αποτελεσματικής συνεργασίας μεταξύ ανθρώπου και αυτοματισμού, όπου αξιοποιούνται τα πλεονεκτήματα κάθε μέρους για την ενίσχυση της παραγωγικότητας, της λήψης αποφάσεων και της συνολικής απόδοσης του συστήματος. Με την κατανόηση της δυναμικής της αλληλεπίδρασης ανθρώπου-αυτοματισμού, οι ερευνητές στοχεύουν στη βελτίωση της χρηστικότητας, της ασφάλειας και της αποδοχής των αυτοματοποιημένων συστημάτων σε διάφορους τομείς, όπως η μεταποίηση, οι μεταφορές, η υγειονομική περίθαλψη και όχι μόνο.

- Η *αλληλεπίδραση ανθρώπου-ρομπότ* (Human-Robot Interaction ή HRI) αναφέρεται στη μελέτη και το σχεδιασμό των αλληλεπιδράσεων μεταξύ ανθρώπων και ρομπότ. Επικεντρώνεται στην ανάπτυξη συστημάτων και τεχνολογιών που επιτρέπουν την αποτελεσματική επικοινωνία, αλληλεπίδραση και συνεργασία μεταξύ ανθρώπων και ρομπότ. Η HRI περιλαμβάνει διάφορες πτυχές, όπως ο σχεδιασμός διεπαφών ρομπότ, τρόποι επικοινωνίας και η ανάπτυξη αλγορίθμων και τεχνικών που επιτρέπουν στα ρομπότ να κατανοούν και να

ανταποκρίνονται στις ανθρώπινες ενέργειες, χειρονομίες και εντολές. Στόχος της HRI είναι να δημιουργήσει διαισθητικούς και φυσικούς τρόπους αλληλεπίδρασης των ανθρώπων με τα ρομπότ, επιτρέποντάς τους να συνεργάζονται απρόσκοπτα σε διάφορους τομείς, όπως η υγειονομική περίθαλψη, η βιομηχανία παραγωγής και οι κλάδοι παροχής υπηρεσιών. Με την κατανόηση της δυναμικής της αλληλεπίδρασης ανθρώπου-ρομπότ, οι ερευνητές στοχεύουν στη βελτίωση της χρηστικότητας, της αποδοχής και της συνολικής αποτελεσματικότητας των ρομποτικών συστημάτων.

- Η *συνεργασία ανθρώπου-ρομπότ* (Human-Robot Collaboration ή HRC) αναφέρεται στις προσπάθειες συνεργασίας μεταξύ ανθρώπων και ρομπότ για την εκτέλεση μιας εργασίας ή την επίτευξη ενός κοινού στόχου. Σε αντίθεση με τα παραδοσιακά βιομηχανικά ρομπότ που λειτουργούν απομονωμένα, η HRC επικεντρώνεται στην ανάπτυξη ρομπότ που μπορούν να εργάζονται μαζί με τους ανθρώπους σε κοινόχρηστους χώρους εργασίας, συμπληρώνοντας τις δεξιότητες και τις ικανότητές τους. Σε ένα συνεργατικό περιβάλλον, τα ρομπότ σχεδιάζονται για να κατανοούν τις ανθρώπινες προθέσεις, να προσαρμόζονται σε δυναμικά περιβάλλοντα και να εκτελούν εργασίες που απαιτούν τόσο ανθρώπινη τεχνογνωσία όσο και ρομποτική ακρίβεια. Η συνεργασία μεταξύ ανθρώπων και ρομπότ περιλαμβάνει πτυχές όπως η κατανομή εργασιών, η ανταλλαγή πληροφοριών, ο συντονισμός και η επικοινωνία. Αξιοποιώντας τα πλεονεκτήματα τόσο των ανθρώπων όσο και των ρομπότ, το HRC στοχεύει στη βελτίωση της παραγωγικότητας, της αποδοτικότητας και της ασφάλειας σε διάφορους τομείς, όπως η βιομηχανία, τα logistics και η υγειονομική περίθαλψη. Η αποτελεσματική συνεργασία ανθρώπου-ρομπότ απαιτεί το σχεδιασμό ρομπότ με τα οποία είναι διαισθητικό να δουλεύει κανείς, τη διασφάλιση διαφανών διαύλων επικοινωνίας και την αντιμετώπιση ζητημάτων ασφάλειας για την ενίσχυση της εμπιστοσύνης και της αποδοχής μεταξύ ανθρώπων και ρομπότ.

- Τα *πιθανοτικά γραφικά μοντέλα* (Probabilistic graphical models ή PGM) παρέχουν ένα ισχυρό πλαίσιο για την κατανόηση και την ανάλυση περίπλοκων σχέσεων μεταξύ μεταβλητών. Με την ενσωμάτωση πιθανολογικής συλλογιστικής, μπορούν να κατασκευαστούν μοντέλα που λαμβάνουν υπόψη την αβεβαιότητα και τις εξαρτήσεις, επιτρέποντας την εξαγωγή ακριβών συμπερασμάτων και την πραγματοποίηση αξιόπιστων προβλέψεων. Επιπλέον, η αποτελεσματικότητα αυτών των μοντέλων στην εξαγωγή συμπερασμάτων και στην ανάλυση επιτρέπει την ταχεία επεξεργασία τεράστιων ποσοτήτων δεδομένων, οδηγώντας σε πολύτιμες γνώσεις.

1.2 Πλαίσιο της διπλωματικής εργασίας

Η εμπιστοσύνη είναι ένας από τους βασικούς παράγοντες για την ανάπτυξη αποδοτικών σχέσεων, συμπεριλαμβανομένων των σχέσεων μεταξύ ανθρώπων και αυτοματισμού. Μπορούμε να αναφέρουμε την εμπιστοσύνη ως μια συνολική ανησυχία που επηρεάζει την αποτελεσματικότητα ενός συστήματος,ιδίως όσον αφορά την ασφάλεια, την απόδοση και το ποσοστό χρήσης [1]. Έχοντας υπόψιν αυτή την ανησυχία, η εμπιστοσύνη έχει γίνει ένα κρίσιμο στοιχείο στο σχεδιασμό και την ανάπτυξη αυτοματοποιημένων συστημάτων [2]. Τα αυτόνομα συστήματα σχεδιάζονται και αναπτύσσονται με αυξημένα επίπεδα ανεξαρτησίας και δυνατοτήτων λήψης αποφάσεων, και αυτές οι ικανότητες θα είναι αποτελεσματικές σε καταστάσεις αβεβαιότητας [3].

Η εμπιστοσύνη ανθρώπου-ρομπότ είναι ένας σημαντικός κλάδος της Αλληλεπίδρασης Ανθρώπου -Ρομπότ (HRI), ο οποίος έχει κερδίσει πρόσφατα ολοένα και μεγαλύτερη προσοχή από μελετητές σε πολλούς κλάδους, όπως η Μηχανική Υπολογιστών [4], η Ψυχολογία [5], η Επιστήμη των Υπολογιστών [6] και η Μηχανολογία [7]. Η εμπιστοσύνη είναι ένας σημαντικός παράγοντας που πρέπει να λαμβάνεται υπόψη όταν τα ρομπότ πρόκειται να εργαστούν ως συμπαίκτες σε ομάδες

ανθρώπων-ρομπότ [8], χρησιμοποιούνται ως αυτόνομοι πράκτορες [9] ή όταν τα ρομπότ πρόκειται να χρησιμοποιηθούν σε πολύπλοκες και επικίνδυνες καταστάσεις [4]. Σε πολλές περιπτώσεις, η εμπιστοσύνη είναι ο κύριος παράγοντας που καθορίζει κατά πόσο ένας ρομποτικός πράκτορας θα γίνει αποδεκτός και θα χρησιμοποιηθεί από τον άνθρωπο [10].

Οι αδύναμες συνεργασίες που προκύπτουν από ακατάλληλη ή μη ισορροπημένη εμπιστοσύνη μεταξύ ανθρώπων και ρομπότ μπορεί να προκαλέσουν κακή χρήση ή αχρηστία ενός ρομποτικού πράκτορα. Η κακή χρήση αναφέρεται σε αστοχίες που προκύπτουν λόγω της υπερβολικής εμπιστοσύνης του χρήστη στον ρομποτικό πράκτορα (π.χ. αποδοχή όλων των λύσεων και των αποτελεσμάτων που παρουσιάζει το ρομπότ χωρίς αμφισβήτηση). Αντίθετα, η αχρηστία αναφέρεται στις αποτυχίες που συμβαίνουν λόγω της ελλιπούς εμπιστοσύνης του ανθρώπου στον ρομποτικό πράκτορα (π.χ. απόρριψη των δυνατοτήτων ενός ρομποτικού πράκτορα)[1]. Για να αποφευχθεί η κακή χρήση και η αχρηστία των ρομπότ από τους ανθρώπινους χειριστές, η εμπιστοσύνη πρέπει να ρυθμιστεί.

1.3 Οργάνωση, κεφαλαίωση, διάρθρωση της εργασίας

Στο κεφάλαιο υπ' αριθμόν 2, όπου και γίνεται η κύρια ανάλυση του θέματος της εργασίας, αναλύονται σε βάθος τα παρακάτω υποκεφάλαια με τίτλους:

1. Η εμπιστοσύνη στην συνεργασία ανθρώπου-ρομπότ
2. Μοντέλα και πλαίσια εμπιστοσύνης
3. Προκλήσεις και προοπτικές στην ανάπτυξη αξιόπιστων ρομπότ
4. Στρατηγικές επιδιόρθωσης της εμπιστοσύνης
5. Αλληλεπίδραση ανθρώπου - κοινωνικού ρομπότ

Ακολουθεί συνοπτική περιγραφή τους:

1. **Η εμπιστοσύνη στην συνεργασία ανθρώπου-ρομπότ** Σε αυτό το υποκεφάλαιο, θα εξεταστεί η σημασία της εμπιστοσύνης στη σχέση ανθρώπου-ρομπότ. Ξεκινώντας, θα επισημανθούν οι διαφορές μεταξύ της εμπιστοσύνης στον αυτοματισμό και της διαπροσωπικής εμπιστοσύνης. Στη συνέχεια, θα εξεταστούν οι παράγοντες που επηρεάζουν την εμπιστοσύνη, όπως η εμπιστοσύνη προδιάθεσης, η περιστασιακή εμπιστοσύνη και η επίκτητη εμπιστοσύνη.

2. **Μοντέλα και πλαίσια εμπιστοσύνης** Σε αυτήν την υποενότητα, ερευνόνται τα ήδη ανεπτυγμένα από την επιστημονική κοινότητα, μοντέλα εμπιστοσύνης με στόχο την αποκωδικοποίηση της διαμόρφωσής της.

3. **Προκλήσεις και προοπτικές στην ανάπτυξη αξιόπιστων ρομπότ** Σε αυτή την ενότητα, θα εξεταστούν οι προκλήσεις που προκύπτουν στην προσπάθεια ανάπτυξης αξιόπιστων ρομπότ καθώς και οι επιμέρους διαδικασίες . Θα εξεταστεί η σημασία του σχεδιασμού με έμφαση στην εμπιστοσύνη, τις μεθόδους απόκτησης, διαφύλαξης και βαθμονόμησης της εμπιστοσύνης και τα υπολογιστικά μοντέλα εμπιστοσύνης.

4. **Στρατηγικές επιδιόρθωσης της εμπιστοσύνης** Σε αυτήν την ενότητα, θα εξεταστούν στρατηγικές που μπορούν να χρησιμοποιηθούν για την ανασύνθεση και ανακατασκευή της εμπιστοσύνης σε περιπτώσεις που αυτή έχει παραβιαστεί ή διαταραχθεί.

5. **Αλληλεπίδραση ανθρώπου - κοινωνικού ρομπότ** Σε αυτήν την τελευταία ενότητα, θα εξεταστεί η αλληλεπίδραση ανθρώπου-κοινωνικού ρομπότ. Θα δωθεί έμφαση στις προδιαθέσεις του ανθρώπου απέναντι στα κοινωνικά ρομπότ, τα ανθρωπομορφικά χαρακτηριστικά προσώπου του ρομπότ και την επιρροή τους στην αξιοπιστία αλλά και την γενικότερη αντίληψη του ανθρώπου για αυτά.

Στο επόμενο και τελευταίο κεφάλαιο, υπ' αριθμόν 3, θα αναλυθούν τα συμπεράσματα της έρευνας που προηγήθηκε, κλείνοντας με μια συζήτηση γύρω από μελλοντικές ερευνητικές επεκτάσεις.

Κεφάλαιο 2

Θεωρητικό μέρος – Βιβλιογραφική έρευνα – Σχετικές προσπάθειες

Η έννοια της εμπιστοσύνης συναντάται σε ποικίλους ερευνητικούς τομείς. Ερευνητές από την ψυχολογία, την κοινωνιολογία, τη φιλοσοφία, τις πολιτικές επιστήμες, τα οικονομικά και τους ανθρώπινους παράγοντες έχουν προσπαθήσει να κατανοήσουν την εμπιστοσύνη και να αναπτύξουν τρόπους εννοιολογικής προσέγγισης του όρου. Οι Mayer, Davis και Schoorman [11] συνέταξαν ένα από τα πιο σημαντικά ερευντικά άρθρα για την εμπιστοσύνη μέχρι σήμερα, εξετάζοντας διεξοδικά τη βιβλιογραφία σχετικά με τα προηγούμενα και τα αποτελέσματα της οργανωσιακής εμπιστοσύνης. Στον τομέα των ανθρώπινων παραγόντων, σημαντική έρευνα έχει επικεντρωθεί στο ρόλο της εμπιστοσύνης στην καθοδήγηση των αλληλεπιδράσεων με διάφορες τεχνολογίες. Για παράδειγμα, οι Corritore, Kracher και Wiedenbeck [12] ανέπτυξαν ένα μοντέλο διαδικτυακής εμπιστοσύνης το οποίο εννοιολογεί τη μεταβλητότητα της εμπιστοσύνης με βάση τις αντιλήψεις για τον κίνδυνο του ιστότοπου, την αξιοπιστία και την ευκολία χρήσης. Οι Gefen, Karahanna και Straub [13] διερευνούν το ρόλο της εμπιστοσύνης και του μοντέλου αποδοχής της τεχνολογίας σε περιβάλλοντα ηλεκτρονικών αγορών. Οι Hoffmann *et al* [14] συζητούν το ρόλο της εμπιστοσύνης, καθώς και της αντιμονοπωλιακής νομοθεσίας, στη δημιουργία και την ελαχιστοποίηση των κινδύνων ασφαλείας στους τομείς του κυβερνοχώρου.

Ένα σημαντικό κοινό στοιχείο στα διάφορα ερευνητικά πεδία είναι ότι σχεδόν κάθε έννοια της εμπιστοσύνης περιλαμβάνει τρεις συνιστώσες. Πρώτον, πρέπει να υπάρχει ένας εμπιστευόμενος για να δώσει εμπιστοσύνη, πρέπει να υπάρχει ένας αποδέκτης εμπιστοσύνης για να δεχτεί εμπιστοσύνη και πρέπει να διακυβεύεται κάτι. Δεύτερον, ο διαχειριστής πρέπει να έχει κίνητρα για να εκτελέσει το έργο. Το κίνητρο μπορεί να ποικίλλει ευρέως, από μια χρηματική ανταμοιβή έως την καλοπροαίρετη επιθυμία να βοηθήσει τους άλλους. Στις τεχνολογικές αλληλεπιδράσεις, το κίνητρο βασίζεται συνήθως στην προβλεπόμενη χρήση ενός συστήματος από τον σχεδιαστή. Τέλος, πρέπει να υπάρχει η πιθανότητα ο εμπιστευόμενος να αποτύχει να εκτελέσει το έργο, προσκαλώντας την αβεβαιότητα και τον κίνδυνο [15]. Αυτά τα στοιχεία σκιαγραφούν την ιδέα ότι η εμπιστοσύνη είναι απαραίτητη όταν ανταλλάσσεται κάτι σε μια σχέση συνεργασίας που χαρακτηρίζεται από αβεβαιότητα. Αυτό ισχύει τόσο για τις διαπροσωπικές σχέσεις όσο και για τις σχέσεις μεταξύ ανθρώπου και αυτοματισμού. Παρόλα αυτά, παρόλο που η σημασία της εμπιστοσύνης στις συνεργατικές σχέσεις είναι γενικά αποδεκτή, παραμένουν ασυνέπειες όσον αφορά τον ακριβή ορισμό της εμπιστοσύνης.

2.1 Η εμπιστοσύνη στην συνεργασία ανθρώπου-ρομπότ

Οι πρώτες έρευνες για την εμπιστοσύνη προσπάθησαν να την ορίσουν με μια γενική έννοια, χωρίς κανένα πλαίσιο. Ο Rotter [16] ξεκίνησε περιγράφοντας την εμπιστοσύνη ως προδιάθεση

προς τον κόσμο και τους ανθρώπους σε αυτόν. Από τότε ο ορισμός αυτός έχει γίνει πιο συγκεκριμένος ως προς το περιεχόμενο και την κατάσταση. Ο Barber [17] θεώρησε τη διαπροσωπική εμπιστοσύνη ως ένα σύνολο κοινωνικά μαθημένων προσδοκιών που ποικίλλουν ανάλογα με την κοινωνική τάξη, ενώ οι Pruitt και Rubin [18] βλέπουν την εμπιστοσύνη ως έναν αριθμό πεποιθήσεων για τους άλλους. Ενώ οι εν λόγω συγγραφείς θεωρούν την εμπιστοσύνη ως πεποίθηση ή στάση, άλλοι συγγραφείς την έχουν ορίσει ως προθυμία αποδοχής της τρωτότητας [11] και ως συμπεριφορική κατάσταση ευαλωτότητας [19]. Είναι σαφές ότι οι μελετητές απέχουν πολύ από το να καταλήξουν σε ομοφωνία για έναν ενιαίο ορισμό της εμπιστοσύνης. Για την παρούσα εργασία, θα βασιστούμε στον ορισμό των Lee και See [1] για την εμπιστοσύνη ως ”τη στάση ότι ένας πράκτορας θα βοηθήσει στην επίτευξη των στόχων ενός ατόμου σε μια κατάσταση που χαρακτηρίζεται από αβεβαιότητα και τρωτότητα”.

Η εμπιστοσύνη βασίζεται πάντα σε τουλάχιστον μία ιδιότητα ή χαρακτηριστικό του εμπιστευόμενου (trustee). Σε μια ενδελεχή ανασκόπηση της βιβλιογραφίας για την εμπιστοσύνη, οι Mayer et al. [11] καθόρισαν τρεις γενικές βάσεις της εμπιστοσύνης: ικανότητα, ακεραιότητα και αγαθότητα. Η σταθερότητα της εμπιστοσύνης ενός ατόμου ποικίλλει αναλόγως σε ποια από τις παραπάνω ιδιότητες αναφέρεται. Για παράδειγμα, εάν η εμπιστοσύνη βασίζεται στην ικανότητα ενός εμπιστευόμενου (trustee), θα ποικίλλει ανάλογα με το πόσο καλά ο εμπιστευόμενος εκτελεί ένα έργο. Η εμπιστοσύνη που βασίζεται στην ακεραιότητα του έμπιστου δεν εξαρτάται από την πραγματική απόδοση του έμπιστου, αλλά από τον βαθμό στον οποίο οι πράξεις του έμπιστου ανταποκρίνονται στις αξίες αυτού που τον εμπιστεύεται (truster). Τέλος, η σταθερότητα της εμπιστοσύνης που βασίζεται στην αγαθότητα εξαρτάται από το κατά πόσον οι ενέργειες του εμπιστευόμενου ταιριάζουν με τους στόχους και τα κίνητρα του ”truster” [1]. Όταν η εμπιστοσύνη βασίζεται κυρίως στην ακεραιότητα ή την καλοσύνη του εμπιστευόμενου, οι κακές επιδόσεις από μόνες τους δεν θα την βλάψουν σημαντικά.

Αν και η εμπιστοσύνη στην τεχνολογία είναι διαφορετική από τη διαπροσωπική εμπιστοσύνη, υπάρχουν ομοιότητες μεταξύ των δύο. Στο πιο θεμελιώδες επίπεδο, οι δύο τύποι εμπιστοσύνης είναι παρόμοιοι στο ότι εκφράζουν στάσεις που σχετίζονται με την κάθε κατάσταση και είναι απαραίτητες μόνο όταν ανταλλάσσεται κάτι σε μια συνεργατική σχέση που χαρακτηρίζεται από αβεβαιότητα. Πέρα από αυτή την εννοιολογική ομοιότητα, οι έρευνες έχουν εντοπίσει πιο συγκεκριμένες ομοιότητες. Για παράδειγμα, αρκετές μελέτες στη δεκαετία του ’90 έδειξαν ότι οι άνθρωποι εφαρμόζουν κοινωνικούς κανόνες που έχουν διδαχθεί, όπως η ευγένεια, στις αλληλεπι-

δράσεις με τις μηχανές [20]. Επιπλέον, νευρολογικές έρευνες υποδεικνύουν ότι ορισμένοι από τους ίδιους νευρικούς μηχανισμούς που χρησιμοποιούν οι συμμετέχοντες στα διαπροσωπικά παιχνίδια εμπιστοσύνης χρησιμοποιούνται και στις αξιολογήσεις με βάση την εμπιστοσύνη των προσφορών στον ιστότοπο του eBay [21]. Ένας πιθανός λόγος για αυτές τις ομοιότητες είναι ότι σε κάποιο βαθμό, η εμπιστοσύνη των ανθρώπων στα τεχνολογικά συστήματα αντιπροσωπεύει την εμπιστοσύνη τους στους σχεδιαστές αυτών των συστημάτων [10]. Με αυτόν τον τρόπο,η εμπιστοσύνη ανθρώπου-αυτοματισμού μπορεί να θεωρηθεί ως ένας ειδικός τύπος διαπροσωπικής εμπιστοσύνης, στον οποίο ο trustee απέχει ένα βήμα από τον truster[22]. Ανεξάρτητα από τον τρόπο με τον οποίο εννοιολογείται η εμπιστοσύνη στον ανθρώπινο αυτοματισμό, υπάρχουν σημαντικές διαφορές μεταξύ αυτής και της διαπροσωπικής εμπιστοσύνης όσον αφορά το σε τι βασίζεται και πώς διαμορφώνεται[23].

2.1.1 Διαφορές μεταξύ της εμπιστοσύνης ανθρώπου-αυτοματισμού και διαπροσωπικής εμπιστοσύνης

Η εμπιστοσύνη ανθρώπου-αυτοματισμού και η διαπροσωπική εμπιστοσύνη εξαρτώνται από διαφορετικά χαρακτηριστικά. Ενώ η διαπροσωπική εμπιστοσύνη μπορεί να βασίζεται στην ικανότητα, την ακεραιότητα ή την καλοσύνη ενός προσώπου που εμπιστεύεται [11], η εμπιστοσύνη ανθρώπου-αυτοματισμού εξαρτάται από την απόδοση, τη διαδικασία ή το σκοπό ενός αυτοματοποιημένου συστήματος [24]. Η εμπιστοσύνη που βασίζεται στην απόδοση, παρόμοια με την εμπιστοσύνη που βασίζεται στην ικανότητα των Mayer κ.ά., ποικίλλει ανάλογα με το πόσο καλά ένα αυτοματοποιημένο σύστημα εκτελεί μια εργασία. Η εμπιστοσύνη με βάση τη διεργασία, ανάλογη με την εμπιστοσύνη με βάση την ακεραιότητα στους ανθρώπους, κυμαίνεται ανάλογα με την κατανόηση του χειριστή των μεθόδων που χρησιμοποιεί ένα αυτόματο σύστημα για την εκτέλεση εργασιών. Τέλος, η εμπιστοσύνη βάσει στόχου εξαρτάται από την προβλεπόμενη χρήση ενός αυτοματοποιημένου συστήματος από τον σχεδιαστή.

Η εξέλιξη του σχηματισμού διαπροσωπικής εμπιστοσύνης διαφέρει επίσης από εκείνη της εμπιστοσύνης ανθρώπου-αυτοματισμού [23]. Οι Rempel, Holmes και Zanna [25] εξηγούν ότι η διαπροσωπική εμπιστοσύνη βασίζεται αρχικά στην απλή προβλεψιμότητα των ενεργειών του έμπιστου, επειδή οι άνθρωποι τείνουν να συνάπτουν σχέσεις με αγνώστους με επιφυλακτικότητα. Καθώς οι διαπροσωπικές σχέσεις εξελίσσονται, η αξιοπιστία ή η ακεραιότητα ενός διαχειριστή γίνεται η βασική βάση της εμπιστοσύνης. Τέλος, οι πλήρως ώριμες διαπροσωπικές σχέσεις βασίζονται στην πίστη ή την καλοσύνη [1]. Από την άλλη πλευρά, η εμπιστοσύνη μεταξύ ανθρώπου και αυτοματισμού συχνά εξελίσσεται με αντίστροφη σειρά. Τα στοιχεία δείχνουν ότι οι άνθρωποι συχνά εμφανίζουν μια θετική προδιάθεση στην εμπιστοσύνη τους προς τα νέα αυτοματοποιημένα συστήματα [26]. Οι άνθρωποι συνήθως υποθέτουν ότι οι μηχανές είναι τέλειες. Ως εκ τούτου, η αρχική τους εμπιστοσύνη βασίζεται στην πίστη. Ωστόσο, αυτή η εμπιστοσύνη διαλύεται γρήγορα μετά από σφάλματα του συστήματος- καθώς οι σχέσεις με τα αυτοματοποιημένα συστήματα εξελίσσονται, η αξιοπιστία και η προβλεψιμότητα αντικαθιστούν την πίστη ως πρωταρχική βάση της εμπιστοσύνης [23].

2.1.2 Παράγοντες επιρροής της εμπιστοσύνης

Σύμφωνα με τους Hoff και Bashir [22] υπάρχουν τρεις βασικές πηγές διαφοροποίησης της εμπιστοσύνης ανθρώπου-αυτοματισμού: τον άνθρωπο-χειριστή, το περιβάλλον και το αυτοματοποιημένο σύστημα. Αυτές οι μεταβλητές αντικατοπτρίζουν αντίστοιχα τα τρία διαφορετικά επίπεδα εμπιστοσύνης που προσδιορίστηκαν από τους Marsh και Dibben [27]: εμπιστοσύνη προδιάθεσης (dispositional trust), εμπιστοσύνη κατάστασης (situational trust) και επίκτητη εμπιστοσύνη (learned trust). Η εμπιστοσύνη προδιάθεσης αντιπροσωπεύει τη διαρκή τάση του ατόμου να εμπιστεύεται τον αυτοματισμό. Η περιστασιακή εμπιστοσύνη, από την άλλη πλευρά, εξαρτάται από το

συγκεκριμένο πλαίσιο μιας αλληλεπίδρασης. Το περιβάλλον επηρεάζει σε μεγάλο βαθμό την περιστασιακή εμπιστοσύνη, αλλά οι εξαρτώμενες από το πλαίσιο μεταβολές στη νοητική κατάσταση ενός χειριστή μπορούν επίσης να μεταβάλουν την εμπιστοσύνη στην κατάσταση. Η τελευταία κατηγορία, η επίκτητη εμπιστοσύνη, βασίζεται σε προηγούμενες εμπειρίες που σχετίζονται με ένα συγκεκριμένο αυτοματοποιημένο σύστημα. Η αποκτηθείσα εμπιστοσύνη συνδέεται στενά με την καταστασιακή εμπιστοσύνη, καθώς καθοδηγείται από προηγούμενες εμπειρίες [27]- η διάκριση μεταξύ των δύο εξαρτάται από το αν η εμπειρία που καθοδηγεί την εμπιστοσύνη είναι σχετική με το αυτοματοποιημένο σύστημα (αποκτηθείσα εμπιστοσύνη) ή με το περιβάλλον (καταστασιακή εμπιστοσύνη). Αν και τα τρία επίπεδα εμπιστοσύνης είναι αλληλένδετα, θα εξεταστούν χωριστά σε αυτό το τμήμα, ξεκινώντας με τη διαθετική εμπιστοσύνη (dispositional trust).

Εμπιστοσύνη προδιάθεσης - Dispositional Trust

Όπως και στον διαπροσωπικό τομέα, οι άνθρωποι παρουσιάζουν μεγάλη ποικιλομορφία στην τάση τους να εμπιστεύονται τον αυτοματισμό. Σε αντίθεση με τα χαρακτηριστικά που εξαρτώνται από το πλαίσιο, όπως η διάθεση και η αυτοπεποίθηση, οι διαφοροποιήσεις της διάθεσης μπορούν να μεταβάλουν το σχηματισμό εμπιστοσύνης σε κάθε κατάσταση. Η προδιάθεση εμπιστοσύνης αντιπροσωπεύει τη συνολική τάση ενός ατόμου να εμπιστεύεται τον αυτοματισμό, ανεξάρτητα από το πλαίσιο ή ένα συγκεκριμένο σύστημα. Ενώ η έρευνα σχετικά με τη βιολογία της εμπιστοσύνης έχει δείξει ότι η γενετική παίζει σημαντικό ρόλο στον καθορισμό της διαπροσωπικής τάσης εμπιστοσύνης [28], χρησιμοποιείται ο όρος προδιάθεση εμπιστοσύνης για να αναφερθούμε σε μακροπρόθεσμες τάσεις που προκύπτουν τόσο από βιολογικές όσο και από περιβαλλοντικές επιδράσεις. Έτσι, το καθοριστικό χαρακτηριστικό της προδιάθεσης εμπιστοσύνης είναι ότι πρόκειται για ένα σχετικά σταθερό χαρακτηριστικό με την πάροδο του χρόνου, σε αντίθεση με την περιστασιακή και την επίκτητη εμπιστοσύνη. Η έρευνά των Hoff & Bashir [22] αποκάλυψε τέσσερις πρωταρχικές πηγές μεταβλητότητας σε αυτό το πιο βασικό επίπεδο εμπιστοσύνης: κουλτούρα, ηλικία, φύλο και προσωπικότητα.

- **Κουλτούρα:** Η κουλτούρα είναι μια ιδιαίτερα σημαντική μεταβλητή, επειδή είναι κάτι με το οποίο σχεδόν όλοι ταυτίζονται. Στον διαπροσωπικό τομέα, σημαντικές έρευνες έχουν δείξει ότι η εμπιστοσύνη διαφέρει μεταξύ χωρών, φυλών, θρησκειών και γενεαλογικών ομάδων [29]. Μέχρι σήμερα, ωστόσο, πολύ λίγες μελέτες έχουν επικεντρωθεί στο ρόλο της κουλτούρας στην εμπιστοσύνη στην αυτοματοποίηση. Σε μια σύγχρονη μελέτη, οι Huerta, Glandon και Petrides [30] διαπίστωσαν ότι οι Μεξικανοί είναι πιο πιθανό να εμπιστεύονται τα αυτοματοποιημένα βοηθήματα λήψης αποφάσεων και λιγότερο πιθανό να εμπιστεύονται τα χειροκίνητα βοηθήματα λήψης αποφάσεων, σε σύγκριση με τους Αμερικανούς. Αρκετές μελέτες έχουν επίσης διαπιστώσει πολιτισμικές διαφορές στον τρόπο με τον οποίο οι άνθρωποι αντιλαμβάνονται τα κοινωνικά ρομπότ [31]. Ενώ οι μελέτες αυτές υποδηλώνουν ότι οι μεταβλητές που βασίζονται στην κουλτούρα μπορεί να έχουν σχέση με την εμπιστοσύνη στην αυτοματοποίηση, χρειάζονται περισσότερες έρευνες.

- **Ηλικία:** Οι ηλικιακές διαφορές στην εμπιστοσύνη στον αυτοματισμό μπορεί να είναι αποτέλεσμα γνωστικών αλλαγών, του φαινομένου της κοορτής (cohort effect) ή κάποιου συνδυασμού και των δύο μεταβλητών [32]. Αρκετές μελέτες έχουν διαπιστώσει ότι η ηλικία αποτελεί σημαντική μεταβλητή. Οι Ho, Wheatley και Scialfa [33] έδειξαν ότι οι μεγαλύτεροι σε ηλικία ενήλικες εμπιστεύονται και βασίζονται στα βοηθήματα λήψης αποφάσεων περισσότερο από τους νεότερους ενήλικες, αλλά δεν διαμορφώνουν διαφορετικά την εμπιστοσύνη τους μετά από σφάλματα αυτοματισμού. Αντίθετα, οι Sanchez, Fisk και Rogers [34] διαπίστωσαν ότι οι μεγαλύτεροι ενήλικες ήταν καλύτεροι στο να ρυθμίζουν την εμπιστοσύνη τους ανάλογα με την μεταβαλλόμενη αξιοπιστία ενός συστήματος υποστήριξης αποφάσεων. Μια

πρόσφατη μελέτη των Pak, Fink, Price, Bass και Sturre [35] έδειξε ότι η προσθήκη της εικόνας ενός γιατρού στη διεπαφή μιας εφαρμογής διαχείρισης διαβήτη οδήγησε τους νεότερους συμμετέχοντες να εμπιστευτούν περισσότερο τις συμβουλές του συστήματος, αλλά δεν είχε καμία επίδραση στην εμπιστοσύνη των μεγαλύτερων σε ηλικία συμμετεχόντων. Συνολικά, αυτή η έρευνα και άλλα ευρήματα [36], [37], [38], [39] υποδηλώνουν ότι οι άνθρωποι διαφορετικών ηλικιών μπορεί να χρησιμοποιούν διαφορετικές στρατηγικές όταν αναλύουν την αξιοπιστία των αυτοματοποιημένων συστημάτων. Ωστόσο, η συγκεκριμένη επίδραση της ηλικίας πιθανότατα ποικίλλει ανάλογα το πλαίσιο.

- **Φύλο:** Συνεκτικές διαφορές μεταξύ των δύο φύλων δεν έχουν εμφανιστεί ακόμη σε μελέτες που επικεντρώνονται αποκλειστικά στην εμπιστοσύνη στον αυτοματισμό. Ωστόσο, έρευνες έχουν δείξει ότι το φύλο μπορεί να παίξει ρόλο στην καθοδήγηση των αλληλεπιδράσεων με άλλους τύπους τεχνολογίας. Για παράδειγμα, η E.J. Lee [40] διαπίστωσε ότι οι γυναίκες είναι επιρρεπείς στην κολακεία των υπολογιστών, ενώ οι άνδρες εμφανίζουν αρνητικές αντιδράσεις σε αυτήν. Αυτή η ασυμφωνία, καθώς και η έρευνα με βάση το φύλο για την αλληλεπίδραση ανθρώπου-ρομπότ [41] υποδηλώνει ότι οι άνδρες και οι γυναίκες μπορεί να αντιδρούν διαφορετικά στο στυλ επικοινωνίας και στην εμφάνιση ενός αυτοματοποιημένου συστήματος. Έτσι, παρόλο που δεν έχουν βρεθεί συνεπείς διαφορές, το δυνητικό φύλο του χειριστή ενός αυτοματοποιημένου συστήματος θα πρέπει να λαμβάνεται υπόψη στη διαδικασία σχεδιασμού ορισμένων συστημάτων

- **Προσωπικότητα:** Τα χαρακτηριστικά της προσωπικότητας ενός χειριστή είναι το τελευταίο συστατικό της εμπιστοσύνης προδιάθεσης. Στον διαπροσωπικό τομέα, η εμπιστοσύνη προδιάθεσης είναι καθαυτή, ένα σταθερό χαρακτηριστικό της προσωπικότητας που αντιπροσωπεύει την τάση ενός ατόμου να εμπιστεύεται άλλους ανθρώπους καθ' όλη τη διάρκεια της ζωής του. Υπάρχουν πολυάριθμες ψυχομετρικές κλίμακες που μετρούν αυτή την τάση [25] και οι έρευνες έχουν δείξει ότι αποτελεί σημαντικό προσδιοριστικό παράγοντα της ανθρώπινης συμπεριφοράς [42].

Αρκετοί ερευνητές έχουν προσπαθήσει να εφαρμόσουν την έννοια της προδιαθετικής εμπιστοσύνης στην αλληλεπίδραση ανθρώπου-αυτοματισμού σε μια προσπάθεια να διαφοροποιήσουν τους ανθρώπους με βάση την τάση τους να εμπιστεύονται τα αυτοματοποιημένα συστήματα. Οι Biros, Fields και Gunsch [43] έδειξαν ότι οι συμμετέχοντες με μεγαλύτερη προδιάθεση εμπιστοσύνης στους υπολογιστές έδειξαν μεγαλύτερη εμπιστοσύνη στις πληροφορίες από ένα μη επανδρωμένο πολεμικό αεροσκάφος. Σε μια μεταγενέστερη μελέτη, οι Merritt και Ilgen [44] διαπίστωσαν ότι η τάση για εμπιστοσύνη (trust prospensity) προέβλεπε την εμπιστοσύνη των συμμετεχόντων μετά την εργασία, έτσι ώστε όταν ένα εργαλείο είχε καλή απόδοση, τα άτομα με υψηλή τάση εμπιστοσύνης είχαν περισσότερες πιθανότητες να εμπιστευτούν περισσότερο το εργαλείο. Αντίθετα, όταν το βοήθημα είχε κακή απόδοση, τα άτομα με χαμηλή τάση εμπιστοσύνης είχαν περισσότερες πιθανότητες να εκφράσουν μεγαλύτερη εμπιστοσύνη στο βοήθημα. Τα αποτελέσματα αυτά υποδηλώνουν ότι τα άτομα με υψηλά επίπεδα προδιαθετικής εμπιστοσύνης στον αυτοματισμό έχουν μεγαλύτερη τάση να εμπιστεύονται αξιόπιστα συστήματα, αλλά η εμπιστοσύνη τους μπορεί να μειωθεί σημαντικότερα μετά από σφάλματα του συστήματος.

Η έρευνα έχει επίσης καταλήξει σε συσχετίσεις μεταξύ διαφόρων πιο συγκεκριμένων χαρακτηριστικών της προσωπικότητας και της εμπιστοσύνης. Όσον αφορά τα πέντε μεγάλα χαρακτηριστικά προσωπικότητας, οι Szalma και Taylor [45] διαπίστωσαν ότι ο νευρωτισμός συσχετίζεται αρνητικά με τη συμφωνία με τις σωστές συμβουλές αυτοματισμού αλλά δεν βρήκαν άλλες συσχετίσεις με άλλα χαρακτηριστικά προσωπικότητας. Οι Merritt και Ilgen [44] έδειξαν ότι οι εξωστρεφείς τείνουν να εμπιστεύονται περισσότερο τις μηχανές από ό,τι οι εσωστρεφείς. Οι Mcbride, Carter και Ntuen [46] χρησιμοποίησαν το Minnesota Multiphasic Personality Inventory (MMPI) για να δείξουν ότι οι νοσηλευτές με διαισθητική προσωπικότητα είχαν περισσότερες πιθανότη-

τες να αποδεχτούν διαγνώσεις από ένα αυτοματοποιημένο βοήθημα λήψης αποφάσεων από ό,τι οι νοσηλευτές με αισθητήρια προσωπικότητα. Αυτές οι τρεις μελέτες υποδηλώνουν ότι οι χειριστές μπορεί να είναι πιο πιθανό να εμπιστευτούν ή να βασιστούν στην αυτοματοποίηση όταν είναι εξωστρεφείς, συναισθηματικά σταθεροί και έχουν διαισθητική παρά αισθητήρια προσωπικότητα

Σε γενικές γραμμές, η έρευνα που συζητήθηκε σε αυτό το τμήμα δείχνει ότι υπάρχουν σημαντικές ατομικές διαφορές στη διάθεση των ανθρώπων να εμπιστεύονται τα αυτοματοποιημένα συστήματα. Ενώ ορισμένες από τις μεταβλητές αυτής της ενότητας μπορούν να μεταβληθούν σταδιακά με την πάροδο του χρόνου (π.χ. πολιτισμικές αξίες, ηλικία και χαρακτηριστικά προσωπικότητας), είναι γενικά σταθερές κατά τη διάρκεια μιας αλληλεπίδρασης. Συνεπώς, η εμπιστοσύνη προδιάθεσης δημιουργεί διακυμάνσεις στην εμπιστοσύνη μεταξύ των αλληλεπιδράσεων με διαφορετικούς χειριστές, δεδομένου ότι η κατάσταση και το αυτοματοποιημένο σύστημα δεν αλλάζουν. Προκειμένου να προσαρμοστούν καλύτερα οι τάσεις εμπιστοσύνης των διαφορετικών χειριστών, οι σχεδιαστές του συστήματος θα πρέπει να εξετάσουν το ενδεχόμενο εφαρμογής χαρακτηριστικών που μπορούν να προσαρμοστούν στις προτιμήσεις και τις τάσεις των διαφορετικών ατόμων.

Περιστασιακή εμπιστοσύνη - Situational Trust

Η ανάπτυξη της εμπιστοσύνης, καθώς και η σημασία της όσον αφορά τη συμπεριφορά, ποικίλλει σε μεγάλο βαθμό ανάλογα με την κατάσταση. Η ανάλυσή των Hoff και Bashir [22] αποκάλυψε δύο ευρείες πηγές μεταβλητότητας στην περιστασιακή εμπιστοσύνη: το εξωτερικό περιβάλλον και τα εσωτερικά, εξαρτώμενα από το πλαίσιο, χαρακτηριστικά του χειριστή.

- **Εξωγενείς διακυμάνσεις:** Η εμπιστοσύνη ενός χειριστή σε ένα αυτοματοποιημένο σύστημα εξαρτάται σε μεγάλο βαθμό από τον τύπο του συστήματος, την πολυπλοκότητά του και τη δυσκολία της εργασίας για την οποία χρησιμοποιείται [47], [48], [49], [50], [51], [52]. Όπως οι άνθρωποι, έτσι και οι μηχανές έχουν ξεχωριστά πλεονεκτήματα και αδυναμίες. Οι ανθρώπινοι χειριστές λαμβάνουν υπόψη τη σχετική δυσκολία των εργασιών όταν αξιολογούν τις δυνατότητες των αυτοματοποιημένων συστημάτων να τις ολοκληρώσουν [49]. Ωστόσο, μπορεί να προκύψουν προβλήματα όταν οι χειριστές δεν αναγνωρίζουν ότι ένα και μόνο σύστημα μπορεί να αποδίδει με ασυνέπεια σε δύο ανεξάρτητες εργασίες ή ότι δύο συστήματα μπορεί να αποδίδουν με διακυμάνσεις σε πανομοιότυπες εργασίες. Για παράδειγμα, σε ένα πείραμα, οι συμμετέχοντες χρησιμοποίησαν δύο βοηθήματα λήψης αποφάσεων είτε με μικτή αξιοπιστία είτε με ομοιόμορφη αξιοπιστία σε ένα έργο αναζήτησης και διάσωσης με βάση βίντεο. Οι συμμετέχοντες αξιολόγησαν την αξιοπιστία του λιγότερο αξιόπιστου βοηθήματος σημαντικά υψηλότερα όταν αυτό συνδυάστηκε με ένα πιο αξιόπιστο βοήθημα, σε σύγκριση με όταν αυτό συνδυάστηκε με ένα βοήθημα της ίδιας χαμηλής αξιοπιστίας [51]. Αυτό είναι ένα παράδειγμα μεροληπτικής αντίληψης που μπορεί να εμφανιστεί λόγω εξωτερικών καταστασιακών παραγόντων (π.χ. παρουσία άλλου βοηθήματος λήψης αποφάσεων)

  Ο φόρτος εργασίας ενός χειριστή καθορίζει συχνά το χρόνο και τη γνώση που μπορεί να δαπανηθεί για την παρακολούθηση της αυτοματοποίησης. Για το λόγο αυτό, ο φόρτος εργασίας είναι μια σημαντική μεταβλητή που μπορεί να μεταβάλει τη δυναμική της αλληλεπίδρασης ανθρώπου-αυτοματισμού. Εμπειρικές έρευνες έχουν δείξει ότι ο φόρτος εργασίας της κύριας εργασίας μπορεί να επηρεάσει τόσο τις συμπεριφορές που βασίζονται στην εμπιστοσύνη ([53]- [39] όσο και την αυτο-αναφερόμενη εμπιστοσύνη [54]- [55]. Ειδικότερα, ο υψηλότερος φόρτος εργασίας φαίνεται να έχει μέτρια επίδραση στη θετική σχέση μεταξύ εμπιστοσύνης και εξάρτησης [54], [53], [55]). Ο λόγος μπορεί να είναι ότι υπό υψηλό εργασιακό φόρτο, οι χειριστές πρέπει να χρησιμοποιούν συχνότερα τον αυτοματισμό για να συμβαδίζουν με τις απαιτήσεις των εργασιών, ανεξάρτητα από το επίπεδο εμπιστοσύνης τους [54]. Άλλες έρευνες έχουν διαπιστώσει ότι ορισμένοι τύποι περισπασμών (π.χ. οπτικο-χωρικός και ακουστικο-λεκτικός) μπορούν να προκαλέσουν μειωμένη εμπιστοσύνη

στα αυτοματοποιημένα βοηθήματα λήψης αποφάσεων, ενώ άλλοι τύποι περισπασμών (π.χ. οπτικο-λεκτικός και ακουστικο-χωρικός) μπορούν στην πραγματικότητα να αυξήσουν την εμπιστοσύνη [56]. Σε μια μελέτη, ωστόσο, οι περισπασμοί δεν είχαν καμία επίδραση στην εμπιστοσύνη των συμμετεχόντων σε ένα σύστημα προειδοποίησης σύγκρουσης [57]. Αυτά τα αντικρουόμενα ευρήματα υποδηλώνουν ότι οι αποσπασματικοί παράγοντες μπορεί να έχουν επιρροή, αλλά η επίδρασή τους, αν υπάρχει, εξαρτάται από τον βαθμό στον οποίο παρεμβαίνουν στην παρακολούθηση του συστήματος. Εάν η παρουσία ενός αντιπερισπασμού κάνει έναν χειριστή να παραβλέψει ένα σφάλμα αυτοματισμού, η εμπιστοσύνη μπορεί να αυξηθεί. Από την άλλη πλευρά, οι αντιπερισπασμοί που δεν παρεμβαίνουν σημαντικά στην παρακολούθηση του συστήματος πιθανόν να μην έχουν καμία επίδραση στην εμπιστοσύνη ή μπορεί να προκαλέσουν ελαφρά μείωση της εμπιστοσύνης (π.χ. αν τα σφάλματα αυτοματισμού γίνουν πιο εμφανή).

Το περιβάλλον είναι επίσης σημαντικό, διότι συμβάλλει στον καθορισμό των πιθανών κινδύνων και οφελών που συνδέονται με τη χρήση του αυτοματισμού. Επειδή η εμπιστοσύνη είναι πάντα σχετική με μια αβεβαιότητα, οι αντιλήψεις για τον κίνδυνο παίζουν κρίσιμο ρόλο στην ανάπτυξη της εμπιστοσύνης. Οι Perkins, Miller, Hashemi και Burns [58] διαπίστωσαν ότι οι συμμετέχοντες εμπιστεύονταν και χρησιμοποιούσαν λιγότερο τις συμβουλές GPS για τον σχεδιασμό διαδρομής όταν το ρίσκο αυξανόταν μέσω της παρουσίας κινδύνων κατά την οδήγηση. Αυτό το συμπέρασμα και άλλες έρευνες[37] υποδηλώνουν ότι οι άνθρωποι τείνουν να μειώνουν την εμπιστοσύνη τους στον αυτοματισμό όταν υπάρχει μεγαλύτερος κίνδυνος. Ωστόσο, το αντίστροφο αποτέλεσμα μπορεί να συμβεί όταν η χρήση του αυτοματισμού προσφέρει μεγαλύτερα οφέλη και ενέχει λιγότερους πιθανούς κινδύνους. Αυτό μπορεί να φανεί στο πείραμα των Lyons και Stokes [59], όπου στους συμμετέχοντες δόθηκε βοήθεια τόσο από έναν άνθρωπο-βοηθό όσο και από ένα αυτοματοποιημένο εργαλείο σε ένα έργο λήψης αποφάσεων. Οι ερευνητές ανακάλυψαν ότι οι συμμετέχοντες βασίζονταν λιγότερο στην ανθρώπινη βοήθεια όταν έπαιρναν αποφάσεις υψηλού κινδύνου, γεγονός που υποδηλώνει μια τάση προς την αυτοματοποίηση

Τα αντιφατικά ευρήματα που συζητήθηκαν παραπάνω μπορεί να οφείλονται σε διαφορές στον τύπο και την πολυπλοκότητα του αυτοματισμού που χρησιμοποιήθηκε στα διάφορα πειράματα. Συγχρόνως, τα αυτοματοποιημένα βοηθήματα στα οποία στηρίχθηκαν λιγότερο σε συνθήκες υψηλού κινδύνου παρείχαν συμβουλές για τη λήψη αποφάσεων ή ελεγχόμενες ενέργειες [37],[58], το αυτοματοποιημένο εργαλείο στο οποίο στηρίχθηκαν περισσότερο σε συνθήκες υψηλού κινδύνου εμφάνιζε πληροφορίες [59]. Έτσι, οι συμμετέχοντες στο τελευταίο πείραμα αλληλεπιδρούσαν με μια πιο βασική μορφή αυτοματοποίησης από ό,τι οι πρώτοι συμμετέχοντες. Αυτό υποδηλώνει ότι υπό συνθήκες υψηλού κινδύνου, οι χειριστές μπορεί να έχουν την τάση να μειώνουν την εξάρτησή τους από τον πολύπλοκο αυτοματισμό αλλά να αυξάνουν την εξάρτησή τους από τον απλό αυτοματισμό

Οι μελέτες που εξετάστηκαν σε αυτή την ενότητα δείχνουν πώς οι εξωγενείς περιβαλλοντικοί παράγοντες μπορούν να μεταβάλουν τη δυναμική της εμπιστοσύνης ανθρώπου-ρομπότ. Ωστόσο, οι εξωτερικοί παράγοντες αποτελούν μόνο ένα μέρος της κατά περίσταση εμπιστοσύνης - οι ενδογενείς μεταβολές σε παράγοντες όπως η αυτοπεποίθηση, η εμπειρογνωμοσύνη, η διάθεση και η ικανότητα προσοχής μπορούν επίσης να μεταβάλουν την κατά περίσταση εμπιστοσύνη στον αυτοματισμό.

- **Ενδογενείς διακυμάνσεις:** Ενώ η προδιαθετική εμπιστοσύνη (dispositional trust) καλύπτει τα μόνιμα χαρακτηριστικά των φορέων, τα άτομα διαφέρουν επίσης σε πιο μεταβατικά χαρακτηριστικά που εξαρτώνται από το εκάστοτε πλαίσιο. Η αυτοπεποίθηση, η οποία συχνά ποικίλλει ανάλογα με τα καθήκοντα, είναι μια μεταβλητή με ιδιαίτερη επιρροή που καθοδηγεί τον σχηματισμό εμπιστοσύνης [1]. Παίζει επίσης σημαντικό ρόλο στις διαδικασίες λήψης αποφάσεων που σχετίζονται με την κατανομή του ελέγχου. Για παράδειγμα, οι de

Vries, Midden και Bouwhuis [60] διαπίστωσαν ότι η σχέση μεταξύ της εμπιστοσύνης και της αυτοπεποίθησης προέβλεπε τις αποφάσεις των συμμετεχόντων να εκτελέσουν ένα έργο σχεδιασμού διαδρομής χειροκίνητα ή με τη βοήθεια ενός αυτοματοποιημένου μέσου. Όταν η εμπιστοσύνη των συμμετεχόντων ήταν υψηλή και η αυτοπεποίθηση χαμηλή, ο αυτοματισμός χρησιμοποιούνταν συχνότερα. Το αντίθετο αποτέλεσμα εμφανίστηκε όταν η εμπιστοσύνη ήταν χαμηλή και η αυτοπεποίθηση υψηλή- αλλά συνολικά, οι συμμετέχοντες εμφάνισαν μια μικρή τάση προς τον χειροκίνητο έλεγχο. Το εύρημα αυτό υποδηλώνει ότι όταν η αυτοπεποίθηση και η εμπιστοσύνη είναι περίπου ίσες, οι χειριστές μπορεί να προτιμούν τον χειροκίνητο έλεγχο [60]. Μια πιο συγκεκριμένη πτυχή της αυτοπεποίθησης είναι η ηλεκτρονική αυτεπάρκεια, η οποία μπορεί να οριστεί ως ”η κρίση της ικανότητας κάποιου να χρησιμοποιεί έναν υπολογιστή” [61]. Οι έρευνες δείχνουν ότι η μεταβλητή αυτή συνδέεται θετικά με την εμπιστοσύνη στην αυτοματοποίηση [62]

Η εμπειρογνωμοσύνη μπορεί επίσης να αλλάξει την εμπιστοσύνη προς τον αυτοματισμό. Η τεχνογνωσία είναι συνήθως αποτέλεσμα εκτεταμένης εμπειρίας σε έναν τομέα και συχνά οδηγεί σε μεγαλύτερη αυτοπεποίθηση. Έρευνες έχουν δείξει ότι τα άτομα με μεγαλύτερη εξειδίκευση στο αντικείμενο είναι λιγότερο πιθανό να βασιστούν στον αυτοματισμό από ό,τι οι αρχάριοι χειριστές [48] [63]. Για παράδειγμα, σε ένα πείραμα, νέοι ενήλικες με εμπειρία στον χειρισμό γεωργικών οχημάτων ήταν πιο απρόθυμοι να βασιστούν σε αυτόματους συναγερμούς κατά τη διάρκεια μιας εργασίας αποφυγής σύγκρουσης από ό,τι ήταν νέοι ενήλικες με μικρή ή καθόλου εμπειρία στον τομέα της γεωργίας [63]

Παρόλο που η τεχνογνωσία συχνά προκύπτει από την εμπειρία, δεν πρέπει να συγχέεται με την εμπειρία που σχετίζεται με ένα συγκεκριμένο αυτοματοποιημένο σύστημα. Η εμπειρογνωμοσύνη εδώ αναφέρεται στην κατανόηση ενός συγκεκριμένου τομέα (π.χ. χειρισμός γεωργικών οχημάτων). Η γνώση που σχετίζεται με έναν συγκεκριμένο τύπο αυτοματοποιημένου συστήματος μπορεί να έχει εντελώς διαφορετική επίδραση στην εμπιστοσύνη και την αξιοπιστία

Το συναίσθημα είναι ένας άλλος παράγοντας που συμβάλλει στην εξήγηση του γιατί η εμπιστοσύνη αναπτύσσεται με ασυνέπεια σε διαφορετικά πλαίσια. Οι Stokes κ.ά. [64] διαπίστωσαν ότι οι συμμετέχοντες με θετική διάθεση εξέφρασαν μεγαλύτερη αρχική εμπιστοσύνη σε ένα αυτοματοποιημένο βοήθημα λήψης αποφάσεων. Οι προϋπάρχουσες διαθέσεις επηρέασαν μόνο την αρχική εμπιστοσύνη σε αυτό το πείραμα, αλλά η Merritt [65] διαπίστωσε ότι οι συμμετέχοντες που είχαν εμμέσως προετοιμαστεί να διαθέτουν το συναίσθημα της ευτυχίας ήταν πιο πιθανό να εμπιστευτούν έναν αυτοματοποιημένο ανιχνευτή όπλων καθ’ όλη τη διάρκεια του πειράματός της. Ενώ τα δύο παραπάνω πειράματα εξέτασαν διαφορετικές παραλλαγές της διάθεσης και κατέληξαν σε ελαφρώς διαφορετικά συμπεράσματα, και τα δύο υποδηλώνουν ότι η αρχική συναισθηματική κατάσταση ενός ατόμου μπορεί να μεταβάλει τη διαδικασία σχηματισμού εμπιστοσύνης

Μια τελευταία εξαρτώμενη από το πλαίσιο μεταβλητή που σχετίζεται με την εμπιστοσύνη είναι η ικανότητα προσοχής ενός χειριστή. Η ικανότητα προσοχής εξαρτάται συχνά από τον φόρτο εργασίας ενός χειριστή, αλλά και άλλους παράγοντες, όπως τα κίνητρα, το άγχος, η απώλεια ύπνου και η πλήξη. Μια πρόσφατη μελέτη των Onnasch, Manzey et al. [66] εξέτασε τις επιδράσεις της στέρησης ύπνου στις επιδόσεις ανθρώπου-αυτοματισμού κατά τη διάρκεια μιας εργασίας εποπτικού ελέγχου διεργασιών. Οι ερευνητές διαπίστωσαν ότι οι συμμετέχοντες με στέρηση ύπνου παρακολουθούσαν το αυτοματοποιημένο σύστημα πιο προσεκτικά και ήταν λιγότερο επιρρεπείς στην προκατάληψη του αυτοματισμού. Ωστόσο, οι εν λόγω συμμετέχοντες είχαν χειρότερες επιδόσεις σε μια δευτερεύουσα εργασία και ήταν πιο επιρρεπείς σε σφάλματα όταν επέστρεφαν στον χειροκίνητο έλεγχο. Αυτό υποδηλώνει ότι, ενώ οι χειριστές με στέρηση ύπνου μπορεί να είναι σε θέση να αντισταθμίσουν την κούρασή τους παρακολουθώντας πιο προσεκτικά την αυτοματοποίηση, είναι λιγότερο ικανοί

να κάνουν ταυτόχρονα πολλαπλές εργασίες.

Εν κατακλείδι, η καταστασιακή εμπιστοσύνη στον αυτοματισμό εξαρτάται τόσο από το εξωτερικό περιβάλλον όσο και από τα εσωτερικά, εξαρτώμενα από το πλαίσιο χαρακτηριστικά του ανθρώπινου χειριστή. Οι εξωτερικοί παράγοντες που μπορούν να επηρεάσουν την εμπιστοσύνη περιλαμβάνουν τον τύπο του συστήματος, την πολυπλοκότητα του συστήματος, την εργασία για την οποία χρησιμοποιείται ένα σύστημα, το οργανωτικό πλαίσιο μιας αλληλεπίδρασης και τον φόρτο εργασίας του χειριστή. Οι εσωτερικοί παράγοντες που μπορούν να επηρεάσουν την εμπιστοσύνη περιλαμβάνουν την αυτοπεποίθηση, την εμπειρογνωμοσύνη, τη διάθεση και την ικανότητα προσοχής.

- **Περιστασιακοί παράγοντες και η σχέση μεταξύ εμπιστοσύνης και εξάρτησης:** Εκτός από το άμεσο αντίκτυπο στην εμπιστοσύνη, οι περιστασιακοί παράγοντες παίζουν πρωταγωνιστικό ρόλο στον προσδιορισμό του βαθμού στον οποίο η εμπιστοσύνη επηρεάζει τη συμπεριφορά απέναντι στον αυτοματισμό. Οι Lee και See [1] υποστήριξαν ότι η εμπιστοσύνη έχει μεγαλύτερο αντίκτυπο στην εξάρτηση όταν η πολυπλοκότητα ενός συστήματος είναι υψηλή και όταν συμβαίνουν απρογραμμάτιστα γεγονότα που απαιτούν από τους χειριστές να προσαρμόσουν γρήγορα τη συμπεριφορά τους.

  Πρώτον, φαίνεται ότι η εμπιστοσύνη ασκεί μεγαλύτερη επιρροή στην εξάρτηση όταν το περιβάλλον παρέχει στους χειριστές μεγαλύτερη δυνατότητα να αξιολογήσουν την απόδοση της αυτοματοποίησης σε σχέση με τη μη υποβοηθούμενη χειροκίνητη απόδοση. Με άλλα λόγια, τα υποκειμενικά επίπεδα εμπιστοσύνης μπορεί να έχουν ασθενέστερη επίδραση στην εξάρτηση όταν οι χειριστές δεν είναι σε θέση να προσδιορίσουν το βαθμό στον οποίο η αυτοματοποίηση τους βοηθάει πραγματικά να εκτελέσουν μια εργασία. Τα ευρήματα από τη διατριβή του Spain [52] απεικονίζουν αυτή την ιδέα. Στο πείραμα, οι συμμετέχοντες χρησιμοποίησαν ένα βοήθημα λήψης αποφάσεων που εμφάνιζε πληροφορίες εμπιστοσύνης του συστήματος, ενώ εκτελούσαν μια εργασία αναγνώρισης πολεμικών επιχειρήσεων με δύο επίπεδα ποιότητας εικόνας: υψηλή και χαμηλή. Η εμπιστοσύνη και η συμμόρφωση μειώθηκαν ως συνάρτηση της εμπιστοσύνης του συστήματος στην κατάσταση υψηλής ποιότητας εικόνας, αλλά όχι στην κατάσταση χαμηλής ποιότητας εικόνας. Ο Spain αποδίδει αυτό το εύρημα στην σημαντικότητα των σφαλμάτων αυτοματισμού στις διαφορετικές συνθήκες ποιότητας εικόνας. Όταν η ποιότητα της εικόνας ήταν υψηλή, οι συμμετέχοντες ανέφεραν ότι τους ήταν ευκολότερο να εντοπίζουν στόχους με το χέρι και επομένως ευκολότερο να παρατηρούν τα σφάλματα του αυτοματισμού. Ωστόσο, όταν η ποιότητα της εικόνας ήταν χαμηλή, οι συμμετέχοντες δυσκολεύονταν περισσότερο να τα παρατηρήσουν [52]. Αυτή η μελέτη υπογραμμίζει τη σημασία της εξέτασης του περιβάλλοντος στο οποίο θα χρησιμοποιηθεί ένα σύστημα, προκειμένου οι σχεδιαστές, οι επόπτες και οι χειριστές να διευκολύνουν αποτελεσματικότερα τις κατάλληλες σχέσεις μεταξύ εμπιστοσύνης και εξάρτησης.

  Η ελευθερία λήψης αποφάσεων είναι μια άλλη περιβαλλοντική συνθήκη που, όταν υπάρχει σε υψηλότερο βαθμό, πιθανόν να προωθεί ισχυρότερες θετικές συσχετίσεις μεταξύ εμπιστοσύνης και εξάρτησης. Η ελευθερία λήψης αποφάσεων αντιπροσωπεύει το βαθμό στον οποίο οι χειριστές είναι σε θέση να λαμβάνουν μελετημένες αποφάσεις σχετικά με τον καλύτερο τρόπο χρήσης του αυτοματισμού. Μπορεί να επηρεαστεί από διάφορους καταστασιακούς παράγοντες, όπως η δυσκολία της εργασίας, ο φόρτος εργασίας, το οργανωτικό περιβάλλον, η εξειδίκευση στο αντικείμενο, η διάθεση και η ικανότητα προσοχής. Σε γενικές γραμμές, η εμπιστοσύνη έχει πιθανώς ασθενέστερη επίδραση στην εξάρτηση παρουσία καταστασιακών παραγόντων που εμποδίζουν την ελευθερία λήψης αποφάσεων. Για παράδειγμα, τρεις μελέτες που εντοπίστηκαν από την παρούσα ανασκόπηση διαπίστωσαν ότι η θετική σχέση μεταξύ εμπιστοσύνης και εξάρτησης μειώθηκε υπό υψηλότερο φόρτο εργασίας [54],[53],[55]. Αυτό μπορεί να οφείλεται στο γεγονός ότι σε συνθήκες υψηλού φόρτου

εργασίας, οι χειριστές δεν έχουν μερικές φορές άλλη επιλογή από το να χρησιμοποιούν αυτοματισμούς προκειμένου να συμβαδίζουν με τις απαιτήσεις των εργασιών [54].

Επίκτητη εμπιστοσύνη - Learned Trust

Οι άνθρωποι είναι πλάσματα της εμπειρίας. Ακριβώς όπως κάνουν οι άνθρωποι στις διαπροσωπικές σχέσεις, οι χειριστές χρησιμοποιούν γνώσεις από προηγούμενες αλληλεπιδράσεις με τον αυτοματισμό όταν αξιολογούν την αξιοπιστία νέων συστημάτων. Η επίκτητη εμπιστοσύνη αντιπροσωπεύει τις εκτιμήσεις ενός χειριστή για ένα σύστημα που προέρχονται από την εμπειρία του παρελθόντος ή την τρέχουσα αλληλεπίδραση. Αυτό το επίπεδο εμπιστοσύνης επηρεάζεται άμεσα από τις προϋπάρχουσες γνώσεις του χειριστή και την απόδοση του αυτοματοποιημένου συστήματος. Τα σχεδιαστικά χαρακτηριστικά του αυτοματισμού μπορούν επίσης να επηρεάσουν τη μαθημένη εμπιστοσύνη, αλλά το κάνουν έμμεσα, μεταβάλλοντας τις αντιλήψεις για την απόδοση του συστήματος.

Κατά τη διάρκεια μιας αλληλεπίδρασης, ένα αυτοματοποιημένο σύστημα μπορεί να έχει ποικίλες επιδόσεις, και η εμπιστοσύνη του χρήστη είναι πιθανό να αυξομειώνεται ανάλογα με τις επιδόσεις του συστήματος σε πραγματικό χρόνο. Για να αποτυπώσουμε τη διαδραστική φύση αυτής της σχέσης, διαιρούμε τη διδαχθείσα εμπιστοσύνη σε δύο κατηγορίες: αρχική και δυναμική. Και οι δύο μορφές μαθημένης εμπιστοσύνης είναι σχετικές με τα χαρακτηριστικά του αυτοματοποιημένου συστήματος- ωστόσο, η αρχικώς μαθημένη εμπιστοσύνη (initial learned trust) αντιπροσωπεύει την εμπιστοσύνη πριν από την αλληλεπίδραση με ένα σύστημα, ενώ η δυναμικά μαθημένη εμπιστοσύνη (dynamic learned trust) αντιπροσωπεύει την εμπιστοσύνη κατά τη διάρκεια μιας αλληλεπίδρασης.

- **Υφιστάμενη γνώση:** Η εμπιστοσύνη ενός χειριστή μπορεί να επηρεαστεί από τη φήμη του συστήματος πριν από την αλληλεπίδραση με αυτό. Πολυάριθμες μελέτες έχουν δείξει ότι οι άνθρωποι τείνουν να εμπιστεύονται περισσότερο τον αυτοματισμό όταν αυτός παρουσιάζεται ως ένα αξιόπιστο ή ”έμπειρο” σύστημα [23] [52]. Ωστόσο, ενώ η φημισμένη αυτοματοποίηση συγκεντρώνει μεγαλύτερη αρχική εμπιστοσύνη από τους χειριστές, η εμπιστοσύνη αυτή μπορεί να υποβαθμιστεί ταχύτερα όταν τα συστήματα κάνουν αισθητά σφάλματα [67]

  Έρευνες έχουν επίσης δείξει ότι οι προϋπάρχουσες νοοτροπίες και προσδοκίες μπορούν να μεταβάλουν τη διαδικασία σχηματισμού εμπιστοσύνης και τις επακόλουθες επιλογές χρήσης [68][36] [69]. Για παράδειγμα, οι Merritt κ.ά. (2012) μελέτησαν την επίδραση των έμμεσων και άμεσων στάσεων απέναντι στον αυτοματισμό στην εμπιστοσύνη σε έναν αυτόματο ανιχνευτή όπλων που είχε μεταβλητή απόδοση σε τρεις συνθήκες (σαφώς καλή, διφορούμενη και σαφώς κακή). Οι έμμεσες στάσεις διαφέρουν από τις άμεσες στάσεις στο ότι λειτουργούν καθαρά μέσω συσχετισμών και οι άνθρωποι συνήθως δεν τις γνωρίζουν [69]. Στο πείραμά τους, οι Merritt et al. διαπίστωσαν ότι η αλληλεπίδραση μεταξύ έμμεσων και άμεσων στάσεων είχε προσθετική επίδραση στην εμπιστοσύνη σε συνθήκες ασάφειας και σαφώς κακής κατάστασης. Όταν τόσο η έμμεση όσο και η άμεση στάση απέναντι στον αυτοματισμό ήταν θετικές, οι συμμετέχοντες ήταν πιο πιθανό να εκφράσουν μεγαλύτερη εμπιστοσύνη στο βοήθημα. Το εύρημα αυτό παρέχει ενδείξεις για έναν ασυνείδητο μηχανισμό που καθοδηγεί τη διαμόρφωση της εμπιστοσύνης στον αυτοματισμό. Ωστόσο, επειδή η έρευνα στον τομέα αυτό είναι περιορισμένη, απαιτούνται μελλοντικές μελέτες για την περαιτέρω εξέταση του ρόλου των έμμεσων στάσεων στην καθοδήγηση της αλληλεπίδρασης ανθρώπου-αυτοματισμού.

  Η προηγούμενη εμπειρία με ένα αυτοματοποιημένο σύστημα ή παρόμοια τεχνολογία, μπορεί να αλλάξει σημαντικά τη διαδικασία σχηματισμού εμπιστοσύνης. Ωστόσο, για να γίνει κατανοητή η συγκεκριμένη επίδραση που έχει η εμπειρία στην εμπιστοσύνη, πρέπει να γίνει

διάκριση μεταξύ της εμπειρογνωμοσύνης στο αντικείμενο (που σχετίζεται με την περιστασιακή εμπιστοσύνη) και της προηγούμενης εμπειρίας με την αυτοματοποίηση (που σχετίζεται με τη μαθημένη εμπιστοσύνη). Μια εικόνα αυτής της διάκρισης προκύπτει από τη σύγκριση των ευρημάτων δύο πειραμάτων που επικεντρώνονται στην επίδραση της προηγούμενης εμπειρίας στην εμπιστοσύνη στον αυτοματισμό. Στην πρώτη μελέτη, οι Yuliver-Gavish και Gopher [70] διαπίστωσαν ότι οι συμμετέχοντες εμπιστεύονταν περισσότερο ένα σύστημα υποστήριξης αποφάσεων αφού είχαν αποκτήσει εμπειρία στη χρήση του. Αρχικά, αυτό το αποτέλεσμα μπορεί να φαίνεται να έρχεται σε αντίθεση με τα ευρήματα των Sanchez κ.ά. [63], οι οποίοι διαπίστωσαν ότι οι έμπειροι αγρότες βασίζονταν λιγότερο στον αυτοματισμό κατά τη διάρκεια μιας εργασίας αποφυγής σύγκρουσης (με ένα γεωργικό όχημα) από ό,τι οι συμμετέχοντες χωρίς γεωργική εμπειρία. Ωστόσο, σε αντίθεση με τους έμπειρους συμμετέχοντες στη μελέτη των Yuliver-Gavish και Gopher, οι έμπειροι συμμετέχοντες στη μελέτη των Sanchez κ.ά. δεν είχαν ποτέ στο παρελθόν χειριστεί τον συγκεκριμένο τύπο αυτοματοποιημένου συστήματος συναγερμού που χρησιμοποιήθηκε κατά τη διάρκεια της εργασίας αποφυγής σύγκρουσης. Συνεπώς, ο τύπος της εμπειρίας που μελετήθηκε στο πείραμα των Sanchez et al. είναι προτιμότερο να ταξινομηθεί ως παράγοντας κατάστασης παρά ως μαθησιακός παράγοντας. Αυτό μπορεί να εξηγήσει γιατί η προηγούμενη εμπειρία οδήγησε σε μειωμένη εξάρτηση στη μελέτη των Sanchez κ.ά., αλλά σε αυξημένη εξάρτηση στο πείραμα των Yuliver-Gavish και Gopher.

Αν και η μελέτη των Yuliver-Gavish & Gopher [70] και πολλά άλλα πειράματα δείχνουν ότι η προηγούμενη εμπειρία με τον αυτοματισμό [71] ή παρόμοια τεχνολογία [72] παρέχει στους χειριστές μεγαλύτερη τάση να εμπιστεύονται ή να βασίζονται στον αυτοματισμό, άλλες έρευνες έχουν δείξει ότι αυτό δεν συμβαίνει πάντα [47]. Στην πραγματικότητα, μπορεί να εμφανιστεί η αντίθετη τάση εάν η προηγούμενη συνεργασία ενός χειριστή με ένα αυτοματοποιημένο σύστημα ήταν μη παραγωγική. Οι Manzey et al. [71] έδειξαν ότι οι αρνητικές εμπειρίες του παρελθόντος οδήγησαν σε μειωμένη εμπιστοσύνη σε ένα αυτοματοποιημένο σύστημα αναγνώρισης σφαλμάτων. Ανεξάρτητα από τη συγκεκριμένη επίδρασή της, η προηγούμενη εμπειρία παίζει σχεδόν πάντα ρόλο στην καθοδήγηση της αλληλεπίδρασης ανθρώπου-αυτοματισμού. Η εμπειρία είναι επίσης σημαντική επειδή μπορεί να ενισχύσει την κατανόηση του χειριστή για τον σκοπό και τη διαδικασία ενός αυτοματοποιημένου συστήματος.

Οι απόψεις και οι γνώσεις σχετικά με τον σκοπό και τις διεργασίες των αυτοματοποιημένων συστημάτων βοηθούν στην καθοδήγηση της διαδικασίας δημιουργίας εμπιστοσύνης. Όταν οι χειριστές στερούνται γνώσεων σχετικά με τον σκοπό ενός συστήματος ή τον τρόπο λειτουργίας του, είναι πιθανό να δυσκολεύονται να εναρμονίσουν με ακρίβεια την εμπιστοσύνη τους με την αξιοπιστία ενός συστήματος σε πραγματικό χρόνο. Αυτό ισχύει ιδιαίτερα όταν οι περιστασιακοί παράγοντες συμβάλλουν στον καθορισμό της απόδοσης ενός συστήματος. Για παράδειγμα, η διαδικασία διαμόρφωσης εμπιστοσύνης που χρησιμοποιούν οι χειριστές εξαρτάται από το βαθμό στον οποίο κατανοούν πώς η απόδοση του αυτοματισμού ποικίλλει σε διαφορετικά πλαίσια και σε διαφορετικές χρονικές φάσεις [73]. Οι λανθασμένες αντιλήψεις σχετικά με αυτές τις μεταβλητές μπορούν να οδηγήσουν σε κακή χρήση, αχρηστία ή/και κατάχρηση του αυτοματισμού. Η εκπαίδευση είναι ένας τρόπος για να μειωθεί η πιθανότητα αυτών των συμπεριφορών. Έρευνες έχουν δείξει ότι με την εκπαίδευση των χειριστών σχετικά με την πραγματική αξιοπιστία ενός βοηθήματος, είναι δυνατόν να μεταβληθούν τα πρότυπα εμπιστοσύνης και εξάρτησης [74], να διευκολυνθεί η καλύτερη εκτέλεση εργασιών [74] και να μειωθεί ο εφησυχασμός [75]. Εκτός από την εκπαίδευση, αρκετές μελέτες έχουν δείξει ότι τα βοηθήματα λήψης αποφάσεων που έχουν σχεδιαστεί για να συμπληρώνουν τη λήψη των αποφάσεών τους με επίπεδα εμπιστοσύνης σε πραγματικό χρόνο βοηθούν τους χρήστες να βαθμονομούν κατάλληλα την εμπιστοσύνη τους [76]. Ενώ οι πληροφορίες

εμπιστοσύνης του συστήματος δεν αποκαλύπτουν ρητά τίποτα σχετικά με το σκοπό ή τη διαδικασία, υπενθυμίζουν στους χειριστές ότι η αυτοματοποίηση είναι ατελής και μπορεί να παρέχει μόνο μια "καλύτερη εικασία" με βάση τις διαθέσιμες πληροφορίες σε μια δεδομένη κατάσταση.

Οι παράγοντες που αναφέρθηκαν στην παρούσα ενότητα μπορούν να επηρεάσουν την εμπιστοσύνη των χειριστών πριν από οποιαδήποτε αλληλεπίδραση με ένα σύστημα. Επειδή η προϋπάρχουσα γνώση δεν αλλάζει συνήθως κατά τη διάρκεια μιας αλληλεπίδρασης, επηρεάζει μόνο την αρχική αποκτηθείσα εμπιστοσύνη και όχι τη δυναμικά αποκτηθείσα εμπιστοσύνη. Μόλις ένας χειριστής αρχίσει να αλληλεπιδρά με ένα σύστημα, η απόδοσή του μπορεί να επηρεάσει τη δυναμικά διδαχθείσα εμπιστοσύνη, η οποία μπορεί να αλλάξει δραστικά κατά τη διάρκεια μιας αλληλεπίδρασης. Ωστόσο, η αντίληψη της απόδοσης εξαρτάται σε μεγάλο βαθμό από τον τρόπο με τον οποίο παρουσιάζονται οι πληροφορίες σε έναν χειριστή. Έτσι, τα χαρακτηριστικά σχεδιασμού του αυτοματισμού είναι σημαντικά, επειδή μπορούν να επηρεάσουν έμμεσα την εμπιστοσύνη μεταβάλλοντας τις αντιλήψεις για την απόδοση του συστήματος.

- **Χαρακτηριστικά σχεδιασμού:** Ουσιαστική έρευνα έχει δείξει ότι τα σχεδιαστικά χαρακτηριστικά μπορούν να μεταβάλουν την εμπιστοσύνη στον αυτοματισμό. Οι διεπαφές ηλεκτρονικών υπολογιστών αποτελούν συχνά το κύριο οπτικό στοιχείο των συστημάτων. Στην περίπτωση αυτή, οι διεπαφές πρέπει να είναι προσεκτικά διαμορφωμένες. Αρκετές μελέτες ηλεκτρονικού εμπορίου έχουν δείξει ότι οι αισθητικά ευχάριστοι ιστότοποι είναι περισσότερο αξιόπιστοι από τους λιγότερο ελκυστικούς ιστότοπους [77]. Αυτά τα ευρήματα ώθησαν τους Weinstock, Oron-Gilad και Parmet [78] να εξετάσουν την επίδραση της αισθητικής του συστήματος στην εμπιστοσύνη σε ένα αυτοματοποιημένο σύστημα εντός του οχήματος προτείνοντάς ότι ο αισθητικός σχεδιασμός μπορεί να είναι λιγότερο σημαντικός για την εμπιστοσύνη στην αυτοματοποίηση απ' ό,τι για την εμπιστοσύνη σε ιστότοπους ηλεκτρονικού εμπορίου (e-commerce).

  Πάραυτα, άλλες έρευνες δείχνουν ότι ο ανθρωπομορφισμός μιας διεπαφής μπορεί να είναι μια σημαντική μεταβλητή. Για παράδειγμα, μια πρόσφατη μελέτη διαπίστωσε ότι η προσθήκη της εικόνας ενός γιατρού στη διεπαφή μιας εφαρμογής διαχείρισης διαβήτη οδήγησε τους νεότερους συμμετέχοντες να εμπιστευτούν περισσότερο τις συμβουλές του συστήματος [35]. Επιπλέον, οι de Visser κ.ά. [79] διαπίστωσαν ότι η αύξηση του ανθρωπομορφισμού ενός αυτοματοποιημένου βοηθήματος έκανε τους συμμετέχοντες να επιδείξουν μεγαλύτερη ανθεκτικότητα στην εμπιστοσύνη (δηλ. η εμπιστοσύνη τους μειώθηκε λιγότερο γρήγορα) μετά από σφάλματα του συστήματος. Στο σύνολό τους, τα ευρήματα αυτά υποδηλώνουν ότι η αύξηση των ανθρώπινων χαρακτηριστικών των συστημάτων μπορεί να συμβάλει στη μείωση της αχρηστίας της αυτοματοποίησης με ορισμένους τύπους αυτοματισμού. Ωστόσο, οι σχεδιαστές πρέπει να λαμβάνουν υπόψη τα αναμενόμενα χαρακτηριστικά των δυνητικών χρηστών (π.χ. ηλικία, φύλο, κουλτούρα), καθώς η ανθρωπομορφοποίηση μιας διεπαφής μπορεί να επηρεάσει διαφορετικά τη διαδικασία σχηματισμού εμπιστοσύνης για διαφορετικά άτομα [35].

  Τα αυτοματοποιημένα συστήματα μπορούν να επικοινωνούν με τους χρήστες τους με διάφορους τρόπους. Οι διαφορετικοί τρόποι επικοινωνίας μπορεί να οδηγήσουν τους χειριστές σε ανομοιογενή επίπεδα εμπιστοσύνης στον αυτοματισμό. Για παράδειγμα, ορισμένα αυτοματοποιημένα συστήματα χρησιμοποιούν λεκτική επικοινωνία από προσωποποιημένους πράκτορες υπολογιστών (computer agents), αντί για απλό κείμενο, προκειμένου να προκαλέσουν αισθήματα εμπιστοσύνης όμοια με αυτά της εμπιστοσύνης μεταξύ ανθρώπων. Ωστόσο, η έρευνα έχει αποκαλύψει ότι οι άνθρωποι προτιμούν μερικές φορές τις διεπαφές κειμένου από τους ανθρωπόμορφους πράκτορες [80]. Προκειμένου να προωθηθεί η εμπιστοσύνη, οι computer agents πρέπει να κατασκευάζονται προσεκτικά ώστε να φαίνονται

τόσο ανθρώπινοι όσο και αξιόπιστοι [80]. Τα αποτελέσματα μιας μελέτης δείχνουν ότι η κίνηση των ματιών ενός πράκτορα, η κανονικότητα της μορφής και το σχήμα του πηγουνιού είναι σημαντικές μεταβλητές που μπορούν να επηρεάσουν την εμπιστοσύνη [81]. Επιπλέον, το φύλο ενός agent μπορεί να επηρεάσει την εμπιστοσύνη. Η E. J. Lee [40] διαπίστωσε ότι οι συμμετέχοντες συμμορφώθηκαν περισσότερο με τους άνδρες πράκτορες υπολογιστών που επικοινωνούσαν μέσω κειμένου από ό,τι με τις γυναίκες πράκτορες. Συνολικά, αυτό το σύνολο των ερευνών υποδηλώνει ότι το στυλ επικοινωνίας ενός αυτοματοποιημένου συστήματος μπορεί να είναι μια σημαντική μεταβλητή. Προκειμένου να προωθηθεί η κατάλληλη εμπιστοσύνη στην αυτοματοποίηση, οι σχεδιαστές θα πρέπει να επιλέξουν έναν τρόπο επικοινωνίας που αντιστοιχεί στις πραγματικές δυνατότητες ενός συστήματος.

Μέσα σε έναν δεδομένο τρόπο επικοινωνίας, τα αυτοματοποιημένα συστήματα μπορούν να παρουσιάσουν ένα ευρύ φάσμα διακριτών ”προσωπικοτήτων”. Η έρευνα έχει δείξει ότι ορισμένα τεχνητά αποδιδόμενα χαρακτηριστικά μπορούν να επηρεάσουν την εμπιστοσύνη των χειριστών. Για παράδειγμα, οι Parasuraman και Miller [82] διαπίστωσαν ότι η ενίσχυση του αυτοματισμού με καλή συμπεριφορά, που επιχειρησιακά ορίζεται ως ”ένα στυλ επικοινωνίας που είναι ”μη παρεμβατικό” και ”υπομονετικό””, οδήγησε σε μεγαλύτερη εμπιστοσύνη και βελτιωμένη διαγνωστική απόδοση (σελ. 54). Σε μια μεταγενέστερη μελέτη, οι Spain και Madhavan [83] όρισαν την εθιμοτυπία του αυτοματισμού ως ”ευγένεια”. Χειριζόμενοι μόνο την επιλογή των λέξεων, οι ερευνητές προκάλεσαν διαφορετικά επίπεδα υποκειμενικής εμπιστοσύνης σε ένα αυτοματοποιημένο βοήθημα. Τα ευρήματα αυτά αποδεικνύουν το ρόλο της ”προσωπικότητας του αυτοματισμού” στην καθοδήγηση της ανάπτυξης εμπιστοσύνης των χειριστών.

Άλλη μια μεταβλητή με σημαντική επιρροή είναι η διαφάνεια της αυτοματοποίησης. Η διαφάνεια αφορά το βαθμό στον οποίο ”οι εσωτερικές λειτουργίες ή η λογική που χρησιμοποιούνται από τα αυτοματοποιημένα συστήματα είναι γνωστές στους χειριστές για να βοηθήσουν στην κατανόηση του συστήματος” [84]. Πολυάριθμες μελέτες έχουν δείξει ότι ο σχεδιασμός συστημάτων που παρέχουν στους χρήστες ακριβή ανατροφοδότηση σχετικά με την αξιοπιστία τους ή τον τρόπο λειτουργίας τους μπορεί να διευκολύνει καλύτερα την κατάλληλη εμπιστοσύνη [85] [86][84] [87] και να βελτιώσουν την απόδοση των εργασιών ανθρώπου-αυτοματισμού [88]. Για παράδειγμα, οι Seong και Bisantz [84] διαπίστωσαν ότι η παροχή στους συμμετέχοντες γνωστικής ανατροφοδότησης από ένα σύστημα (με τη μορφή μετα-πληροφοριών) οδήγησε σε υψηλότερες βαθμολογίες εμπιστοσύνης σε αυτοματισμούς χαμηλής αξιοπιστίας, αλλά σε χαμηλότερες βαθμολογίες εμπιστοσύνης σε αυτοματισμούς υψηλής αξιοπιστίας. Οι συγγραφείς αποδίδουν αυτό το γεγονός στο αποτέλεσμα της μεταπληροφόρησης να βοηθά τους χειριστές στη συνθήκη χαμηλής αξιοπιστίας να αγνοούν το βοήθημα όταν είναι απαραίτητο και να υπενθυμίζουν στους χειριστές στη συνθήκη υψηλής αξιοπιστίας ότι το βοήθημα ήταν ατελές [84]. Σε παρόμοια μελέτη, οι Wang, Jamieson και Hollands [87] διαπίστωσαν ότι η παροχή πληροφοριών αξιοπιστίας του συστήματος στους χειριστές μπορεί να βελτιώσει την καταλληλότητα της εμπιστοσύνης. Ωστόσο, η προβολή αυτών των πληροφοριών με διαφορετικούς τρόπους μπορεί να μεταβάλει τις στρατηγικές εμπιστοσύνης και την απόδοση ανθρώπου-αυτοματισμού [89]. Τέλος, οι Dzindolet κ.ά. [26] έδειξαν ότι η παροχή εξηγήσεων στους χειριστές σχετικά με τους λόγους για τους οποίους συμβαίνουν αποτυχίες του αυτοματισμού μπορεί να οδηγήσει σε αυξημένη εμπιστοσύνη. Συνολικά, αυτή η έρευνα υποδηλώνει ότι ο σχεδιασμός διαφανών συστημάτων που παρέχουν ακριβή, χρήσιμη ανατροφοδότηση μπορεί να μειώσει τη συχνότητα της κακής χρήσης και της αχρησίας των αυτοματισμών.

Η εμπιστοσύνη στον αυτοματισμό εξαρτάται επίσης από το επίπεδο ελέγχου των λειτουργιών του συστήματος από τον χειριστή [90] [91]. Για παράδειγμα, μια πρόσφατη μελέτη έδειξε ότι οι αυτοματισμοί που αναλαμβάνουν καθήκοντα παρέχοντας ταυτόχρονα πληρο-

φορίες στον χειριστή θεωρούνται πιο αξιόπιστοι από τους αυτοματισμούς που αναλαμβάνουν ρόλους χωρίς να παρέχουν καμία πληροφορία στον χειριστή [91]. Αυτό το εύρημα υποδηλώνει ότι οι χειριστές μπορεί να έχουν την τάση να εμπιστεύονται τα χαμηλότερα επίπεδα σύνθετης αυτοματοποίησης περισσότερο από τα υψηλότερα επίπεδα. Δυστυχώς, η αυτοματοποίηση χαμηλού επιπέδου συχνά μειώνει την αποδοτικότητα μέσω των πρόσθετων καθυστερήσεων που προκύπτουν όταν ένα σύστημα παρέχει πληροφορίες στον χειριστή του και στη συνέχεια τον περιμένει. Ακόμα, αν και οι υψηλού επιπέδου αυτοματισμοί μπορούν να εκτελέσουν εργασίες ταχύτερα, οι ανθρώπινοι χειριστές βρίσκονται ”εκτός της διαδικασίας” (”out of the loop”). Αυτό σημαίνει ότι οι χειριστές δεν μπορούν να αποτρέψουν τα σφάλματα στα συστήματα, οπότε, αντίθετα, πρέπει να αντιμετωπιστούν μετά την εμφάνισή τους. Επιπλέον, οι χειριστές που βγαίνουν ”εκτός λούπας” μπορεί να εξαρτώνται από την αυτοματοποίηση για την εκτέλεση των καθηκόντων τους. Εάν παρουσιαστούν σφάλματα, είναι πιθανό να δυσκολευτούν περισσότερο να ολοκληρώσουν τις εργασίες τους χειροκίνητα. Προκειμένου να καθοριστεί το βέλτιστο επίπεδο ελέγχου, οι σχεδιαστές θα πρέπει να εξετάζουν τις πιθανές απαιτήσεις του περιβάλλοντος στο οποίο πιθανόν θα χρησιμοποιηθεί ένα σύστημα.

Σε κάποια πλαίσια, ο προσαρμόσιμος αυτοματισμός μπορεί να αποτελέσει μια αποτελεσματική λύση για το συμβιβασμό όσον αφορά τα διαφορετικά επίπεδα ελέγχου αφού προσφέρει τη δυνατότητα βελτίωσης τόσο της ασφάλειας όσο και της αποτελεσματικότητας των συστημάτων εργασίας ανθρώπου-αυτοματισμού, προσαρμόζοντας το επίπεδο ελέγχου του στις ανάγκες της τρέχουσας κατάστασης [90]. Ο προσαρμοστικός αυτοματισμός μπορεί έτσι να είναι χρήσιμος επειδή μπορεί να αλλάξει τη συμπεριφορά του με βάση τις προτιμήσεις του χρήστη. Παρά τα δυνητικά οφέλη, αυτή η μορφή αυτοματισμού δεν είναι πάντα πρακτική για χρήση στον πραγματικό κόσμο.

- **Απόδοση:** Η εμπιστοσύνη εξαρτάται άμεσα από τα αποτελέσματα. Αξιόλογη έρευνα έχει δείξει ότι οι ανθρώπινοι χειριστές προσαρμόζουν την εμπιστοσύνη τους στον αυτοματισμό ανάλογα με την απόδοσή της σε πραγματικό χρόνο. Έχουν μελετηθεί διάφορες πτυχές της απόδοσης. Αρχικά, πολυάριθμες μελέτες έχουν δείξει ότι η αξιοπιστία και η εγκυρότητα των λειτουργιών ενός αυτοματοποιημένου συστήματος είναι σημαντικά προγενέστερα στοιχεία της εμπιστοσύνης [47] [90] [23] [84] [51] [34] [55]. Η αξιοπιστία αναφέρεται στη συνέπεια των λειτουργιών ενός αυτοματοποιημένου συστήματος, ενώ η εγκυρότητα αναφέρεται στο βαθμό στον οποίο ένα αυτοματοποιημένο σύστημα εκτελεί την προβλεπόμενη εργασία. Η προβλεψιμότητα και η αξιοπιστία της αυτοματοποίησης είναι επίσης σημαντικές. Η προβλεψιμότητα αναφέρεται στο βαθμό στον οποίο η αυτοματοποίηση εκτελεί με τρόπο που να συνάδει με τις προσδοκίες του χειριστή, ενώ η αξιοπιστία αναφέρεται στη συχνότητα των βλαβών ή των μηνυμάτων σφάλματος της αυτοματοποίησης [44]. Έρευνες έχουν δείξει ότι οι χειριστές εμπιστεύονται περισσότερο τα αυτοματοποιημένα συστήματα όταν αυτά είναι ιδιαίτερα προβλέψιμα και αξιόπιστα [54] [44].

  Όταν συμβαίνουν αστοχίες στον αυτοματισμό, διαφορετικοί τύποι σφαλμάτων μπορεί να έχουν διαφορετικές επιπτώσεις στην εμπιστοσύνη και στις μεταγενέστερες συμπεριφορές εμπιστοσύνης. Πιο συγκεκριμένα, η έρευνα έχει δείξει ότι οι ψευδείς συναγερμοί (false alarms) (όταν τα συστήματα προειδοποιούν εσφαλμένα τους χειριστές για την παρουσία ενός σήματος) και οι αστοχίες (όταν ο αυτοματισμός αποτυγχάνει να ανιχνεύσει ένα πραγματικό σήμα) έχουν γενικά διαφορετικές επιπτώσεις στις εξαρτώμενες από την εμπιστοσύνη συμπεριφορές [92]. Σημαντικό είναι ότι οι ψευδείς συναγερμοί απαιτούν συμμόρφωση (οι χειριστές πρέπει να υποθέσουν ότι ένα σήμα είναι υπαρκτό), ενώ οι αστοχίες απαιτούν εμπιστοσύνη (οι χειριστές πρέπει να υποθέσουν ότι ένα σήμα απουσιάζει). Αυτή η διάκριση είναι ουσιώδης, διότι πολυάριθμες μελέτες έχουν διαπιστώσει ότι αυτοματισμοί με ψευδείς συναγερμούς μειώνουν τη συμμόρφωση των χειριστών περισσότερο από ό,τι την εξάρτηση, ενώ

αυτοματισμοί με αστοχίες μειώνουν την εξάρτηση περισσότερο από ό,τι τη συμμόρφωση [93]. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η συμμόρφωση και η εξάρτηση δεν είναι εντελώς ανεξάρτητες η μία από την άλλη [94].

Πέραν των διακριτών επιρροών τους στην εξαρτώμενη από την εμπιστοσύνη συμπεριφορά, οι ψευδείς συναγερμοί και οι αστοχίες μπορεί να επηρεάζουν διαφορετικά τα υποκειμενικά αισθήματα εμπιστοσύνης. Ένας πιθανός λόγος γι' αυτό είναι ότι οι ψευδείς συναγερμοί είναι συνήθως πιο εμφανείς από τις αστοχίες και απαιτούν από τον χειριστή να καταβάλει προσπάθεια για περιττές έρευνες. Κατά συνέπεια, οι ψευδείς συναγερμοί μπορεί να έχουν μεγαλύτερο αρνητικό αντίκτυπο στην εμπιστοσύνη από ό,τι οι αστοχίες. Ενώ ορισμένες έρευνες έχουν υποστηρίξει αυτή την υπόθεση [95], άλλες έρευνες δείχνουν ότι οι ψευδείς συναγερμοί και οι αστοχίες έχουν παρόμοιες επιπτώσεις [49]. Επιπροσθέτως, τουλάχιστον δύο έρευνες διαπίστωσαν ότι οι συμμετέχοντες εμπιστεύονταν περισσότερο τον αυτοματισμό με ψευδείς συναγερμούς από τον αυτοματισμό με αστοχίες [93]. Σε γενικές γραμμές, ο συγκεκριμένος αντίκτυπος που έχουν οι ψευδείς συναγερμοί και οι αστοχίες στην εμπιστοσύνη εξαρτάται πιθανώς από τις αρνητικές συνέπειες που συνδέονται με κάθε τύπο σφάλματος σε ένα συγκεκριμένο πλαίσιο. Για παράδειγμα, ενώ μια λανθασμένη ειδοποίηση από έναν ανιχνευτή μονοξειδίου του άνθρακα είναι μια μικρή ενόχληση, μια αστοχία μπορεί να οδηγήσει σε θάνατο. Σε αυτή την περίπτωση, οι αστοχίες θα προκαλέσουν πιθανώς μεγαλύτερη ζημία στην εμπιστοσύνη από ό,τι οι ψευδείς συναγερμοί.

Η χρησιμότητα ενός αυτοματοποιημένου συστήματος είναι η τελευταία βασισμένη στις επιδόσεις μεταβλητή που μπορεί να επηρεάσει την εμπιστοσύνη. Η εμπιστοσύνη στον αυτοματισμό είναι πάντα σχετική με μια εργασία που ο χειριστής επιθυμεί να εκτελεστεί. Εάν ένας χειριστής συνειδητοποιήσει ότι η χρήση ενός αυτόματου συστήματος για την εκτέλεση μιας εργασίας καθιστά την εργασία πιο απαιτητική, πιθανότατα δεν θα δει καμία ανάγκη να χρησιμοποιήσει και, επομένως, να εμπιστευτεί τον αυτοματισμό. Συνεπώς, ο αυτοματισμός πρέπει πρώτα να αποδειχθεί χρήσιμος στους χειριστές για να διακυβεύεται η εμπιστοσύνη. Λίγες εμπειρικές έρευνες έχουν μελετήσει άμεσα τη χρησιμότητα. Ωστόσο, ο Parkes [50] διαπίστωσε ότι οι συμμετέχοντες βασίζονταν περισσότερο στις συμβουλές ενός συστήματος υποστήριξης αποφάσεων όταν το θεωρούσαν χρήσιμο. Επιπλέον, σε δύο ξεχωριστές μελέτες, οι Abe και Richardson [68] έδειξαν ότι οι οδηγοί εμπιστεύονταν ένα σύστημα προειδοποίησης σύγκρουσης σημαντικά λιγότερο όταν το σύστημα παρείχε συναγερμούς αφού είχε ήδη ξεκινήσει το φρενάρισμα. Η μείωση της εμπιστοσύνης μπορεί να ήταν αποτέλεσμα του ότι οι καθυστερημένοι συναγερμοί παρείχαν ασθενέστερα οφέλη στους οδηγούς

Η μελέτη που συζητήθηκε σε αυτή την ενότητα δείχνει πώς οι χειριστές προσαρμόζουν την εμπιστοσύνη τους στην αυτοματοποίηση ώστε να ανταποκρίνεται στις τρέχουσες επιδόσεις της. Παρόλο που η εμπιστοσύνη εξαρτάται σε μεγάλο βαθμό από την απόδοση, η διαδικασία διαμόρφωσης αυτής εξαρτάται επίσης από τους περιστασιακούς παράγοντες, τα σχεδιαστικά χαρακτηριστικά του αυτοματισμού, την εμπιστοσύνη προδιάθεσης (dispositional trust) του χειριστή και τις προϋπάρχουσες γνώσεις. Στο πλαίσιο μιας μεμονωμένης αλληλεπίδρασης, οι περισσότερες από αυτές τις μεταβλητές είναι σταθερές, ενώ η απόδοση δεν είναι.

- **Περιστασιακοί παράγοντες που δεν σχετίζονται με την εμπιστοσύνη:** Παρόλο που τόσο η αρχική όσο και η δυναμικά αποκτηθείσα εμπιστοσύνη επηρεάζουν την εξάρτηση από τον αυτοματισμό, δεν είναι οι μόνοι παράγοντες που συμβάλλουν. Οι Lee και See [1] εξηγούν ότι "η εμπιστοσύνη καθοδηγεί -αλλά δεν καθορίζει πλήρως- την εμπιστοσύνη" (σ. 51). Επιπρόσθετοι καταστατικοί παράγοντες, όπως το επίπεδο της προσπάθειας που απαιτείται για την εμπλοκή με ένα σύστημα, οι εναλλακτικές λύσεις για τη χρήση του αυτοματισμού, οι χρονικοί περιορισμοί , η αντίληψη της κατάστασης και η σωματική ευεξία του χειριστή, μπορούν να καθοδηγήσουν την εξάρτηση χωρίς απαραίτητα να επηρεάζουν την εμπιστοσύνη. Για παράδειγμα, οι Rice και Keller [96] έδειξαν ότι οι συμμετέχοντες ήταν πιο πιθανό να

συμμορφωθούν με ένα διαγνωστικό βοήθημα όταν ο χρόνος λήψης αποφάσεων μειωνόταν. Έτσι, ο βαθμός ελευθερίας λήψης αποφάσεων ενός χειριστή, καθώς και άλλοι καταστασιακοί παράγοντες, μπορούν να επηρεάσουν τις τάσεις της αλληλοεξαρτώμενης σχέσης μεταξύ εμπιστοσύνης, εξάρτησης και απόδοσης.

Οι Hoff και Bashir [22], μετά από ενδελεχή έρευνα, διαπίστωσαν ότι η εμπιστοσύνη ενός χειριστή σε ένα αυτοματοποιημένο σύστημα είναι μια σύνθεση των τάσεων του για εμπιστοσύνη, της κατάστασης και των αντιλήψεών του για το σύστημα. Εντόπισαν επίσης καταστασιακούς παράγοντες που επηρεάζουν τον βαθμό επιρροής της εμπιστοσύνης στην εμπιστοσύνη σε ένα αυτοματοποιημένο σύστημα. Τονίζουν επίσης τη σημασία των ατομικών διαφορών στη εμπιστοσύνη προδιάθεσης καταλήγοντας στο συμπέρασμα ότι ο σχηματισμός εμπιστοσύνης είναι μια δυναμική διαδικασία που καθοδηγείται από μια σύνθετη αλληλεπίδραση παραγόντων που προέρχονται από τρία αλληλοεξαρτώμενα στρώματα μεταβλητότητας και ότι η κατανόηση αυτών των παραγόντων είναι ζωτικής σημασίας για την προώθηση της ορθής χρήσης του αυτοματισμού και την ελαχιστοποίηση της συχνότητας των σχετικών ατυχημάτων.

2.2 Μοντέλα και πλαίσια εμπιστοσύνης

Καθώς η εμπιστοσύνη είναι μια από τις απαραίτητες προϋποθέσεις για την ανάπτυξη μιας επιτυχημένης αλληλεπίδρασης ανθρώπου-ρομπότ, υπάρχει ανάγκη από μεθόδους για τη μοντελοποίηση, τη μέτρηση και τη βαθμονόμηση της εμπιστοσύνης. Το πρώτο βήμα για τη μοντελοποίηση της εμπιστοσύνης είναι ένας σαφής ορισμός της εμπιστοσύνης. Ωστόσο, παρά τις ευρείες προσπάθειες και τον αριθμό των μελετών που επικεντρώνονται στην εμπιστοσύνη, δεν υπάρχει ιδιαίτερη ομοφωνία σε έναν ενιαίο ορισμό, καθώς ο ορισμός της εμπιστοσύνης εξαρτάται σε μεγάλο βαθμό από το πλαίσιο στο οποίο συζητείται η εμπιστοσύνη [97]. Αυτό τονίζεται, για παράδειγμα, στο [98], όπου η εμπιστοσύνη ορίζεται ως ”η εμπιστοσύνη ενός πράκτορα ότι ενέργειες που βλάπτουν την ευημερία του πράκτορα αυτού δεν θα πραγματοποιηθούν από άλλους πράκτορες με επιρροή”. Για κάθε εφαρμογή και τομέα που χρησιμοποιούνται τα ρομπότ, η εμπιστοσύνη πρέπει να ορίζεται, να μετράται και να διερευνάται ρητά. Για παράδειγμα, σε ρομποτικές εφαρμογές υψηλού κινδύνου, όπως τα ρομπότ εκκένωσης έκτακτης ανάγκης [4], ο ορισμός της εμπιστοσύνης μπορεί να διαφέρει σημαντικά από ρομποτικές εφαρμογές χαμηλού κινδύνου, όπως τα ρομπότ εντοπισμού ποδιών [99]. Ένας από τους πιο εμπεριστατωμένους ορισμούς της εμπιστοσύνης, ο οποίος χρησιμοποιείται από πολλές άλλες μελέτες που επικεντρώνονται στην εμπιστοσύνη ανθρώπου-ρομπότ, είναι των Lee και See [1]. Ορίζουν την εμπιστοσύνη από τη σκοπιά του αυτοματισμού. Ο ορισμός αυτός προέκυψε από την ανασκόπηση πολλών άλλων μελετών που επικεντρώθηκαν στον ορισμό της εμπιστοσύνης και ήταν συμπληρωματικός σε πολλές άλλες εργασίες. Ορίζουν την εμπιστοσύνη ως εξής: ”η άποψη ότι ένας πράκτορας θα βοηθήσει στην επίτευξη των στόχων ενός ατόμου σε μια κατάσταση που χαρακτηρίζεται από αβεβαιότητα και τρωτότητα”. Αυτός ο ορισμός της εμπιστοσύνης είναι αποδεκτός και χρησιμοποιείται από πολλές μελέτες σχετικά με την εμπιστοσύνη στην HRI (Human-Robot Interaction). Οι Wagner κ.α. [100] παρείχαν επίσης έναν ολοκληρωμένο ορισμό για την εμπιστοσύνη: ”μια πεποίθηση, που έχει ο εμπιστευόμενος, ότι ο έμπιστος θα ενεργήσει με τρόπο που θα μετριάσει τον κίνδυνο του εμπιστευόμενου σε μια κατάσταση στην οποία ο εμπιστευόμενος έχει διακινδυνεύσει τα αποτελέσματά του”. Παρείχαν επίσης ένα μοντέλο για τον προσδιορισμό του κατά πόσον μια αλληλεπίδραση απαιτεί εμπιστοσύνη ή όχι. Όλοι αυτοί οι ορισμοί έχουν ένα κοινό σημείο, αυτό είναι: ”αν οι ενέργειες και οι συμπεριφορές ενός ρομπότ ανταποκρίνονται στο ανθρώπινο συμφέρον ή όχι”. Για να αντιμετωπιστεί αυτή η ανησυχία σε κάθε ρομποτική εφαρμογή η εμπιστοσύνη πρέπει να μοντελοποιηθεί με βάση τα ανθρώπινα συμφέροντα στον συγκεκριμένο τομέα.

Τα μοντέλα εμπιστοσύνης εκφράζουν την επίδραση των διαφόρων παραγόντων στη διαμόρφωση και τη μεταβολή της εμπιστοσύνης στα ρομπότ. Στην πραγματικότητα, τα μοντέλα εμπι-

στοσύνης χρησιμοποιούν παράγοντες που επηρεάζουν την εμπιστευτικότητα για την εκτίμηση της εμπιστοσύνης. Δεδομένου ότι οι παράγοντες αυτοί ποικίλλουν σε τομείς και περιβάλλοντα, οι παράγοντες εισόδου στα μοντέλα εμπιστοσύνης διαφέρουν ανάλογα με τον τομέα εφαρμογής. Για παράδειγμα, οι Robinette κ.ά. [4] μοντελοποιούν την εμπιστοσύνη στην εκκένωση έκτακτης ανάγκης με βάση τον κίνδυνο κατάστασης (π.χ. το μέγεθος του κινδύνου που αντιλαμβάνεται ο άνθρωπος στο περιβάλλον γύρω του) και τον κίνδυνο του πράκτορα (π.χ. τη συμπεριφορά και την εμφάνιση του πράκτορα) για να μοντελοποιήσουν την αντιλαμβανόμενη εμπιστοσύνη από τον άνθρωπο και την απόφαση του ανθρώπου να εμπιστευτεί ή όχι την καθοδήγηση του ρομπότ. Αντίθετα, το [101] προτείνει ένα μοντέλο εμπιστοσύνης για την εποπτευόμενη συνεργασία και διατυπώνει την εμπιστοσύνη ως συνάρτηση της επιτυχίας και της αποτυχίας του ρομπότ στην εκτέλεση της εργασίας. Η έξοδος αυτού του μοντέλου εμπιστοσύνης είναι το κλείσιμο του βρόχου μεταξύ της ανθρώπινης εμπιστοσύνης και της λειτουργίας του ρομπότ με την προσαρμογή των ενεργειών του ρομπότ για τη βελτίωση της αποτελεσματικότητας της συνεργασίας. Τέλος, το [102] προτείνει ένα γενικότερο μοντέλο εμπιστοσύνης που βασίζεται στη διαμόρφωση της ομάδας, την εργασία, το σύστημα, το πλαίσιο και τις διαδικασίες της ομάδας για την διαβάθμιση της εμπιστοσύνης.

Πολλές από τις μελέτες σχετικά με τη μοντελοποίηση της εμπιστοσύνης στην HRI θεωρούν την απόδοση της συνεργασίας ως ένα από τα κύρια στοιχεία εισόδου για το μοντέλο τους [99] [103] [7]. Τα περισσότερα από αυτά τα μοντέλα εξετάζουν την επίδραση της απόδοσης σε συνδυασμό με κάποιους άλλους παράγοντες. Για παράδειγμα, το πιθανοτικό μοντέλο εμπιστοσύνης OPTIMo [99] χρησιμοποιεί ποσοστά αποτυχιών του ρομπότ και ανθρώπινες παρεμβάσεις σε συνδυασμό με την απόδοση της εργασίας ως δεδομένα εισόδου στο μοντέλο για την εκτίμηση του βαθμού εμπιστοσύνης του ανθρώπου σε έναν ρομποτικό συνεργάτη. Εν τω μεταξύ, το [103] χρησιμοποιεί την αντίληψη του χειριστή για τις δυνατότητες του συστήματος, την προηγούμενη εμπειρία και την εκπαίδευση για την εκτίμηση της αρχικής εμπιστοσύνης. Η εμπιστοσύνη ενημερώνεται σε έναν βρόχο με βάση την απόδοση του συστήματος, το γνωστικό φόρτο εργασίας και τη συχνότητα των αλλαγών από τηλεχειρισμό σε αυτόνομη λειτουργία. Η έξοδος αυτού του μοντέλου εμπιστοσύνης είναι ένα μέτρο του κέρδους και της απώλειας εμπιστοσύνης και του αντίκτυπου αυτών των αλλαγών στην εμπιστοσύνη στην απόδοση της συνεργασίας. Οι Sadrfaridpour κ.ά. [7] μοντελοποιούν την εμπιστοσύνη με βάση την ανθρώπινη απόδοση (δηλ. την κόπωση των μυών και τη δυναμική της αποκατάστασης), την απόδοση του ρομπότ (δηλ. την ταχύτητα του ρομπότ που εκτελεί τη συγκεκριμένη εργασία), τον φόρτο εργασίας και τις ανθρώπινες προσδοκίες για την απόδοση της εργασίας. Η έξοδος αυτού του μοντέλου είναι η ανατροφοδότηση προς το ρομπότ ώστε να προσαρμόζει την απόδοσή του σύμφωνα με τις επιθυμίες του χειριστή.

Η εμπιστοσύνη στην HRI έχει πολλά κοινά με την εμπιστοσύνη στην HAI (Human-Automation Interaction), η οποία έχει μελετηθεί εκτενώς. Οι Muir κ.ά. [104] διαπίστωσαν ότι οι διαθέσιμοι ορισμοί για την εμπιστοσύνη μεταξύ ανθρώπων δεν συνάδουν με τη φύση των HAI με βάση την πολυδιάστατη κατασκευή της εμπιστοσύνης. Όρισε ένα μοντέλο εμπιστοσύνης για την HAI, το οποίο βασίστηκε σε ένα μοντέλο ανθρώπινων προσδοκιών για τον αυτοματισμό που προτάθηκε από τους Barber κ.α. [17]. Οι Lee και Moray [24] βασίστηκαν στη στρατηγική της Muir για τη μοντελοποίηση της εμπιστοσύνης, η οποία ήταν ο εντοπισμός ανεξάρτητων μεταβλητών που επηρεάζουν την εμπιστοσύνη, και εισήγαγαν ένα άλλο μοντέλο εμπιστοσύνης. Αργότερα, άλλοι ερευνητές μοντελοποίησαν την εμπιστοσύνη του χειριστή στον αυτοματισμό, λαμβάνοντας υπόψη περισσότερους παράγοντες που επηρεάζουν την εμπιστοσύνη [105] [1]. Αυτά τα μοντέλα ταξινομήθηκαν τελικά σε πέντε ομάδες [106]: μοντέλα που βασίζονται στην παλινδρόμηση [104] [24], μοντέλα χρονοσειρών [107], ποιοτικά μοντέλα [108], πιθανολογικά μοντέλα που βασίζονται σε επιχειρήματα [109] και μοντέλα νευρωνικών δικτύων [105].

Παρά τις ομοιότητες, τα περισσότερα από τα μοντέλα που δημιουργήθηκαν για τη μοντελοποίηση της εμπιστοσύνης στην HRI δεν συνάδουν με τις ανάγκες της HAI. Σύμφωνα με τον Desai [110]"τα μοντέλα αυτά δεν λαμβάνουν υπόψη ορισμένους παράγοντες που εμφανίζονται κατά την συνεργασία με τα ρομπότ, όπως η επίγνωση της κατάστασης, η χρηστικότητα της διεπαφής, η

φυσική παρουσία των ρομπότ (που βρίσκονται μαζί με τον άνθρωπο ή απομακρυσμένα), οι περιορισμοί και η πολυπλοκότητα του περιβάλλοντος εργασίας, ο φόρτος εργασίας, η δυσκολία της εργασίας κ.λπ. που επηρεάζουν σημαντικά την HRI". Οι Desai κ.ά. [110] παρουσίασαν σχηματικά ένα μοντέλο που εξετάζει ορισμένους παράγοντες που επηρεάζουν την εμπιστοσύνη ανθρώπου-ρομπότ σε συνδυασμό με παράγοντες που επηρεάζουν την εμπιστοσύνη ανθρώπου-αυτοματισμού. Οι Yagoda κ.ά. [102] εισήγαγαν ένα από τα πρώτα μοντέλα για την εμπιστοσύνη ανθρώπου-ρομπότ με βάση τις διάφορες διαστάσεις μιας εργασίας αλληλεπίδρασης ανθρώπου-ρομπότ και την αξιολόγηση της εγκυρότητας καθεμιάς από αυτές τις κατευθύνσεις από ειδικούς σε θέματα στην HRI. Οι Desai κ.α. [6] ήταν επίσης κάποιοι από τους πρωτοπόρους στη μοντελοποίηση της εμπιστοσύνης στην HRI. Δημιούργησαν ένα πιο λεπτομερές μοντέλο για την εμπιστοσύνη στην τηλεργασία ανθρώπου και αυτόνομου ρομπότ. Το μοντέλο αυτό χρησιμοποίησε το μέτρο Area Under Trust Curve (AUTC) για να λάβει υπόψη του ολόκληρη τη διαδραστική εμπειρία ενός ατόμου με το ρομπότ.

Σύμφωνα με μελέτες, υπάρχει ισχυρή συσχέτιση μεταξύ του επιπέδου εμπιστοσύνης στους συνεργάτες ανθρώπων-ρομπότ και της απόδοσης της εργασίας των ρομποτικών πρακτόρων, επηρεάζοντας παράλληλα την ποιότητα της αλληλεπίδρασής τους [1] [6]. Σύμφωνα με τον Xu [111] , τα υψηλά επίπεδα εμπιστοσύνης μεταξύ συναδέλφων ανθρώπων-ρομπότ συχνά επιδεικνύουν μεγάλη συνέργεια, κατά την οποία οι συνδυαστικές δεξιότητες λήψης αποφάσεων του ανθρώπινου μέλους της ομάδας συμπληρώνουν τις εξαντλητικές δυνατότητες ελέγχου και εκτέλεσης του ρομποτικού πράκτορα. Αντίθετα, ένα χαμηλό επίπεδο εμπιστοσύνης μεταξύ των συναδέλφων ανθρώπων - ρομπότ μπορεί να οδηγήσει τον άνθρωπο να αρνηθεί να αναθέσει καθήκοντα στον ρομποτικό πράκτορα ή μερικές φορές να αποφασίσει να απενεργοποιήσει τον ρομποτικό πράκτορα [111]. Δεδομένου ότι υπάρχει συχνά υψηλή συσχέτιση μεταξύ της εμπιστοσύνης και της απόδοσης της εργασίας στη συνεργασία ανθρώπου-ρομπότ, η εμπιστοσύνη έχει μοντελοποιηθεί με βάση την απόδοση σε πολλές έρευνες [101] [7]. Τα περισσότερα από αυτά τα μοντέλα εμπιστοσύνης με βάση την απόδοση χρησιμοποιούνται ως βρόχος ανατροφοδότησης για την προσαρμογή της πραγματικής απόδοσης του ρομπότ στις προσδοκίες του ανθρώπου, ώστε να πεισθεί ο ίδιος να ενεργήσει με εμπιστοσύνη προς το ρομπότ. Υπάρχουν επίσης ορισμένα μοντέλα εμπιστοσύνης που βασίζονται στην απόδοση των λειτουργιών των ρομπότ, τα οποία δεν αποσκοπούν στην τροποποίηση της απόδοσης της συνεργασίας. Για παράδειγμα, στο [112] ένα μοντέλο εμπιστοσύνης με βάση την απόδοση για εργασίες πολλαπλών ρομπότ έχει σχεδιαστεί για να ανιχνεύει ρομποτικούς πράκτορες που δεν είναι αξιόπιστοι. Στη συνέχεια, στους λιγότερο αξιόπιστους πράκτορες ανατίθενται λιγότερο κρίσιμες αρμοδιότητες ή μερικές φορές αγνοούνται κατά την ανάθεση καθηκόντων. Από την άλλη πλευρά, η απόδοση της συνεργασίας ανθρώπου-ρομπότ μπορεί επίσης να μοντελοποιηθεί με βάση την εμπιστοσύνη [113] [103]. Αυτά τα μοντέλα χρησιμοποιούνται για την τροποποίηση της συμπεριφοράς του ρομπότ που αφορά την εμπιστοσύνη, ώστε να διαχειρίζονται τη συνολική απόδοση της συνεργασίας.

Μια διαδεδομένη κατηγορία συνεργασιών ανθρώπου-ρομπότ είναι η εποπτική συνεργασία στην οποία ο άνθρωπος παίζει το ρόλο του επόπτη και το ρομπότ το ρόλο του εργάτη. Ο επόπτης αναθέτει καθήκοντα στον εργαζόμενο και επιβλέπει την εκτέλεση της εργασίας. Ο επόπτης έχει επίσης την εξουσία να αναλαμβάνει τον έλεγχο του ρομπότ όταν το ρομπότ κάνει κάτι λάθος. Το μοντέλο εμπιστοσύνης για την εποπτική συνεργασία που παρουσιάζεται στο [101] βασίζεται στην εμπιστοσύνη στη συνεργασία μεταξύ ανθρώπων. Δημιουργεί μια ποσότητα που δείχνει τη συμβατότητα της απόδοσης του ρομπότ με τις ανθρώπινες προσδοκίες, επιτρέποντας στο ρομπότ να τροποποιήσει την απόδοσή του ώστε να ικανοποιήσει τις ανθρώπινες προσδοκίες και να βελτιώσει την εμπιστοσύνη. Αργότερα αυτό το μοντέλο εμπιστοσύνης βελτιώθηκε [114] και στο σχεδιασμό του μοντέλου εμπιστοσύνης συμπεριλήφθηκαν περισσότεροι παράγοντες που επηρεάζουν την εμπιστοσύνη στην εποπτική συνεργασία, όπως το ποσοστό αποτυχίας στον αυτόνομο πράκτορα και το ποσοστό παρέμβασης του επόπτη. Το Online Probabilistic Trust Inference Model (OPTIMo) [99] είναι ένα άλλο μοντέλο εμπιστοσύνης στην εποπτική συνεργασία που εισήχθη από την ίδια

ερευνητική ομάδα. Αυτό το μοντέλο διαμορφώνει Μπεϋζιανές (Bayesian) πεποιθήσεις σχετικά με την κατάσταση εμπιστοσύνης του ανθρώπου με βάση την απόδοση του ρομπότ στην εργασία με την πάροδο του χρόνου για να δημιουργήσει μια εκτίμηση της εμπιστοσύνης του ανθρώπου σε πραγματικό χρόνο.

Όταν η εμπιστοσύνη μπορεί να μοντελοποιηθεί και να μετρηθεί σε πραγματικό χρόνο στη συνεργασία ανθρώπου-ρομπότ, μπορεί να βοηθήσει το ρομπότ να αποκαταστήσει την εμπιστοσύνη κάθε φορά που ο άνθρωπος αρχίζει να μην εμπιστεύεται επαρκώς το ρομπότ [99]. Ένα μοντέλο εμπιστοσύνης σε πραγματικό χρόνο (trust-POMDP) για τη συνεργασία ανθρώπου-ρομπότ με ισότιμους συνεργάτες εισάγεται στο [115], το οποίο ενσωματώνει τη μετρούμενη εμπιστοσύνη στη λήψη αποφάσεων του ρομπότ. Το μοντέλο trust-POMDP κλείνει το βρόχο μεταξύ της μετρούμενης εμπιστοσύνης από το μοντέλο εμπιστοσύνης πραγματικού χρόνου και της διαδικασίας λήψης αποφάσεων του ρομπότ για τη μεγιστοποίηση της απόδοσης της συνεργασίας. Αυτό το μοντέλο παρέχει σε ένα ρομπότ τη δυνατότητα να επηρεάζει συστημικά την ανθρώπινη εμπιστοσύνη, ώστε να μειώνει και να αυξάνει την εμπιστοσύνη σε καταστάσεις υπερβολικής και υπο-εμπιστοσύνης, αντίστοιχα.

Ορισμένες σύγχρονες μελέτες στη μοντελοποίηση της εμπιστοσύνης χρησιμοποιούν τεχνικές μέτρησης της υποκειμενικής εμπιστοσύνης σε συνδυασμό με τεχνικές μέτρησης της αντικειμενικής εμπιστοσύνης. Οι τεχνικές αυτές χρησιμοποιούνται τόσο στην HAI όσο και στην HRI για την αύξηση της ακρίβειας και της αξιοπιστίας των μετρήσεων εμπιστοσύνης. Υπάρχουν μερικές μελέτες στην εμπιστοσύνη ανθρώπου-αυτοματισμού, στην αλληλεπίδραση ανθρώπου-υπολογιστή και στην ανθρώπινη εμπιστοσύνη στην τεχνητή νοημοσύνη που χρησιμοποιούν ψυχοφυσιολογικές μετρήσεις για τη μοντελοποίηση της εμπιστοσύνης [116] [117]. Οι Khalid κ.ά. [118] εισάγουν επίσης ένα μοντέλο εμπιστοσύνης στην ανθρώπινη τεχνητή νοημοσύνη, το οποίο χρησιμοποιεί εκφράσεις προσώπου, χαρακτηριστικά φωνής και εξαγόμενα χαρακτηριστικά καρδιακού ρυθμού σε συνδυασμό με την αυτοαναφερόμενη εμπιστοσύνη των ανθρώπων για τη μοντελοποίηση της εμπιστοσύνης. Αυτό το μοντέλο εμπιστοσύνης ταξινομεί το επίπεδο εμπιστοσύνης σε χαμηλό, φυσικό και υψηλό επίπεδο εμπιστοσύνης χρησιμοποιώντας έναν ταξινομητή εμπιστοσύνης με νευρο-ασαφή (Neuro-fuzzy) χαρακτήρα.

2.3 Προκλήσεις και προοπτικές στην ανάπτυξη αξιόπιστων ρομπότ

Η εμπιστοσύνη ανθρώπου-ρομπότ είναι ζωτικής σημασίας στον σημερινό κόσμο όπου τα σύγχρονα κοινωνικά ρομπότ αναπτύσσονται όλο και περισσότερο. Στην υγειονομική περίθαλψη, τα ρομπότ χρησιμοποιούνται για την αποθεραπεία των ασθενών [119] και για την παροχή βοήθειας πρώτης γραμμής κατά τη διάρκεια της πανδημίας COVID-19 [120]. Στο πλαίσιο της εκπαίδευσης, τα προσωποποιημένα κοινωνικά ρομπότ χρησιμοποιούνται ως δάσκαλοι για να βοηθήσουν τη μάθηση των παιδιών [121]. Συγχρόνως, λογικό είναι να αναδειχθούν και πρόβληματα όσον αφορά την εμπιστοσύνη ενός χειριστή σε ένα ρομποτ είτε αυτό είναι το πρόβλημα της έλλειψης εμπιστοσύνης, είτε η μεγιστοποίηση της που μπορεί να μην οδηγεί απαραίτητα σε θετικά αποτελέσματα αλληλεπίδρασης. Η υπερβολική εμπιστοσύνη [122] [123] είναι επίσης εξαιρετικά ανεπιθύμητη, ιδίως στα παραπάνω περιβάλλοντα υγειονομικής περίθαλψης και εκπαίδευσης. Για παράδειγμα, το [122] έδειξε ότι οι άνθρωποι ήταν πρόθυμοι να αφήσουν ένα άγνωστο ρομπότ να εισέλθει σε περιορισμένους χώρους (π.χ. κρατώντας του την πόρτα), εγείροντας έτσι ανησυχίες σχετικά με την ασφάλεια, το απόρρητο και την προστασία της ιδιωτικής ζωής. Ίσως ακόμη πιο δραματικά, μια μελέτη του [123] έδειξε ότι οι άνθρωποι αγνόησαν πρόθυμα την πινακίδα εξόδου κινδύνου για να ακολουθήσουν ένα ρομπότ εκκένωσης που έκανε λάθος στροφή κατά τη διάρκεια μιας (προσομοιωμένης αλλά ρεαλιστικής) έκτακτης ανάγκης πυρκαγιάς, ακόμη και όταν το εν λόγω ρομπότ λειτουργούσε αναποτελεσματικά πριν από την έναρξη της έκτακτης ανάγκης. Αυτά τα παραδείγματα οδηγούν σε ένα βασικό μήνυμα: η λανθασμένη εμπιστοσύνη μπορεί να οδηγήσει σε κακή χρήση των ρομπότ.

Η σημασία της εμπιστοσύνης στα κοινωνικά ρομπότ σίγουρα δεν έχει περάσει απαρατήρητη στη βιβλιογραφία - υπάρχουν αρκετές ενδιαφέρουσες αναλύσεις σχετικά με την εμπιστοσύνη στα ρομπότ σε ένα ευρύ φάσμα θεμάτων, όπως η αποκατάσταση της εμπιστοσύνης [124] [125], η εμπιστοσύνη στην αυτοματοποίηση [5] [1] [3], η εμπιστοσύνη στη ρομποτική της υγειονομικής περίθαλψης [119], η μέτρηση της εμπιστοσύνης [126] [127] και η πιθανοτική μοντελοποίηση της εμπιστοσύνης (probabilistic trust modeling) [128]. Ακριβώς, όμως, όπως οι εξελίξεις στη μηχανολογία μας έφεραν στο κατώφλι μιας ρομποτικής επανάστασης, έτσι σωστό είναι να εξεταστεί αν οι πρόσφατες μεθοδολογικές ανακαλύψεις μπορούν ομοίως να βοηθήσουν στην απάντηση ενός θεμελιώδους ανθρώπινου ερωτήματος: εμπιστευόμαστε τα ρομπότ να ζουν ανάμεσά μας και, αν όχι, μπορούμε να δημιουργήσουμε ρομπότ που να είναι σε θέση να κερδίσουν την εμπιστοσύνη μας;

Για να ληφθεί μια ουσιαστική απάντηση στο ερώτημα που αναφέρθηκε, πρέπει να εξακριβώσουμε τι ακριβώς εννοούμε με τον όρο ”εμπιστοσύνη σε ένα ρομπότ”. Είναι η έννοια της εμπιστοσύνης σε αυτοματοποιημένα συστήματα (π.χ. ένα σύστημα προειδοποίησης για ανεμοθώρακα) ισοδύναμη με την εμπιστοσύνη σε ένα κοινωνικό ρομπότ; Πώς διαφέρει η εμπιστοσύνη από έννοιες όπως η αξιοπιστία και η φήμη;

Ιστορικά, η εμπιστοσύνη έχει μελετηθεί όσον αφορά τα αυτοματοποιημένα συστήματα [5], [3]. Έκτοτε, έχει καταβληθεί μεγάλη προσπάθεια για την επέκταση της μελέτης της εμπιστοσύνης στην αλληλεπίδραση ανθρώπου-ρομπότ (HRI). Χρησιμοποιούμε τον όρο ”ρομπότ” για να αναφερθούμε σε ενσαρκωμένους πράκτορες με φυσική εκδήλωση που προορίζονται να λειτουργούν σε θορυβώδη, δυναμικά περιβάλλοντα [124]. Από την άλλη πλευρά, ένα αυτοματοποιημένο σύστημα θα μπορούσε να είναι μια ηλεκτρονική διαδικασία χωρίς συγκεκριμένη φυσική μορφή. Ενώ η έρευνα σχετικά με την εμπιστοσύνη στην αυτοματοποίηση μπορεί να ενημερώσει την κατανόηση της εμπιστοσύνης στα ρομπότ [5], αυτή η διάκριση μεταξύ ρομπότ και αυτοματισμού (η οποία, ομολογουμένως, δεν είναι αρκετά ευκρινής) έχει σημαντικές επιπτώσεις στον τρόπο με τον οποίο εννοιολογούμε την εμπιστοσύνη. Πρώτον, η φυσική ενσάρκωση ενός ρομπότ καθιστά το σχεδιασμό του βασικό παράγοντα στη διαμόρφωση της εμπιστοσύνης. Επιπλέον, οραματιζόμαστε ότι τα κοινωνικά ρομπότ θα αναπτύσσονται συνήθως σε δυναμικά και αδόμητα περιβάλλοντα και θα πρέπει να εργάζονται παράλληλα με τους ανθρώπους. Αυτό υποδηλώνει ότι η ικανότητα πλοήγησης στην αβεβαιότητα και στα κοινωνικά πλαίσια παίζει μεγαλύτερο ρόλο στη διαμόρφωση και διατήρηση της ανθρώπινης εμπιστοσύνης.

Η εμπιστοσύνη θα πρέπει επίσης να διακρίνεται από τις συναφείς έννοιες της αξιοπιστίας και της φήμης. Η εμπιστοσύνη (σε έναν πράκτορα) είναι η ιδιότητα του ανθρώπινου χρήστη σε σχέση με τον εν λόγω πράκτορα [129]. Αντίθετα, η αξιοπιστία είναι ιδιότητα του πράκτορα και όχι του ανθρώπινου χρήστη [129]. Ως εκ τούτου, ένας ανθρώπινος χρήστης μπορεί να μην εμπιστεύεται ένα αξιόπιστο ρομπότ (και το αντίστροφο). Η διάκριση μεταξύ εμπιστοσύνης και φήμης είναι ελαφρώς πιο διαφοροποιημένη. Ενώ και οι δύο μπορούν να θεωρηθούν ως ”γνώμη” σχετικά με τον εν λόγω πράκτορα, η φήμη δεν περιλαμβάνει μόνο τη γνώμη του μεμονωμένου ανθρώπινου χρήστη (όπως στην εμπιστοσύνη) αλλά και τη συλλογική γνώμη άλλων ανθρώπων [130].

Πέρα από τις διακρίσεις, δυστυχώς δεν υπάρχει στη βιβλιογραφία ένας ενιαίος ορισμός της εμπιστοσύνης [1]. Αυτό έχει οδηγήσει στον πληθωρισμό ποιοτικά διαφορετικών τρόπων ορισμού της εμπιστοσύνης. Για παράδειγμα, η εμπιστοσύνη έχει θεωρηθεί ως πεποίθηση [1], στάση [1], συναισθηματική αντίδραση [131], αίσθηση προθυμίας [132], μορφή αμοιβαίας κατανόησης [133] και πράξη εμπιστοσύνης [134].

Για αυτής της ασάφειας όσον αφορά τους ορισμούς, ορίσθηκε η εμπιστοσύνη σε τρία μέρη, καθένα από τα οποία βασίζεται σε προηγούμενες δουλειές:

- Πρώτον, η έννοια της εμπιστοσύνης μπορεί να προκύψει μόνο σε μια κατάσταση που περιλαμβάνει αβεβαιότητα και ευπάθεια [11].

- Δεύτερον, η εμπιστοσύνη είναι μια πολύπλευρη κατασκευή που είναι κρυφή και δεν μπορεί

να παρατηρηθεί άμεσα [135] και

- Τρίτον, η εμπιστοσύνη διαμεσολαβεί στη σχέση μεταξύ της ιστορίας των παρατηρούμενων γεγονότων και της επακόλουθης πράξης εμπιστοσύνης του πράκτορα [1] [3].

2.3.1 Σχεδιασμός με έμφαση στην ανάπτυξη της εμπιστοσύνης

Ένα πρώτο βήμα για την εδραίωση της εμπιστοσύνης μεταξύ ανθρώπου και ρομπότ είναι ο σχεδιασμός του ρομπότ με τρόπο που να προτρέπει τους ανθρώπους να το εμπιστεύονται κατάλληλα. Συμβατικά, ένας τρόπος για να γίνει αυτό είναι η διαμόρφωση της φυσικής εμφάνισης του ρομπότ [5]. Με τους ανθρώπους, η απόφασή μας να εμπιστευτούμε ένα άτομο εξαρτάται συχνά από τις πρώτες εντυπώσεις [136]. Το ίδιο ισχύει και για τα ρομπότ - οι άνθρωποι συχνά κρίνουν την αξιοπιστία ενός ρομπότ με βάση τη φυσική του εμφάνιση [5] [137] [138] [139]. Για παράδειγμα, τα ρομπότ που έχουν χαρακτηριστικά που μοιάζουν με τα ανθρώπινα τείνουν να θεωρούνται πιο αξιόπιστα [137], αλλά μόνο μέχρι ενός σημείου. Τα ρομπότ που φαίνονται να μοιάζουν πολύ, αλλά εξακολουθούν να διακρίνονται αρκετά από τους ανθρώπους, θεωρούνται λιγότερο αξιόπιστα [138] [139]. Αυτή η πτώση στην αντιλαμβανόμενη αξιοπιστία ανακάμπτει στη συνέχεια για τα ρομπότ που φαίνονται δυσδιάκριτα από τους ανθρώπους. Αυτό το φαινόμενο -που αποκαλείται "κοιλάδα του αλλόκοτου" [124] [138] [139] λόγω της συνάρτησης σχήματος U που συνδέει την αντιλαμβανόμενη αξιοπιστία με τον ανθρωπομορφισμό των ρομπότ- έχει σημαντικές επιπτώσεις για το σχεδιασμό κοινωνικών ρομπότ.

Πιο πρόσφατα, αρκετές έρευνες αποκάλυψαν ότι ο αντίκτυπος του σχεδιασμού στην εμπιστοσύνη ξεπερνά την απλή φυσική εμφάνιση. Για παράδειγμα, ο τρόπος με τον οποίο απεικονίζεται το ρομπότ μπορεί να επηρεάσει σε μεγάλο βαθμό την αντιληπτή αξιοπιστία του ρομπότ. Όταν ένα κοινωνικό ρομπότ παρουσιάζεται απλώς ως έχει, εμπειρικά στοιχεία δείχνουν ότι οι ανθρώπινοι χρήστες τείνουν να υπερεκτιμούν τις ικανότητες του ρομπότ [140] πριν εμπλακούν μαζί του. Αυτή η αναντιστοιχία μεταξύ προσδοκιών και πραγματικότητας -το χάσμα προσδοκιών- οδηγεί στη συνέχεια σε μια περιττή απώλεια εμπιστοσύνης μετά την αλληλεπίδραση των χρηστών με το ρομπότ [140]. Σε μια προσπάθεια να ενθαρρυνθεί ο σχηματισμός εμπιστοσύνης στο ρομπότ, αρκετές μελέτες έχουν έκτοτε προσπαθήσει να καλύψουν αυτό το κενό παρέχοντας πρόσθετες πληροφορίες μαζί με το ρομπότ. Αυτό μπορεί να γίνει χειροκίνητα με την κατάλληλη διαμόρφωση των δυνατοτήτων του ρομπότ πριν από οποιαδήποτε αλληλεπίδραση [141].

Εναλλακτικά, μπορεί επίσης να γίνει αυτόματα με τον αλγοριθμικό εντοπισμό αρκετών "κρίσιμων καταστάσεων" - εκείνων των καταστάσεων όπου η ενέργεια που λαμβάνεται έχει ισχυρό αντίκτυπο στο τελικό αποτέλεσμα - από την πολιτική του ρομπότ και την παρουσίαση της συμπεριφοράς του ρομπότ σε αυτές τις καταστάσεις στο χρήστη [142]. Οι μελέτες αυτές αποδεικνύουν ότι οι κατάλληλα σχεδιασμένες συμπληρωματικές πληροφορίες μπορούν να βοηθήσουν στην ενίσχυση της εμπιστοσύνης των ανθρώπων στο ρομπότ, προσαρμόζοντας τις αρχικές προσδοκίες των ανθρώπων σε ένα κατάλληλο επίπεδο. Με άλλα λόγια, ο σχεδιασμός αξιόπιστων ρομπότ δεν σταματά απλώς σε εκτιμήσεις σχετικά με το μέγεθος, το ύψος και τη φυσική σύσταση του ρομπότ. Αντίθετα, ο επιτυχημένος σχεδιασμός απαιτεί προσεκτική σκέψη σχετικά με τον τρόπο με τον οποίο το ρομπότ παρουσιάζεται στον χρήστη.

2.3.2 Απόκτηση, διαφύλαξη και βαθμονόμηση της εμπιστοσύνης

Ενώ ο σχεδιασμός ενός ρομπότ μπορεί να προετοιμάσει τον χρήστη να υιοθετήσει ένα ορισμένο επίπεδο εμπιστοσύνης σε αυτό, η αποτελεσματικότητά του συχνά εξαρτάται από το πλαίσιο [143] και τις ατομικές διαφορές μεταξύ των ανθρώπων-χρηστών [144]. Ως εκ τούτου, ο σχεδιασμός από μόνος του μπορεί να μην προκαλεί το κατάλληλο επίπεδο εμπιστοσύνης. Το ρομπότ εξακολουθεί να πρέπει να κερδίζει, να διατηρεί και να βαθμονομεί ενεργά την εμπιστοσύνη του χρήστη για να υπάρξει επιτυχής συνεργασία [145]. Αυτό είναι ιδιαίτερα δύσκολο, αν αναλογιστεί κανείς

Εικόνα 1. *Τα συνεργατικά ρομπότ επιτρέπουν σε ανθρώπους και μηχανές να εργάζονται δίπλα-δίπλα σε γραμμές συναρμολόγησης. Φωτογραφία: ευγενική προσφορά της Universal Robots A/S*

ότι οι ουσιαστικές κοινωνικές αλληλεπιδράσεις συχνά λαμβάνουν χώρα για μεγάλο χρονικό διάστημα. Η εμπιστοσύνη του χρήστη στο ρομπότ δεν είναι ένα στατικό φαινόμενο - η εμπιστοσύνη κυμαίνεται δυναμικά καθώς η αλληλεπίδραση εξελίσσεται με την πάροδο του χρόνου [146]. Για να το αντιμετωπίσει αυτό, το εν λόγω ρομπότ πρέπει να αναπτύξει αποτελεσματικές στρατηγικές για την πλοήγηση στο μεταβαλλόμενο τοπίο εμπιστοσύνης. Στη συνέχεια, έχουμε οργανώσει τις υπάρχουσες στρατηγικές στη βιβλιογραφία σε τέσσερις μεγάλες ομάδες κατά σειρά αυξανόμενης πολυπλοκότητας του μοντέλου, ξεκινώντας από (i) ευρετικές (heuristics) τεχνικές και (ii) τεχνικές που εκμεταλλεύονται τη διαδικασία αλληλεπίδρασης και προχωρώντας σε (iii) υπολογιστικά μοντέλα εμπιστοσύνης.

Ευρετικές μέθοδοι (Heuristics)

Μια σημαντική κατηγορία στρατηγικών βαθμονόμησης της εμπιστοσύνης στη βιβλιογραφία έχει τη μορφή ευρετικών μεθόδων. Ο σχεδιασμός αυτών των ευρετικών μεθόδων βασίζεται συχνά σε εμπειρικά στοιχεία από την ψυχολογία. Οι ευρετικές μέθοδοι έχουν προταθεί για την αντιμετώπιση δύο διαφορετικών καταστάσεων: για την καταπολέμηση της υπερβολικής εμπιστοσύνης και για την αποκατάσταση της εμπιστοσύνης. Ως παράδειγμα της πρώτης, το [134] πρότεινε τη χρήση οπτικών προτροπών για να ωθήσει τους χρήστες να επανεκτιμήσουν την εμπιστοσύνη τους στο ρομποτικό σύστημα όταν ο χρήστης έχει αφήσει το αυτοματοποιημένο σύστημα να λειτουργεί χωρίς επίβλεψη για πολύ καιρό.

Σε ορισμένες περιπτώσεις, τα ρομπότ ενδέχεται να αντιμετωπίσουν σφάλματα που θα μπορούσαν ενδεχομένως να βλάψουν την εμπιστοσύνη του χρήστη [147] [148]. Αυτό δεν είναι ασυνήθιστο φαινόμενο, καθώς τα ρομπότ είναι ασταθή, ιδίως όταν χρησιμοποιούνται για μεγάλο χρονικό διάστημα. Ως εκ τούτου, είναι ζωτικής σημασίας για το ρομπότ να αντιδράσει κατάλληλα μετά από ένα σφάλμα για να αποκαταστήσει την εμπιστοσύνη του χρήστη. Αυτή η διαδικασία είναι κοινώς γνωστή ως αποκατάσταση της εμπιστοσύνης [125] και έχουν εντοπιστεί διάφορες στρατηγικές από τους ερευνητές για την επίτευξη αυτού του στόχου [124] [125]. Για παράδειγμα, το ρομπότ μπορεί να δώσει μια εξήγηση για την αποτυχία ή να προτείνει εναλλακτικά σχέδια [125]. Ωστόσο, είναι

κρίσιμο να ληφθεί υπόψη το πλαίσιο της κατάστασης πριν από την επιλογή μιας συγκεκριμένης στρατηγικής επιδιόρθωσης. Πρόσφατες μελέτες [149] έχουν δείξει ότι οι συγγνώμες μπορούν να αποτελέσουν αποτελεσματικό μέσο αποκατάστασης της εμπιστοσύνης όταν η έλλειψη ικανότητας του ρομπότ ευθύνεται για την απώλεια εμπιστοσύνης. Αντίθετα, όταν οι ενέργειες του ρομπότ γίνονται αντιληπτές ως σκόπιμες, η άρνηση οποιασδήποτε πρόθεσης μπορεί να είναι πιο αποτελεσματική στρατηγική αποκατάστασης από τη συγγνώμη. Χρησιμοποιώντας αυτές τις στρατηγικές με σύνεση, μπορεί να εξασφαλιστεί η επιτυχής αλληλεπίδραση ανθρώπου-ρομπότ.

Αξιοποιώντας τη διαδικασία της αλληλεπίδρασης

Όταν επιδιώκεται η δημιουργία εμπιστοσύνης μεταξύ των ανθρώπων και των ρομπότ, είναι σημαντικό να εξεταστούν όχι μόνο οι ευρετικές αρχές που καθοδηγούν τις ενέργειές τους, αλλά και η συμπεριφορά τους. Έρευνες έχουν δείξει ότι τα άτομα τείνουν να εμπιστεύονται περισσότερο τα ρομπότ που μεταφέρουν τους περιορισμούς τους μέσω κατατοπιστικών κινήσεων των χεριών τους [150], παρέχουν σαφείς αιτιολογήσεις για τις ενέργειές τους [151] [152] ή προσφέρουν πιο λεπτομερείς πληροφορίες [153]. Σε γνωστικό επίπεδο, οι άνθρωποι τείνουν να μην εμπιστεύονται τα ρομπότ που αναλαμβάνουν κινδύνους σε αβέβαια περιβάλλοντα [154], αν και αυτό μπορεί να διαφέρει ανάλογα με τις ατομικές προτιμήσεις τους για τον κίνδυνο [155]. Αντίθετα, τα ρομπότ που εκφράζουν ευπάθεια [2] ή συναισθήματα [156] μέσω της φυσικής γλώσσας τείνουν να θεωρούνται πιο αξιόπιστα. Επιπλέον, η χρήση της γλώσσας για την επικοινωνία με τους χρήστες μπορεί να συμβάλει στον μετριασμό της πιθανής απώλειας εμπιστοσύνης που προκύπτει από αποτυχίες απόδοσης [157]. Τελικά, η οικοδόμηση εμπιστοσύνης με τα ρομπότ είναι μια προσπάθεια προσανατολισμένη στη διαδικασία, με έμφαση στην επίτευξη μιας εργασίας με αξιόπιστο και συνεπή τρόπο.

Υπολογιστικά μοντέλα εμπιστοσύνης

Οι τεχνικές που εξετάστηκαν παραπάνω βασίζονται σε προ-προγραμματισμένες στρατηγικές, οι οποίες μπορεί να είναι δύσκολο να προσαρμοστούν για ρομπότ που πρέπει να λειτουργούν σε πολλά διαφορετικά πλαίσια. Μια πιο γενική προσέγγιση είναι η άμεση μοντελοποίηση της δυναμικής εμπιστοσύνης του ανθρώπου στο ρομπότ. Οι εργασίες σε αυτόν τον τομέα έχουν επικεντρωθεί σε δύο προβλήματα:

- την εκτίμηση της εμπιστοσύνης με βάση παρατηρήσεις της ανθρώπινης συμπεριφοράς [128] [158] [159] [101] [160][161][99] και
- τη χρήση της εκτίμησης της εμπιστοσύνης για την καθοδήγηση της συμπεριφοράς του ρομπότ [158] [101] [115][162]

Μια σημαντική σειρά εργασιών στον τομέα αυτό ξεκίνησε με την εισαγωγή του Online Proba--bilistic Trust Inference Model (OPTIMo) [99], το οποίο αποτυπώνει την εμπιστοσύνη ως λανθάνουσα μεταβλητή σε ένα δυναμικό πιθανοτικό γραφικό μοντέλο (Probabilistic Graphical Model/ PGM) [163]. Ενώ υπήρξαν και άλλες πρωτοποριακές προσπάθειες μοντελοποίησης της εμπιστοσύνης, περιορίστηκαν σε απλές συναρτήσεις [161] ή απέτυχαν να λάβουν υπόψη την αβεβαιότητα στην εκτίμηση της εμπιστοσύνης [160]. Συγκριτικά, η πιθανολογική γραφική προσέγγιση που παρουσιάζεται στο OPTIMo αξιοποιεί ισχυρές τεχνικές εξαγωγής συμπερασμάτων που επιτρέπουν την εκτίμηση της εμπιστοσύνης σε πραγματικό χρόνο. Επιπλέον, αυτή η προσέγγιση λαμβάνει υπόψη τόσο την αβεβαιότητα της εκτίμησης όσο και τη δυναμική φύση της εμπιστοσύνης με τρόπο που βασίζεται σε αρχές μέσω της Μπεϋζιανής συμπερασματολογίας [163]. Αυτή η προσέγγιση γίνεται ίσως ακόμη πιο ενδιαφέρουσα από τα στοιχεία της γνωστικής επιστήμης ότι οι άνθρωποι ενεργούν με έναν ορθολογικό τρόπο κατά Bayes [164], γεγονός που υποδηλώνει ότι τα ρομπότ που είναι εξοπλισμένα με αυτή την παραλλαγή του μοντέλου εμπιστοσύνης συλλογίζονται στην

πραγματικότητα με μια έγκυρη προσέγγιση της εμπιστοσύνης του ανθρώπινου χρήστη. Με αυτό το πλαίσιο γραφικού μοντέλου, μπορούμε να μετατρέψουμε τα εννοιολογικά διαγράμματα της δυναμικής της εμπιστοσύνης σε ένα ποσοτικό πρόγραμμα που μπορεί να ελεγχθεί μέσω υποθέσεων. [165].

Από την ανάπτυξη του OPTIMo, άλλα άρθρα συνέβαλαν σε σημαντικές επεκτάσεις. Για παράδειγμα, το [159] μοντελοποίησε την εμπιστοσύνη με ένα β-διωνυμικό (αντί για γραμμική Γκαουσιανή) και διερεύνησε πώς το μοντέλο μπορεί να χρησιμοποιηθεί για την ομαδοποίηση ατόμων με βάση το προφίλ ”εμπιστοσύνης” τους. Δυναμικά Μπαγιεσιανά δίκτυα αυτού του είδους έχουν επίσης χρησιμοποιηθεί για την καθοδήγηση των ενεργειών του ρομπότ. Σε αρκετές εργασίες [158], η εκτιμώμενη εμπιστοσύνη χρησιμοποιήθηκε ως μηχανισμός για να αποφασίσει το ρομπότ εάν ο έλεγχος του ρομπότ θα πρέπει να παραχωρηθεί στον άνθρωπο. Στις εργασίες [115] [162], μια παραλλαγή OPTIMo ενσωματώθηκε στο σχεδιασμό POMDP, επιτρέποντας έτσι στο ρομπότ να αποκτήσει μια τακτική που αιτιολογεί τη λανθάνουσα εμπιστοσύνη του ανθρώπινου χρήστη. Αυτή η Μπεϋζιανή προσέγγιση της αιτιολόγησης σχετικά με την εμπιστοσύνη έχει επίσης διερευνηθεί μη παραμετρικά χρησιμοποιώντας διαδικασίες Gauss [166]. Τέλος, αυτό το πλαίσιο έχει επίσης επεκταθεί για τη μοντελοποίηση της εμπιστοσύνης ενός χρήστη σε πολλαπλά ρομπότ [167], ανοίγοντας έτσι το δρόμο για τη δυναμική μοντελοποίηση της εμπιστοσύνης σε περιβάλλοντα πολλαπλών πρακτόρων.

2.3.3 Ερευνητικές προκλήσεις

Καθώς η επιστημονική πρόοδος προοδεύει, φέρνει νέες προκλήσεις και συναρπαστικές δυνατότητες. Ο τομέας της έρευνας σχετικά με την αξιοπιστία των ρομπότ δεν διαφέρει. Στο πλαίσιο αυτό, εντοπίζονται τρεις μεγάλες προκλήσεις:

Μέτρηση της εμπιστοσύνης σε πραγματικές συνθήκες

Η εμπιστοσύνη του ανθρώπου σε ένα ρομπότ είναι ένα μη παρατηρήσιμο φαινόμενο. Ως εκ τούτου, την τελευταία δεκαετία έχουν γίνει σημαντικές πρωτοποριακές προσπάθειες για την ανάπτυξη εργαλείων που μετρούν την εμπιστοσύνη στην αλληλεπίδραση ανθρώπου-ρομπότ, κυρίως με τη μορφή κλιμάκων αυτο-αναφοράς [146] [69] [168] [169] [170]. Πιο πρόσφατα, υπήρξε μια τάση απομάκρυνσης από τα μέτρα αυτοαναφοράς προς πιο ”αντικειμενικά” μέτρα εμπιστοσύνης [126] [127] [171] [172]. Αυτά περιλαμβάνουν φυσιολογικά μέτρα όπως η παρακολούθηση των ματιών [173], κοινωνικές ενδείξεις που εξάγονται από βίντεο/κάμερες [174], ήχο [175], δερματική απόκριση [176] [177] και νευρωνικά μέτρα [176] [177] [178], καθώς και τη συμπεριφορά παιχνιδιού σε συμπεριφορικά οικονομικά παιχνίδια [174].

Παρόλο που έχει καταβληθεί σημαντική προσπάθεια για τη βελτίωση της μέτρησης της εμπιστοσύνης, υπάρχουν ακόμη ορισμένα δύσκολα ζητήματα που πρέπει να αντιμετωπιστούν. Ενώ υπάρχουν διαθέσιμες επικυρωμένες κλίμακες, είναι ανησυχητικό το γεγονός ότι δεν υπάρχουν αρκετές επιβεβαιωτικές δοκιμές έναντι της διερευνητικής ανάλυσης παραγόντων. Εξαίρεση αποτελεί η κλίμακα εμπιστοσύνης που αναπτύχθηκε στο [168] και προοριζόταν να εξετάσει την εμπιστοσύνη στην αυτοματοποίηση (όχι ειδικά για τα ρομπότ), της οποίας η παραγοντική δομή επιβεβαιώθηκε ξεχωριστά στο [179].

Επιπλέον, η βιβλιογραφία είναι σχετικά σιωπηλή σχετικά με το θέμα της αναλλοίωτης μέτρησης [180]: βρέθηκε μόνο μια αναφορά στο [179]. Εν συντομία, μια κλίμακα που παρουσιάζει αναλλοίωτη μέτρηση μετρά την ίδια κατασκευή (π.χ. εμπιστοσύνη) σε διαφορετικές ομάδες σύγκρισης ή περιστάσεις [180]. Εάν η αναλλοίωτη μέτρηση δεν ικανοποιείται, τότε οι διαφορές στις παρατηρούμενες βαθμολογίες μεταξύ δύο πειραματικών ομάδων ή δύο χρονικών στιγμών, ακόμη και αν είναι στατιστικά σημαντικές, μπορεί να μην αντανακλούν πραγματικές διαφορές στην εμπιστοσύνη. Τέλος, υπάρχει έλλειψη πληροφοριών σχετικά με τις ψυχομετρικές ιδιότητες, όπως η

αξιοπιστία [181], των "αντικειμενικών" μέτρων. Παρά τις θετικές ιδιότητες που αναφέρθηκαν παραπάνω, τα "αντικειμενικά" μέτρα δεν είναι απαλλαγμένα από ψυχομετρικούς προβληματισμούς [182]. Αντίθετα, τα "αντικειμενικά" μέτρα μπορούν να θεωρηθούν ως εκδηλώσεις της υποκείμενης εμπιστοσύνης και θα πρέπει να εξετάζονται εξονυχιστικά για την αξιοπιστία και την εγκυρότητά τους, όπως ακριβώς συμβαίνει και με τα ερωτηματολόγια αυτοαναφοράς.

Γεφυρώνοντας τα αξιόπιστα ρομπότ και την ανθρώπινη εμπιστοσύνη

Μέχρι στιγμής, έχει διερευνηθεί λεπτομερώς ο τρόπος με τον οποίο οι ανθρώπινοι χρήστες αναπτύσσουν εμπιστοσύνη στα ρομπότ. Ωστόσο, η βιβλιογραφία σχετικά με την ανθρώπινη εμπιστοσύνη στα ρομπότ μπορεί επίσης να εξεταστεί σε σχέση με την έρευνα για τα αξιόπιστα συστήματα (και πιο πρόσφατα, την αξιόπιστη τεχνητή νοημοσύνη [183]), όπου έχουν αναπτυχθεί πλαίσια και μέθοδοι για να διασφαλιστεί (ή να αξιολογηθεί αν) ένα δεδομένο σύστημα ικανοποιεί τις επιθυμητές ιδιότητες. Αυτές οι μέθοδοι έχουν τις ρίζες τους στον τομέα της μηχανικής λογισμικού (software engineering) και παραδοσιακά επικεντρώνονται στην ικανοποίηση μετρικών που βασίζονται σε μη λειτουργικές ιδιότητες του συστήματος (π.χ. αξιοπιστία) [184] καθώς και στην ποιότητα των υπηρεσιών (π.χ. ενσυναίσθηση ενός κοινωνικού ρομπότ) [184] [185].

Πιο πρόσφατα, υπάρχει μια τάση υιοθέτησης αυτών των τεχνικών για τη μοντελοποίηση πτυχών του ανθρώπινου χρήστη. Μια σημαντική κατηγορία τεχνικών είναι η τυπική επαλήθευση [186], οι οποίες αποτελούν ισχυρά εργαλεία που επιτρέπουν στους σχεδιαστές συστημάτων να παρέχουν εγγυήσεις για την απόδοση του συστήματος. Πρόσφατες εργασίες έχουν επεκτείνει τις τυπικές μεθόδους για να χειριστούν έννοιες όπως η δικαιοσύνη [187], η ιδιωτικότητα [188], ακόμη και το γνωστικό φορτίο [189]. Οι τεχνικές τυπικής επαλήθευσης έχουν επίσης εφαρμοστεί σε προβλήματα στην αλληλεπίδραση ανθρώπου-μηχανισμού [190], γεγονός που υποδηλώνει ότι αυτές οι μέθοδοι μπορούν επίσης να εφαρμοστούν αποδοτικά στην HRI.

Μέχρι σήμερα, πολύ λίγες εργασίες έχουν διερευνήσει τον τρόπο με τον οποίο οι τυπικές μέθοδοι μπορούν να εφαρμοστούν σε συνδυασμό με την ανθρώπινη εμπιστοσύνη στα ρομπότ. Η πρωτοποριακή εργασία [191] παρουσίασε ένα πλαίσιο επαλήθευσης και επικύρωσης που επιτρέπει στα βοηθητικά ρομπότ να αποδείξουν την αξιοπιστία τους σε μια εργασία παράδοσης. Παρομοίως, η μελέτη [192] έχει διερευνήσει τυπικές μεθόδους επαλήθευσης για τη γνωστική εμπιστοσύνη σε ένα περιβάλλον πολλαπλών πρακτόρων, στο οποίο η εμπιστοσύνη τυποποιείται ως τελεστής στη λογική. Συνδυάζοντας τη λογική με ένα πιθανοτικό μοντέλο του περιβάλλοντος, μπορεί κανείς να χρησιμοποιήσει τεχνικές τυπικής επαλήθευσης για να εκτιμήσει αν το συγκεκριμένο σύστημα ανθρώπου-ρομπότ ικανοποιεί έναν κατάλογο απαιτούμενων ιδιοτήτων που σχετίζονται με την εμπιστοσύνη. Ωστόσο, μένουν πολλά να γίνουν. Μια βασική τεχνική πρόκληση είναι η κλιμάκωση των υφιστάμενων εργαλείων στην πολυπλοκότητα της HRI. Πρέπει επίσης να διασφαλιστεί ότι τα μοντέλα που χρησιμοποιούνται στην τυπική επαλήθευση αντιπροσωπεύουν επαρκώς το HRI αντικείμενο, ώστε οι εγγυήσεις να είναι ουσιαστικές

Μια σημαντική πτυχή των αξιόπιστων συστημάτων που συνδέεται στενά με την εμπιστοσύνη και χρήζει πρόσθετης προσοχής είναι η ιδιωτικότητα. Στο [193], οι συγγραφείς εξέφρασαν την ανησυχία ότι όταν τα κοινωνικά ρομπότ αναπτύσσονται για να εργαστούν με ευάλωτες ομάδες (π.χ. παιδιά), υπάρχει η πιθανότητα τα ρομπότ να χρησιμοποιηθούν για να παραβιάσουν την ιδιωτική τους ζωή. Για παράδειγμα, αν το ευάλωτο άτομο αναπτύξει στοργή ή εμπιστοσύνη στο ρομπότ, μπορεί να "εμπιστευτεί" ακατάλληλα το ρομπότ. Αυτό μπορεί να αποτελέσει μια οδό για κατάχρηση από άλλους κακόβουλους παράγοντες, εάν το ρομπότ έχει δυνατότητες καταγραφής. Η ιδέα αυτή αναπτύχθηκε περαιτέρω στο [194], όπου έδειξαν πειραματικά ότι η εμπιστοσύνη στα ρομπότ μπορεί να αξιοποιηθεί για να πειστούν οι ανθρώπινοι χρήστες να προβούν σε επικίνδυνες πράξεις (π.χ. τυχερά παιχνίδια) και να αποκαλύψουν ευαίσθητες πληροφορίες (π.χ. πληροφορίες που χρησιμοποιούνται συχνά στην επαναφορά κωδικών πρόσβασης σε τράπεζες). Αυτά τα ζητήματα ιδιωτικότητας που σχετίζονται με την εμπιστοσύνη ξεπερνούν την εσκεμμένη εισβολή από εξωτε-

ρικούς πράκτορες όταν εξετάζουμε τον προγραμματισμό πολλαπλών πρακτόρων, όπου υπάρχει η πιθανότητα διαρροής ατομικών πληροφοριών [195]. Το πρόβλημα αυτό γίνεται ακόμη πιο έντονο αν πρόκειται για ιατρικές πληροφορίες (π.χ. διάγνωση ασθενειών που σχετίζονται με την κίνηση, όπως η νόσος του Πάρκινσον) [196]. Ευρύτερα, αυτό το ζήτημα της ιδιωτικής ζωής μπορεί να εξεταστεί από την άποψη της ανάπτυξης ασφαλών (και συνεπώς αξιόπιστων) ρομπότ, όπου η έννοια της ασφάλειας περιλαμβάνει όχι μόνο τη φυσική ασφάλεια [197] αλλά και την ασφάλεια από ανεπιθύμητες παραβιάσεις της ιδιωτικής ζωής.

Πρόσφατες αναλύσεις δείχνουν ότι υπάρχουν τρόποι επίλυσης αυτής της έντασης μεταξύ ιδιωτικότητας και εμπιστοσύνης. Ειδικότερα, τεχνικές από τη βιβλιογραφία της διαφορικής ιδιωτικότητας [198] μπορούν να χρησιμοποιηθούν για την καταπολέμηση της διαρροής ιδιωτικών πληροφοριών [196]. Όσον αφορά την παραβίαση της ιδιωτικής ζωής μέσω ρομπότ, ο τομέας της ρομποτικής ηθικής μπορεί να μας ενημερώσει σχετικά με τους κανονισμούς και τους κώδικες συμπεριφοράς που πρέπει να ισχύουν πριν από την ανάπτυξη κοινωνικών ρομπότ μεταξύ ευάλωτων ομάδων [199]. Θα πρέπει επίσης να αναγνωριστεί ότι η επίδραση της εμπιστοσύνης στην ιδιωτικότητα μπορεί στην πραγματικότητα να αξιοποιηθεί για το κοινωνικό καλό. Στο [200], η ανάπτυξη εμπιστοσύνης στους εικονικούς ανθρώπους ενθάρρυνε τους ασθενείς να αποκαλύψουν ιατρικές πληροφορίες που διαφορετικά θα απέκρυπταν παρουσία ενός πραγματικού γιατρού λόγω του φόβου της κοινωνικής κρίσης κατά τη διάρκεια ενός υγειονομικού ελέγχου. Από αυτή την άποψη, ένα ισχυρό ηθικό και νομικό πλαίσιο για τη ρύθμιση της χρήσης των κοινωνικών ρομπότ είναι σημαντικό για την αντιμετώπιση των ανησυχιών σχετικά με την ιδιωτικότητα και την ασφάλεια που συνδέονται με τη χρήση των κοινωνικών ρομπότ στην καθημερινή μας ζωή [199].

Πολύπλοκα μοντέλα εμπιστοσύνης για πραγματικά σενάρια

Η εμπιστοσύνη περιγράφεται συχνότερα ως μια πλούσια, πολυδιάστατη κατασκευή [135] και σε ένα ρεαλιστικό περιβάλλον, ποικίλα στοιχεία συνδυάζονται για να επηρεάσουν την εμπιστοσύνη. Αν και δεν υπάρχει αμφιβολία ότι τα υπάρχοντα μοντέλα έχουν αποδώσει εντυπωσιακά αποτελέσματα, υπάρχει ακόμη πολλή δουλειά που πρέπει να γίνει για την πλήρη αποτύπωση της εμπιστοσύνης στα ρομπότ. Ορισμένες πρόσφατες εργασίες έχουν αρχίσει να διερευνούν αυτόν τον τομέα. Για παράδειγμα, το [155] εξέτασε δύο διαφορετικές πτυχές της εμπιστοσύνης - εμπιστοσύνη στην πρόθεση ενός ρομπότ και εμπιστοσύνη στις ικανότητες ενός ρομπότ - και απέδειξε ότι αυτοί οι δύο παράγοντες αλληλεπιδρούν για να προκαλέσουν εμπιστοσύνη στο ρομπότ. Ομοίως, το [166] έδειξε ότι η εμπιστοσύνη μπορεί να μοντελοποιηθεί ως λανθάνουσα συνάρτηση (δηλαδή, ένα άπειρο-διάστατο μοντέλο εμπιστοσύνης) για να ενσωματώσει διαφορετικά πλαίσια και έδειξε ότι αυτή η προσέγγιση μπορεί να καταγράψει τον τρόπο με τον οποίο η εμπιστοσύνη μεταφέρεται σε διαφορετικά περιβάλλοντα εργασιών.

Τα δύο αυτά παραδείγματα αναδεικνύουν δύο διαφορετικές αλλά συμπληρωματικές προσεγγίσεις για τη μελέτη της πολυδιάστατης φύσης της εμπιστοσύνης. Η πρώτη είναι ότι, λαμβάνοντας σοβαρά υπόψη την πολυδιάστατη φύση, μπορούμε να αναπτύξουμε μια πλουσιότερη και ακριβέστερη κατανόηση όχι μόνο του τι κάνει τους ανθρώπους να συνεργάζονται (ή όχι) με τα ρομπότ, αλλά και πώς προκύπτει αυτή η συνεργασία. Με άλλα λόγια, θα μπορούσε να μας δώσει εικόνα για τους μηχανισμούς, και όχι μόνο για τα προηγούμενα, μέσω των οποίων η εμπιστοσύνη επηρεάζει τη συνεργασία ανθρώπου-ρομπότ. Σύμφωνα με αυτή την άποψη, μερικές πρωτοποριακές εργασίες έχουν αρχίσει να διερευνούν τον διαμεσολαβητικό ρόλο της (μονοδιάστατης) εμπιστοσύνης μέσω τυπικής ανάλυσης αιτιώδους διαμεσολάβησης [201]. Στόχος αυτής της προσέγγισης είναι να ελεγχθεί εμπειρικά αν ένας συγκεκριμένος μηχανισμός που βασίζεται στην εμπιστοσύνη υποστηρίζεται από τα δεδομένα [153] [202]. Ένας δυνητικά πολύτιμος αλλά ανεξερεύνητος τομέας είναι η κατανόηση των μηχανισμών εμπιστοσύνης σε πολυδιάστατες περιπτώσεις. Είναι σημαντικό να αξιολογηθεί η βιωσιμότητα των πραγματικών μηχανισμών για τη μοντελοποίηση της εμπιστοσύνης στο πλαίσιο PGM πέρα από την επιστημονική περιέργεια. Συχνά, σε τέτοια μοντέλα, η δομή

Εικόνα 2. *Αυτό το ρομπότ χρησιμοποιεί κάμερες και αισθητήρες πίεσης για να εντοπίζει και να πιάνει αντικείμενα χωρίς να προσκρούει σε ανθρώπους. Φωτογραφία προσφορά της Rethink Robotics Inc.*

του γραφήματος υποτίθεται ότι είναι αληθής εξ αρχής, αλλά δεν είναι εγγυημένο ότι η δεδομένη δομή του γραφήματος αντιστοιχεί στο πραγματικό αιτιώδες γράφημα των φαινομένων που μας ενδιαφέρουν [163]. Η απόκτηση λεπτομερών μηχανισμών που διέπουν την εμπιστοσύνη στα ρομπότ είναι κρίσιμη για πιο ισχυρά μοντέλα.

Μια δεύτερη πιθανή προσέγγιση είναι η αξιοποίηση των εξελίξεων στα βαθιά πιθανοτικά μοντέλα για τη μοντελοποίηση της λανθάνουσας εμπιστοσύνης στο πλαίσιο δεδομένων εισόδου υψηλής διάστασης, κάτι που επικρατεί σε πραγματικές συνθήκες. Αυτό είναι ιδιαίτερα σημαντικό υπό το πρίσμα του πρόσφατου ενδιαφέροντος για την ενσωμάτωση βίντεο, ήχου και ψυχοφυσιολογικών μέτρων στη μοντελοποίηση της εμπιστοσύνης [172]. Η πρόσφατη ενσωμάτωση των βαθιών νευρωνικών δικτύων με την πιθανολογική μοντελοποίηση [203] κατέστησε δυνατή τη διαχείριση των πολυδιάστατων μη δομημένων δεδομένων στο πλαίσιο των PGMs. Αυτές οι μέθοδοι χρησιμοποιούν νευρωνικά δίκτυα για να αντιστοιχίσουν τα υψηλής διάστασης ακατέργαστα δεδομένα σε έναν μειωμένο χώρο που χαρακτηρίζεται από ένα λανθάνον τυχαίο πραγματικό διάνυσμα. Πρόσφατες εργασίες έχουν ως στόχο την εκμάθηση ” αποσαφηνισμένων ” λανθάνουσων αναπαραστάσεων με επιμέρους διαστάσεις που παρέχουν ουσιαστικές πληροφορίες για τα μοντελοποιημένα δεδομένα. [204]. Αν και η ερμηνευσιμότητα είναι σημαντική, η προσέγγιση αυτή είναι πολύτιμη στη ρομποτική. Τα κοινωνικά ρομπότ βασίζονται σε ακατέργαστες αισθητήριες πληροφορίες για να κατανοήσουν το περιβάλλον τους. Η εξαγωγή συμπερασμάτων εμπιστοσύνης με βάση αυτά τα δεδομένα επιτρέπει στα ρομπότ να ενεργούν κατάλληλα σε κοινωνικά περιβάλλοντα.

Αυτές οι δύο προσεγγίσεις για τη μελέτη της πολυδιάστατης εμπιστοσύνης δεν είναι ασύμβατες. Η πρώτη παρέχει την απαραίτητη κατανόηση της δομής της εμπιστοσύνης. Η δεύτερη προσέγγιση επιτρέπει στα ρομπότ να χειρίζονται αποτελεσματικά τις πλούσιες πηγές αισθητηριακών δεδομένων για να αιτιολογούν την εμπιστοσύνη σε πραγματικό χρόνο. Αυτές οι δύο προσεγγίσεις μπορούν να συνδυαστούν [205] για να πάρουμε το καλύτερο και από τους δύο κόσμους, επιτρέποντάς έτσι ενδεχομένως να αναπτυχθούν ρομπότ που μπορούν να εκτελούν δομημένη συμπερασματολογία εμπιστοσύνης σε έναν χαώδη, πολυδιάστατο κόσμο.

2.4 Στρατηγικές επιδιόρθωσης της εμπιστοσύνης

Οι ομάδες ανθρώπου-ρομπότ έχουν μεγάλες προοπτικές να επεκτείνουν τις δυνατότητες της ανθρώπινης ευελιξίας, προσαρμοστικότητας και δημιουργικότητας, συνδυάζοντας τις με την ικανότητα ενός ρομπότ για ακρίβεια, ταχύτητα και συνέπεια [206]. Ως αποτέλεσμα, οι ομάδες ανθρώπου-ρομπότ αρχίζουν να εμφανίζονται σε διάφορα εργασιακά περιβάλλοντα [207]. Ωστόσο, για την επιτυχία αυτών των ομάδων ανθρώπου-ρομπότ απαιτείται εμπιστοσύνη [208], [209], [207], [210]. Η εμπιστοσύνη, ή η αποδοχή της ευπάθειας που συνδέεται με την εμπιστοσύνη σε άλλους, είναι ζωτικής σημασίας για τις ομάδες ανθρώπου-ρομπότ, αλλά συχνά μειώνεται από τις παραβιάσεις της εμπιστοσύνης [211]. Η εμπιστοσύνη είναι δυναμική και μεταβάλλεται με την πάροδο του χρόνου, με σημαντικές μειώσεις στην εμπιστοσύνη να συμβαίνουν όταν τα ρομπότ μοιραία σφάλλουν (δηλαδή παραβιάζουν την εμπιστοσύνη) [124],[135]. Οι ερευνητές έχουν μελετήσει διάφορες προσεγγίσεις για τον μετριασμό των αρνητικών επιπτώσεων των παραβιάσεων της εμπιστοσύνης μέσω διαφόρων στρατηγικών αποκατάστασης της εμπιστοσύνης, όπως συγγνώμες, αρνήσεις, εξηγήσεις ή υποσχέσεις [212] ωστόσο δεν είναι σαφές εάν ή πότε μια συγκεκριμένη στρατηγική αποκατάστασης της εμπιστοσύνης είναι αποτελεσματική.

- Οι **παραβιάσεις της εμπιστοσύνης** (trust violations) είναι γεγονότα που μειώνουν την αντίληψη ενός ατόμου (trustor) για την αξιοπιστία και την εμπιστοσύνη του σε έναν εμπιστευόμενο (trustee) [213], [214]. Αυτή η μείωση της εμπιστοσύνης έχει αποδειχθεί ότι επηρεάζει αρνητικά τις δυνατότητες συνεργασίας μιας ομάδας, ανεξάρτητα από το αν η ομάδα αυτή αποτελείται από ανθρώπους [215] ή περιλαμβάνει ρομπότ [124], [135]. στην HRI, έχουν οριστεί τρεις κατηγορίες παραβιάσεων εμπιστοσύνης: παραβιάσεις της ικανότητας, της ακεραιότητας και της καλοπιστίας [216].

  Οι παραβιάσεις της εμπιστοσύνης με βάση τις ικανότητες συμβαίνουν όταν ένα ρομπότ παραβιάζει τις προσδοκίες του ανθρώπου σχετικά με τις επιδόσεις του ρομπότ. Παραδείγματα αυτού του τύπου παραβίασης στη βιβλιογραφία της HRI είναι περιπτώσεις όπου ένα ρομπότ κάνει ένα λάθος χωρίς να το επιδιώκει [149]. Οι παραβιάσεις εμπιστοσύνης με βάση την ακεραιότητα συμβαίνουν όταν ένα ρομπότ παραβιάζει τις προσδοκίες του ανθρώπου σχετικά με την ειλικρίνεια και την ηθική συνέπεια τού ίδιου. Παραδείγματα παραβιάσεων που βασίζονται στην ακεραιότητα στη βιβλιογραφία του HRI έχουν τη μορφή ενός ρομπότ που ενεργεί εναντίον ενός συμπαίκτη του κατά τη διάρκεια συνεργατικών εργασιών, ακόμη και όταν υποσχέθηκε να μην το κάνει [149]. Οι παραβιάσεις εμπιστοσύνης με βάση την καλοπιστία συμβαίνουν όταν ένα ρομπότ αποτυγχάνει να ανταποκριθεί στις προσδοκίες του ανθρώπου σχετικά με τον σκοπό του ρομπότ (δηλαδή για ποιον έχει σχεδιαστεί το ρομπότ). Οι παραβιάσεις που βασίζονται στην καλοσύνη διαφέρουν από τις παραβιάσεις που βασίζονται στην ακεραιότητα στο ότι οι παραβιάσεις που βασίζονται στην καλοσύνη υποδηλώνουν ένα βαθμό κακοβουλίας ή κακής θέλησης, ενώ οι παραβιάσεις που βασίζονται στην ακεραιότητα δεν το κάνουν. Η βιβλιογραφία δεν έχει εξετάσει ρητά τις παραβιάσεις της καλοσύνης. Πιθανές μορφές αυτού του τύπου παραβίασης, ωστόσο, θα μπορούσαν να περιλαμβάνουν ένα ρομπότ που υπονομεύει έναν ανθρώπινο συμπαίκτη στην υπηρεσία ενός άλλου, ένα ρομπότ που απορρίπτει ή αγνοεί την ανατροφοδότηση ή ένα ρομπότ που ενεργεί σε άμεση σύγκρουση με τις επιθυμίες ενός ανθρώπου. Τελικά, όλοι αυτοί οι τύποι παραβιάσεων οδηγούν σε μείωση της εμπιστοσύνης.

- Η **αποκατάσταση της εμπιστοσύνης** (trust repairs) μπορεί να οριστεί ως η προσπάθεια που καταβάλλεται για την αποκατάσταση της εμπιστοσύνης μετά από μια πραγματική ή φαινομενική παραβίαση της εμπιστοσύνης [217], [218]. Οι στρατηγικές επιδιόρθωσης της εμπιστοσύνης λαμβάνουν συνήθως μία από τις ακόλουθες τέσσερις μορφές: απολογίες, αρνήσεις, εξηγήσεις ή υποσχέσεις [212],[219]. Συνολικά κάθε μία από αυτές τις στρατηγικές έχει θεωρηθεί γενικά αποτελεσματική, αλλά οι διάφορες θεωρητικές αιτιολογήσεις για το

πώς αυτές οι στρατηγικές αποκαθιστούν την εμπιστοσύνη είναι λιγότερο συζητημένες και υπόκεινται σε συνεχή συζήτηση [213].

Οι *απολογίες* είναι προσπάθειες επαναπροσδιορισμού του εμπιστευόμενου (trustee) μετά την παραβίαση της εμπιστοσύνης [220]. Συγκεκριμένα, οι συγγνώμες είναι ένας τύπος λεκτικής στρατηγικής αποκατάστασης της εμπιστοσύνης που επιδιώκει να εκφράσει τη μεταμέλεια για μια σχεσιακή ή κοινωνική παράβαση σε συνδυασμό με μια ρητή ή σιωπηρή παραδοχή ενοχής [212],[220]. Για παράδειγμα, η φράση ”λυπάμαι που το έκανα αυτό” είναι μια συγγνώμη. Από την άλλη πλευρά, η *άρνηση* είναι η απόρριψη της ενοχής σε συνδυασμό με έναν ή περισσότερους εξωτερικούς λόγους για τους οποίους διαπράχθηκε η παραβίαση της εμπιστοσύνης [124].Ο στόχος της άρνησης είναι να μετατοπιστεί η ευθύνη από τον εμπιστευόμενο σε κάποιον άλλο [221], [220]. Για παράδειγμα, η φράση ”Δεν τα θαλάσσωσα εγώ... κάτι άλλο πρέπει να συνέβη” είναι μια άρνηση. Οι *εξηγήσεις* είναι ρητές προφορικές δηλώσεις που γίνονται με σκοπό να δώσουν τους λόγους για τους οποίους έγινε μια ενέργεια [222]. Ένα παράδειγμα εξήγησης θα μπορούσε να είναι: ”Βλέπω, οι αισθητήρες μου δεν ήταν βαθμονομημένοι και έτσι έχασα το αντικείμενο”. Οι εξηγήσεις παρέχουν διαφάνεια, η οποία βοηθά τους άλλους να κατανοήσουν την εσωτερική λειτουργία ή τη λογική πίσω από το γιατί συνέβη ένα γεγονός [221], [212]. Οι *υποσχέσεις* είναι ισχυρισμοί ενός διαχειριστή που αποσκοπούν στη μετάδοση θετικών προθέσεων για μελλοντικές πράξεις [219]. Ένα παράδειγμα υπόσχεσης είναι η δήλωση: ”Υπόσχομαι ότι θα το κάνω σωστά την επόμενη φορά”.

Η πλειοψηφία των ερευνητικών εκθέσεων που σχετίζονται με την επιδιόρθωση της εμπιστοσύνης στην HRI επικεντρώθηκε στην εμπιστοσύνη. Λόγω αυτού, εξετάστηκε η υπάρχουσα βιβλιογραφία για την αποκατάσταση της εμπιστοσύνης στην HRI με στόχο να προσδιοριστεί ο τρόπος με τον οποίο οι επανορθώσεις εμπιστοσύνης επηρέασαν την συνολική εμπιστοσύνη στην αλληλεπίδραση ανθρώπου-ρομπότ με έμφαση στις στρατηγικές επιδιόρθωσης της απολογίας, της άρνησης, των εξηγήσεων και των υποσχέσεων. Δεδομένης αυτής της εστίασης, οι ακόλουθες υποενότητες συνοψίζουν τον τρόπο με τον οποίο οι στρατηγικές αυτές έχουν βρεθεί να επηρεάζουν την εμπιστοσύνη.

1. Ο αντίκτυπος της *απολογίας* στην εμπιστοσύνη στην HRI υπήρξε μάλλον αμφιλεγόμενος, καθώς τρεις μελέτες διαπίστωσαν ότι η απολογία αποκατέστησε την εμπιστοσύνη [223], [224], [225], δύο διαπίστωσαν ότι η απολογία δεν αποκατέστησε την εμπιστοσύνη [226], [227] και μία διαπίστωσε ότι η απολογία έβλαψε την εμπιστοσύνη [228]. Μια πιθανή εξήγηση για αυτά τα μεικτά αποτελέσματα θα μπορούσε να σχετίζεται με διαφορετικούς παράγοντες μετριασμού ή μεσολαβητές. Ειδικότερα, ο χρόνος έχει εξεταστεί και έχει βρεθεί ότι έχει επιρροή. Σε τέτοιες περιπτώσεις, οι συγγνώμες που δόθηκαν πιο κοντά στο γεγονός κατά το οποίο παραβιάστηκε η εμπιστοσύνη βρέθηκαν να είναι πιο αποτελεσματικές από τις συγγνώμες που παρακρατήθηκαν και δόθηκαν μετά την πάροδο του χρόνου [229], [230]. Αξίζει να σημειωθεί ότι μέχρι σήμερα δεν έχει διεξαχθεί καμία εξέταση των συγγνώμων που δόθηκαν μετά από πολλαπλές παραβιάσεις και επιδιορθώσεις της εμπιστοσύνης.

2. Ο αντίκτυπος των *αρνήσεων* στην αποκατάσταση της εμπιστοσύνης φαίνεται να είναι μικτός, με μία μελέτη να δείχνει ότι οι αρνήσεις ήταν αποτελεσματικές [225] και μία να δείχνει ότι οι αρνήσεις ήταν αναποτελεσματικές [227]. Για άλλη μια φορά, οι παράγοντες μετριασμού θα μπορούσαν να εξηγήσουν αυτά τα μικτά αποτελέσματα. Μέχρι σήμερα, ωστόσο, έχει διερευνηθεί μόνο ένας παράγοντας, δηλαδή, ο τύπος της παραβίασης της εμπιστοσύνης που συνέβη. Για το σκοπό αυτό, οι [231] και [149] εξέτασαν αυτόν τον πιθανό ρυθμιστή και δεν διαπίστωσαν σημαντικές διαφορές στην εμπιστοσύνη μεταξύ των αρνήσεων που δόθηκαν μετά από παραβιάσεις εμπιστοσύνης με βάση την ικανότητα έναντι της ακεραιότητας

στην περίπτωση του [149] και των αρνήσεων που δόθηκαν μετά από παραβιάσεις της λογικής, της σημασιολογίας ή του συντακτικού στην περίπτωση του [231]. Αυτό έρχεται σε σύγκρουση με τα ευρήματα από τη βιβλιογραφία για την εμπιστοσύνη μεταξύ ανθρώπων, σύμφωνα με τα οποία ο τύπος της παραβίασης είναι καθοριστικός για το πότε μια άρνηση είναι πιθανό να είναι αποτελεσματική [214], [212]. Ενώ αυτό μπορεί να υποδεικνύει μια θεμελιώδη διαφορά μεταξύ ανθρώπων και ρομπότ όσον αφορά τις αρνήσεις και την αποκατάσταση της εμπιστοσύνης, τα αποτελέσματα αυτά θα μπορούσαν επίσης να επηρεαστούν από τη σχετικά σαφή αιτία των παραβιάσεων εμπιστοσύνης που εξετάστηκαν. Αυτό θα μπορούσε να συμβαίνει δεδομένου ότι η αιτία των παραβιάσεων εμπιστοσύνης που χρησιμοποιούνται στις περισσότερες μελέτες HRI παρουσιάζεται με σαφήνεια στα άτομα μειώνοντας την εγκυρότητα των αρνήσεων. Αυτό οφείλεται στο γεγονός ότι οι αρνήσεις βασίζονται στην αμφιβολία του εξεταζόμενου να δώσει μια εναλλακτική ερμηνεία των γεγονότων [214]. Δεδομένου ότι καμία μελέτη στη βιβλιογραφία HRI μέχρι σήμερα δεν έχει λάβει ρητά υπόψη την αμφιβολία του χρήστη για τα γεγονότα ή δεν έχει χρησιμοποιήσει σκόπιμα θολές αιτίες για τις παραβιάσεις της εμπιστοσύνης, πολλά παραμένουν άγνωστα όσον αφορά την πραγματική καταλληλότητα των αρνήσεων ως στρατηγική αποκατάστασης της εμπιστοσύνης.

3. Οι επιπτώσεις των *εξηγήσεων* στην εμπιστοσύνη ήταν επίσης ανάμεικτες, με μία μελέτη να δείχνει ότι οι εξηγήσεις αποκαθιστούν την εμπιστοσύνη [224] και πέντε μελέτες να δείχνουν ότι οι εξηγήσεις δεν την αποκαθιστούν [230], [227], [226], [228]. Επιπλέον, δύο παράγοντες φαίνεται να επηρεάζουν την αποτελεσματικότητα των εξηγήσεων, συγκεκριμένα, η σοβαρότητα της παραβίασης της εμπιστοσύνης [232] και ο χρόνος της εξήγησης [229], [230]. Όσον αφορά τη σοβαρότητα, τα αποτελέσματα ήταν αρκετά απλά, καθώς όσο πιο σοβαρή ήταν η παραβίαση, τόσο λιγότερο αποτελεσματικές ήταν οι εξηγήσεις [232]. Ωστόσο, όσον αφορά τον χρόνο των εξηγήσεων, οι μελέτες έδωσαν αντικρουόμενα αποτελέσματα- συγκεκριμένα, μια μελέτη διαπίστωσε ότι ο χρόνος ήταν σημαντική μεταβλητή [229] και μια άλλη διαπίστωσε ότι δεν ήταν [230]. Μια προσεκτικότερη εξέταση αυτών των μελετών δείχνει διαφορές στα μεγέθη των δειγμάτων που μπορεί να εξηγούν αυτή τη διαφωνία. Συγκεκριμένα, η μελέτη που βρήκε σημαντικά αποτελέσματα [229] για το χρονισμό χρησιμοποίησε αρκετά περισσότερα άτομα (n=39) από τη μελέτη που βρήκε μη σημαντικά αποτελέσματα (n=18) [230]. Ανεξάρτητα από αυτό, η διαφωνία αυτή δικαιολογεί περαιτέρω εξέταση και απαιτούνται περισσότερες μελέτες σχετικά με τις επιδράσεις του χρονισμού.

4. Οι *υποσχέσεις* στη βιβλιογραφία για την αποκατάσταση της εμπιστοσύνης στην HRI έχουν ως επί το πλείστον εξεταστεί σε συνδυασμό με απολογίες ή εξηγήσεις [149], [223], [233], [234]. Όπου οι υποσχέσεις έχουν εξεταστεί ανεξάρτητα, τα αποτελέσματα ήταν ανομοιογενή, με μια μελέτη να δείχνει ότι οι υποσχέσεις αποκαθιστούν αποτελεσματικά την εμπιστοσύνη [235] και μια άλλη ότι δεν το κάνουν [233]. Ένας λόγος για αυτά τα ευρήματα μπορεί και πάλι να προέρχεται από διάφορους παράγοντες. Για παράδειγμα, έχει υποστηριχθεί ότι οι υποσχέσεις επηρεάζονται από τον χρόνο, όπου οι υποσχέσεις που δόθηκαν αμέσως μετά από μια παραβίαση βρέθηκαν να είναι πιο αποτελεσματικές από εκείνες που δόθηκαν με καθυστέρηση [229]. Τελικά, οι υποσχέσεις έχουν δυνατότητες ως στρατηγική επιδιόρθωσης, αλλά υπάρχει έλλειψη ισχυρής εξέτασης αυτής της στρατηγικής επιδιόρθωσης. Ειδικότερα, χρειάζονται περισσότερες μελέτες που να εστιάζουν σε υποσχέσεις ανεξάρτητες από απολογίες. Τέτοιες μελέτες θα πρέπει να λαμβάνουν υπόψη όχι μόνο το χρονοδιάγραμμα αλλά και τον τύπο της παραβίασης.

Η τρέχουσα βιβλιογραφία σχετικά με την αποκατάσταση της εμπιστοσύνης στην HRI δίνει μια σύνθετη και διασπασμένη εικόνα για το αν και πώς τα ρομπότ μπορούν να αποκαταστήσουν την εμπιστοσύνη, δεδομένου ότι τα ευρήματα σχετικά με τις επιπτώσεις των απολογιών, των αρ-

νήσεων, των εξηγήσεων και των υποσχέσεων είναι σε μεγάλο βαθμό ανάμεικτα. Επιπλέον, ενώ οι παράγοντες που θα μπορούσαν να συμβάλουν στην εξήγηση αυτών των μεικτών αποτελεσμάτων, η βιβλιογραφία που εξετάζει αυτούς τους παράγοντες είναι συχνά μπερδεμένη, ελλιπής ή απούσα. Ως αποτέλεσμα, είναι δύσκολο να γνωρίζουμε ποιες στρατηγικές επιδιόρθωσης είναι αποτελεσματικές και πότε. Επιπλέον, οι σχετικές έρευνες δεν έχουν εξετάσει διεξοδικά τον τρόπο με τον οποίο λειτουργούν οι επιδιορθώσεις εμπιστοσύνης (δηλαδή τους μηχανισμούς μέσω των οποίων δρουν). Ως εκ τούτου, δικαιολογούνται μελλοντικές μελέτες. Ειδικότερα, απαιτούνται μελέτες που να εξετάζουν ανεξάρτητα τις στρατηγικές επιδιόρθωσης- να συγκρίνουν τις στρατηγικές αυτές μεταξύ τους- να εξετάζουν θεωρητικούς μηχανισμούς (δηλαδή μεσολαβητές)- και να λαμβάνουν επίσης υπόψη το χρόνο, τον τύπο της παραβίασης και ενδεχομένως άλλους επιδραστικούς παράγοντες.

2.5 Αλληλεπίδραση ανθρώπου - κοινωνικού ρομπότ

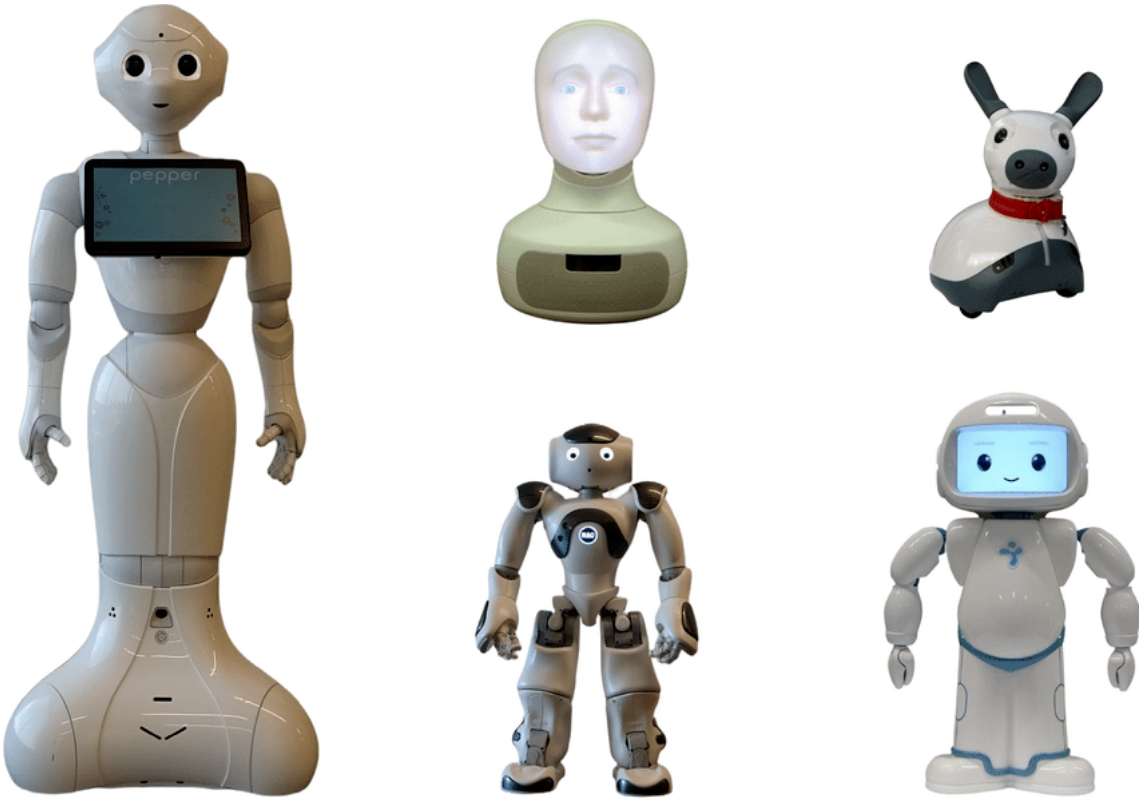
2.5.1 Προδιαθέσεις απέναντι στα κοινωνικά ρομπότ

Σύμφωνα με μια ευρέως αναφερόμενη έρευνα μεγάλης κλίμακας [236], ένα σημαντικό ποσοστό των πολιτών της ΕΕ έχει αρνητική στάση απέναντι στη χρήση ρομπότ στην υγειονομική περίθαλψη και σε άλλους τομείς που παραδοσιακά κυριαρχούνται από ανθρώπους. Έχουν επίσης υπάρξει ενδείξεις για αυξανόμενη ανησυχία μεταξύ του κοινού ότι η αυτοματοποίηση, η οποία διευκολύνεται από τη ρομποτική, θα οδηγήσει σε σημαντική απώλεια θέσεων εργασίας [97]. Ενώ οι στάσεις απέναντι στα ρομπότ εμφανίζονται ανομοιογενείς, πιθανόν να εξαρτώνται από το περιβάλλον και την ερώτηση που τίθεται και σε ορισμένες περιπτώσεις είναι κάπως αποκομμένες από την πραγματικότητα (π.χ. υπάρχουν ενδείξεις ότι οι στάσεις βασίζονται στην επιστημονική φαντασία και όχι στην αντικειμενική πραγματικότητα, [237]). Ενώ οι στάσεις δεν προβλέπουν σταθερά τη συμπεριφορά, θεωρείται ότι επηρεάζουν τις προθέσεις των ανθρώπων [238]. Ως εκ τούτου, μπορούν να προβλέψουν την αποδοχή και τη χρήση των ρομπότ μαζί με άλλες μεταβλητές, όπως το άγχος, η εμπιστοσύνη και η πρόθεση χρήσης και συνεργασίας με τα ρομπότ. Η καλύτερη κατανόηση των στάσεων των ανθρώπων απέναντι στα ρομπότ θα πρέπει επομένως να βοηθήσει στην ενημέρωση της μελλοντικής έρευνας, ανάπτυξης και ανάπτυξης της ρομποτικής σε διάφορους τομείς της δημόσιας και ιδιωτικής ζωής.

Η παρούσα ενότητα επικεντρώνεται στα κοινωνικά ρομπότ, λόγω της αυξανόμενης χρήσης τους σε διάφορα περιβάλλοντα, όπως η υγειονομική περίθαλψη, η ψυχαγωγία και η εξυπηρέτηση πελατών [5]. Ενώ η ιδέα των ρομπότ που μπορούν να αλληλεπιδρούν κοινωνικά με τους ανθρώπους υπάρχει εδώ και αρκετό καιρό, η χρήση τους ήταν σχετικά περιορισμένη και λιγότερο διαδεδομένη σε σύγκριση, για παράδειγμα, με τα ρομπότ κατασκευής [239]. Παρ' όλα αυτά, τα κοινωνικά ρομπότ συγκεντρώνουν την προσοχή τόσο των μέσων ενημέρωσης όσο και του κοινού και έχουν προκαλέσει συζητήσεις σχετικά με τον πιθανό αντίκτυπό τους στην κοινωνία [240]. Ως κοινωνικό ρομπότ ορίζεται ένας φυσικά ενσαρκωμένος τεχνικός πράκτορας (δηλαδή κάτι που έχει φυσική δομή που μιμείται τη συμπεριφορά, την εμφάνιση ή την κίνηση ενός ζωντανού όντος - συνήθως ανθρώπου, αλλά θα μπορούσε να είναι και ζώο ή φυτό) που:

- έχει χαρακτηριστικά που επιτρέπουν στους ανθρώπους να αντιληφθούν τον πράκτορα ως κοινωνική οντότητα (π.χ. μάτια),
- είναι ικανό να αλληλεπιδρά με τους ανθρώπους μέσω μιας κοινωνικής διεπαφής [241],
- και μπορεί να επικοινωνεί λεκτικές ή/και μη λεκτικές πληροφορίες στους ανθρώπους.

Εν ολίγοις, ένα κοινωνικό ρομπότ είναι ένα ενσαρκωμένο σύστημα που μπορεί να γίνει αντιληπτό ως κοινωνική οντότητα και είναι ικανό να επικοινωνεί με τον χρήστη [242].



Εικόνα 3. *Εικόνες των κοινωνικών ρομπότ (από αριστερά προς τα δεξιά, από πάνω προς τα κάτω): Pepper, Furhat, Miro, NAO και QT Robot.*

Στάσεις απέναντι στα κοινωνικά ρομπότ

Τα τρέχοντα δεδομένα σχετικά με τη στάση των ανθρώπων απέναντι στα κοινωνικά ρομπότ αποκαλύπτουν μια κάπως διφορούμενη εικόνα που καθιστά δύσκολο να πούμε αν οι άνθρωποι, γενικά, έχουν αρνητική ή θετική άποψη για τα κοινωνικά ρομπότ. Αυτό είναι, τουλάχιστον σε κάποιο βαθμό, πιθανό να οφείλεται στην ποικιλία των πλαισίων στα οποία χρησιμοποιούνται τα κοινωνικά ρομπότ. Οι άνθρωποι γενικά συμφωνούν ότι, ενώ η εργασία δίπλα στα ρομπότ δεν είναι εκτός συζήτησης, τα ρομπότ δεν θα πρέπει να αντικαταστήσουν εξ ολοκλήρου τους ανθρώπους σε θέσεις εργασίας που απαιτούν σημαντικές κοινωνικές δεξιότητες (π.χ. νοσηλευτική, [243]). Ταυτόχρονα, ορισμένες μελέτες έχουν διαπιστώσει θετική στάση απέναντι στα ρομπότ που εκτελούν εργασίες που απαιτούν περισσότερες κοινωνικές δεξιότητες [236] [243].

Άγχος για τα κοινωνικά ρομπότ

Ορισμένες μελέτες παρέχουν στοιχεία ότι το άγχος, μαζί με τις στάσεις, προβλέπει την πρόθεση χρήσης κοινωνικών ρομπότ και την ποιότητα της αλληλεπίδρασης των ανθρώπων με τα κοινωνικά ρομπότ [244] [41] [245]. Το άγχος απέναντι στα ρομπότ μετράται συχνά με τη χρήση μέτρων αυτοαναφοράς, όπως η Κλίμακα Άγχους Ρομπότ (”Robot Anxiety Scale”/ RAS; [245]) ή με άμεση παρατήρηση της συμπεριφοράς κατά τη διάρκεια της αλληλεπίδρασης ανθρώπου-ρομπότ (HRI). Παρά τη δυνητική σημασία του άγχους στη διαμόρφωση του τρόπου με τον οποίο οι άνθρωποι αλληλεπιδρούν με τα ρομπότ, τα τρέχοντα δεδομένα παρουσιάζουν μια μικτή εικόνα ως προς το πόσο αγχωμένοι είναι οι άνθρωποι για τα κοινωνικά ρομπότ. Για παράδειγμα, οι Nomura, Shintani, Fujii και Hokabe [246] διαπίστωσαν ότι τόσο το άγχος όσο και οι στάσεις μπορούν να επηρεάσουν τον τρόπο με τον οποίο οι άνθρωποι συμπεριφέρονται κατά τη διάρκεια της HRI με παρόμοιους τρό-

πους, ενώ οι de Graaf και Allouch [247] διαπίστωσαν ότι οι συμμετέχοντες που αλληλεπιδρούσαν με ένα ρομπότ παρουσίασαν αλλαγή στο άγχος τους αλλά όχι στις στάσεις τους.

Εμπιστοσύνη στα κοινωνικά ρομπότ

Η εμπιστοσύνη έχει επίσης αναγνωριστεί ως ένας παράγοντας που, τουλάχιστον εν μέρει, προβλέπει όχι μόνο την ποιότητα των HRI αλλά και το πόσο πρόθυμοι είναι οι άνθρωποι να χρησιμοποιήσουν κοινωνικά ρομπότ για ορισμένες εργασίες [148]. Η εμπιστοσύνη είναι πιθανό να είναι ιδιαίτερα σημαντική σε σχέση με τα κοινωνικά ρομπότ, ιδίως στην υγειονομική περίθαλψη, όπου η εμπιστοσύνη έχει συνδεθεί με την ικανοποίηση των ασθενών και τη θεραπευτική αποτελεσματικότητα [248]. Μέχρι στιγμής, οι ανασκοπήσεις έχουν επικεντρωθεί στον αντίκτυπο της εμπιστοσύνης στα ρομπότ στην αλληλεπίδραση ανθρώπου-ρομπότ, δείχνοντας ότι η κύρια παράγοντες που επηρεάζουν την εμπιστοσύνη σχετίζονται με πτυχές του ρομπότ (π.χ. το σχεδιασμό και τις επιδόσεις του ρομπότ), ενώ οι περιβαλλοντικοί παράγοντες διαδραματίζουν έναν πιο μέτριο ρόλο στο κατά πόσο οι άνθρωποι εμπιστεύονται τα ρομπότ. Ωστόσο, ο αντίκτυπος της εμπιστοσύνης σε σχέση με τα κοινωνικά ρομπότ συγκεκριμένα, δεν έχει επανεξεταστεί [249].

Αποδοχή των κοινωνικών ρομπότ

Η αποδοχή γενικά ορίζεται ως η πρόθεση χρήσης και, σε ορισμένες περιπτώσεις, ως η πραγματική χρήση των ρομπότ [250]. Σε σύγκριση με το άγχος και την εμπιστοσύνη, υπάρχουν σημαντικά περισσότερα στοιχεία για το βαθμό αποδοχής των κοινωνικών ρομπότ από τους ανθρώπους, ιδίως στους τομείς της υγειονομικής περίθαλψης και της φροντίδας ηλικιωμένων. Η αποδοχή των ρομπότ στην υγειονομική περίθαλψη έχει βρεθεί ότι είναι μικτή και μπορεί να ποικίλλει σημαντικά ανάλογα με τη λειτουργία και την εμφάνιση του ρομπότ [251]. Παρά τις δυνατότητες που έχουν τα κοινωνικά ρομπότ να ανακουφίσουν τις συνεχώς αυξανόμενες απαιτήσεις των επαγγελματιών υγείας [251], τα χαμηλά επίπεδα αποδοχής μπορεί να αποδειχθούν επιζήμια για την ανάπτυξη και τη χρήση της εν λόγω τεχνολογίας [251].

2.5.2 Παράγοντες επιρροής της ανθρώπινης αντίληψης

Διάφοροι παράγοντες είναι πιθανό να σχετίζονται με τη στάση των ανθρώπων απέναντι στα κοινωνικά ρομπότ, την εμπιστοσύνη σε αυτά, την αποδοχή τους και το άγχος τους. Για παράδειγμα, οι πεποιθήσεις των ανθρώπων μπορεί να διαφέρουν ανάλογα με το αν έχουν εκτεθεί πρόσφατα σε κοινωνικά ρομπότ (π.χ. οι μελέτες που παρουσιάζουν άμεση HRI μπορεί να αναφέρουν διαφορετικές στάσεις σε σχέση με τις μελέτες όπου οι συμμετέχοντες δεν αλληλεπιδρούν με ένα ρομπότ), τον προβλεπόμενο τομέα εφαρμογής (π.χ. συντροφικότητα και οικιακή βοήθεια, εκπαίδευση ή υγειονομική περίθαλψη) και το σχεδιασμό του ρομπότ (π.χ. ανθρωποειδές ή ανθρωπόμορφο).

Είδος έκθεσης στα ρομπότ

Ο τρόπος με τον οποίο οι άνθρωποι σκέφτονται για τα ρομπότ μπορεί να επηρεαστεί από το αν τους δίνεται η ευκαιρία να αλληλεπιδράσουν με ένα ρομπότ, άμεσα ή έμμεσα, πριν από τη διαπίστωση της στάσης τους. Οι μελέτες συνήθως προσδίδουν στους συμμετέχοντες τουλάχιστον έναν από τους τρεις τύπους έκθεσης σε ρομπότ (δηλ. HRI):

- **Χωρίς HRI:** Οι συμμετέχοντες δεν κλήθηκαν να αλληλεπιδράσουν, να δουν ή να να φανταστούν ένα κοινωνικό ρομπότ ή ρομπότ (π.χ., οι συμμετέχοντες ρωτήθηκαν μόνο για τη γενική στάση τους απέναντι στα κοινωνικά ρομπότ [252]),
- **Έμμεση HRI:** Οι συμμετέχοντες παρατήρησαν μια άμεση αλληλεπίδραση ή είδαν (ή τους ζητήθηκε να φανταστούν) μια αναπαράσταση ενός κοινωνικού ρομπότ ή ρομπότ (π.χ. οι συμμετέχοντες διάβασαν μια εικονογραφημένη περιγραφή ενός ρομπότ NAO, [253]),

- **Άμεση HRI:** Οι συμμετέχοντες αλληλεπιδρούσαν με ένα κοινωνικό ρομπότ που ήταν φυσικά παρόν στον ίδιο τόπο και χρόνο (π.χ. οι συμμετέχοντες συμμετείχαν σε μια εικονική συνέντευξη με ένα ρομπότ Geminoid HI-2, [254]).

Τομέας εφαρμογής

Τα στοιχεία δείχνουν ότι η στάση των ανθρώπων απέναντι στα ρομπότ μπορεί, σε κάποιο βαθμό, να εξαρτώνται από τον τομέα στον οποίο το ρομπότ χρησιμοποιείται (ή προορίζεται να χρησιμοποιηθεί) [255] [256]. Για τους σκοπούς του παρόντος ανασκόπησης, προσδιορίστηκαν έξι ευρείς τομείς εφαρμογής:

- **Ρομποτικής συντροφιάς και οικιακής βοήθειας:** Ρομπότ σχεδιασμένα ειδικά και αποκλειστικά για να αλληλεπιδρούν κοινωνικά με τον άνθρωπο για παρατεταμένο χρονικό διάστημα και να παρέχουν συντροφιά (π.χ., μια μελέτη διερευνά τις στάσεις απέναντι στα ρομπότ NAO και Darwin; [257])- ή ρομπότ που έχουν σχεδιαστεί για να βοηθούν στις οικιακές εργασίες, καθώς και να παρέχουν κοινωνική αλληλεπίδραση (π.χ., μια μελέτη που διερευνά την αξιολόγηση ενός κοινωνικά βοηθητικού ρομπότ σε ένα έξυπνο σπίτι [258]),

- **Εκπαιδευτικά ρομπότ:** Σχεδιασμένα για να βοηθούν τους εκπαιδευτικούς στη διδασκαλία και την κοινωνική αλληλεπίδραση με τους μαθητές (π.χ., μια μελέτη που διερευνά τον τρόπο με τον οποίο οι μαθητές αξιολογούν τη χρήση του NAO για την διδασκαλία των Αγγλικών [259]).

- **Υγειονομική περίθαλψη:** Ρομπότ που σχεδιάζονται για την βοήθεια με ασθενείς, γιατρούς και υγειονομικούς φορείς γενικότερα (π.χ. μια μελέτη που διερευνά τις στάσεις και τις προτιμήσεις του προσωπικού, των κατοίκων και των συγγενών των κατοίκων ενός συνταξιοδοτικού κέντρου απέναντι σε ένα ρομπότ υγειονομικής περίθαλψης [260]).

- **Παιδιατρική φροντίδα:** Ρομπότ που χρησιμοποιούνται στην υγειονομική περίθαλψη, αλλά ειδικά σχεδιασμένα για να βοηθούν τα παιδιά και τους παρόχους υγειονομικής περίθαλψης που τα θεραπεύουν (π.χ. αξιολόγηση της αποδοχής των φυσικοθεραπευτών των βοηθητικών ρομπότ ως θεραπευτικό βοήθημα για παιδιά στην αποκατάσταση, [261]).

- **HRI:** Ρομπότ που έχουν σχεδιαστεί πρωτίστως για να αλληλεπιδρούν με τους ανθρώπους, με οποιαδήποτε πρόσθετη λειτουργία (π.χ. παροχή φροντίδας) να είναι δευτερεύουσα. Για παράδειγμα, παίζοντας παιχνίδια ή συνομιλώντας (π.χ. μια μελέτη που εξετάζει την επίδραση του μεγέθους της ομάδας στις στάσεις και τις συμπεριφορές των ανθρώπων απέναντι σε ρομπότ ως συνεργάτες αλληλεπίδρασης- [262]).

- **Γενική εφαρμογή:** Η μελέτη δεν προσδιορίζει ούτε υπονοεί ένα πεδίο εφαρμογής για το ρομπότ ή τα ρομπότ που διερευνώνται. (π.χ., μια μελέτη που διερευνά την αποτελεσματικότητα των εκθέσεων ρομπότ ως μέσο διαμόρφωσης των πεποιθήσεων των ανθρώπων για αυτά, [263]).

Σχεδίαση του ρομπότ

Τα σχεδιαστικά χαρακτηριστικά των ρομπότ, όπως ο βαθμός ομοιότητας με τον άνθρωπο, είναι πιθανό να επηρεάζουν τη στάση των ανθρώπων απέναντι στα ρομπότ [249] [264]:

- **Ανθρωποειδές:** Ένα ρομπότ που μοιάζει με άνθρωπο (π.χ., το NAO; [265]).

- **Ανθρωπόμορφο:** Ένα ρομπότ που μιμείται ορισμένα μέρη του ανθρώπινου σώματος και μπορεί να υποστεί ανθρωπομορφοποίηση από τον χρήστη (π.χ. ένα ρομπότ με πρόσωπο που μοιάζει με ανθρώπινο, [266]).

Εικόνα 4. *Moxie, ένα κοινωνικό ρομπότ που βοηθά τα παιδιά με την κοινωνικο- συναισθηματική εκμάθηση*

- **Μη ανθρωποειδές:** Ρομπότ που μοιάζει με οποιοδήποτε άλλο ζωντανό οργανισμό εκτός από τον άνθρωπο ή δεν μιμείται έναν ζωντανό οργανισμό. (π.χ. Aibo, ένα ρομπότ που μοιάζει με σκύλο, [267]).

Γεωγραφική τοποθεσία

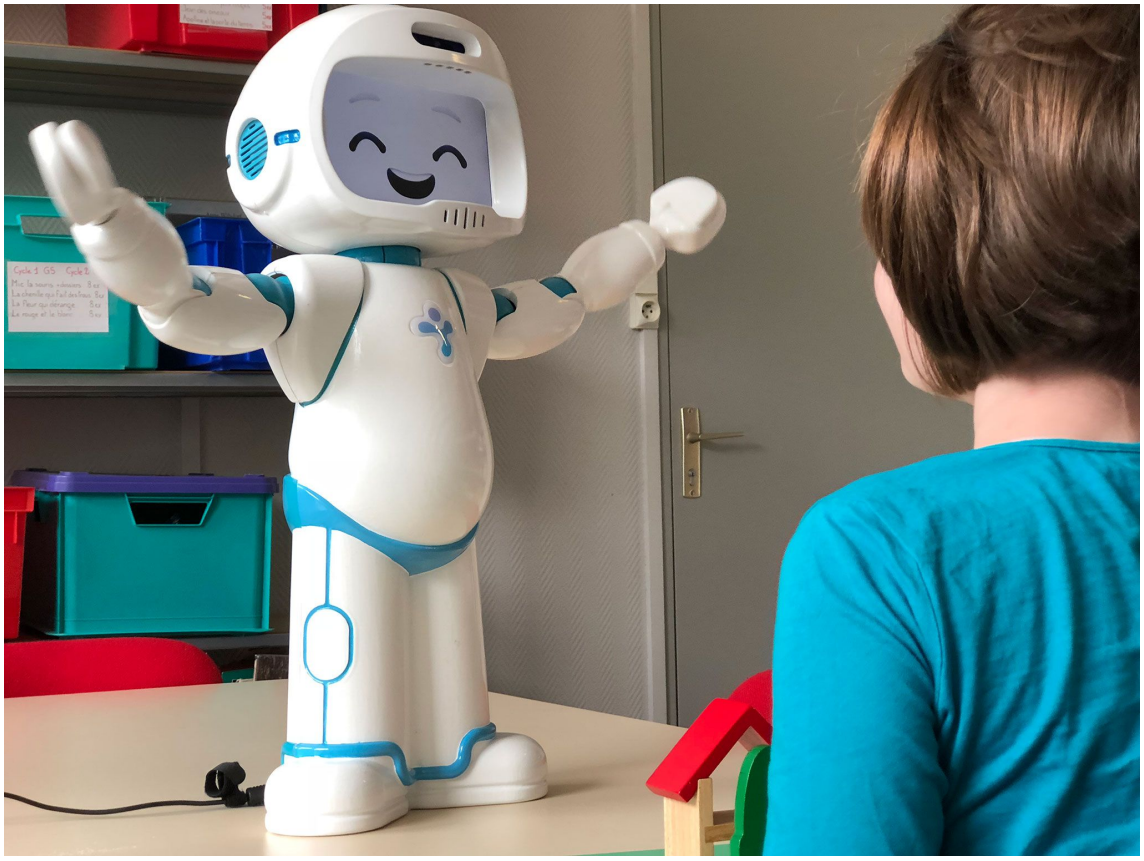
Το πολιτισμικό υπόβαθρο και η εθνικότητα των χρηστών μπορεί να συμβάλλουν στη διαφοροποίηση της στάσης των ανθρώπων απέναντι στα [268], την εμπιστοσύνη στα [31] και την αποδοχή των [269] κοινωνικών ρομπότ.

Χαρακτηριστικά του δείγματος

Η στάση απέναντι στα ρομπότ πιθανόν να διαφέρει και ανάλογα με δημογραφικούς παράγοντες, όπως η ηλικία και το φύλο των χρηστών [264]. Για παράδειγμα, οι άνδρες τείνουν γενικά να έχουν πιο θετική στάση απέναντι στα ρομπότ από ό,τι οι γυναίκες [255]. Ομοίως, οι νέοι ενήλικες τείνουν να έχουν πιο θετική στάση απέναντι στα ρομπότ από τους ηλικιωμένους και είναι πιο πρόθυμοι να κάνουν χρήση των ρομπότ [255]. Επιπλέον, ορισμένες μελέτες έχουν αναφέρει ότι η προηγούμενη εμπειρία με μακροχρόνια έκθεση σε ρομπότ επηρεάζει επίσης τις στάσεις των ανθρώπων [270].

2.5.3 Ανθρωπομορφικά χαρακτηριστικά προσώπου του ρομπότ και αξιοπιστία

Δεδομένου ότι η τεχνολογία έχει εξελιχθεί και εφαρμόζεται σε διάφορα καθημερινά πλαίσια [271], το κοινωνικό ρομπότ, ως ένας από τους εκπροσώπους της τελευταίας καινοτομίας, είναι ένα



Εικόνα 5. *Ρομπότ QTrobot*

σύστημα τεχνητής νοημοσύνης που μπορεί να επικοινωνεί και να αλληλεπιδρά κοινωνικά με τον άνθρωπο [272]. Διαφορετικά από τα παραδοσιακά ανθρωποειδή ρομπότ (π.χ. ρομποτικό προϊόν, Zora Robot) που ενσωματώνονται φυσικά με συγκεκριμένα ανθρώπινα χαρακτηριστικά, ορισμένα πιο πρόσφατα κοινωνικά ρομπότ (π.χ. ρομποτικά προϊόντα, Jibo, Welbo, Misa, QTrobot, Hub, Mykie και Buddy Robot) έχουν σχεδιαστεί με μια οθόνη, η οποία διασυνδέεται με ένα κινούμενο πρόσωπο που μοιάζει με άνθρωπο, για να επικοινωνούν και να αλληλεπιδρούν με τους ανθρώπους [273]. Μπορεί να είναι απαραίτητο να σχεδιαστεί μια διεπαφή που μοιάζει με κεφάλι για τα κοινωνικά ρομπότ, ώστε να βοηθηθεί η επικοινωνία, καθώς οι άνθρωποι τείνουν να χρησιμοποιούν τις γνωστικές και αντιληπτικές τους διαδικασίες για να διαμορφώνουν προσδοκίες σχετικά με τον τρόπο αλληλεπίδρασης μαζί τους.

Μεταξύ των διαφόρων αντιληπτών χαρακτηριστικών, όπως η κυριαρχία, η φιλικότητα και η ελκυστικότητα, η αξιοπιστία προς ένα κοινωνικό ρομπότ παίζει καθοριστικό ρόλο στην αλληλεπίδραση ανθρώπου-ρομπότ (HRI) για δύο λόγους [274]. Από τη μία πλευρά, η αξιοπιστία είναι ζωτικής σημασίας σε κοινωνικά πλαίσια, καθώς έχει σημαντικό αντίκτυπο στην πειθώ και θα μπορούσε να επηρεάσει άμεσα την πρόθεση των ανθρώπων να ακολουθήσουν τις προτάσεις των άλλων [148]. Από την άλλη πλευρά, τα κοινωνικά ρομπότ λειτουργούν ως ”επικοινωνούντες και αντιδρώντες”, παρέχοντας όχι μόνο φυσική βοήθεια αλλά και συναισθηματική υποστήριξη στους ανθρώπους, επομένως θα πρέπει αρχικά να θεωρούνται ασφαλή για να τα εμπιστεύονται [275].

Επιπλέον, οι Gompei και Umemuro [276] ανέφεραν ότι υπάρχουν διάφορα ζητήματα που διαδραματίζουν σημαντικό ρόλο στον καθορισμό της αξιοπιστίας των κοινωνικών ρομπότ: ζητήματα που σχετίζονται με το ρομπότ (π.χ, τα χαρακτηριστικά και οι επιδόσεις του ρομπότ), θέματα σχετικά με τον άνθρωπο (η συγκεκριμένη ανάγκη, η τάση για εμπιστοσύνη, η προσωπικότητα, η άνεση, η αυτοπεποίθηση, η στάση, η μνήμη, η προσοχή, η εμπειρογνωμοσύνη, η ικανότητα, ο

φόρτος εργασίας, η προηγούμενη εμπειρία και η επίγνωση της κατάστασης), και θέματα σχετικά με το περιβάλλον (η εφαρμογή της εργασίας, η πολυπλοκότητα της εργασίας, η απαίτηση πολλαπλών εργασιών, το φυσικό περιβάλλον, η συμμετοχή στην ομάδα, η κουλτούρα, η επικοινωνία, η ομαδική συνεργασία κ.λπ.) Μεταξύ αυτών, τα θέματα που σχετίζονται με το ρομπότ είναι οι σημαντικότεροι παράγοντες που επηρεάζουν την αξιολόγηση της αξιοπιστίας των ανθρώπων απέναντι στην αλληλεπίδραση ανθρώπου-ρομπότ [5] [277]. Για να ακριβολογηθεί, μπορεί να σχετίζονται με τη συμπεριφορά του ρομπότ, την επάρκεια, την αξιοπιστία, την προβλεψιμότητα, την αυτοματοποίηση, το ποσοστό αποτυχίας, τη διαφάνεια, την εγγύτητα, την προσωπικότητα, την προσαρμοστικότητα, τον τύπο και τον ανθρωπομορφισμό [5]. Για παράδειγμα, προηγούμενες μελέτες έχουν υποδείξει ότι τα προβαλλόμενα ανθρωπομορφικά πρόσωπα τείνουν να κάνουν τους ανθρώπους να αισθάνονται μεγαλύτερη διέγερση και μεγαλύτερη συμπάθεια [278] [279], οδηγώντας τελικά σε υψηλότερο επίπεδο αντιλαμβανόμενης αξιοπιστίας για τα κοινωνικά ρομπότ (σε σύγκριση με ένα μηχανικό πρόσωπο των κοινωνικών ρομπότ) [280].

Εικόνα 6. *Ρομπότ Mykie*

Πράγματι, οι άνθρωποι μπορούσαν να αξιολογήσουν τα πρόσωπα ανθρώπων και άψυχων αντικειμένων, όπως ρομπότ και προϊόντων, σε απίστευτα σύντομο χρονικό διάστημα [281]. Προηγούμενες έρευνες έδειξαν ότι 100 ms ήταν αρκετά για να μπορέσουν οι άνθρωποι να κρίνουν πολλαπλά χαρακτηριστικά της προσωπικότητας, όπως η αξιοπιστία, η ικανότητα και η επιθετικότητα [282]. Ο λόγος για τον οποίο οι άνθρωποι φαίνεται να είναι έτοιμοι να αντιληφθούν και να επεξεργαστούν τα πρόσωπα στα αντικείμενα έγκειται στην ανθρώπινη εξελικτική προσαρμογή: το ανθρώπινο πρόσωπο είναι ένα εξέχων εξελικτικά και αξιοσημείωτο ερέθισμα που προσελκύει την προσοχή το οποίο μπορεί να επεξεργαστεί ταυτόχρονα [283]. Όταν πρόκειται για την αξιολόγηση ενός ρομπότ, σε αντίθεση με την απλή πρόθεση να αναζητήσουν την ομοιότητα με ένα ανθρώπινο πρόσωπο, οι άνθρωποι θα μπορούσαν να αντιληφθούν συγκεκριμένα χαρακτηριστικά του προσώπου ή εκφράσεις στο ρομπότ ταυτίζοντας συγκεκριμένα χαρακτηριστικά του ρομπότ με ανθρώπινα χαρακτηριστικά και κάνοντας την αναλογία [278] [284].

Τα χαρακτηριστικά του προσώπου σε ένα κοινωνικό ρομπότ μπορεί επίσης να έχουν αντίκτυπο στην εμπιστευτικότητα τέτοιων τεχνητών πρακτόρων [276] [277]. Προηγούμενες σχετικά άρθρα σχετικά με την αξιοπιστία επικεντρώνονται σε μεγάλο βαθμό στη σύνοψη των χαρακτηριστικών εμπιστοσύνης του ανθρώπινου προσώπου και στη συζήτηση της γενικής αξιοπιστίας στη σχέση ανθρώπου-υπολογιστή/ανθρώπου-μηχανής. Για παράδειγμα, οι Hancock κ.ά. [5] αξιολογούν τις επιδράσεις των ανθρώπινων, ρομποτικών και περιβαλλοντικών παραγόντων στην αντιλαμβανόμενη εμπιστοσύνη σε HRI γενικά. Ωστόσο, ο όρος της εμπιστοσύνης είναι πράγματι μια έννοια πολλαπλών δομών που περιείχε διάφορα στάδια αξιολόγησης, όπως η αρχική αξιολόγηση στην πρώτη εντύπωση και η εκ των υστέρων αξιολόγηση στα τελευταία στάδια [282]. Εξάλλου, αν και η

προηγούμενη βιβλιογραφία προσπάθησε να αξιολογήσει το ρόλο των διαφόρων χαρακτηριστικών του προσώπου στην επεξεργασία της εμπιστοσύνης [285], δεν θα μπορούσε να εισαχθεί απλά στην HRI λόγω της ανθρωπομορφικής φύσης των κοινωνικών ρομπότ [280]. Πράγματι, η έρευνα σχετικά με τον σχεδιασμό του προσώπου του κοινωνικού ρομπότ είναι ένα διεπιστημονικό πεδίο που σπάνια αναλύεται συστηματικά και μελετάται σποραδικά από διάφορους τομείς.Συγκεκριμένα, υπάρχουν τουλάχιστον τρεις σημαντικές επιστημονικές προσεγγίσεις σχετικά με την ανθρωπομορφική αξιοπιστία του προσώπου των κοινωνικών ρομπότ [286]:

1. Δεδομένου ότι οι άνθρωποι και τα κοινωνικά ρομπότ μπορεί να μοιράζονται παρόμοια χαρακτηριστικά του προσώπου, όπως τα μάτια και το στόμα, η ανθρώπινη αξιοπιστία του προσώπου από την ψυχολογία, η οποία έχει συζητήσει εδώ και καιρό τα ειδικά για τον άνθρωπο χαρακτηριστικά της αξιοπιστίας του προσώπου [287], θα μπορούσε ενδεχομένως να συμβάλει στη γνώση της ανθρωπομορφικής αξιοπιστίας του προσώπου του κοινωνικού ρομπότ.

2. Ως ρομποτικό προϊόν στην αγορά, το κοινωνικό ρομπότ θα μπορούσε να αντλήσει έμπνευση από προηγούμενη βιβλιογραφία σχετικά με την εμφάνιση προϊόντων από το μάρκετινγκ και τον μηχανολογικό σχεδιασμό, η οποία έχει συζητηθεί σχετικά με τον τρόπο κατασκευής μιας αξιόπιστης εμφάνισης για ανθρωπόμορφα προϊόντα [288]. Η ανθρωπόμορφη εμφάνιση ενός προϊόντος αναφέρεται στη φυσική εμφάνιση ενός προϊόντος με ανθρώπινα χαρακτηριστικά ή γνωρίσματα του προσώπου, όπως ο προβολέας ενός αυτοκινήτου ή ο ακροδέκτης τροφοδοσίας μιας πρίζας [289]. Για παράδειγμα, οι Maeng και Aggarwal [284] πρότειναν ότι το μπροστινό μέρος ενός αυτοκινήτου με χαμηλότερο λόγο πλάτους-ύψους προσώπου (”facial width-height-ratio”/ fWHR) μπορεί να αξιολογηθεί με μεγαλύτερη αξιοπιστία. Παρόλο που ο ανθρωπόμορφος σχεδιασμός προϊόντων δεν σχετίζεται άμεσα με τον σχεδιασμό προσώπου ρομπότ, θα μπορούσε επίσης να παρέχει, τουλάχιστον, κάποιες διαισθήσεις για τον σχεδιασμό ενός αξιόπιστου ρομπότ, δεδομένου ότι μπορεί όλα να μοιράζονται παρόμοια ανθρωπόμορφα χαρακτηριστικά στην επικοινωνία της αξιοπιστίας.

3. Αν και η προηγούμενη βιβλιογραφία για τα κοινωνικά ρομπότ έχει εξετάσει την αξιοπιστία του προσώπου των κοινωνικών ρομπότ, επικεντρώθηκε κυρίως στη γενική επίδραση της ανθρωπόμορφης αξιολόγησης της αξιοπιστίας, όπως η διαφορά αξιοπιστίας μεταξύ ανθρωπόμορφων προσώπων και μηχανικών προσώπων στα κοινωνικά ρομπότ [280]. Πράγματι, πρόκειται για ένα διεπιστημονικό ερευνητικό πεδίο, ενώ ένα μόνο ερευνητικό πεδίο δύσκολα θα μπορούσε να παράσχει συγκεκριμένες οδηγίες για να βοηθήσει τους σχεδιαστές και τους μηχανικούς των κοινωνικών ρομπότ να βελτιώσουν την αξιοπιστία στο πρόσωπο του ρομπότ.

Παρόλο που τόσο η έρευνα για τα ρομπότ όσο και η έρευνα για τη συμπεριφορά έχουν αναγνωρίσει τη σημασία του σχεδιασμού των κοινωνικών ρομπότ για την επιτυχία τους στην αγορά και το σχετικό κοινωνικό όφελος για τους χρήστες τους [273] [281], τα συγκεκριμένα χαρακτηριστικά του προσώπου στην ανάδειξη της ανθρωπόμορφης αξιοπιστίας των κοινωνικών ρομπότ εξακολουθούν να λαμβάνουν περιορισμένη προσοχή.

Ανάλυση χαρακτηριστικών

Η έρευνα σχετικά με τα αξιόπιστα χαρακτηριστικά του προσώπου για τους ανθρώπους, τα προϊόντα και τα ρομπότ χωρίζεται σε τέσσερα ρεύματα: εσωτερικά, εξωτερικά, συνδυασμοί καθώς και δυναμικά χαρακτηριστικά και συναισθήματα. Όπως ανέφεραν οι Santos και Young [290], τα εσωτερικά χαρακτηριστικά ενός προσώπου περιλαμβάνουν το μέγεθος, το χρώμα, το σχήμα, το βλέμμα, τα φρύδια, την αντίθεση φωτεινότητας, το μάγουλο, τη μύτη, τα χείλη και το στόμα. Τα εξωτερικά χαρακτηριστικά περιλαμβάνουν τον λόγο πλάτους-ύψους προσώπου (fWHR), τον

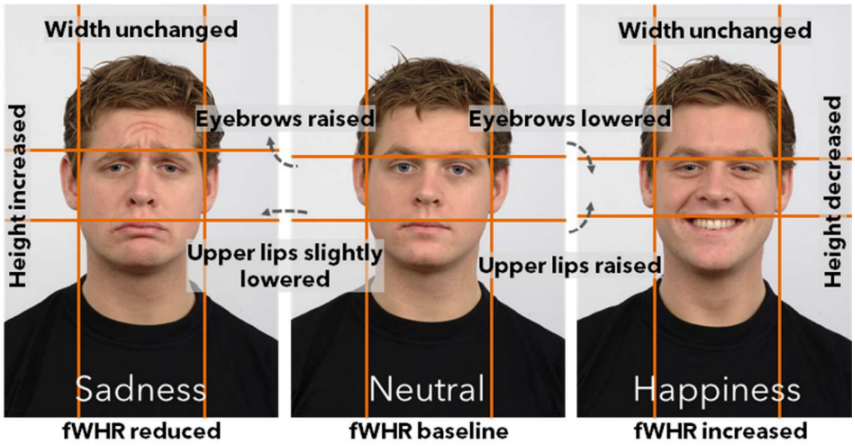
λόγο φρυδιών-μύτης-πηγουνιού, τα μαλλιά, το μέτωπο, τα αυτιά, τα γένια, το πηγούνι, τα γυαλιά, το τατουάζ, την ηλικία και την εθνικότητα. Διαφορετικοί συνδυασμοί αυτών των χαρακτηριστικών μπορούν να δημιουργήσουν ορισμένα χαρακτηριστικά, όπως ομορφιά, συμμετρία και αρρενωπότητα. Τα δυναμικά χαρακτηριστικά αναφέρονται στην κίνηση συγκεκριμένων χαρακτηριστικών του προσώπου, ενώ οι συναισθηματικές εκφράσεις αναφέρονται σε ένα σύνολο χαρακτηριστικών του προσώπου, που κινητοποιούν τους ανθρώπους να αντιληφθούν τα συναισθήματα που προκαλούν.

- **Εσωτερικά χαρακτηριστικά:** Η περιοχή των ματιών είναι ένας σημαντικός παράγοντας που επηρεάζει την αξιολόγηση της αξιοπιστίας τόσο για τους ανθρώπους όσο και για τα προϊόντα [281] [291] [292] [290]. Αυτή η περιοχή έχει ποικίλα χαρακτηριστικά που θα μπορούσαν να μεταδώσουν την αξιοπιστία, όπως το μέγεθος των ματιών, το σχήμα των ματιών, το βλέμμα, το χρώμα καθώς και τα φρύδια [290] [293]. Μελέτες σχετικά με το σχήμα και το μέγεθος των ματιών δείχνουν ότι οι άνθρωποι με στρογγυλά μάτια (έναντι στενών) [294] και μεγαλύτερα μάτια (έναντι μικρότερων) [292] [295] θεωρούνται πιο αξιόπιστοι, δεδομένου ότι όλα αυτά τα χαρακτηριστικά μοιράζονται και απολαμβάνουν τα χαρακτηριστικά εμφάνισης του παιδικού προσώπου από εξελικτική άποψη [296]. Μελέτες έχουν επίσης διαπιστώσει ότι η διατήρηση άμεσης οπτικής επαφής και η ύπαρξη λεπτών, ανασηκωμένων φρυδιών μπορεί να ενισχύσει την αντιληπτή αξιοπιστία και τη φυσική ελκυστικότητα ενός ατόμου [291] [297] [294] [290] [293] [298]. Στο πεδίο των κοινωνικών ρομπότ, μπορεί να υπάρχει μια διαφοροποιημένη σχέση μεταξύ του βλέμματος και της αξιοπιστίας: Οι Stanton και Stevens [299] πρότειναν ότι το σταθερό βλέμμα, σε σύγκριση με το αποστρεφόμενο βλέμμα, μπορεί να υποδηλώνει κυριαρχία, παρά αξιοπιστία, και αυτό το αποτέλεσμα ήταν ιδιαίτερα σημαντικό όταν οι γυναίκες συμμετέχουσες προσπαθούσαν να αξιολογήσουν το ρομπότ. Όπως ανέφερε ο συγγραφέας, ένας από τους περιορισμούς της εργασίας τους είναι το σχετικά μικρό μέγεθος του δείγματος και το μη ισορροπημένο φύλο [299]. Σε αντίθεση με άλλα εσωτερικά χαρακτηριστικά, το χρώμα των ματιών δεν είναι ένα απομονωμένο χαρακτηριστικό αλλά ένα χαρακτηριστικό που σχετίζεται με την εθνική ομάδα και εμφανίζεται με άλλα χαρακτηριστικά του προσώπου εντός της πολιτισμικής προέλευσης [300]. Παρόλο που ο Kleisner και οι συνεργάτες του [292] ανέφεραν ότι τα πρόσωπα με καστανά μάτια θεωρούνται πιο αξιόπιστα από τα πρόσωπα με μπλε μάτια, εξήγησαν περαιτέρω ότι η διαφορά στην αντίληψη της αξιοπιστίας μπορεί να σχετίζεται με τα χαρακτηριστικά του προσώπου που σχετίζονται.

  Οι περιοχές της μύτης και του στόματος είναι σημαντικά χαρακτηριστικά που επηρεάζουν την αξιολόγηση της αξιοπιστίας από τους ανθρώπους. Προηγούμενες μελέτες έχουν υποδείξει ότι τα κεντρικά χαρακτηριστικά του προσώπου (μύτη και στόμα) [301] [302] συσχετίζονται θετικά με την προσοχή και την αξιοπιστία. Σύμφωνα με προηγούμενη βιβλιογραφία, υπάρχουν τρεις τύποι σχημάτων στόματος: προς τα πάνω (χαμογελαστό), προς τα κάτω (λυπημένο) και ουδέτερο. Τα αντιληπτά κοινωνικά χαρακτηριστικά διαφέρουν σημαντικά μεταξύ αυτών των τριών σεναρίων. Ένα προϊόν ή ένα ανθρώπινο πρόσωπο με ανεστραμμένο στόμα ή χαμόγελο θεωρείται πιο αξιόπιστο, φιλικό και ελκυστικό σε σύγκριση με ένα ουδέτερο ή ανάποδα στραμμένο στόμα [301] [303] [281] [284] [292]. Ορισμένα χαρακτηριστικά του προσώπου μπορούν να επηρεάσουν την αντίληψη της αξιοπιστίας, όπως τα έντονα ζυγωματικά, το φαρδύ πηγούνι και τα λεπτά χείλη χωρίς να λείπουν τα μπροστινά δόντια. Αντίθετα, τα άτομα με ρηχά ζυγωματικά, λεπτά πηγούνια και γεμάτα χείλη με ελλείποντα μπροστινά δόντια μπορεί να φαίνονται λιγότερο αξιόπιστα [293]. Προηγούμενες εμπειρικές έρευνες δείχνουν αντιφατικά αποτελέσματα σχετικά με το πώς τα χαρακτηριστικά της μύτης επηρεάζουν την αξιολόγηση της αξιοπιστίας. Ορισμένοι ερευνητές συμφωνούν ότι μια μικρότερη μύτη μπορεί να έχει ως αποτέλεσμα την αντίληψη λιγότερης στιβαρότητας και αξιοπιστίας στους άνδρες. [292], ενώ η επικρατούσα βιβλιογραφία υποστηρίζει ότι η κο-

ντή και ρηχή μύτη αποτελούν βασικό παράγοντα για τη διαμόρφωση της αξιοπιστίας [295] [293].

Συνοψίζοντας, ένα κοινωνικό ρομπότ μπορεί να θεωρείται πιο αξιόπιστο με βάση ορισμένα εσωτερικά χαρακτηριστικά ή συνδυασμούς αυτών, όπως στρογγυλά ή μεγάλα μάτια, άμεσο βλέμμα, κοντή μύτη και προς τα πάνω στραμμένο στόμα.



- **Εξωτερικά χαρακτηριστικά:** Οι αναλογίες προσώπου, συμπεριλαμβανομένης της fWHR, επηρεάζουν την αξιολόγηση της αξιοπιστίας [284]. Συγκεκριμένα, ένα μεγαλύτερο fWHR γίνεται αντιληπτό ως πιο κυρίαρχο, επιθετικό, μη ελκυστικό και αναξιόπιστο στην ανθρώπινη αντίληψη [287] [295]. Ωστόσο, στον τομέα της αξιολόγησης προϊόντων, αυτό που μπορεί να είναι αντιφατικό είναι ότι ένα μεγάλο fWHR του προϊόντος θα άρεσε περισσότερο, καθώς λειτουργεί ως ένδειξη της κυρίαρχης θέσης του χρήστη [284]. Σύμφωνα με έρευνες, η αναλογία μετώπου-μύτης-πιγουνιού επηρεάζει την αξιοπιστία, αλλά η επίδραση μπορεί να ποικίλλει ανάλογα με το πλαίσιο [297]. Για παράδειγμα, η αναλογία αποτελεί σημαντικό προγνωστικό παράγοντα για την αξιολόγηση της αξιοπιστίας του προσώπου ενός 12χρονου αγοριού, αλλά όχι για άλλες ηλικίες ή φύλα.

  Υπάρχουν αρκετές μελέτες που προσπαθούν να διερευνήσουν άλλα εξωτερικά χαρακτηριστικά που επηρεάζουν την εκτίμηση της αξιοπιστίας [294] [290] [298].Προηγούμενες μελέτες υποδεικνύουν μια ασαφή συσχέτιση μεταξύ του μεγέθους του μετώπου και της αντίληψης της αξιοπιστίας. Ωστόσο, η εξελικτική ψυχολογία προτείνει ότι τα βρέφη με προεξέχον μέτωπο, μικρό πηγούνι και κοντά αυτιά θεωρούνται αξιόπιστα, ενώ μια προηγούμενη μελέτη διαπίστωσε ότι το μέγεθος του μετώπου μπορεί να επηρεάσει την αντίληψη της αξιοπιστίας [298] [284] [294] [292]. Ο λόγος για τον οποίο εμφανίζεται αυτή η αντίφαση μπορεί να οφείλεται στους διαφορετικούς ορισμούς της ίδιας λέξης. Συγκεκριμένα, η λέξη ”ψηλότερο και μικρότερο μέτωπο”, που αναφέρεται σε προηγούμενη έρευνα [298], αναφέρεται σε μια σχετικά μικρή περιοχή του μετώπου με σχετικό μεγάλο ύψος. Ωστόσο, το μέγεθος του μετώπου που αναφέρεται στο [287] είναι στην πραγματικότητα η απόσταση από τα μάτια έως τα μαλλιά. Το μέγεθος του μετώπου και το ύψος χρειάζονται σαφέστερη εξήγηση σε διαφορετικά πλαίσια. Οι Hellström και Tekle [304] διαπίστωσαν ότι τα γυαλιά και η γενειάδα αυξάνουν την αντιλαμβανόμενη εξυπηρετικότητα και αξιοπιστία. Επιπλέον, τα μαλλιά και η απουσία τατουάζ στο πρόσωπο μπορούν να επηρεάσουν την αξιολόγηση της καλής εμφάνισης, της αξιοπιστίας, της ακεραιότητας και της ηγεσίας [305]. Ωστόσο, αυτό το αποτέλεσμα βασίζεται σε διαφορετικά επαγγέλματα. Η μείωση των πωλήσεων συνδέθηκε με στερεότυπα για άτομα με υψηλή μόρφωση, όπως οι καθηγητές, οι οποίοι πιστεύεται ότι είναι αξιόπιστοι, έξυπνοι και εξυπηρετικοί και αναμένεται να φορούν γυαλιά και να έχουν γένια αλλά όχι μαλλιά [304].

Έτσι, ένα κοινωνικό ρομπότ μπορεί να θεωρηθεί πιο αξιόπιστο αν έχει μεγάλο fWHR, μικρή αναλογία φρυδιών-μύτης-πηγουνιού, ψηλό μέτωπο, κοντά αυτιά και μικρό πηγούνι.

- **Συνδυασμοί χαρακτηριστικών:** Όσον αφορά τους συνδυασμούς χαρακτηριστικών, οι αντιλαμβανόμενοι τείνουν να εμπιστεύονται εκείνους που τους μοιάζουν, ανήκουν στην πολιτιστική τους ομάδα, έχουν παρουσιαστεί στο παρελθόν ή έχουν συμμετρικά πρόσωπα [306] [307] [302] [308]. Οι λόγοι έγκεινται στο ότι τόσο η ομοιότητα όσο και η τυπικότητα θα μπορούσαν να αυξήσουν την εξοικείωση, η οποία θα μπορούσε τελικά να ενισχύσει τη θετική αξιολόγηση της αξιοπιστίας [306] [302]. Η έκθεση σε κοινωνικά σχετικές πληροφορίες μπορεί να διαμορφώσει τις αντιλήψεις μας για τους ξένους επηρεάζοντας την επεξεργασία άγνωστων πληροφοριών του προσώπου, η οποία εξαρτάται από τις προσδοκίες μας από τις εμπειρίες της πραγματικής ζωής [307]. Επιπλέον, η συμμετρία του προσώπου είναι ένας καλά τεκμηριωμένος δείκτης τόσο της ελκυστικότητας όσο και της αξιοπιστίας [308]. Όταν υπάρχουν ασυμμετρίες του προσώπου, το αριστερό μισό του προσώπου επικοινωνεί την αξιοπιστία πιο αποτελεσματικά στις χαρούμενες εκφράσεις από ό,τι το δεξί μισό, επειδή η αριστερή πλευρά συνδέεται με τη συναισθηματική πλευρά του εγκεφάλου (το δεξί ημισφαίριο) και έχει πλεονέκτημα στην απόκρυψη αντικοινωνικών προθέσεων [302].

- **Δυναμικά χαρακτηριστικά και συναισθήματα:** Προηγούμενες μελέτες σχετικά με την επίδραση των κινήσεων του προσώπου στην αξιοπιστία έχουν επικεντρωθεί κυρίως στις κινήσεις των ματιών, του στόματος και του κεφαλιού. Ωστόσο, το άμεσο βλέμμα μπορεί να διαδραματίσει κρίσιμο ρόλο στην αξιολόγηση της ελκυστικότητας και της αξιοπιστίας, καθώς μπορεί να επηρεάσει την προσοχή και να υποδηλώσει κοινωνικό ενδιαφέρον, όπως συζητήθηκε σε προηγούμενη ενότητα σχετικά με τα εσωτερικά χαρακτηριστικά. [303] [291].

  Όπως τα μάτια, έτσι και οι κινήσεις του στόματος μπορούν να υποδηλώνουν την ειλικρίνεια και την αξιοπιστία. Αυτό συμβαίνει επειδή συνδέεται στενά με θετικές ή αρνητικές συναισθηματικές εκφράσεις, όπως το χαμόγελο. [301] [309] [297] [290]. Το χαμόγελο συνδέεται συνήθως με ένα στόμα σε σχήμα U και ανασηκωμένα φρύδια, υποδηλώνοντας θετικό συναίσθημα. Αντίθετα, ένα στόμα σε σχήμα ανεστραμμένου U και χαμηλωμένα φρύδια υποδηλώνουν θλίψη ή θυμό [301] [309]. Πράγματι, το συναίσθημα και η αντιλαμβανόμενη αξιοπιστία συνδέονται μεταξύ τους: ενώ το χαρούμενο πρόσωπο θεωρείται πιο αξιόπιστο, το αξιόπιστο πρόσωπο πιστεύεται επίσης ότι είναι πιο χαρούμενο [310]. Δεδομένου ότι η κρίση της αξιοπιστίας συνδέεται συχνά με την ευτυχία [303], η κίνηση του στόματος φαίνεται στη συνέχεια να είναι ένα σημαντικό σήμα της κοινωνικής αντίληψης [301]. Ενώ το χαμόγελο θεωρείται γενικά ως θετικός συναισθηματικός δείκτης, οι άνθρωποι μπορούν να διακρίνουν μεταξύ διαφορετικών τύπων χαμόγελου, συμπεριλαμβανομένων των γνήσιων και μη γνήσιων χαμόγελων, καθώς μπορεί να συνδέονται με συγκεκριμένες κοινωνικές έννοιες. Οι άνθρωποι είναι πιο πιθανό να εμπιστευτούν και να συνεργαστούν με εκείνους που έχουν αυθεντικά χαμόγελα σε αντίθεση με τα ψεύτικα. [309] [311].

  Τα αρνητικά συναισθήματα, όπως ο θυμός, η αηδία ή ο φόβος, μπορεί να υποδηλώνουν αξιοπιστία σε ορισμένα πλαίσια [310]. Για παράδειγμα, η έκφραση φόβου χαρακτηρίζεται από ανασηκωμένα εσωτερικά και εξωτερικά φρύδια, διευρυμένα μάτια, ένα τράβηγμα των γωνιών των χειλιών προς τα έξω και ένα πεσμένο σαγόνι. Οι Reed και DeScioli [312] έδειξαν ότι οι άνθρωποι ήταν πιο πιθανό να πιστέψουν ένα μήνυμα ειδοποίησης που παραδιδόταν με φοβισμένη έκφραση, γεγονός που υποδηλώνει ότι τα αρνητικά συναισθήματα μπορούν να ενισχύσουν την αξιοπιστία. Ο Flowe [313] διαπίστωσε ότι μια θυμωμένη έκφραση γίνεται αντιληπτή ως λιγότερο αξιόπιστη, πιο κυρίαρχη και πιο ποινικοφανής στις αξιολογήσεις της εγκληματικής εμφάνισης. Ωστόσο, σε ένα σενάριο χωρίς πλαίσιο, η έκφραση φόβου μπορεί να μην επηρεάζει σημαντικά την αξιολόγηση της αξιοπιστίας, ενώ οι θυμωμένες ή αηδιασμένες εκφράσεις εξακολουθούν να συμβάλλουν αποτελεσματικά στις αναξιόπιστες

αντιλήψεις [310].

Ο Engell και οι συνεργάτες του [311] μελέτησαν τον τρόπο με τον οποίο οι άνθρωποι αξιολογούν την αξιοπιστία των ουδέτερων προσώπων μετά την προσαρμογή τους σε χαρούμενες ή θυμωμένες εκφράσεις. Τα αποτελέσματα έδειξαν ότι η αρχική προσαρμογή σε μια χαρούμενη ή θυμωμένη έκφραση επηρέασε την αντιλαμβανόμενη αξιοπιστία ενός ουδέτερου προσώπου σε μεταγενέστερο στάδιο. Οι φοβισμένες εκφράσεις δεν είχαν την ίδια επίδραση, γεγονός που υποδηλώνει ένα φαινόμενο γενίκευσης, σύμφωνα με το οποίο μπορεί να εμπλέκεται ένα κοινό σύστημα ουδετερότητας κατά την αξιολόγηση της αξιοπιστίας του προσώπου σε χαρούμενες ή θυμωμένες εκφράσεις.

Κεφάλαιο 3

Αποτελέσματα – Ευρήματα / Επιτεύγματα

Στο παρακάτω κεφάλαιο θα αναφερθούν συνοπτικά τα κυριότερα αποτελέσματα κάθε ενότητας.

3.1 Κυριότερα ευρήματα / αποτελέσματα

3.1.1 Η εμπιστοσύνη στην συνεργασία ανθρώπου-ρομπότ

- **Ορισμός της εμπιστοσύνης:** Πρώτες έρευνες περί εμπιστοσύνης διερεύνησαν την έννοια της χωρίς συγκεκριμένο πλαίσιο. Στην εξέλιξη των ερευνών, έχουν προταθεί διάφοροι ορισμοί, όπως προδιάθεση προς τον κόσμο, κοινωνικά μαθημένες προσδοκίες, πεποιθήσεις, στάσεις, και προθυμία αποδοχής της τρωτότητας.

- **Βάσεις της Εμπιστοσύνης:** Η εμπιστοσύνη βασίζεται σε κοινά χαρακτηριστικά. Ορίστηκαν τρεις γενικές βάσεις: ικανότητα, ακεραιότητα και αγαθότητα. Αυτές εφαρμόζονται διαφορετικά ανάλογα με τον τύπο της εμπιστοσύνης.

- **Εξέλιξη της Εμπιστοσύνης:** Η εξέλιξη της εμπιστοσύνης μεταξύ ανθρώπου και αυτοματισμού διαφέρει από αυτή μεταξύ ανθρώπων. Στην εμπιστοσύνη ανθρώπου-αυτοματισμού, η προδιάθεση προς εμπιστοσύνη προς τα νέα αυτοματοποιημένα συστήματα μπορεί να είναι υψηλή, αλλά μειώνεται γρήγορα με σφάλματα. Αυτό αντικαθίσταται με την αξιοπιστία και την προβλεψιμότητα του συστήματος.

- **Διαφορές ανάμεσα στην εμπιστοσύνη ανθρώπου-αυτοματισμού και τη διαπροσωπική εμπιστοσύνη:** Οι δύο τύποι εμπιστοσύνης βασίζονται σε διαφορετικά χαρακτηριστικά. Η εμπιστοσύνη ανθρώπου-αυτοματισμού εξαρτάται από την απόδοση, τη διαδικασία ή το σκοπό του αυτοματοποιημένου συστήματος, ενώ η διαπροσωπική εμπιστοσύνη εξαρτάται από την ικανότητα, την ακεραιότητα και την αγαθότητα των ανθρώπων.

- **Περιγραφή της Έννοιας της Εμπιστοσύνης Προδιάθεσης:** Η εμπιστοσύνη προδιάθεσης αντιπροσωπεύει τη συνολική τάση ενός ατόμου να εμπιστεύεται τον αυτοματισμό, ανεξάρτητα από το πλαίσιο ή ένα συγκεκριμένο σύστημα.

- **Πηγές Μεταβλητότητας της Εμπιστοσύνης Προδιάθεσης:** Οι κύριες πηγές μεταβλητότητας στην εμπιστοσύνη προδιάθεσης είναι η κουλτούρα, η ηλικία, το φύλο και οι χαρακτηριστικά της προσωπικότητας:

  - Η κουλτούρα αποτελεί σημαντική μεταβλητή που επηρεάζει την εμπιστοσύνη στον αυτοματισμό. Έρευνες έχουν δείξει διαφορές στην εμπιστοσύνη μεταξύ διαφόρων κοινωνικών ομάδων, όπως χώρες, φύλα, θρησκείες και γενεαλογικές ομάδες.
  - Οι ηλικιακές διαφορές επηρεάζουν την εμπιστοσύνη στον αυτοματισμό, με διαφορετικές στρατηγικές ανάλυσης της αξιοπιστίας των συστημάτων ανάλογα με την ηλικία.
  - Το φύλο μπορεί να επηρεάσει την αντίδραση των ατόμων στον αυτοματισμό, με παρατηρούμενες διαφορές στον τρόπο αλληλεπίδρασης και εμπιστοσύνης με τεχνολογίες.
  - Τα χαρακτηριστικά της προσωπικότητας ενός ατόμου αποτελούν σταθερό παράγοντα της εμπιστοσύνης προδιάθεσης. Η εμπιστοσύνη αυτή είναι η τάση του ατόμου να εμπιστεύεται άλλους ανθρώπους καθ' όλη τη διάρκεια της ζωής του.

- **Περιστασιακή Εμπιστοσύνη και Εξωγενείς παράγοντες:** Η εμπιστοσύνη σε αυτοματοποιημένα συστήματα διαφέρει ανάλογα με την κατάσταση. Εξωτερικοί παράγοντες, όπως η δυσκολία της εργασίας και η πολυπλοκότητα του συστήματος, επηρεάζουν την εμπιστοσύνη ως εξής:

  - Ο υψηλός φόρτος εργασίας μπορεί να αυξήσει την εξάρτηση από τον αυτοματισμό και να επηρεάσει την αλληλεπίδραση,
  - Η αντίληψη ότι ο αυτοματισμός προσφέρει μεγαλύτερα οφέλη μπορεί να αυξήσει την εμπιστοσύνη, ενώ ο κίνδυνος μπορεί να μειώσει την εξάρτηση από τον αυτοματισμό.
  - Τέλος, παράγοντες όπως η πολυπλοκότητα, η είδηση, η παρουσία άλλων βοηθημάτων, και η φύση των περιστασιακών κινδύνων επηρεάζουν την εμπιστοσύνη, πιθανώς κάνοντας τις αντιφατικές αντιδράσεις λογικές.

- **Περιστασιακή Εμπιστοσύνη και Ενδογενείς παράγοντες:** Η προδιαθετική εμπιστοσύνη καλύπτει τα μόνιμα χαρακτηριστικά των φορέων. Ωστόσο, άτομα διαφέρουν σε μεταβατικά χαρακτηριστικά που εξαρτώνται από το εκάστοτε πλαίσιο, όπως η αυτοπεποίθηση και η εμπειρογνωμοσύνη:

  - Η αυτοπεποίθηση επηρεάζει την εμπιστοσύνη και τις αποφάσεις λήψης ελέγχου. Υψηλή αυτοπεποίθηση και χαμηλή εμπιστοσύνη μπορεί να οδηγήσει σε προτίμηση του χειροκίνητου ελέγχου, ενώ αντίθετες συνθήκες μπορεί να οδηγήσουν σε προτίμηση του αυτοματισμού.
  - Η τεχνογνωσία επηρεάζει την εμπιστοσύνη. Πιο έμπειροι χειριστές μπορεί να είναι λιγότερο πρόθυμοι να βασιστούν στον αυτοματισμό από αρχάριους.
  - Το συναίσθημα επηρεάζει την αρχική εμπιστοσύνη. Θετική διάθεση μπορεί να οδηγήσει σε μεγαλύτερη αρχική εμπιστοσύνη στον αυτοματισμό.
  - Η ικανότητα προσοχής του χειριστή επηρεάζει τον σχηματισμό εμπιστοσύνης. Στέρηση ύπνου μπορεί να οδηγήσει σε πιο προσεκτική παρακολούθηση του αυτοματισμού, αλλά και σε μεγαλύτερο κίνδυνο σφαλμάτων σε άλλες εργασίες.

- **Περιστασιακοί παράγοντες και η σχέση μεταξύ εμπιστοσύνης και εξάρτησης:**

  - Οι Lee και See [1] υποστηρίζουν ότι η εμπιστοσύνη έχει μεγαλύτερη επίδραση όταν η πολυπλοκότητα του συστήματος είναι υψηλή και όταν συμβαίνουν απρογραμμάτιστα γεγονότα που απαιτούν γρήγορη προσαρμογή.
  - Η εμπιστοσύνη επηρεάζει περισσότερο την εξάρτηση όταν το περιβάλλον επιτρέπει την αξιολόγηση της απόδοσης της αυτοματοποίησης έναντι της χειροκίνητης απόδοσης.

  - Υποκειμενικά επίπεδα εμπιστοσύνης μπορεί να έχουν ασθενέστερη επίδραση όταν οι χειριστές δεν μπορούν να αξιολογήσουν εύκολα την απόδοση του αυτοματισμού.
  - Η ελευθερία λήψης αποφάσεων επηρεάζει τη θετική συσχέτιση μεταξύ εμπιστοσύνης και εξάρτησης. Η ελευθερία λήψης αποφάσεων αναφέρεται στην ικανότητα των χειριστών να λαμβάνουν μελετημένες αποφάσεις για τη χρήση του αυτοματισμού.
  - Σε περιβάλλοντα με υψηλή ελευθερία λήψης αποφάσεων, η θετική σχέση μεταξύ εμπιστοσύνης και εξάρτησης είναι πιθανότερη.
  - Υψηλότερο φορτίο εργασίας μπορεί να μειώσει τη θετική σχέση μεταξύ εμπιστοσύνης και εξάρτησης. Οι χειριστές σε υψηλό φορτίο εργασίας μπορεί να χρησιμοποιούν αυτοματισμούς για να ανταποκριθούν στις απαιτήσεις, αλλά αυτό μπορεί να μειώσει την επίδραση της εμπιστοσύνης στην εξάρτηση.

- **Επίκτητη εμπιστοσύνη και υφιστάμενη γνώση:** Οι χειριστές βασίζονται στις προηγούμενες εμπειρίες τους με τον αυτοματισμό για να αξιολογήσουν την αξιοπιστία νέων συστημάτων. Η επίκτητη εμπιστοσύνη διαιρείται σε δύο κατηγορίες: αρχική (initial) και δυναμική (dynamic) εμπιστοσύνη. Η αρχικά μαθημένη εμπιστοσύνη αφορά την εμπιστοσύνη πριν την αλληλεπίδραση με ένα σύστημα, ενώ η δυναμικά μαθημένη εμπιστοσύνη αντικατοπτρίζει την εμπιστοσύνη κατά τη διάρκεια της αλληλεπίδρασης.

  - Οι επιδόσεις του αυτοματοποιημένου συστήματος επηρεάζουν την εμπιστοσύνη του χρήστη και μπορεί να οδηγήσουν σε αυξομειώσεις της εμπιστοσύνης κατά τη διάρκεια της αλληλεπίδρασης.
  - Η εμπιστοσύνη ενός χειριστή μπορεί να επηρεαστεί από τη φήμη του συστήματος πριν από την αλληλεπίδραση με αυτό. Οι άνθρωποι τείνουν να εμπιστεύονται περισσότερο τον αυτοματισμό όταν αυτός παρουσιάζεται ως αξιόπιστος ή ”έμπειρος” [23] [52].
  - Η προηγούμενη εμπειρία με ένα αυτοματοποιημένο σύστημα μπορεί να επηρεάσει τον σχηματισμό εμπιστοσύνης. Οι έμπειροι χρήστες έχουν την τάση να εμπιστεύονται περισσότερο το σύστημα μετά από προηγούμενη θετική εμπειρία, αλλά αυτό δεν ισχύει πάντα και μπορεί να εξαρτάται από τον τύπο της εμπειρίας [71] [72].
  - Οι προϋπάρχουσες νοοτροπίες και προσδοκίες του χειριστή μπορούν να επηρεάσουν τη διαδικασία σχηματισμού εμπιστοσύνης και τις επιλογές χρήσης. Οι έμμεσες στάσεις (που δρουν μέσω συσχετίσεων) μπορούν να επηρεάσουν την εμπιστοσύνη σε συνθήκες ασάφειας και κακής κατάστασης [69]. Η εμπειρία μπορεί να ενισχύσει την κατανόηση του χειριστή για τον σκοπό και τη διαδικασία ενός αυτοματοποιημένου συστήματος
  - Η προηγούμενη εμπειρία με τον αυτοματισμό μπορεί να επηρεάσει τον βαθμό εξάρτησης του χειριστή από το σύστημα. Εάν η εμπειρία ήταν αρνητική, μπορεί να οδηγήσει σε μειωμένη εμπιστοσύνη και εξάρτηση από τον αυτοματισμό [71].

- **Επίκτητη εμπιστοσύνη και σχεδιασμός του αυτοματισμού:**

  - Οι διεπαφές ηλεκτρονικών υπολογιστών αποτελούν οπτικό στοιχείο και επηρεάζουν την εμπιστοσύνη στον αυτοματισμό.
  - Αισθητικά ευχάριστοι ιστότοποι είναι πιο αξιόπιστοι, και αυτό ισχύει και για αυτοματοποιημένα συστήματα.
  - Ο ανθρωπομορφισμός μιας διεπαφής μπορεί να επηρεάσει την εμπιστοσύνη. Προσθήκη εικόνας γιατρού αυξάνει την εμπιστοσύνη σε μια εφαρμογή διαχείρισης διαβήτη [35].
  - Διακριτά χαρακτηριστικά ”προσωπικότητας” του αυτοματισμού, όπως ενίσχυση με καλή συμπεριφορά, επηρεάζουν την εμπιστοσύνη και τη διαγνωστική απόδοση.

- Ο σχεδιασμός της επικοινωνίας πρέπει να λαμβάνει υπόψη τις προτιμήσεις και δυνατότητες των χρηστών για την ενίσχυση της εμπιστοσύνης.
- Η διαφάνεια αναφέρεται στο βαθμό που οι εσωτερικές λειτουργίες ή η λογική πίσω από τα αυτοματοποιημένα συστήματα είναι κατανοητές και γνωστές στους χειριστές. Αυτή η διαφάνεια μπορεί να επηρεάσει την εμπιστοσύνη των χειριστών προς τα αυτοματοποιημένα συστήματα.
- Οι μελέτες έχουν δείξει ότι η παροχή ακριβούς ανατροφοδότησης σχετικά με την αξιοπιστία και τη λειτουργία των αυτοματοποιημένων συστημάτων μπορεί να βελτιώσει την εμπιστοσύνη των χρηστών προς αυτά και να βελτιώσει την απόδοση των εργασιών ανθρώπου-αυτοματισμού.
- Η εμπιστοσύνη στα αυτοματοποιημένα συστήματα εξαρτάται από το επίπεδο ελέγχου που διατηρεί ο χειριστής. Τα αυτοματοποιημένα συστήματα που παρέχουν πληροφορίες στον χειριστή θεωρούνται πιο αξιόπιστα.
- Ο προσαρμοστικός αυτοματισμός μπορεί να είναι αποτελεσματικός στο να βελτιώνει την ασφάλεια και την αποτελεσματικότητα των αυτοματοποιημένων συστημάτων, προσαρμόζοντας το επίπεδο ελέγχου σύμφωνα με τις προτιμήσεις του χρήστη.
- Τα χαμηλότερα επίπεδα αυτοματοποίησης μπορεί να επιφέρουν πρόσθετες καθυστερήσεις λόγω της ανάγκης για πληροφορίες από τον χειριστή, ενώ οι υψηλότερες επίπεδα μπορεί να απομακρύνουν τον χειριστή από τη διαδικασία.
- Οι χειριστές που βγαίνουν ”εκτός λούπας” μπορεί να εξαρτώνται περισσότερο από τον αυτοματισμό, και αυτό μπορεί να δυσκολέψει την αντιμετώπιση σφαλμάτων.

- **Επίκτητη εμπιστοσύνη και απόδοση:**
  - Η αξιοπιστία και εγκυρότητα των λειτουργιών ενός αυτοματοποιημένου συστήματος είναι βασικοί παράγοντες της εμπιστοσύνης των χειριστών. Η αξιοπιστία αφορά τη συνέπεια των λειτουργιών, ενώ η εγκυρότητα αναφέρεται στον βαθμό που η αυτοματοποίηση εκτελεί την προβλεπόμενη εργασία.
  - Η προβλεψιμότητα της αυτοματοποίησης, δηλαδή η συμμόρφωση με τις προσδοκίες του χειριστή, και η αξιοπιστία, που αφορά τη συχνότητα σφαλμάτων, επηρεάζουν την εμπιστοσύνη των χρηστών. Η προβλεψιμότητα μπορεί να ενισχύσει την εμπιστοσύνη, ενώ χαμηλή αξιοπιστία μπορεί να μειώσει την εξάρτηση από τον αυτοματισμό.
  - Οι ψευδείς συναγερμοί και οι αστοχίες έχουν διαφορετικές επιπτώσεις στην εμπιστοσύνη. Οι ψευδείς συναγερμοί μπορεί να μειώνουν τη συμμόρφωση, ενώ οι αστοχίες μπορεί να μειώνουν την εξάρτηση από την αυτοματοποίηση.
  - Η εμπιστοσύνη στον αυτοματισμό είναι σχετική με την χρησιμότητά του για τους χειριστές. Αν ο αυτοματισμός προσφέρει πραγματικά οφέλη και βοηθά στην εκτέλεση των καθηκόντων, οι χειριστές είναι πιθανότερο να τον εμπιστευτούν.

3.1.2 Μοντέλα και πλαίσια εμπιστοσύνης

- Τα μοντέλα εμπιστοσύνης αναπαριστούν την επίδραση διάφορων παραγόντων στη διαμόρφωση και αλλαγή της εμπιστοσύνης. Αυτά τα μοντέλα μπορεί να βασίζονται σε παράγοντες όπως η απόδοση του ρομπότ, η απόδοση του ανθρώπου, τον κίνδυνο κατάστασης, το γνωστικό φόρτο εργασίας και οι ανθρώπινες προσδοκίες.

- Οι παράγοντες εισόδου στα μοντέλα εμπιστοσύνης διαφέρουν ανάλογα με τον τομέα εφαρμογής. Οι μελέτες εξετάζουν την επίδραση παραγόντων όπως η απόδοση της συνεργασίας, η αντίληψη του χειριστή για τις δυνατότητες του συστήματος, τον φόρτο εργασίας και οι ανθρώπινες προσδοκίες.

- Οι στόχοι της εμπιστοσύνης στις μελέτες περιλαμβάνουν τη βελτίωση της απόδοσης της συνεργασίας, την αντιμετώπιση της αβεβαιότητας, και την προσαρμογή των ενεργειών του ρομπότ για τη βελτίωση της αποτελεσματικότητας της αλληλεπίδρασης.

- Η εποπτεία στην HRI έχει μελετηθεί με τη χρήση διάφορων μοντέλων εμπιστοσύνης. Οι Muir [104], Barber [17], Lee και Moray [24] ανέπτυξαν μοντέλα εμπιστοσύνης που βασίζονται σε ανθρώπινες προσδοκίες για τον αυτοματισμό.

- Τα μοντέλα εμπιστοσύνης στην HRI χωρίζονται σε πέντε κύριες κατηγορίες: μοντέλα παλινδρόμησης, μοντέλα χρονοσειρών, ποιοτικά μοντέλα, πιθανολογικά μοντέλα βασισμένα σε επιχειρήματα, και μοντέλα νευρωνικών δικτύων.

- Οι παράγοντες που επηρεάζουν την εμπιστοσύνη στην HRI περιλαμβάνουν την κατάσταση, τη χρηστικότητα της διεπαφής, τη φυσική παρουσία των ρομπότ, τους περιορισμούς και την πολυπλοκότητα του περιβάλλοντος εργασίας και άλλους.

- Στην εποπτική συνεργασία, οι ρόλοι του επόπτη και του ρομπότ είναι σημαντικοί. Υπάρχουν μοντέλα που εξετάζουν την εμπιστοσύνη στην εποπτική συνεργασία και προσπαθούν να βελτιώσουν την απόδοση της εργασίας.

- Το μοντέλο ”trust-POMDP” συνδυάζει τη μετρούμενη εμπιστοσύνη με τη διαδικασία λήψης αποφάσεων του ρομπότ. Ταυτόχρονα επιτρέπει στο ρομπότ να επηρεάζει την ανθρώπινη εμπιστοσύνη για τη βελτίωση της συνεργασίας.

- Οι μελέτες στη μοντελοποίηση της εμπιστοσύνης συνδυάζουν τεχνικές μέτρησης υποκειμενικής και αντικειμενικής εμπιστοσύνης.Κάποιες μελέτες χρησιμοποιούν ψυχοφυσιολογικές μετρήσεις για τη μοντελοποίηση της εμπιστοσύνης.

3.1.3 Προκλήσεις και προοπτικές στην ανάπτυξη αξιόπιστων ρομπότ

- Ο ρόλος της εμπιστοσύνης στην αλληλεπίδραση ανθρώπου-ρομπότ, ιδίως σε περιβάλλοντα όπου τα ρομπότ έχουν ενσωματωθεί ως βοηθητικοί πράκτορες, είναι ζωτικής σημασίας.

- Προβλήματα σχετικά με την εμπιστοσύνη περιλαμβάνουν την υπερβολική εμπιστοσύνη, την έλλειψη εμπιστοσύνης και την κακή χρήση των ρομπότ λόγω λανθασμένης εμπιστοσύνης.

- Η εμπιστοσύνη διαδραματίζει σημαντικό ρόλο στην ασφάλεια, το απόρρητο και την προστασία της ιδιωτικής ζωής σε περιβάλλοντα με ρομπότ.

- Ο ρόλος της εμπιστοσύνης διακρίνεται από την αξιοπιστία και τη φήμη, με την αξιοπιστία να είναι ιδιότητα του πράκτορα και τη φήμη να συμπεριλαμβάνει τη γνώμη τρίτων.

- **Σχεδιασμός με έμφαση στην ανάπτυξη της εμπιστοσύνης:**

  - Η εμπιστοσύνη σε ένα ρομπότ συχνά βασίζεται στη φυσική του εμφάνιση.
  - Οι πρώτες εντυπώσεις επηρεάζουν την απόφαση εμπιστοσύνης, όπως και στην ανθρώπινη αλλά και ρομποτική αλληλεπίδραση.
  - Τα ρομπότ με ανθρωπόμορφα χαρακτηριστικά αποκτούν αυξημένη αξιοπιστία, αλλά υπάρχει μια οριακή τιμή μετά την οποία η υπερβολική ανθρωπομορφία μειώνει την αξιοπιστία.
  - Ο σχεδιασμός επηρεάζει επίσης τις κρίσιμες καταστάσεις, όπου η αντίδραση του ρομπότ έχει σημαντικό αντίκτυπο.

  - Η ενίσχυση της εμπιστοσύνης μπορεί να επιτευχθεί με συμπληρωματικές πληροφορίες, είτε χειροκίνητα ενσωματώνοντας τις δυνατότητες του ρομπότ εκ των προτέρων, είτε αυτόματα με αλγοριθμικό εντοπισμό κρίσιμων καταστάσεων.

- **Απόκτηση, διαφύλαξη και βαθμονόμηση της εμπιστοσύνης:**

  - Η έννοια των ευρετικών μεθόδων, βασίζεται σε εμπειρικά δεδομένα από την ψυχολογία. Ο σχεδιασμός τέτοιων μεθόδων αποσκοπεί στο να επηρεάσει την αντίληψη του ανθρώπου-χρήστη για το ρομπότ. Προτείνονται δύο προσεγγίσεις: (α) για την αντιμετώπιση της υπερβολικής εμπιστοσύνης και (β) για την αποκατάσταση της εμπιστοσύνης μετά από σφάλματα.
  - Η εμπιστοσύνη δεν είναι στατικό φαινόμενο, αλλά μεταβάλλεται δυναμικά κατά τη διάρκεια της αλληλεπίδρασης. Είναι ανάγκη το ρομπότ να αναπτύξει στρατηγικές που θα επιτρέπουν την προσαρμογή στις μεταβαλλόμενες επιπέδου εμπιστοσύνης κατά τη διάρκεια της αλληλεπίδρασης.
  - Ως στρατηγική αποκατάστασης της εμπιστοσύνης προτείνεται σε περιπτώσεις όπου το ρομπότ δημιουργεί αρνητική εμπειρία, να μπορεί να παράσχει εξηγήσεις για τα σφάλματα ή να προτείνει εναλλακτικές λύσεις. Προτείνεται, επίσης, η ανάπτυξη προσαρμοστικών στρατηγικών που θα λαμβάνουν υπόψη το πλαίσιο και τις ατομικές διαφορές, λόγω των διαφορετικών ατομικών αντιδράσεων των χρηστών.
  - Ερευνητικά αποτελέσματα δείχνουν ότι τα άτομα τείνουν να εμπιστεύονται περισσότερο τα ρομπότ που εκφράζουν τους περιορισμούς τους μέσω κατατοπιστικών κινήσεων των χεριών τους, παρέχουν αιτιολογήσεις για τις ενέργειές τους ή παρέχουν λεπτομερείς πληροφορίες.
  - Η χρήση της φυσικής γλώσσας για την επικοινωνία με τους χρήστες μπορεί να συμβάλει στον μετριασμό της πιθανής απώλειας εμπιστοσύνης που προκύπτει από αποτυχίες απόδοσης.
  - Η εμπιστοσύνη στα ρομπότ μπορεί να επηρεαστεί από την στάση τους απέναντι στον κίνδυνο. Τα ρομπότ που αναλαμβάνουν κινδύνους σε αβέβαια περιβάλλοντα μπορεί να μην θεωρούνται αξιόπιστα, ενώ αυτά που εκφράζουν ευπάθεια ή συναισθήματα μπορούν να θεωρούνται πιο αξιόπιστα.
  - Η εμπιστοσύνη μπορεί να εκτιμηθεί μέσω της παρατήρησης κινήσεων και ενεργειών του ανθρώπου [128] [158] [159].
  - Η εκτιμώμενη εμπιστοσύνη χρησιμοποιείται για την απόφαση του ρομπότ αν θα παραχωρήσει τον έλεγχο στον άνθρωπο [158].
  - Το μοντέλο OPTIMo (Online Probabilistic Trust Inference Model) αναπαριστά την εμπιστοσύνη ως λανθάνουσα μεταβλητή σε ένα δυναμικό πιθανοτικό γραφικό μοντέλο. Το άρθρο [159] χρησιμοποίησε διαφορετική μοντελοποίηση με β-διωνυμική κατανομή για την ομαδοποίηση ανθρώπων βάσει του προφίλ εμπιστοσύνης τους. Επιπλέον, το OPTIMo ενσωματώθηκε σε διάφορες εφαρμογές. Χρησιμοποιήθηκε ως μηχανισμός για την απόφαση εάν το ρομπότ θα πρέπει να παραχωρήσει τον έλεγχο στον άνθρωπο [158] και ενσωματώθηκε στο σχεδιασμό POMDP για την αιτιολόγηση της συμπεριφοράς του ρομπότ [115] [162].

- **Ερευνητικές προκλήσεις:**

  - *Μέτρηση της εμπιστοσύνης σε πραγματικές συνθήκες:* Η εμπιστοσύνη του ανθρώπου σε ένα ρομπότ είναι δύσκολο να μετρηθεί, καθώς αποτελεί ένα μη παρατηρήσιμο φαινόμενο. Προσπάθειες έχουν γίνει για την ανάπτυξη εργαλείων μέτρησης εμπιστοσύνης,

με τη χρήση κλιμάκων αυτο-αναφοράς και αντικειμενικών μέτρων όπως η παρακολούθηση των ματιών, κοινωνικές ενδείξεις από βίντεο/κάμερες, ήχο, δερματική απόκριση και νευρωνικά μέτρα. Παρ' όλα αυτά, υπάρχουν προκλήσεις όπως η αναλλοίωτη μέτρηση, όπου διαφορετικά μέτρα ενδέχεται να μην αντικατοπτρίζουν πραγματικές διαφορές στην εμπιστοσύνη και εκφράζουν γενικές ασυνέπειες μεταξύ τους.

– *Μοντελοποίηση της εμπιστοσύνης:* Η μοντελοποίηση της εμπιστοσύνης είναι μια προκλητική πτυχή, με προσπάθειες να εκτιμηθεί η εμπιστοσύνη μέσω της ανθρώπινης συμπεριφοράς. Επίσης, εξετάζεται πώς η εκτιμώμενη εμπιστοσύνη μπορεί να καθοδηγήσει τη συμπεριφορά του ρομπότ, ειδικά όσον αφορά την παραχώρηση του ελέγχου στον άνθρωπο.

– *Συνδυασμός αντικειμενικών και υποκειμενικών μέτρων:* Ο συνδυασμός των αντικειμενικών και υποκειμενικών μέτρων αποτελεί πρόκληση, καθώς απαιτεί την ανάπτυξη συνολικών πλαισίων για την εκτίμηση της εμπιστοσύνης. Η υποκειμενική εμπειρία του ανθρώπου προς το ρομπότ μπορεί να επηρεάζεται από πολλούς παράγοντες, ενώ τα αντικειμενικά μετρήσιμα στοιχεία παρέχουν στο ρομπότ πληροφορίες για τη συνολική κατάσταση.

– Η έρευνα σχετικά με την ανθρώπινη εμπιστοσύνη στα ρομπότ μπορεί να εξεταστεί σε συνάρτηση με την έρευνα για τα αξιόπιστα συστήματα και την τεχνητή νοημοσύνη.

– Τεχνικές τυπικής επαλήθευσης έχουν αναπτυχθεί για να εξασφαλίσουν την απόδοση και την ικανοποίηση επιθυμητών ιδιοτήτων στα συστήματα.

– Η ανάπτυξη αξιόπιστων ρομπότ προκαλεί ερωτήματα σχετικά με την ασφάλεια και την ιδιωτικότητα, ιδίως όταν τα ρομπότ αλληλεπιδρούν με ευάλωτες ομάδες.

– Το πεδίο της διαφορικής ιδιωτικότητας προσφέρει τεχνικές για την αντιμετώπιση της διαρροής ιδιωτικών πληροφοριών σε ρομποτικά συστήματα.

– Οι προκλήσεις της ιδιωτικότητας μπορούν να αντιμετωπιστούν μέσω τεχνικών από τη διαφορική ιδιωτικότητα και με ηθικά και νομικά πλαίσια για τη ρύθμιση της χρήσης κοινωνικών ρομπότ.

– Για την κατανόηση της δομής της εμπιστοσύνης υποστηρίζεται ότι η ανάλυση αιτιότητας μπορεί να διαδραματίσει καθοριστικό ρόλο στην ανακάλυψη των μηχανισμών που την διέπουν. Αυτή η προσέγγιση επιδιώκει να διαπιστώσει πώς η εμπιστοσύνη επηρεάζει τη συνεργασία ανθρώπου-ρομπότ μέσω μιας λανθάνουσας συνάρτησης εμπιστοσύνης.

– Για την αντιμετώπιση υψηλής διάστασης δεδομένων κατά την μοντελοποίηση της εμπιστοσύνης, όπως βίντεο, ήχος και ψυχοφυσιολογικές μετρήσεις, επισημαίνεται ότι η συνδυασμένη χρήση βαθιών νευρωνικών δικτύων και πιθανοτικών μοντέλων επιτρέπει τη διαχείριση πολυδιάστατων, μη δομημένων δεδομένων σαν αυτά και συνεπώς, την εξαγωγή εμπιστοσύνης.

3.1.4 Στρατηγικές επιδιόρθωσης της εμπιστοσύνης

- Υπάρχουν ομάδες ανθρώπου-ρομπότ που συνδυάζουν τις ανθρώπινες δυνατότητες με τις ικανότητες των ρομπότ για ακρίβεια και ταχύτητα. Αυτές οι ομάδες εμφανίζονται σε διάφορα εργασιακά περιβάλλοντα.

- Η εμπιστοσύνη είναι ζωτικής σημασίας για τις ομάδες ανθρώπου-ρομπότ, αλλά μερικές φορές μειώνεται λόγω παραβιάσεων της εμπιστοσύνης. Οι παραβιάσεις μπορεί να βασίζονται στις ικανότητες, την ακεραιότητα και την καλοσύνη των ρομπότ.

- Οι παραβιάσεις της εμπιστοσύνης χωρίζονται σε τρεις κατηγορίες: παραβιάσεις της ικανότητας, της ακεραιότητας και της καλοπιστίας. Κάθε κατηγορία περιλαμβάνει διαφορετικούς τύπους παραβιάσεων.

- Η αποκατάσταση της εμπιστοσύνης είναι η προσπάθεια να αποκατασταθεί η εμπιστοσύνη μετά από παραβίαση. Οι στρατηγικές αποκατάστασης περιλαμβάνουν απολογίες, αρνήσεις, εξηγήσεις και υποσχέσεις.

- Υπάρχουν αμφιλεγόμενα αποτελέσματα όσον αφορά το αν οι παραπάνω στρατηγικές αποκατάστασης της εμπιστοσύνης είναι αποτελεσματικές ή όχι. Τα αποτελέσματα φαίνεται να επηρεάζονται από παράγοντες όπως ο χρόνος και ο τύπος της παραβίασης της εμπιστοσύνης.

3.1.5 Αλληλεπίδραση ανθρώπου - κοινωνικού ρομπότ

- Με βάση την έρευνα του Eurobarometer το 2012 [236], υπάρχει αρνητική στάση απέναντι στη χρήση ρομπότ σε τομείς όπου παραδοσιακά δραστηριοποιούνται άνθρωποι, όπως η υγειονομική περίθαλψη. Υπάρχει ανησυχία ότι η αυτοματοποίηση που συνδέεται με τη ρομποτική μπορεί να οδηγήσει σε σημαντική απώλεια θέσεων εργασίας.

- Οι στάσεις απέναντι στα ρομπότ διαφέρουν ανάλογα με το περιβάλλον και τον τομέα, και σε ορισμένες περιπτώσεις είναι απομακρυσμένες από την πραγματικότητα.

- Οι στάσεις απέναντι στα ρομπότ επηρεάζουν τις προθέσεις των ανθρώπων όσον αφορά τη χρήση και τη συνεργασία με τα ρομπότ

- **Ορισμός Κοινωνικού Ρομπότ**: Ένα κοινωνικό ρομπότ είναι ένα ενσαρκωμένο σύστημα που αντιλαμβάνεται ως κοινωνική οντότητα, επικοινωνεί με τους ανθρώπους μέσω μιας κοινωνικής διεπαφής και μπορεί να επικοινωνεί λεκτικές ή μη λεκτικές πληροφορίες.

- **Στάσεις απέναντι στα κοινωνικά ρομπότ**: Η στάση των ανθρώπων απέναντι στα κοινωνικά ρομπότ είναι διφορούμενη, καθώς εξαρτάται από το πλαίσιο χρήσης των ρομπότ. Αν και η εργασία δίπλα σε ρομπότ δεν αμφισβητείται, η αντικατάσταση των ανθρώπων σε θέσεις εργασίας που απαιτούν κοινωνικές δεξιότητες είναι αμφιλεγόμενη. Ορισμένες μελέτες υποδεικνύουν θετική στάση απέναντι στα ρομπότ που απαιτούν περισσότερες κοινωνικές δεξιότητες.

- **Άγχος για τα κοινωνικά ρομπότ**: Το άγχος παίζει σημαντικό ρόλο στις αλληλεπιδράσεις με τα κοινωνικά ρομπότ και προβλέπει την πρόθεση χρήσης και την ποιότητα της αλληλεπίδρασης. Ο τρόπος μέτρησης του άγχους συμπεριλαμβάνει αυτοαναφορικές κλίμακες και παρατηρήσεις κατά τη διάρκεια της αλληλεπίδρασης.

- **Εμπιστοσύνη στα κοινωνικά ρομπότ**: Η εμπιστοσύνη αναγνωρίζεται ως σημαντικός παράγοντας στην ποιότητα των αλληλεπιδράσεων ανθρώπου-ρομπότ και στην πρόθεση χρήσης των ρομπότ. Ο αντίκτυπος της εμπιστοσύνης στα κοινωνικά ρομπότ δεν έχει εξεταστεί επαρκώς, αλλά φαίνεται ότι συνδέεται με την ικανοποίηση των ανθρώπων και την θεραπευτική αποτελεσματικότητα, με έμφαση στην υγειονομική περίθαλψη.

- **Αποδοχή των κοινωνικών ρομπότ**:

  – Η αποδοχή ορίζεται ως η πρόθεση χρήσης και, σε ορισμένες περιπτώσεις, ως η πραγματική χρήση των ρομπότ.

  – Οι χαμηλές στάσεις αποδοχής μπορεί να αποδειχθούν επιζήμιες για την ανάπτυξη και τη χρήση της τεχνολογίας των κοινωνικών ρομπότ.

- **Παράγοντες επίδρασης στην ανθρώπινη αντίληψη**:
  - Διάφοροι παράγοντες συσχετίζονται με τη στάση των ανθρώπων απέναντι στα κοινωνικά ρομπότ, την εμπιστοσύνη σε αυτά, την αποδοχή τους και το άγχος τους.
  - Οι πεποιθήσεις των ανθρώπων μπορεί να διαφέρουν ανάλογα με τον τύπο της έκθεσης στα ρομπότ (χωρίς HRI, έμμεση HRI, άμεση HRI).
  - Ο τομέας εφαρμογής (ρομποτική συντροφιά, εκπαιδευτικά ρομπότ, υγειονομική περίθαλψη, παιδιατρική φροντίδα, HRI, γενική εφαρμογή) επηρεάζει επίσης τη στάση των ανθρώπων προς τα ρομπότ.
- Οι σχεδιαστικοί παράγοντες των ρομπότ, όπως ο βαθμός ομοιότητάς τους με τον άνθρωπο, επηρεάζουν τη στάση των ανθρώπων απέναντι σε αυτά. Συγκεκριμένα, διακρίνονται τρεις κατηγορίες σχεδιασμού: ανθρωποειδές, ανθρωπόμορφο, και μη ανθρωποειδές ρομπότ.
- Η γεωγραφική τοποθεσία των χρηστών επίσης μπορεί να επηρεάσει τη στάση τους απέναντι στα ρομπότ. Το πολιτισμικό υπόβαθρο και η εθνικότητα μπορούν να διαδραματίσουν ρόλο στην εμπιστοσύνη, την αποδοχή και τη στάση των ανθρώπων απέναντι σε κοινωνικά ρομπότ.
- Η στάση των ανθρώπων απέναντι στα ρομπότ επίσης μπορεί να εξαρτάται από δημογραφικούς παράγοντες, όπως η ηλικία και το φύλο τους. Γενικά, οι άνδρες και οι νεότεροι ενήλικες τείνουν να έχουν πιο θετική στάση απέναντι στα ρομπότ από τις γυναίκες και τους ηλικιωμένους.
- **Ανθρωπομορφικά χαρακτηριστικά προσώπου του ρομπότ και αξιοπιστία**:
  - Τα κοινωνικά ρομπότ αποτελούν ένα σύστημα τεχνητής νοημοσύνης που είναι σχεδιασμένο να επικοινωνεί και να αλληλεπιδρά κοινωνικά με τους ανθρώπους. Μπορούν να σχεδιαστούν με ανθρωπομορφικά χαρακτηριστικά προσώπου για να βελτιώσουν την επικοινωνία τους με τους ανθρώπους.
  - Η αξιοπιστία παίζει καίριο ρόλο στην αλληλεπίδραση ανθρώπου-ρομπότ. Η αξιοπιστία επηρεάζει την πειθώ και την πρόθεση των ανθρώπων να ακολουθήσουν τις οδηγίες των ρομπότ.
  - Οι παράγοντες που επηρεάζουν την αξιοπιστία των κοινωνικών ρομπότ περιλαμβάνουν τα χαρακτηριστικά και τις επιδόσεις του ρομπότ, αλλά και τις ανθρώπινες παράμετρους όπως η ανάγκη, η επιθυμία για εμπιστοσύνη, η προσωπικότητα, και η επίγνωση της κατάστασης.
  - Τα ανθρωπομορφικά χαρακτηριστικά και οι εκφράσεις στα κοινωνικά ρομπότ μπορούν να επηρεάσουν την αντίληψη της αξιοπιστίας από τους ανθρώπους.
  - Η αντίληψη της αξιοπιστίας των κοινωνικών ρομπότ μπορεί να επηρεαστεί από παράγοντες όπως η συμπεριφορά του ρομπότ, η επάρκεια, η προβλεψιμότητα, και ο ανθρωπομορφισμός.
  - Υπάρχουν τρεις βασικές επιστημονικές προσεγγίσεις σχετικά με την ανθρωπόμορφη αξιοπιστία του προσώπου των κοινωνικών ρομπότ. Αυτές περιλαμβάνουν την εξέταση της ψυχολογικής πτυχής της αξιοπιστίας του προσώπου, την εφαρμογή των αρχών του μάρκετινγκ και του μηχανολογικού σχεδιασμού για τον σχεδιασμό αξιόπιστης εμφάνισης, και την εξέταση της επίδρασης της ανθρωπόμορφης αξιολόγησης στην αξιοπιστία των ρομπότ.
  - Οι περιοχές των ματιών, το σχήμα και το μέγεθος των ματιών, το βλέμμα, τα φρύδια, και η αντίθεση φωτεινότητας αναδεικνύονται ως σημαντικοί παράγοντες που επηρεάζουν την αξιολόγηση της αξιοπιστίας του προσώπου.

- Στοιχεία όπως ο λόγος πλάτους-ύψους προσώπου, ο λόγος φρυδιών-μύτης-πηγουνιού, τα μαλλιά, το μέτωπο, τα αυτιά, τα γένια, το πηγούνι, τα γυαλιά, τα τατουάζ, η ηλικία και η εθνικότητα επίσης επηρεάζουν την αξιοπιστία.
- Είναι σημαντικό να σημειωθεί ότι διάφοροι συνδυασμοί αυτών των χαρακτηριστικών μπορούν να δημιουργήσουν εντυπώσεις όπως ομορφιά, συμμετρία και αρρενωπότητα, οι οποίες επίσης επηρεάζουν την αξιοπιστία.
- Η κίνηση των χαρακτηριστικών του προσώπου και οι συναισθηματικές εκφράσεις επηρεάζουν την αντιληπτή αξιοπιστία, καθώς μπορούν να κινητοποιήσουν τους ανθρώπους να αντιληφθούν τα συναισθήματα που προκαλούν.
- Οι συνδυασμοί χαρακτηριστικών, όπως η ομοιότητα με τον αξιολογητή, το αν ανήκουν στην ίδια πολιτιστική ομάδα και η συμμετρία του προσώπου, επίσης επηρεάζουν την αξιοπιστία του ατόμου.

Κεφάλαιο 4

Συζήτηση – Συμπεράσματα – Μελλοντικές επεκτάσεις

4.1 Ανακεφαλαίωση

Μέσω της παρούσας διπλωματικής εργασίας, επιτυγχάνεται η ενδελεχής αναφορά ποικίλων ερευνητικών προσπαθειών ανασκόπησης της συνεργασίας ανθρώπου-ρομπότ και την εμπιστοσύνη που ορίζει την σχέση τους. Συγκεκριμένα, έπειτα από έρευνα, αναφέρθηκαν τα εξής: Η εμπιστοσύνη στην HRI και οι διαφορές μεταξύ αυτής και της διαπροσωπικής εμπιστοσύνης, παράγοντες που επηρεάζουν την διαμόρφωση της εμπιστοσύνης χωρισμένους σε τρεις μεγάλες κατηγορίες (Εμπιστοσύνη προδιάθεσης, περιστασιακή εμπιστοσύνη και επίκτητη), διάφορα μοντέλα και πλαίσια εμπιστοσύνης αλλά και προκλήσεις στην ανάπτυξη αξιόπιστων ρομπότ. Επίσης, συζητήθηκαν μέθοδοι για τον σχεδιασμό τον ρομπότ, για την απόκτηση, διαφύλαξη και βαθμονόμηση της εμπιστοσύνης αλλά και πιθανές ερευνητικές προκλήσεις. Έπειτα, μετά από τις διάφορες στρατηγικές επιδιόρθωσης της εμπιστοσύνης, αναπτύχθηκε εις βάθος η αλληλεπίδραση ανθρώπου και κοινωνικού ρομπότ, με έμφαση στις ανθρώπινες προδιαθέσεις και τις διάφορες μεταβλητές που επηρεάζουν την ανθρώπινη αντίληψη καταλήγοντας στα ανθρωπόμορφα χαρακτηριστικά ενός ρομπότ που μπορεί να ωθήσουν τον ανθρώπινο χειριστή στην ανάπτυξη θετικών συναισθημάτων.

Ερευνητικά η εργασία ακολούθησε τον εξής τρόπο συλλογής επιστημονικών άρθρων:

1. Εντοπισμός της βασικής θεματολογίας, που είναι η αλληλεπίδραση ανθρώπου-μηχανής με έμφαση στην εμπιστοσύνη του χειριστή.

2. Αναζήτηση με λέξεις κλειδιά σε αντίστοιχες μηχανές αναζήτησης: ”human-robot interaction”, ”HRI”, ”trust in HRI”, ”operator’s trust”, ”robot trust”, ”human-automation interaction”, ”trustworthy robots” και τα λοιπά.

3. Σχετικά ερευνητικά άρθρα ανακαλήφθηκαν και από παρεμφερή προσπάθειες.

4. Μελέτη αρχικά της περίληψης κάθε άρθρου και έπειτα το κεντρικό τους θέμα με περισσότερες λεπτομέρειες.

5. Και τέλος την ένταξη τους στο αντίστοιχο κεφάλαιο, μέσα από το πλάνο οργάνωσης των κεφαλαίων και σημειώσεις κάθε άρθρου.

4.2 Μελλοντικές επεκτάσεις / Πρακτικές Προεκτάσεις της Έρευνας

Κατά την έρευνα τα προβλήματα που αντιμετώπιστηκαν ήταν κατά κύριο λόγο τα ακόλουθα:

- Η ύπαρξη μεγάλου όγκου πληροφορίας καθιστά δύσκολη την συλλογή της και συνεπώς, την ανάπτυξη της στα πλαίσια μιας διπλωματικής εργασίας.
- Πρόκειται, λοιπόν για ένα θέμα το οποίο, πέραν της απαιτούμενης έντονης νοητικής προσπάθειας για την συγκέντρωση της πληροφορίας, χρειάζεται και την ανάλογη χρονική δέσμευση.
- Ένα επιπλέον, πρόβλημα θα μπορούσε επίσης να είναι η έλλειψη επιστημονικών άρθρων γύρω από μοντέλα και πλαίσια εμπιστοσύνης αλλά και στρατηγικές επιδιόρθωσης αυτής.

Με βάση αυτά, είναι σημαντικό να αναφερθεί πως πρόκειται για έναν τομέα που παραλληλίζει με ένα χρυσωρυχείο και εύκολα κάποιος με τον ανάλογο ζήλο, μπορεί να βγάλει υποκειμενικά συμπεράσματα βασιζόμενος σε αυτόν ή ακόμα και καινοτόμα αποτελέσματα με τον αντίστοιχο αντίκτυπο στην επιστημονική κοινότητα.

Βιβλιογραφικές Αναφορές

[1] J. D. Lee και K. A. See. «Trust in Automation: Designing for Appropriate Reliance». Στο: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1 Ιαν. 2004), σσ. 50–80. ISSN: 0018-7208. DOI: 10.1518/hfes.46.1.50_30392. URL: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392.

[2] Nikolas Martelaro, Victoria C Nneji, Wendy Ju και Pamela Hinds. «Tell me more designing hri to encourage more trust, disclosure, and companionship». Στο: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, σσ. 181–188.

[3] Kristin E Schaefer, Jessie YC Chen, James L Szalma και Peter A Hancock. «A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems». Στο: *Human factors* 58.3 (2016), σσ. 377–400.

[4] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard και Alan R Wagner. «Overtrust of robots in emergency evacuation scenarios». Στο: *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE. 2016, σσ. 101–108.

[5] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser και Raja Parasuraman. «A meta-analysis of factors affecting trust in human-robot interaction». Στο: *Human factors* 53.5 (2011), σσ. 517–527.

[6] Munjal Desai. «Modeling trust to improve human-robot interaction». Διδακτορική διατρ. University of Massachusetts Lowell, 2012.

[7] Behzad Sadrfaridpour, Hamed Saeidi, Jenny Burke, Kapil Madathil και Yue Wang. «Modeling and control of trust in human-robot collaborative manufacturing». Στο: *Robust intelligence and trust in autonomous systems* (2016), σσ. 115–141.

[8] Victoria Groom και Clifford Nass. «Can robots be teammates?: Benchmarks in human–robot teams». Στο: *Interaction studies* 8.3 (2007), σσ. 483–500.

[9] Anthony Selkowitz, Shan Lakhmani, Jessie YC Chen και Michael Boyce. «The effects of agent transparency on human interaction with an autonomous robotic agent». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 59. 1. SAGE Publications Sage CA: Los Angeles, CA. 2015, σσ. 806–810.

[10] Raja Parasuraman και Victor Riley. «Humans and automation: Use, misuse, disuse, abuse». Στο: *Human factors* 39.2 (1997), σσ. 230–253.

[11] Roger C Mayer, James H Davis και F David Schoorman. «An integrative model of organizational trust». Στο: *Academy of management review* 20.3 (1995), σσ. 709–734.

[12] Cynthia L Corritore, Beverly Kracher και Susan Wiedenbeck. «On-line trust: concepts, evolving themes, a model». Στο: *International journal of human-computer studies* 58.6 (2003), σσ. 737–758.

[13] David Gefen, Elena Karahanna και Detmar W Straub. «Trust and TAM in online shopping: An integrated model». Στο: *MIS quarterly* (2003), σσ. 51–90.

[14] Robert R Hoffman, John D Lee, David D Woods, Nigel Shadbolt, Janet Miller και Jeffrey M Bradshaw. «The dynamics of trust in cyberdomains». Στο: *IEEE Intelligent Systems* 24.6 (2009), σσ. 5–11.

[15] Russell Hardin. *Trust*. Τόμ. 10. Polity, 2006.

[16] Julian B Rotter. «A new scale for the measurement of interpersonal trust.» Στο: *Journal of personality* (1967).

[17] Bernard Barber. «The logic and limits of trust». Στο: (1983).

[18] Dean G Pruit και Jeffrey Z Rubin. «Social conflict: escalation, stalemate and settlement». Στο: *A textbook of social Psychology* (1986), σσ. 268–269.

[19] Morton Deutsch. «The effect of motivational orientation upon trust and suspicion». Στο: *Human relations* 13.2 (1960), σσ. 123–139.

[20] Clifford Nass, Jonathan Steuer και Ellen R Tauber. «Computers are social actors». Στο: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1994, σσ. 72–78.

[21] Angelika Dimoka. «What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study». Στο: *Mis Quarterly* (2010), σσ. 373–396.

[22] Kevin Anthony Hoff και Masooda Bashir. «Trust in Automation». Στο: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (3 Μάι. 2015), σσ. 407–434. ISSN: 0018-7208. DOI: 10.1177/0018720814547570. URL: http://journals.sagepub.com/doi/10.1177/0018720814547570.

[23] Poornima Madhavan και Douglas A Wiegmann. «Similarities and differences between human–human and human–automation trust: an integrative review». Στο: *Theoretical Issues in Ergonomics Science* 8.4 (2007), σσ. 277–301.

[24] John Lee και Neville Moray. «Trust, control strategies and allocation of function in human-machine systems». Στο: *Ergonomics* 35.10 (1992), σσ. 1243–1270.

[25] John K Rempel, John G Holmes και Mark P Zanna. «Trust in close relationships.» Στο: *Journal of personality and social psychology* 49.1 (1985), σ. 95.

[26] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce και Hall P Beck. «The role of trust in automation reliance». Στο: *International journal of human-computer studies* 58.6 (2003), σσ. 697–718.

[27] Stephen Marsh και Mark R Dibben. «The role of trust in information science and technology». Στο: *Annual Review of Information Science and Technology (ARIST)* 37 (2003), σσ. 465–98.

[28] René Riedl και Andrija Javor. «The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging.» Στο: *Journal of Neuroscience, Psychology, and Economics* 5.2 (2012), σ. 63.

[29] Michael Naef, Ernst Fehr, Urs Fischbacher, Jürgen Schupp και Gert G Wagner. «Decomposing trust: Explaining national trust differences». Στο: *International Journal of Psychology*. Τόμ. 43. 3-4. PSYCHOLOGY PRESS 27 CHURCH RD, HOVE BN3 2FA, EAST SUSSEX, ENGLAND. 2008, σσ. 577–577.

[30] Esperanza Huerta, TerryAnn Glandon και Yanira Petrides. «Framing, decision-aid systems, and culture: Exploring influences on fraud investigations». Στο: *International Journal of Accounting Information Systems* 13.4 (2012), σσ. 316–333.

[31] Dingjun Li, PL Patrick Rau και Ye Li. «A cross-cultural study: Effect of robot appearance and task». Στο: *International Journal of Social Robotics* 2 (2010), σσ. 175–186.

[32] Geoffrey Ho, Liana Maria Kiff, Tom Plocher και Karen Zita Haigh. «A Model of Trust and Reliance of Automation Technology for Older Users.» Στο: *AAAI Fall Symposium: Caring Machines*. 2005, σσ. 45–50.

[33] Geoffrey Ho, Dana Wheatley και Charles T Scialfa. «Age differences in trust and reliance of a medication management system». Στο: *Interacting with Computers* 17.6 (2005), σσ. 690–710.

[34] Julian Sanchez, Arthur D Fisk και Wendy A Rogers. «Reliability and age-related effects on trust and reliance of a decision support aid». Στο: *proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 48. 3. Sage Publications Sage CA: Los Angeles, CA. 2004, σσ. 586–589.

[35] Richard Pak, Nicole Fink, Margaux Price, Brock Bass και Lindsay Sturre. «Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults». Στο: *Ergonomics* 55.9 (2012), σσ. 1059–1072.

[36] Neta Ezer, Arthur D Fisk και Wendy A Rogers. «Reliance on automation as a function of expectation of reliability, cost of verification, and age». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 51. 1. SAGE Publications Sage CA: Los Angeles, CA. 2007, σσ. 6–10.

[37] Neta Ezer, Arthur D Fisk και Wendy A Rogers. «Age-related differences in reliance behavior attributable to costs within a human-decision aid system». Στο: *Human factors* 50.6 (2008), σσ. 853–863.

[38] Sara E McBride, Wendy A Rogers και Arthur D Fisk. «Do younger and older adults differentially depend on an automated system?» Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 54. 2. SAGE Publications Sage CA: Los Angeles, CA. 2010, σσ. 175–179.

[39] Sara E McBride, Wendy A Rogers και Arthur D Fisk. «Understanding the effect of workload on automation use for younger and older adults». Στο: *Human factors* 53.6 (2011), σσ. 672–686.

[40] Eun-Ju Lee. «Flattery may get computers somewhere, sometimes: The moderating role of output modality, computer gender, and user gender». Στο: *International Journal of Human-Computer Studies* 66.11 (2008), σσ. 789–800.

[41] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki και Kensuke Kato. «Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots». Στο: *IEEE transactions on robotics* 24.2 (2008), σσ. 442–451.

[42] Jason A Colquitt, Brent A Scott και Jeffery A LePine. «Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance.» Στο: *Journal of applied psychology* 92.4 (2007), σ. 909.

[43] David P Biros, Gregory Fields και Gregg Gunsch. «The effect of external safeguards on human-information system trust in an information warfare environment». Στο: *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE. 2003, 10–pp.

[44] Stephanie M Merritt και Daniel R Ilgen. «Not all trust is created equal: Dispositional and history-based trust in human-automation interactions». Στο: *Human factors* 50.2 (2008), σσ. 194–210.

[45] James L Szalma και Grant S Taylor. «Individual differences in response to automation: the five factor model of personality.» Στο: *Journal of experimental psychology: Applied* 17.2 (2011), σ. 71.

[46] Maranda McBride, Lemuria Carter και Celestine Ntuen. «The impact of personality on nurses' bias towards automated decision aid acceptance». Στο: *International Journal of Information Systems and Change Management* 6.2 (2012), σσ. 132–146.

[47] Nathan R Bailey και Mark W Scerbo. «Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust». Στο: *Theoretical Issues in Ergonomics Science* 8.4 (2007), σσ. 321–348.

[48] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater και Mica R Endsley. «The influence of agent reliability on trust in human-agent collaboration». Στο: *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*. 2008, σσ. 1–8.

[49] Poornima Madhavan, Douglas A Wiegmann και Frank C Lacson. «Automation failures on tasks easily performed by operators undermine trust in automated aids». Στο: *Human factors* 48.2 (2006), σσ. 241–256.

[50] Alison Parkes. «Persuasive decision support: Improving reliance on decision aids». Στο: *Pacific Asia Journal of the Association for Information Systems* 4.3 (2012), σ. 2.

[51] Jennifer Marie Ross. *Moderators of trust and reliance across multiple decision aids*. university of central florida, 2008.

[52] Randall D Spain. «The effects of automation expertise, system confidence, and image quality on trust, compliance, and performance». Στο: (2009).

[53] Mark A Daly. *Task load and automation use in an uncertain environment*. Αδημοσίευτη ερευνητική εργασία. AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGINEERING AND …, 2002.

[54] David P Biros, Mark Daly και Gregg Gunsch. «The influence of task load and automation trust on deception detection». Στο: *Group Decision and Negotiation* 13.2 (2004), σσ. 173–189.

[55] Jacob M Wetzel. *Driver trust, annoyance, and compliance for an automated calendar system*. Central Michigan University, 2005.

[56] Rachel R Phillips και Poornima Madhavan. «The effect of distractor modality and processing code on human-automation interaction». Στο: *Cognition, Technology & Work* 13 (2011), σσ. 233–244.

[57] Monica N Lees και John D Lee. «The influence of distraction and driving context on driver response to imperfect collision warning systems». Στο: *Ergonomics* 50.8 (2007), σσ. 1264–1286.

[58] LeeAnn Perkins, Janet E Miller, Ali Hashemi και Gary Burns. «Designing for human-centered systems: Situational risk as a factor of trust in automation». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 54. 25. SAGE Publications Sage CA: Los Angeles, CA. 2010, σσ. 2130–2134.

[59] Joseph B Lyons και Charlene K Stokes. «Human–human reliance in the context of automation». Στο: *Human factors* 54.1 (2012), σσ. 112–121.

[60] Peter De Vries, Cees Midden και Don Bouwhuis. «The effects of errors on system trust, self-confidence, and the allocation of control in route planning». Στο: *International Journal of Human-Computer Studies* 58.6 (2003), σσ. 719–735.

[61] Mark Dishaw, Diane Strong και D Brent Bandy. «Extending the task-technology fit model with self-efficacy constructs». Στο: *AMCIS 2002 proceedings* (2002), σ. 143.

[62] Poornima Madhavan και Rachel R Phillips. «Effects of computer self-efficacy and system reliability on user interaction with decision support systems». Στο: *Computers in Human Behavior* 26.2 (2010), σσ. 199–204.

[63] Julian Sanchez, Wendy A Rogers, Arthur D Fisk και Ericka Rovira. «Understanding reliance on automation: effects of error type, error distribution, age and experience». Στο: *Theoretical issues in ergonomics science* 15.2 (2014), σσ. 134–160.

[64] Charlene K Stokes, Joseph B Lyons, Kenneth Littlejohn, Joseph Natarian, Ellen Case και Nicholas Speranza. «Accounting for the human in cyberspace: Effects of mood on trust in automation». Στο: *2010 International Symposium on Collaborative Technologies and Systems*. IEEE. 2010, σσ. 180–187.

[65] Stephanie M Merritt. «Affective processes in human–automation interactions». Στο: *Human Factors* 53.4 (2011), σσ. 356–370.

[66] Linda Onnasch, Christopher D Wickens, Huiyang Li και Dietrich Manzey. «Human performance consequences of stages and levels of automation: An integrated meta-analysis». Στο: *Human factors* 56.3 (2014), σσ. 476–488.

[67] Poornima Madhavan και Douglas A Wiegmann. «Effects of Information Source, Pedigree, and Reliability on Operators' Utilizaton of Diagnostic Advice». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 49. 3. SAGE Publications Sage CA: Los Angeles, CA. 2005, σσ. 487–491.

[68] Genya Abe και John Richardson. «Alarm timing, trust and driver expectation for forward collision warning systems». Στο: *Applied ergonomics* 37.5 (2006), σσ. 577–586.

[69] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell και Deborah Lee. «I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system». Στο: *Human factors* 55.3 (2013), σσ. 520–534.

[70] Nirit Yuviler-Gavish και Daniel Gopher. «Effect of descriptive information and experience on automation reliance». Στο: *Human factors* 53.3 (2011), σσ. 230–244.

[71] Dietrich Manzey, Juliane Reichenbach και Linda Onnasch. «Human performance consequences of automated decision aids: The impact of degree of automation and system experience». Στο: *Journal of Cognitive Engineering and Decision Making* 6.1 (2012), σσ. 57–87.

[72] Mary L Cummings, Andrew Clare και Christin Hart. «The role of human-automation consensus in multiple unmanned vehicle scheduling». Στο: *Human Factors* 52.1 (2010), σσ. 17–27.

[73] Marvin S Cohen, Jared T Freeman και Bryan Thompson. «Critical thinking skills in tactical decision making: A model and a training strategy.» Στο: (1998).

[74] Arnaud Koustanaï, Viola Cavallo, Patricia Delhomme και Arnaud Mas. «Simulator training with a forward collision warning system: Effects on driver-system interactions and driver trust». Στο: *Human factors* 54.5 (2012), σσ. 709–721.

[75] Dietrich Manzey, J Elin Bahner και Anke-Dorothea Hueper. «Misuse of automated aids in process control: Complacency, automation bias and possible training interventions». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 50. 3. Sage Publications Sage CA: Los Angeles, CA. 2006, σσ. 220–224.

[76] Stavros Antifakos, Nicky Kern, Bernt Schiele και Adrian Schwaninger. «Towards improving trust in context-aware systems by displaying system confidence». Στο: *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. 2005, σσ. 9–14.

[77] Robin L Wakefield, Morris H Stocks και W Mark Wilder. «The role of web site characteristics in initial trust formation». Στο: *Journal of Computer Information Systems* 45.1 (2004), σσ. 94–103.

[78] Alona Weinstock, Tal Oron-Gilad και Yisrael Parmet. «The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system». Στο: *Work* 41.Supplement 1 (2012), σσ. 258–265.

[79] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk και Raja Parasuraman. «The world is not enough: Trust in cognitive agents». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 56. 1. Sage Publications Sage CA: Los Angeles, CA. 2012, σσ. 263–267.

[80] Li Gong. «How social is social responses to computers? The function of the degree of anthropomorphism in computer representations». Στο: *Computers in Human Behavior* 24.4 (2008), σσ. 1494–1509.

[81] Brian Daniel Green. *Applying human characteristics of trust to animated anthropomorphic software agents*. State University of New York at Buffalo, 2010.

[82] Raja Parasuraman και Christopher A Miller. «Trust and etiquette in high-criticality automated systems». Στο: *Communications of the ACM* 47.4 (2004), σσ. 51–55.

[83] Randall D Spain και Poornima Madhavan. «The role of automation etiquette and pedigree in trust and dependence». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 53. 4. SAGE Publications Sage CA: Los Angeles, CA. 2009, σσ. 339–343.

[84] Younho Seong και Ann M Bisantz. «The impact of cognitive feedback on judgment performance and trust with decision aids». Στο: *International Journal of Industrial Ergonomics* 38.7-8 (2008), σσ. 608–625.

[85] Ji Gao και John D Lee. «Effect of shared information on trust and reliance in a demand forecasting task». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 50. 3. SAGE Publications Sage CA: Los Angeles, CA. 2006, σσ. 215–219.

[86] Greg A Jamieson, Lu Wang και Heather F Neyedli. *Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems (Developpement d'Interfaces Homme-Machine Pour Appuyer la Confiance dans les Systemes Automatises d'Identification au Combat)*. Αδημοσίευτη ερευνητική εργασία. TORONTO UNIV (ONTARIO) COGNITIVE ENGINEERING LAB, 2008.

[87] Lu Wang, Greg A Jamieson και Justin G Hollands. «Trust and reliance on an automated combat identification system». Στο: *Human factors* 51.3 (2009), σσ. 281–291.

[88] Mary Dzindolet, Linda Pierce, Scott Peterson, Lori Purcell, Hall Beck και Hall Beck. «The influence of feedback on automation use, misuse, and disuse». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 46. 3. SAGE Publications Sage CA: Los Angeles, CA. 2002, σσ. 551–555.

[89] Frank C Lacson, Douglas A Wiegmann και Poornima Madhavan. «Effects of attribute and goal framing on automation reliance and compliance». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 49. 3. SAGE Publications Sage CA: Los Angeles, CA. 2005, σσ. 482–486.

[90] Ewart de Visser και Raja Parasuraman. «Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload». Στο: *Journal of Cognitive Engineering and Decision Making* 5.2 (2011), σσ. 209–231.

[91] Frank MF Verberne, Jaap Ham και Cees JH Midden. «Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars». Στο: *Human factors* 54.5 (2012), σσ. 799–810.

[92] Julian Sanchez. *Factors that affect trust and reliance on an automated aid*. Georgia Institute of Technology, 2006.

[93] Randy B Davenport και Ernesto A Bustamante. «Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 54. 19. SAGE Publications Sage CA: Los Angeles, CA. 2010, σσ. 1513–1517.

[94] Stephen R Dixon, Christopher D Wickens και Jason S McCarley. «On the independence of compliance and reliance: Are automation false alarms worse than misses?» Στο: *Human factors* 49.4 (2007), σσ. 564–572.

[95] Jason D Johnson, Julian Sanchez, Arthur D Fisk και Wendy A Rogers. «Type of automation failure: The effects on trust and reliance in automation». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 48. 18. SAGE Publications Sage CA: Los Angeles, CA. 2004, σσ. 2163–2167.

[96] Stephen Rice και David Keller. «Automation reliance under time pressure.» Στο: *Cognitive Technology* (2009).

[97] David Cameron, J Aitken, E Collins, Luke Boorman, Adriel Chua, Samuel Fernando, Owen McAree, Uriel Martinez Hernandez και James Law. «Framing factors: The importance of context and the individual in understanding trust in human-robot interaction». Στο: (2015).

[98] Kristin E Oleson, Deborah R Billings, Vivien Kocsis, Jessie YC Chen και Peter A Hancock. «Antecedents of trust in human-robot collaborations». Στο: *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE. 2011, σσ. 175–178.

[99] Anqi Xu και Gregory Dudek. «Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations». Στο: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 2015, σσ. 221–228.

[100] Alan R Wagner και Ronald C Arkin. «Recognizing situations that demand trust». Στο: *2011 RO-MAN*. IEEE. 2011, σσ. 7–14.

[101] Anqi Xu και Gregory Dudek. «Trust-driven interactive visual navigation for autonomous robots». Στο: *2012 IEEE International Conference on Robotics and Automation*. IEEE. 2012, σσ. 3922–3929.

[102] Rosemarie E. Yagoda και Douglas J. Gillan. «You Want Me to Trust a ROBOT? The Development of a Human–Robot Interaction Trust Scale». Στο: *International Journal of Social Robotics* 4 (3 Αύγ. 2012), σσ. 235–248. ISSN: 1875-4791. DOI: 10.1007/s12369-012-0144-0. URL: http://link.springer.com/10.1007/s12369-012-0144-0.

[103] Fei Gao, Andrew S Clare, Jamie C Macbeth και Missy L Cummings. «Modeling the impact of operator trust on performance in multiple robot control». Στο: AAAI. 2013.

[104] Bonnie M Muir. «Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems». Στο: *Ergonomics* 37.11 (1994), σσ. 1905–1922.

[105] Simon Farrell και Stephan Lewandowsky. «A connectionist model of complacency and adaptive recovery under automation.» Στο: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26.2 (2000), σ. 395.

[106] Neville Moray και Toshiyuki Inagaki. «Laboratory studies of trust between humans and machines in automated systems». Στο: *Transactions of the Institute of Measurement and Control* 21.4-5 (1999), σσ. 203–211.

[107] John D Lee και Neville Moray. «Trust, self-confidence, and operators' adaptation to automation». Στο: *International journal of human-computer studies* 40.1 (1994), σσ. 153–184.

[108] Victor Riley. «Operator reliance on automation: Theory and data». Στο: *Automation and human performance: Theory and applications*. CRC Press, 2018, σσ. 19–35.

[109] Marvin S Cohen, Raja Parasuraman και JT Freeman. «Trust in decision aids: A model and its training implications». Στο: *Proceedings of the 1998 Command and Control Research and Technology Symposium*. CCRP Washington, DC. 1998, σσ. 1–37.

[110] Munjal Desai, Kristen Stubbs, Aaron Steinfeld και Holly Yanco. «Creating trustworthy robots: Lessons and inspirations from automated systems». Στο: (2009).

[111] Anqi Xu και Gregory Dudek. «Maintaining efficient collaboration with trust-seeking robots». Στο: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, σσ. 3312–3319.

[112] Charles Pippin και Henrik Christensen. «Trust modeling in multi-robot patrolling». Στο: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, σσ. 59–66.

[113] Amos Freedy, Ewart DeVisser, Gershon Weltman και Nicole Coeyman. «Measurement of trust in human-robot collaboration». Στο: *2007 International symposium on collaborative technologies and systems*. Ieee. 2007, σσ. 106–114.

[114] Anqi Xu και Gregory Dudek. «Towards modeling real-time trust in asymmetric human–robot collaborations». Στο: *Robotics Research: The 16th International Symposium ISRR*. Springer. 2016, σσ. 113–129.

[115] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu και Siddhartha Srinivasa. «Planning with trust for human-robot collaboration». Στο: *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 2018, σσ. 307–315.

[116] Siddharth Gulati, Sonia Sousa και David Lamas. «Modelling trust: An empirical assessment». Στο: *Human-Computer Interaction–INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV 16*. Springer. 2017, σσ. 40–61.

[117] Siddharth Gulati, Sonia Sousa και David Lamas. «Design, development and evaluation of a human-computer trust scale». Στο: *Behaviour & Information Technology* 38.10 (2019), σσ. 1004–1015.

[118] Halimahtun M Khalid, Liew Wei Shiung, Parham Nooralishahi, Zeeshan Rasool, Martin G Helander, Loo Chu Kiong και Chin Ai-vyrn. «Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 60. 1. SAGE Publications Sage CA: Los Angeles, CA. 2016, σσ. 697–701.

[119] Allison Langer, Ronit Feingold-Polak, Oliver Mueller, Philipp Kellmeyer και Shelly Levy-Tzedek. «Trust in socially assistive robots: Considerations for use in rehabilitation». Στο: *Neuroscience & Biobehavioral Reviews* 104 (2019), σσ. 231–239.

[120] ”Statt N.” «”Boston dynamics’spot robot is helping hospitals remotely treat coronavirus patients”». Στο: *”The Verge”* (2020). URL: https://www.theverge.com/2020/4/23/21231855/boston-dynamics-spot-robot-covid-19-coronavirus-telemedicine.

[121] B Scassellati, J Kennedy, T Belpaeme, A Ramachandran και F Tanaka. «Social robots for education: A review». Στο: *Sci. Robot* 3.21 (2018).

[122] Robinette Paul, Li Wenchen και Allen Robert. «Howard Ayanna M., and Wagner Alan R.. 2016». Στο: *Overtrust of robots in emergency evacuation scenarios. In Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction, IEEE, Christchurch, New Zealand*. 2016, σσ. 101–108.

[123] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos και Radhika Nagpal. «Piggybacking robots: Human-robot overtrust in university dormitory security». Στο: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 2017, σσ. 426–434.

[124] Anthony L Baker, Elizabeth K Phillips, Daniel Ullman και Joseph R Keebler. «Toward an understanding of trust repair in human-robot interaction: Current research and future directions». Στο: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), σσ. 1–30.

[125] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon και Myrthe L Tielman. «Taxonomy of trust-relevant failures and mitigation strategies». Στο: *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*. 2020, σσ. 3–12.

[126] Chandrayee Basu και Mukesh Singhal. «Trust dynamics in human autonomous vehicle interaction: a review of trust models». Στο: *2016 AAAI spring symposium series*. 2016.

[127] Matthew Brzowski και Dan Nathan-Roberts. «Trust measurement in human–automation interaction: A systematic review». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 63. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, σσ. 1595–1599.

[128] Bin Liu. «A survey on trust modeling from a Bayesian perspective». Στο: *Wireless Personal Communications* 112.2 (2020), σσ. 1205–1227.

[129] Christiano Castelfranchi και Rino Falcone. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons, 2010.

[130] WT Luke Teacy, Jigar Patel, Nicholas R Jennings και Michael Luck. «Travos: Trust and reputation in the context of inaccurate information sources». Στο: *Autonomous Agents and Multi-Agent Systems* 12 (2006), σσ. 183–198.

[131] Mark Coeckelbergh. «Can we trust robots?» Στο: *Ethics and information technology* 14 (2012), σσ. 53–60.

[132] Roger C. Mayer, James H. Davis και F. David Schoorman. «An Integrative Model of Organizational Trust». Στο: *The Academy of Management Review* 20.3 (1995), σσ. 709–734. ISSN: 03637425. URL: http://www.jstor.org/stable/258792 (επίσκεψη 16/07/2023).

[133] Carlos RB Azevedo, Klaus Raizer και Ricardo Souza. «A vision for human-machine mutual understanding, trust establishment, and collaboration». Στο: *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE. 2017, σσ. 1–3.

[134] Kazuo Okamura και Seiji Yamada. «Adaptive trust calibration for human-AI collaboration». Στο: *Plos one* 15.2 (2020), e0229132.

[135] Michael Lewis, Katia Sycara και Phillip Walker. «The role of trust in human-robot interaction». Στο: *Foundations of trusted autonomy* (2018), σσ. 135–159.

[136] Michael Yu, Muniba Saleem και Cleotilde Gonzalez. «Developing trust: First impressions and experience». Στο: *Journal of Economic Psychology* 43 (2014), σσ. 16–29.

[137] Manisha Natarajan και Matthew Gombolay. «Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction». Στο: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, 2020, σσ. 33–42. ISBN: 9781450367462. DOI: 10.1145/3319502.3374839. URL: https://doi.org/10.1145/3319502.3374839.

[138] Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck και Hiroshi Ishiguro. «Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy». Στο: *Paladyn, Journal of Behavioral Robotics* 7.1 (2016), σ. 000010151520160005.

[139] Maya B. Mathur και David B. Reichling. «Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley». Στο: *Cognition* 146 (2016), σσ. 22–32. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2015.09.008. URL: https://www.sciencedirect.com/science/article/pii/S0010027715300640.

[140] Minae Kwon, Malte F Jung και Ross A Knepper. «Human expectations of social robots». Στο: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, σσ. 463–464.

[141] Auriel Washburn, Akanimoh Adeleye, Thomas An και Laurel D Riek. «Robot errors in proximate HRI: how functionality framing affects perceived reliability and trust». Στο: *ACM Transactions on Human-Robot Interaction (THRI)* 9.3 (2020), σσ. 1–21.

[142] Sandy H Huang, Kush Bhatia, Pieter Abbeel και Anca D Dragan. «Establishing appropriate trust via critical states». Στο: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, σσ. 3929–3936.

[143] De'Aira Bryant, Jason Borenstein και Ayanna Howard. «Why should we gender? The effect of robot gendering and occupational stereotypes on human trust and perceived competency». Στο: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 2020, σσ. 13–21.

[144] Jasmin Bernotat, Friederike Eyssel και Janik Sachse. «The (fe) male robot: how robot body shape impacts first impressions and trust towards robots». Στο: *International Journal of Social Robotics* 13 (2021), σσ. 477–489.

[145] Maartje Ma De Graaf, Somaya Ben Allouch και Tineke Klamer. «Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot». Στο: *Computers in human behavior* 43 (2015), σσ. 1–14.

[146] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld και Holly Yanco. «Impact of robot failures and feedback on real-time trust». Στο: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, σσ. 251–258.

[147] Nicole Salomons, Michael Van Der Linden, Sarah Strohkorb Sebo και Brian Scassellati. «Humans conform to robots: Disambiguating trust, truth, and conformity». Στο: *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*. 2018, σσ. 187–195.

[148] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian και Kerstin Dautenhahn. «Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust». Στο: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, σσ. 141–148.

[149] Sarah Strohkorb Sebo, Priyanka Krishnamurthi και Brian Scassellati. «"I don't believe you": Investigating the effects of robot trust violation and repair». Στο: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, σσ. 57–65.

[150] Minae Kwon, Sandy H Huang και Anca D Dragan. «Expressing robot incapability». Στο: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, σσ. 87–95.

[151] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li και Julie A Shah. «Evaluating effects of user experience and system transparency on trust in automation». Στο: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 2017, σσ. 408–416.

[152] Ning Wang, David V Pynadath και Susan G Hill. «Trust calibration within a human-robot team: Comparing automatically generated explanations». Στο: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, σσ. 109–116.

[153] Aya Hussein, Sondoss Elsawah και Hussein A Abbass. «The reliability and transparency bases of trust in human-swarm interaction: principles and implications». Στο: *Ergonomics* 63.9 (2020), σσ. 1116–1132.

[154] Tom Bridgwater, Manuel Giuliani, Anouk van Maris, Greg Baker, Alan Winfield και Tony Pipe. «Examining profiles for robotic risk assessment: Does a robot's approach to risk affect user trust?» Στο: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, σσ. 23–31.

[155] Yaqi Xie, Indu P Bodala, Desmond C Ong, David Hsu και Harold Soh. «Robot capability and intention in trust-based decisions across tasks». Στο: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, σσ. 39–47.

[156] Adriana Hamacher, Nadia Bianchi-Berthouze, Anthony G Pipe και Kerstin Eder. «Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction». Στο: *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE. 2016, σσ. 493–500.

[157] Stefan-Dan Ciocirlan, Roxana Agrigoroaie και Adriana Tapus. «Human-robot team: effects of communication in analyzing trust». Στο: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2019, σσ. 1–7.

[158] Yue Wang, Laura R Humphrey, Zhanrui Liao και Huanfei Zheng. «Trust-based multi-robot symbolic motion planning with a human-in-the-loop». Στο: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), σσ. 1–33.

[159] Yaohui Guo, Chongjie Zhang και X Jessie Yang. «Modeling trust dynamics in human-robot teaming: A bayesian inference approach». Στο: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, σσ. 1–7.

[160] Kumar Akash, Wan-Lin Hu, Tahira Reid και Neera Jain. «Dynamic modeling of trust in human-machine interactions». Στο: *2017 American Control Conference (ACC)*. IEEE. 2017, σσ. 1542–1548.

[161] Michael W Floyd, Michael Drinkwater και David W Aha. «Adapting autonomous behavior using an inverse trust estimation». Στο: *Computational Science and Its Applications–ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30–July 3, 2014, Proceedings, Part I 14*. Springer. 2014, σσ. 728–742.

[162] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu και Siddhartha Srinivasa. «Trust-Aware Decision Making for Human-Robot Collaboration». Στο: *ACM Transactions on Human-Robot Interaction* 9 (2 Ιούν. 2020), σσ. 1–23. ISSN: 2573-9522. DOI: 10.1145/3359616. URL: https://dl.acm.org/doi/10.1145/3359616.

[163] Daphne Koller και Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[164] Adam N Sanborn και Nick Chater. «Bayesian brains without probabilities». Στο: *Trends in cognitive sciences* 20.12 (2016), σσ. 883–893.

[165] Edoardo M Airoldi. «Getting started in probabilistic graphical models». Στο: *PLoS Computational Biology* 3.12 (2007), e252.

[166] Harold Soh, Yaqi Xie, Min Chen και David Hsu. «Multi-task trust transfer for human–robot interaction». Στο: *The International Journal of Robotics Research* 39.2-3 (2020), σσ. 233–249.

[167] Huanfei Zheng, Zhanrui Liao και Yue Wang. «Human-robot trust integrated task allocation and symbolic motion planning for heterogeneous multi-robot systems». Στο: *Dynamic systems and control conference*. Τόμ. 51913. American Society of Mechanical Engineers. 2018, V003T30A010.

[168] Jiun-Yin Jian, Ann M Bisantz και Colin G Drury. «Foundations for an empirically determined scale of trust in automated systems». Στο: *International journal of cognitive ergonomics* 4.1 (2000), σσ. 53–71.

[169] Kristin Schaefer. «The perception and measurement of human-robot trust». Στο: (2013).

[170] Kristin E Schaefer. «Measuring trust in human robot interactions: Development of the "trust perception scale-HRI"». Στο: *Robust intelligence and trust in autonomous systems*. Springer, 2016, σσ. 191–218.

[171] Gale Lucas, Giota Stratou, Shari Lieblich και Jonathan Gratch. «Trust me: multimodal signals of trustworthiness». Στο: *Proceedings of the 18th ACM international conference on multimodal interaction*. 2016, σσ. 5–12.

[172] Saeid Nahavandi. «Trust in autonomous systems-iTrust lab: Future directions for analysis of trust with autonomous systems». Στο: *IEEE Systems, Man, and Cybernetics Magazine* 5.3 (2019), σσ. 52–59.

[173] Quaneisha Jenkins και Xiaochun Jiang. «Measuring trust and application of eye tracking in human robotic interaction». Στο: *IIEAnnual Conference and Expo*. 2010.

[174] Jin Joo Lee, Brad Knox και Cynthia Breazeal. «Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions». Στο: *2013 AAAI Spring Symposium Series*. 2013.

[175] Aaron C Elkins και Douglas C Derrick. «The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents». Στο: *Group decision and negotiation* 22.5 (2013), σσ. 897–913.

[176] Kumar Akash, Wan-Lin Hu, Neera Jain και Tahira Reid. «A classification model for sensing human trust in machines using EEG and GSR». Στο: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), σσ. 1–20.

[177] Wan-Lin Hu, Kumar Akash, Neera Jain και Tahira Reid. «Real-time sensing of trust in human-machine interactions». Στο: *IFAC-PapersOnLine* 49.32 (2016), σσ. 48–53.

[178] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner και Mark Billinghurst. «Measuring human trust in a virtual assistant using physiological sensing in virtual reality». Στο: *2020 IEEE Conference on virtual reality and 3D user interfaces (VR)*. IEEE. 2020, σσ. 756–765.

[179] Randall D Spain, Ernesto A Bustamante και James P Bliss. «Towards an empirically developed scale for system trust: Take two». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 52. 19. SAGE Publications Sage CA: Los Angeles, CA. 2008, σσ. 1335–1339.

[180] Diane L Putnick και Marc H Bornstein. «Measurement invariance conventions and reporting: The state of the art and future directions for psychological research». Στο: *Developmental review* 41 (2016), σσ. 71–90.

[181] Maxwell L Elliott, Annchen R Knodt, David Ireland, Meriwether L Morris, Richie Poulton, Sandhya Ramrakha, Maria L Sison, Terrie E Moffitt, Avshalom Caspi και Ahmad R Hariri. «What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis». Στο: *Psychological science* 31.7 (2020), σσ. 792–806.

[182] Keith B Burt και Jelena Obradović. «The construct of psychophysiological reactivity: Statistical and psychometric issues». Στο: *Developmental Review* 33.1 (2013), σσ. 29–57.

[183] HLEG AI. *High-level expert group on artificial intelligence*. 2019.

[184] Arunkumar Ramaswamy, Bruno Monsuez και Adriana Tapus. «Modeling non-functional properties for human-machine systems». Στο: *2014 AAAI Spring Symposium Series*. 2014.

[185] Arunkumar Ramaswamy, Bruno Monsuez και Adriana Tapus. «SafeRobots: A model-driven Framework for developing Robotic Systems». Στο: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, σσ. 1517–1524.

[186] James Bret Michael, Doron Drusinsky, Thomas W Otani και Man-Tak Shing. «Verification and validation for trustworthy software systems». Στο: *IEEE software* 28.6 (2011), σσ. 86–92.

[187] Yuanjie Si, Jun Sun, Yang Liu, Jin Song Dong, Jun Pang, Shao Jie Zhang και Xiaohu Yang. «Model checking with fairness assumptions using PAT». Στο: *Frontiers of Computer Science* 8 (2014), σσ. 1–16.

[188] Rezvan Joshaghani, Stacy Black, Elena Sherman και Hoda Mehrpouyan. «Formal specification and verification of user-centric privacy policies for ubiquitous systems». Στο: *Proceedings of the 23rd International Database Applications & Engineering Symposium*. 2019, σσ. 1–10.

[189] Rimvydas Rukšėnas, Jonathan Back, Paul Curzon και Ann Blandford. «Verification-guided modelling of salience and cognitive load». Στο: *Formal Aspects of Computing* 21 (2009), σσ. 541–569.

[190] Matthew L Bolton, Ellen J Bass και Radu I Siminiceanu. «Using formal verification to evaluate human-automation interaction: A review». Στο: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43.3 (2013), σσ. 488–503.

[191] Matt Webster, David Western, Dejanira Araiza-Illan, Clare Dixon, Kerstin Eder, Michael Fisher και Anthony G Pipe. «A corroborative approach to verification and validation of human–robot teams». Στο: *The International Journal of Robotics Research* 39.1 (2020), σσ. 73–99.

[192] Xiaowei Huang, Marta Kwiatkowska και Maciej Olejnik. «Reasoning about cognitive trust in stochastic multiagent systems». Στο: *ACM Transactions on Computational Logic (TOCL)* 20.4 (2019), σσ. 1–64.

[193] Amanda JC Sharkey. «Should we welcome robot teachers?» Στο: *Ethics and Information Technology* 18 (2016), σσ. 283–297.

[194] Alexander Mois Aroyo, Francesco Rea, Giulio Sandini και Alessandra Sciutti. «Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?» Στο: *IEEE Robotics and Automation Letters* 3.4 (2018), σσ. 3701–3708.

[195] Michal Štolba, Jan Tožička και Antonín Komenda. «Quantifying privacy leakage in multi-agent planning». Στο: *ACM Transactions on Internet Technology (TOIT)* 18.3 (2018), σσ. 1–21.

[196] Thomas Given-Wilson, Axel Legay και Sean Sedwards. «Information security, privacy, and trust in social robotic assistants for older adults». Στο: *Human Aspects of Information Security, Privacy and Trust: 5th International Conference, HAS 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings 5*. Springer. 2017, σσ. 90–109.

[197] Inaki Maurtua, Aitor Ibarguren, Johan Kildal, Loreto Susperregi και Basilio Sierra. «Human–robot collaboration in industrial applications: Safety, interaction and trust». Στο: *International Journal of Advanced Robotic Systems* 14.4 (2017), σ. 1729881417716010.

[198] Cynthia Dwork και Aaron Roth. «The Algorithmic Foundations of Differential Privacy». Στο: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), σσ. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: http://dx.doi.org/10.1561/0400000042.

[199] Alan F. T. Winfield και Marina Jirotka. «Ethical governance is essential to building trust in robotics and artificial intelligence systems». Στο: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), σ. 20180085. DOI: 10.1098/rsta.2018.0085. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2018.0085. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2018.0085.

[200] Gale M Lucas, Jonathan Gratch, Aisha King και Louis-Philippe Morency. «It's only a computer: Virtual humans increase willingness to disclose». Στο: *Computers in Human Behavior* 37 (2014), σσ. 94–100.

[201] Bengt Muthén και Tihomir Asparouhov. «Causal effects in mediation modeling: An introduction with applications to latent variables». Στο: *Structural equation modeling: a multidisciplinary journal* 22.1 (2015), σσ. 12–23.

[202] Eric T Chancey, James P Bliss, Alexandra B Proaps και Poornima Madhavan. «The role of trust as a mediator between system characteristics and response behaviors». Στο: *Human factors* 57.6 (2015), σσ. 947–958.

[203] Tan Zhi-Xuan, Harold Soh και Desmond Ong. «Factorized inference in deep markov models for incomplete multimodal time series». Στο: *Proceedings of the AAAI Conference on Artificial Intelligence*. Τόμ. 34. 06. 2020, σσ. 10334–10341.

[204] Wei-Ning Hsu, Yu Zhang και James Glass. «Unsupervised learning of disentangled and interpretable representations from sequential data». Στο: *Advances in neural information processing systems* 30 (2017).

[205] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams και Sandeep R Datta. «Composing graphical models with neural networks for structured representations and fast inference». Στο: *Advances in neural information processing systems* 29 (2016).

[206] Sangseok You, Jeong-Hwan Kim, SangHyun Lee, Vineet Kamat και Lionel P Robert Jr. «Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments». Στο: *Automation in Construction* 96 (2018), σσ. 161–170.

[207] Jin Xu και Ayanna Howard. «Would you take advice from a robot? Developing a framework for inferring human-robot trust in time-sensitive scenarios». Στο: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2020, σσ. 814–820.

[208] Sangseok You και Lionel Robert. «Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction». Στο: *You, S. and Robert, LP (2019). Trusting Robots in Teams: Examining the Impacts of Trusting Robots on Team Performance and Satisfaction, Proceedings of the 52th Hawaii International Conference on System Sciences, Jan*. 2018, σσ. 8–11.

[209] Murat Kirtay, Erhan Oztop, Minoru Asada και Verena V Hafner. «Trust me! I am a robot: an affective computational account of scaffolding in robot-robot interaction». Στο: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, σσ. 189–196.

[210] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay και Michael L Walters. «How social robots influence people's trust in critical situations». Στο: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2020, σσ. 1020–1025.

[211] Connor Esterwood και Lionel P Robert. «Do you still trust me? human-robot trust repair strategies». Στο: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, σσ. 183–188.

[212] Roy J Lewicki και Chad Brinsfield. «Trust repair». Στο: *Annual review of organizational psychology and organizational behavior* 4 (2017), σσ. 287–313.

[213] Nicole Gillespie, Steve Lockey, Matthew Hornsey και Tyler Okimoto. «Trust repair: A multilevel framework». Στο: *Understanding Trust in Organizations*. Routledge, 2021, σσ. 143–176.

[214] Peter H Kim, Donald L Ferrin, Cecily D Cooper και Kurt T Dirks. «Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations.» Στο: *Journal of applied psychology* 89.1 (2004), σ. 104.

[215] Roy J Lewicki, Barbara B Bunker κ.ά. «Developing and maintaining trust in work relationships». Στο: *Trust in organizations: Frontiers of theory and research* 114 (1996), σ. 139.

[216] Steven L Grover, Markus C Hasel, Caroline Manville και Carolina Serrano-Archimi. «Follower reactions to leader trust violations: A grounded theory of violation types, likelihood of recovery, and recovery process». Στο: *European Management Journal* 32.5 (2014), σσ. 689–702.

[217] Ana Cristina Costa, Donald L Ferrin και C Ashley Fulmer. «Trust at work.» Στο: (2018).

[218] Roderick M Kramer και Roy J Lewicki. «Repairing and enhancing trust: Approaches to reducing organizational trust deficits». Στο: *The Academy of Management Annals* 4.1 (2010), σσ. 245–277.

[219] Maurice E Schweitzer, John C Hershey και Eric T Bradlow. «Promises and lies: Restoring violated trust». Στο: *Organizational behavior and human decision processes* 101.1 (2006), σσ. 1–19.

[220] Edward C Tomlinson και Roger C Mryer. «The role of causal attribution dimensions in trust repair». Στο: *Academy of management review* 34.1 (2009), σσ. 85–104.

[221] Robert J Bies και Debra L Shapiro. «Interactional fairness judgments: The influence of causal accounts». Στο: *Social justice research* 1 (1987), σσ. 199–218.

[222] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K Pradhan, X Jessie Yang και Lionel P Robert Jr. «Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload». Στο: *Transportation research part C: emerging technologies* 104 (2019), σσ. 428–442.

[223] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck και Emil Coman. «Investigating the effects of (empty) promises on human-automation interaction and trust repair». Στο: *Proceedings of the 8th International Conference on Human-Agent Interaction*. 2020, σσ. 6–14.

[224] Manisha Natarajan και Matthew Gombolay. «Effects of anthropomorphism and accountability on trust in human robot interaction». Στο: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 2020, σσ. 33–42.

[225] Spencer C Kohn, Ali Momen, Eva Wiese, Yi-Ching Lee και Tyler H Shaw. «The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles». Στο: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Τόμ. 63. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, σσ. 222–226.

[226] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa και Paul Rybski. «Gracefully mitigating breakdowns in robotic services». Στο: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, σσ. 203–210.

[227] Spencer C Kohn, Daniel Quinn, Richard Pak, Ewart J De Visser και Tyler H Shaw. «Trust repair strategies with self-driving vehicles: An exploratory study». Στο: *Proceedings of the human factors and ergonomics society annual meeting*. Τόμ. 62. 1. Sage Publications Sage CA: Los Angeles, CA. 2018, σσ. 1108–1112.

[228] David Cameron, Stevienna de Saille, Emily C Collins, Jonathan M Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh και James Law. «The effect of social-cognitive recovery strategies on likability, capability and trust in social robots». Στο: *Computers in human behavior* 114 (2021), σ. 106561.

[229] Paul Robinette, Ayanna M Howard και Alan R Wagner. «Timing is key for robot trust repair». Στο: *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*. Springer. 2015, σσ. 574–583.

[230] Esther S Kox, José H Kerstholt, Tom F Hueting και Peter W de Vries. «Trust repair in human-agent teams: the effectiveness of explanations and expressing regret». Στο: *Autonomous Agents and Multi-Agent Systems* 35.2 (2021), σ. 30.

[231] Xinyi Zhang. «"Sorry, It Was My Fault": Repairing Trust in Human-Robot Interactions». Στο: (2021).

[232] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo και Ana Paiva. «Exploring the impact of fault justification in human-robot trust». Στο: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. 2018, σσ. 507–513.

[233] Ning Wang, David V Pynadath, Ericka Rovira, Michael J Barnes και Susan G Hill. «Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams». Στο: *Persuasive Technology: 13th International Conference, PERSUASIVE 2018, Waterloo, ON, Canada, April 18-19, 2018, Proceedings 13*. Springer. 2018, σσ. 56–69.

[234] Russell Perkins, Zahra Rezaei Khavas και Paul Robinette. «Trust calibration and trust respect: A method for building team cohesion in human robot teams». Στο: *arXiv preprint arXiv:2110.06809* (2021).

[235] Samantha Reig, Elizabeth J Carter, Terrence Fong, Jodi Forlizzi και Aaron Steinfeld. «Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies». Στο: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 2021, σσ. 158–167.

[236] Special Eurobarometer. «Public attitudes towards robots». Στο: *European Commission* (2012).

[237] Sarah Kriz, Toni D Ferro, Pallavi Damera και John R Porter. «Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations». Στο: *19th international symposium in robot and human interactive communication*. IEEE. 2010, σσ. 458–463.

[238] Icek Ajzen. *The theory of planned behaviour. Organizational behaviour and human decision processes, 50 (2), 179-211*. 1991.

[239] Tatsuya Nomura, Takayuki Kanda και Tomohiro Suzuki. «Experimental investigation into influence of negative attitudes toward robots on human–robot interaction». Στο: *Ai & Society* 20 (2006), σσ. 138–150.

[240] Marco Nørskov. «Technological dangers and the potential of human–robot interaction: a philosophical investigation of fundamental epistemological mechanisms of discrimination». Στο: *Social Robots*. Routledge, 2017, σσ. 99–121.

[241] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend και Gerhard Sagerer. «Understanding social robots». Στο: *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE. 2009, σσ. 169–174.

[242] Joost Broekens, Marcel Heerink, Henk Rosendal κ.ά. «Assistive social robots in elderly care: a review». Στο: *Gerontechnology* 8.2 (2009), σσ. 94–103.

[243] Sibylle Enz, Martin Diruf, Caroline Spielhagen, Carsten Zoll και Patricia A Vargas. «The social role of robots in the future—explorative measurement of hopes and fears». Στο: *International Journal of Social Robotics* 3 (2011), σσ. 263–271.

[244] Tatsuya Nomura, Takayuki Kanda, Sachie Yamada και Tomohiro Suzuki. «Exploring influences of robot anxiety into HRI». Στο: *Proceedings of the 6th international conference on Human-robot interaction*. 2011, σσ. 213–214.

[245] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda και Kensuke Kato. «Measurement of anxiety toward robots». Στο: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, σσ. 372–377.

[246] Tatsuya Nomura, Takuya Shintani, Kazuki Fujii και Kazumasa Hokabe. «Experimental investigation of relationships between anxiety, negative attitudes, and allowable distance of robots». Στο: *Proceedings of the 2nd IASTED international conference on human computer interaction, Chamonix, France. ACTA Press*. Citeseer. 2007, σσ. 13–18.

[247] Maartje MA de Graaf και Somaya Ben Allouch. «The relation between people's attitude and anxiety towards robots in human-robot interaction». Στο: *2013 IEEE RO-MAN*. IEEE. 2013, σσ. 632–637.

[248] Mark A Hall, Elizabeth Dugan, Beiyao Zheng και Aneil K Mishra. «Trust in physicians and medical institutions: what is it, can it be measured, and does it matter?» Στο: *The milbank quarterly* 79.4 (2001), σσ. 613–639.

[249] Peter A Hancock, Deborah R Billings και Kristen E Schaefer. «Can you trust your robot?» Στο: *Ergonomics in Design* 19.3 (2011), σσ. 24–29.

[250] Marcel Heerink, Ben Kröse, Vanessa Evers και Bob Wielinga. *Assessing acceptance of assistive social agent technology by older adults: the almere model*. 2010.

[251] Elizabeth Broadbent, Rebecca Stafford και Bruce MacDonald. «Acceptance of healthcare robots for the older population: Review and future directions». Στο: *International journal of social robotics* 1 (2009), σσ. 319–330.

[252] Maartje MA de Graaf, Somaya Ben Allouch και Shariff Lutfi. «What are people's associations of domestic robots?: Comparing implicit and explicit measures». Στο: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2016, σσ. 1077–1083.

[253] Natalia Reich-Stiebert, Friederike Eyssel και Charlotte Hohnemann. «Involve the user! Changing attitudes toward robots by user participation in a robot prototyping process». Στο: *Computers in Human Behavior* 91 (2019), σσ. 290–296.

[254] Jakub A Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck και Hiroshi Ishiguro. «Persistence of the uncanny valley: the influence of repeated interactions and a robot's attitude on its perception». Στο: *Frontiers in psychology* 6 (2015), σ. 883.

[255] David C May, Kristie J Holler, Cindy L Bethel, Lesley Strawderman, Daniel W Carruth και John M Usher. «Survey of factors for the prediction of human comfort with a non-anthropomorphic robot in public spaces». Στο: *International Journal of Social Robotics* 9 (2017), σσ. 165–180.

[256] Nina Savela, Tuuli Turja και Atte Oksanen. «Social acceptance of robots in different occupational fields: a systematic literature review». Στο: *International Journal of Social Robotics* 10.4 (2018), σσ. 493–502.

[257] S Maryam Fakhr Hosseini, Dylan Lettinga, Eric Vasey, Zhi Zheng, Myounghoon Jeon, Chung Hyuk Park και Ayanna M Howard. «Both "look and feel" matter: Essential factors for robotic companionship». Στο: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, σσ. 150–155.

[258] Elena Torta, Franz Werner, David O Johnson, James F Juola, Raymond H Cuijpers, Marco Bazzani, Johannes Oberzaucher, John Lemberger, Hadas Lewy και Joseph Bregman. «Evaluation of a small socially-assistive humanoid robot in intelligent homes for the care of the elderly». Στο: *Journal of Intelligent & Robotic Systems* 76 (2014), σσ. 57–71.

[259] Minoo Alemi, Ali Meghdari και Maryam Ghazisaedy. «The effect of employing humanoid robots for teaching English on students' anxiety and attitude». Στο: *2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)*. IEEE. 2014, σσ. 754–759.

[260] Elizabeth Broadbent, Rie Tamagawa, Anna Patience, Brett Knock, Ngaire Kerse, Karen Day και Bruce A MacDonald. «Attitudes towards health-care robots in a retirement village». Στο: *Australasian journal on ageing* 31.2 (2012), σσ. 115–120.

[261] Felip Martí Carrillo, Joanna Butchart, Nicholas Kruse, Adam Scheinberg, Lisa Wise και Chris McCarthy. «Physiotherapists' acceptance of a socially assistive robot in ongoing clinical deployment». Στο: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, σσ. 850–855.

[262] Wan-Ling Chang, Jeremy P White, Joohyun Park, Anna Holm και Selma Šabanović. «The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay». Στο: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2012, σσ. 845–850.

[263] Min-Gyu Kim, Jaeryoung Lee, Yuusuke Aichi, Hiroki Morishita και Munehiro Makino. «Effectiveness of robot exhibition through visitors experience: A case study of Nagoya Science Hiroba exhibition in Japan». Στο: *2016 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*. IEEE. 2016, σσ. 1–5.

[264] Maartje MA De Graaf και Somaya Ben Allouch. «Exploring influencing variables for the acceptance of social robots». Στο: *Robotics and autonomous systems* 61.12 (2013), σσ. 1476–1486.

[265] Sofia Serholt, Christina Anne Basedow, Wolmet Barendregt και Mohammad Obaid. «Comparing a humanoid tutor to a human tutor delivering an instructional task to children». Στο: *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE. 2014, σσ. 1134–1141.

[266] Carl J Dunst, Carol M Trivette, Jeremy Prior, Deborah W Hamby και Davon Embler. «Parents' Judgments of the Acceptability and Importance of Socially Interactive Robots for Intervening with Young Children with Disabilities. Social Robots Research Reports, Number 1.» Στο: *Orelena Hawks Puckett Institute* (2013).

[267] Christoph Bartneck, Tomohiro Suzuki, Takayuki Kanda και Tatsuya Nomura. «The influence of people's culture and prior experiences with Aibo on their attitude towards robots». Στο: *Ai & Society* 21 (2007), σσ. 217–230.

[268] Christoph Bartneck, Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki και Kennsuke Kato. «Cultural differences in attitudes towards robots». Στο: AISB. 2005.

[269] Jasmin Bernotat και Friederike Eyssel. «Can ('t) Wait to Have a Robot at Home?-Japanese and German Users' Attitudes Toward Service Robots in Smart Homes». Στο: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, σσ. 15–22.

[270] Iolanda Leite, Carlos Martinho και Ana Paiva. «Social robots for long-term interaction: a survey». Στο: *International Journal of Social Robotics* 5 (2013), σσ. 291–308.

[271] Nazaret Gómez-del-Río, Carina S González-González, Pedro A Toledo-Delgado, Vanesa Muñoz-Cruz και Francisco García-Peñalvo. «Health promotion for childhood obesity: An approach based on self-tracking of data». Στο: *Sensors* 20.13 (2020), σ. 3778.

[272] Carina S González-González, Erika Herrera-González, Lorenzo Moreno-Ruiz, Nuria Reyes-Alonso, Selene Hernández-Morales, María D Guzmán-Franco και Alfonso Infante-Moro. «Computational thinking and down syndrome: An exploratory study using the KIBO robot». Στο: *Informatics*. Τόμ. 6. 2. MDPI. 2019, σ. 25.

[273] Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Fardad Faridi, Jesse Gray, Matt Berlin, Harald Quintus-Bosz, Robert Hartmann, Mike Hess, Stacy Dyer κ.ά. «Tega: a social robot». Στο: (2016).

[274] Maya B Mathur και David B Reichling. «Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley». Στο: *Cognition* 146 (2016), σσ. 22–32.

[275] Pay Ling Yu, MS Balaji και Kok Wei Khong. «Building trust in internet banking: a trustworthiness perspective». Στο: *Industrial Management & Data Systems* 115.2 (2015), σσ. 235–252.

[276] Takayuki Gompei και Hiroyuki Umemuro. «Factors and development of cognitive and affective trust on social robots». Στο: *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*. Springer. 2018, σσ. 45–54.

[277] Shane Saunderson και Goldie Nejat. «How robots influence humans: A survey of nonverbal communication in social human–robot interaction». Στο: *International Journal of Social Robotics* 11 (2019), σσ. 575–608.

[278] Conor McGinn. «Why do robots need a head? The role of social interfaces on service robots». Στο: *International Journal of Social Robotics* 12.1 (2020), σσ. 281–295.

[279] Lee Sproull, Mani Subramani, Sara Kiesler, Janet H Walker και Keith Waters. «When the interface is a face». Στο: *Human-computer interaction* 11.2 (1996), σσ. 97–124.

[280] Akanksha Prakash και Wendy A Rogers. «Why some humanoid faces are perceived more positively than others: effects of human-likeness and task». Στο: *International journal of social robotics* 7.2 (2015), σσ. 309–331.

[281] Jan R Landwehr, Ann L McGill και Andreas Herrmann. «It's got the look: The effect of friendly and aggressive "facial" expressions on product liking and sales». Στο: *Journal of marketing* 75.3 (2011), σσ. 132–146.

[282] Janine Willis και Alexander Todorov. «First impressions: Making up your mind after a 100-ms exposure to a face». Στο: *Psychological science* 17.7 (2006), σσ. 592–598.

[283] Francesca Simion και Elisa Di Giorgio. «Face perception and processing in early infancy: inborn predispositions and developmental changes». Στο: *Frontiers in psychology* 6 (2015), σ. 969.

[284] Ahreum Maeng και Pankaj Aggarwal. «Facing dominance: Anthropomorphism and the effect of product face ratio on consumer preference». Στο: *Journal of Consumer Research* 44.5 (2018), σσ. 1104–1122.

[285] Sara Santos, Ines Almeida, Barbara Oliveiros και Miguel Castelo-Branco. «The role of the amygdala in facial trustworthiness processing: A systematic review and meta-analyses of fMRI studies». Στο: *PloS one* 11.11 (2016), e0167276.

[286] Yao Song. «Building a 'deeper'trust: Mapping the facial anthropomorphic trustworthiness in social robot design through multidisciplinary approaches». Στο: *The Design Journal* 23.4 (2020), σσ. 639–649.

[287] Michael Stirrat και David I Perrett. «Valid facial cues to cooperation and trust: Male facial width and trustworthiness». Στο: *Psychological science* 21.3 (2010), σσ. 349–354.

[288] Marielle EH Creusen και Jan PL Schoormans. «The different roles of product appearance in consumer choice». Στο: *Journal of product innovation management* 22.1 (2005), σσ. 63–81.

[289] Linda Miesler, Helmut Leder και Andreas Herrmann. «Isn't it cute: An evolutionary perspective of baby-schema effects in visual product designs». Στο: *International Journal of Design* 5.3 (2011).

[290] Isabel M Santos και Andrew W Young. «Inferring social attributes from different face regions: Evidence for holistic processing». Στο: *Quarterly Journal of Experimental Psychology* 64.4 (2011), σσ. 751–766.

[291] Raphaela E Kaisler και Helmut Leder. «Trusting the looks of others: Gaze effects of faces in social settings». Στο: *Perception* 45.8 (2016), σσ. 875–892.

[292] Karel Kleisner, Lenka Priplatova, Peter Frost και Jaroslav Flegr. «Trustworthy-looking face meets brown eyes». Στο: *PloS one* 8.1 (2013), e53285.

[293] Alexander Todorov, Sean G Baron και Nikolaas N Oosterhof. «Evaluating face trustworthiness: a model based approach». Στο: *Social cognitive and affective neuroscience* 3.2 (2008), σσ. 119–127.

[294] Jaume Masip, Eugenio Garrido και Carmen Herrero. «Facial appearance and impressions of 'credibility': The effects of facial babyishness and age on person perception». Στο: *International journal of psychology* 39.4 (2004), σσ. 276–289.

[295] Lenka Linke, S Adil Saribay και Karel Kleisner. «Perceived trustworthiness is associated with position in a corporate hierarchy». Στο: *Personality and Individual Differences* 99 (2016), σσ. 22–27.

[296] Sheila Brownlow. «Seeing is believing: Facial appearance, credibility, and attitude change». Στο: *Journal of nonverbal behavior* 16 (1992), σσ. 101–115.

[297] Fengling Ma, Fen Xu και Xianming Luo. «Children's and adults' judgments of facial trustworthiness: the relationship to facial attractiveness». Στο: *Perceptual and Motor Skills* 121.1 (2015), σσ. 179–198.

[298] Fen Xu, Dingcheng Wu, Rie Toriyama, Fengling Ma, Shoji Itakura και Kang Lee. «Similarities and differences in Chinese and Caucasian adults' use of facial cues for trustworthiness judgments». Στο: *PLoS One* 7.4 (2012), e34859.

[299] Christopher John Stanton και Catherine J Stevens. «Don't stare at me: the impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task». Στο: *International Journal of Social Robotics* 9 (2017), σσ. 745–753.

[300] Damian A Stanley, Peter Sokol-Hessner, Mahzarin R Banaji και Elizabeth A Phelps. «Implicit race attitudes predict trustworthiness judgments and economic trust decisions». Στο: *Proceedings of the National Academy of Sciences* 108.19 (2011), σσ. 7710–7715.

[301] Manuel G Calvo, Patricia Álvarez-Plaza και Andrés Fernández-Martín. «The contribution of facial regions to judgements of happiness and trustworthiness from dynamic expressions». Στο: *Journal of Cognitive Psychology* 29.5 (2017), σσ. 618–625.

[302] Matia Okubo, Kenta Ishikawa και Akihiro Kobayashi. «No trust on the left side: Hemifacial asymmetries for trustworthiness and emotional expressions». Στο: *Brain and Cognition* 82.2 (2013), σσ. 181–186.

[303] Mathieu Arminjon, Amer Chamseddine, Vladimir Kopta, Aleksandar Paunović και Christine Mohr. «Are We Modular Lying Cues Detectors? The Answer Is "Yes, Sometimes".» Στο: *Plos one* 10.9 (2015), e0136418.

[304] Åke Hellström και Joseph Tekle. «Person perception through facial photographs: Effects of glasses, hair, and beard on judgments of occupation and personal qualities». Στο: *European Journal of Social Psychology* 24.6 (1994), σσ. 693–705.

[305] Arman Bakmazian. «The man behind the beard: Perception of men's trustworthiness as a function of facial hair». Στο: *Psychology* 2014 (2014).

[306] Harry Farmer, Ryan McKay και Manos Tsakiris. «Trust in me: Trustworthy others are seen as more physically similar to the self». Στο: *Psychological science* 25.1 (2014), σσ. 290–292.

[307] Ferenc Kocsor και Tamas Bereczkei. «First impressions of strangers rely on generalization of behavioral traits associated with previously seen facial features». Στο: *Current Psychology* 36 (2017), σσ. 385–391.

[308] Leslie A Zebrowitz, Luminita Voinescu και Mary Ann Collins. «" Wide-eyed" and" crooked-faced": Determinants of perceived and real honesty across the life span». Στο: *Personality and social psychology bulletin* 22.12 (1996), σσ. 1258–1269.

[309] Lucy Johnston, Lynden Miles και C Neil Macrae. «Why are you smiling at me? Social functions of enjoyment and non-enjoyment smiles». Στο: *British Journal of Social Psychology* 49.1 (2010), σσ. 107–127.

[310] Aida Gutiérrez-García και Manuel G Calvo. «Social anxiety and trustworthiness judgments of dynamic facial expressions of emotion». Στο: *Journal of behavior therapy and experimental psychiatry* 52 (2016), σσ. 119–127.

[311] Eva Krumhuber, Antony SR Manstead, Darren Cosker, Dave Marshall, Paul L Rosin και Arvid Kappas. «Facial dynamics as indicators of trustworthiness and cooperative behavior.» Στο: *Emotion* 7.4 (2007), σ. 730.

[312] JA Desor και Gary K Beauchamp. «The human capacity to transmit olfactory information». Στο: *Perception & Psychophysics* 16.3 (1974), σσ. 551–556.

[313] Heather D Flowe. «Do characteristics of faces that convey trustworthiness and dominance underlie perceptions of criminality?» Στο: *PloS one* 7.6 (2012), e37253.