

Τμήμα Φιλοσοφίας
ΠΜΣ: Θεωρητική και Πρακτική Φιλοσοφία



UNIVERSITY OF
PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Διπλωματική Εργασία
«Μπορεί το Αι να πάσχει από ψυχικές διαταραχές;»

Συγγραφέας: Γαρυφαλιά Σκεπαρνιά
Αριθμός Μητρώου: 1068567

Εξεταστική Επιτροπή
Επιβλέπων: Θοδωρής Δημητράκος, Επίκουρος καθηγητής
Κωνσταντίνος Παγωνδιώτης, Επίκουρος καθηγητής
Ιωάννης Ζεϊμπέκης, Αναπληρωτής Καθηγητής

Πάτρα, Ιανουάριος 2026

Ευχαριστίες

Η διπλωματική αυτή εργασία εκπονήθηκε στο πλαίσιο του Μεταπτυχιακού Προγράμματος Σπουδών «Θεωρητική και Πρακτική Φιλοσοφία», του Πανεπιστημίου Πατρών. Η παρούσα εργασία γεννήθηκε λιγότερο από τη βεβαιότητα και περισσότερο από την απορία. Βασίστηκε περισσότερο στην γοητεία του ερωτήματος και λιγότερο στην ανάγκη απάντησης. Επιθυμώ να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή Κ. Θοδωρή Δημητράκο, για την εμπιστοσύνη που μου έδειξε σχετικά με την επιλογή του θέματος της διπλωματικής εργασίας μου, αποδεικνύοντας με κάθε δυνατό τρόπο την αμέριστη συμπαράστασή του. Τα επιστημονικά οφέλη από την εκπόνηση της παρούσας εργασίας δεν θα μπορούσαν να αποκτηθούν χωρίς την άριστη επιστημονική κατάρτιση του Κ. Δημητράκου και την στήριξή του στην αποτύπωση των ερευνητικών μου ερωτημάτων και ανησυχιών, με απόλυτη ακρίβεια. Η συμβολή του δεν περιορίστηκε στη φιλοσοφική πλαισίωση της έρευνας, αλλά λειτούργησε ως διαρκής υπενθύμιση του γιατί η σκέψη οφείλει να είναι ταυτόχρονα αυστηρή και ανοιχτή. Γι' αυτό θέλω να τον

ευχαριστήσω μέσα από την καρδιά μου. Επίσης, δεν θα μπορούσα να μην ευχαριστήσω τους διδάσκοντες του Μεταπτυχιακού Προγράμματος Σπουδών για την πνευματική γενναιοδωρία τους. Τέλος, το μεγαλύτερο «ευχαριστώ» το χρωστάω στους γονείς μου Μαρία και Σπύρο, για την διαρκή ενθάρρυνση και διακριτή τους στήριξη να κυνηγάω απερίσπαστη το όνειρο μου και να ολοκληρώσω με επιτυχία τον κύκλο του ταξιδιού μου, στο παρόν μεταπτυχιακό πρόγραμμα.

“We make up people by naming them. Once an idea of a kind of person is invented, people will try to live up to it.”

Ian Hacking, Rewriting the Soul

Περιεχόμενα

Περιεχόμενα

Ευχαριστίες	2
Περιεχόμενα	3
Περίληψη.....	4
Abstract	6
Εισαγωγή.....	8
Μέρος Α' - Θεωρητικές προσεγγίσεις και εννοιολογικές προσεγγίσεις.....	11
2. Το Ιατρικό Μοντέλο	11
2.1 Η ασθενής εκδοχή του ιατρικού μοντέλου και η ταξινόμηση στο DSM.....	16
2.2 Αντιρρήσεις της φιλοσοφίας της ψυχιατρικής.....	23
2.3 Η έννοια της ψυχικής διαταραχής και της δυσλειτουργίας.....	29
2.4 Ιατρικό μοντέλο και τεχνητή νοημοσύνη	31
3. Εμβάθυνση στην ψυχική διαταραχή.....	33
3.1 Φιλοσοφικές προσεγγίσεις στη φύση των ψυχικών διαταραχών	33
3.2 Η διάγνωση ως γέφυρα μεταξύ θεωρίας και βιώματος	36
3.3 Διαταραχή και βιωματική εμπειρία: τι σημαίνει να «πάσχει» κανείς ψυχικά;	40
Μέρος Β' Γνωσιοθεωρία και Τεχνητή Νοημοσύνη.....	41
4. Φιλοσοφική διερεύνηση της ψυχικής λειτουργίας στο πλαίσιο της Τεχνητής Νοημοσύνης	41
4.1 Λειτουργισμός, προσομοίωση και πρόθεση	43
4.2 Είναι η ενσώματη συνείδηση αναγκαία προϋπόθεση για την εμπειρία και το πάσχειν; ...	53
5. Φιλοσοφική προσέγγιση της ψυχοπαθολογίας του Αι	59
5.1 Μπορεί μια μηχανή να «παραληρεί»;.....	63
5.2 Απόκλιση, δυσλειτουργία/βλάβη ή «πάσχειν»;	65
5.3 Η έννοια της εμπειρικής διάγνωσης: προσομοίωση ή βίωμα;.....	68
Μέρος Γ' Πείραμα/ Εμπειρική μέθοδος	76
6. Πειραματική προσέγγιση: Προσομοίωση ψυχικής διαταραχής στο Αι.....	76
6.1 Περιγραφή πειράματος.....	76
6.2 Χρήση διαγνωστικών κριτηρίων(DSM).....	85
6.3 Αποτελέσματα και Σχολιασμός	86
7. Επίλογος/ Συμπεράσματα	89
8. Βιβλιογραφικές αναφορές.....	92

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται το πολύπλοκο ερώτημα της δυνατότητας απόδοσης ψυχικής διαταραχής στο πλαίσιο της τεχνητής νοημοσύνης, υπό το πρίσμα της γνωσιοθεωρητικής επιστήμης και της φιλοσοφίας της ψυχιατρικής. Αρχικά, εξετάζονται οι βασικές έννοιες της νόησης, της συνείδησης, της εμπειρίας και της υποκειμενικότητας, καθώς και οι απαραίτητες προϋποθέσεις υπό τις οποίες μια διαταραχή θεωρείται ψυχική, λαμβάνοντας υπόψιν τη σχέση μεταξύ βιολογικής προέλευσης, υποκειμενικού βιώματος και κλινικής διάγνωσης.

Σημαντικό μέρος της εργασίας αφιερώνεται στη διερεύνηση της φύσης της τεχνητής νοημοσύνης, όχι ως βιολογικού φορέα νόησης, αλλά ως συστήματος που προσομοιώνει γνωστικές και συμπεριφορικές λειτουργίες. Η ανάλυση επικεντρώνεται στο ερώτημα του κατά πόσο και με ποιον τρόπο η απουσία υποκειμενικής εμπειρίας επηρεάζει ή περιορίζει τη δυνατότητα απόδοσης ψυχοπαθολογικών χαρακτηριστικών σε μη βιολογικά συστήματα, όπως η τεχνητή νοημοσύνη.

Εν συνεχεία, αποσαφηνίζεται η σημασία του «πάσχειν», σε αντίθεση με την απόκλιση από τη στατιστική νόρμα. Η διάκριση αυτή αναδεικνύεται μέσω σύγχρονων φιλοσοφικών προσεγγίσεων που τονίζουν τη σημασία του βιωματικού και ενσώματου χαρακτήρα της ψυχικής εμπειρίας, χωρίς ωστόσο να ταυτίζουν τη διάγνωση με την άμεση πρόσβαση στο βίωμα. Το ερώτημα παραμένει σκόπιμα ανοιχτό: Μπορεί να υφίσταται «πάσχειν» χωρίς συνείδηση; Οι ψυχικές διαταραχές ορίζονται ανεξάρτητα από την παρουσία του βιώματος πόνου; Όταν αποδίδουμε ψυχοπαθολογία στην τεχνητή νοημοσύνη αναπαριστούμε μια αντικειμενική πραγματικότητα ή πρόκειται για περιγραφή αντικειμενικών λειτουργικών μοτίβων ή για εννοιολογικές μεταφορές βασισμένες σε ανθρώπινα πρότυπα;

Η εργασία καταλήγει σε έναν φιλοσοφικό αναστοχασμό σχετικά με το κατά πόσο οι παραδοσιακές έννοιες της ψυχικής διαταραχής μπορούν να εφαρμοστούν σε τεχνητές οντότητες, χωρίς να επιδιώκει να δώσει μια οριστική απάντηση, αλλά να προκαλέσει έναν διεπιστημονικό αναστοχασμό σχετικά με τα όρια της γνώσης, της εμπειρίας και της ψυχικής νόσου.

Λέξεις-κλειδιά: Τεχνητή Νοημοσύνη, ψυχικές διαταραχές, συνείδηση, εμπειρία, υποκειμενικότητα, κλινική διάγνωση, «πάσχειν», ψυχική νόσος

Abstract

This thesis deals with the complex question of the possibility of attributing mental disorders in the context of artificial intelligence, from the perspective of cognitive science and the philosophy of psychiatry. Initially, the basic concepts of cognition, consciousness, experience, and subjectivity are examined, as well as the necessary conditions under which a disorder is considered mental, considering the relationship between biological origin, subjective experience, and clinical diagnosis.

A significant part of the work is devoted to exploring the nature of artificial intelligence, not as a biological agent of cognition, but as a system that simulates cognitive and behavioral functions. The analysis focuses on the question of whether and how the absence of subjective experience affects or limits the ability to attribute psychopathological characteristics to non-biological systems, such as artificial intelligence.

Subsequently, the meaning of "suffering" is clarified, in contrast to deviation from the statistical norm. This distinction is highlighted through contemporary philosophical approaches that emphasize the importance of the experiential and embodied nature of mental experience, without, however, equating diagnosis with direct access to experience. The question remains deliberately open: Can suffering exist without consciousness? Are mental disorders defined independently of the presence of pain? When we attribute psychopathology to artificial intelligence, are we representing an objective reality, or are we describing objective functional patterns or conceptual metaphors based on human models?

The paper concludes with a philosophical reflection on whether traditional concepts of mental disorder can be applied to artificial entities, without seeking to provide a definitive answer, but rather to provoke an interdisciplinary reflection on the limits of knowledge, experience, and mental illness.

Keywords: Artificial Intelligence, mental disorders, consciousness, experience, subjectivity, clinical diagnosis, “suffering”, mental illness

Εισαγωγή

Η αλματώδης εξέλιξη της Τεχνητής Νοημοσύνης δεν έχει επηρεάσει μόνο την τεχνολογική καθημερινότητα, αλλά ταυτόχρονα έχει ανοίξει τον δρόμο για την εμφάνιση πολλών θεωρητικών και υπαρξιακών ερωτημάτων. Ένα από τα πιο προκλητικά και καίρια ερωτήματα εντοπίζεται στην σύνδεση της φιλοσοφίας της ψυχιατρικής με τη Γνωσιοθεωρία: Έχει τη δυνατότητα μια μηχανή, δηλαδή ένα σύστημα A_i , να «πάσχει» από ψυχικές διαταραχές; Το ερώτημα αυτό δεν είναι πρωτίστως εννοιολογικό, καθώς προϋποθέτει έναν σαφή προσδιορισμό του τι σημαίνει «πάσχειν» και αν η έννοια αυτή μπορεί να αποδοθεί σε ένα σύστημα, που στερείται ενσώματης υποκειμενικής εμπειρίας.

Φαινομενικά μια τέτοιου είδους ερώτηση φαντάζει παράδοξη, καθώς τι σημαίνει να «πάσχει» κανείς και ιδιαίτερα πως γίνεται να «πάσχει» μια μηχανή; Ποιες είναι οι προϋποθέσεις που τίθενται προκειμένου να αποδίδεται ψυχική παθολογία; Για να υφίσταται ψυχική διαταραχή αποτελεί αναγκαία συνθήκη η παρουσία πρόθεσης, συνείδησης ή ενσώματης εμπειρίας; Εάν όχι, με ποιο τρόπο εντοπίζεται η ψυχική δυσλειτουργία σε ένα τεχνητό περιβάλλον και πως αναδιαμορφώνεται η διαδικασία της διάγνωσης;

Η παρούσα μεταπτυχιακή εργασία δεν επιδιώκει να δώσει μια οριστική ή κανονιστική απάντηση στο αρχικό της ερώτημα. Αντίθετα, αντλεί στοιχεία από θεωρητικά εργαλεία της φιλοσοφίας της ψυχιατρικής και τη γνωσιοθεωρητική προβληματική γύρω από τη νόηση και το ιατρικό μοντέλο της ψυχιατρικής, όπως αυτό αποτυπώνεται στις κατηγοριοποιήσεις του DSM. Με λίγα λόγια, επιχειρεί να διερευνήσει τα όρια και τις προϋποθέσεις της έννοιας του «πάσχειν», όταν αυτή αποδίδεται σε μη ανθρώπινες οντότητες, όπως τα συστήματα Τεχνητής Νοημοσύνης. Μέσω αυτής της διερεύνησης, επιχειρεί να φωτίσει τις υποβόσκουσες παραδοχές που συνδέονται με τη διάγνωση, τη δυσλειτουργία και την εμπειρία, χωρίς να προδικάζει αν και πώς μια μηχανή μπορεί να θεωρηθεί ότι «πάσχει». Παράλληλα, συμπληρωματικός στόχος είναι να εξετασθεί, αν κάτω από κατάλληλες συνθήκες οι μηχανές μπορούν να εμφανίσουν ψυχοπαθολογία και αν η συμπτωματολογία που παρουσιάζουν, απλώς προσομοιάζει τις ανθρώπινες διαταραχές, με τη μορφή αλγοριθμικού σφάλματος.

Στο πλαίσιο αυτό, εξετάζεται κριτικά η διάκριση ανάμεσα στην ψυχική διαταραχή ως βιωματική κατάσταση και στη συμπτωματολογία ως λειτουργική απόκλιση από έναν κανόνα. Ιδιαίτερη έμφαση δίδεται στο ερώτημα, αν τα συμπτώματα μπορούν να αποκοπούν εννοιολογικά από το πάσχον υποκείμενο ή αν η ψυχική παθολογία προϋποθέτει κατ' ανάγκην την ύπαρξη υποκειμενικής εμπειρίας.

Το θεωρητικό μέρος της εργασίας συμπληρώνεται με την παρουσία ενός πειραματικού σκέλους, το οποίο εστιάζει στην εφαρμογή διαγνωστικών εργαλείων της ψυχιατρικής (DSM-5), σε ένα σύστημα Τεχνητής Νοημοσύνης [OnLine chatbot (ChatGPT)]. Ωστόσο, δεν αποτελεί αυτοσκοπό του «πειράματος» η εξαγωγή τεκμηρίων για την απόδειξη αν όντως το Αί δύναται να «πάσχει», αλλά να αναδείξει αν η ψυχιατρική ερμηνεία κατορθώνει να ξεφύγει από τα παραδοσιακά όριά της και κατά ποιον τρόπο επιτυγχάνεται η μετατόπιση της από ανθρώπινα όντα, σε τεχνητά μη ενσώματα όντα.

Στο πρώτο μέρος της εργασίας, και συγκεκριμένα στα κεφάλαια 2 και 3 επιχειρείται να αποδοθεί μια φιλοσοφική θεμελίωση της έννοια της ψυχικής διαταραχής. Ως πρότυπο, χρησιμοποιείται το παραδοσιακό ιατρικό μοντέλο διάγνωσης και οι κατηγοριοποιήσεις του DSM-5, με σκοπό να αναδειχθεί αν αποτελεί επαρκές σύστημα για την θεμελίωση των ψυχικών διαταραχών ή εναλλακτικά να γίνει μια στροφή προς ένα περισσότερο βιωματικό μοντέλο διάγνωσης, με επίκεντρο το ίδιο το υποκείμενο. Ιδιαίτερη έμφαση δίνεται στον ρόλο της διάγνωσης, στην έννοια της ενσώματης εμπειρίας και στον τρόπο που τα σύγχρονα ψυχιατρικά συστήματα ταξινόμησης διαμορφώνουν την έννοια του «παθολογικού». Τα κεφάλαια 2 και 3 αφιερώνονται στην ανάλυση του ιατρικού μοντέλου στην ψυχιατρική, στις βασικές του υποθέσεις και στις φιλοσοφικές του κριτικές, ιδίως σε σχέση με τη δυνατότητα εφαρμογής του σε μη ανθρώπινα όντα, όπως τα συστήματα Αί.

Στο κεφάλαιο 4, αναλύεται το πώς η πρόθεση, η εμπειρία, η ενσώματη συνείδηση και η γνωστική δυνατότητα σχετίζονται με την έννοια του «πάσχειν». Το ερώτημα της ενσώματης εμπειρίας τίθεται ως κρίσιμος παράγοντας για την κατανόηση της ψυχικής παθολογίας. Το κεφάλαιο 5 εστιάζει ειδικά στο ερώτημα αν και πώς θα μπορούσαμε να μιλήσουμε για «ψυχοπαθολογία» σε ένα μηχανικό σύστημα. Πότε μπορούμε να πούμε ότι μια μηχανή «παραληρεί»; Υπάρχει προσομοίωση συμπτώματος χωρίς υποκείμενο βίωσης;

Στο τρίτο μέρος της εργασίας, παρουσιάζεται η πειραματική μέθοδος σύμφωνα με την οποία ένα σύστημα Αί εκτίθεται σε ερεθίσματα που έχουν συνδεθεί με συγκεκριμένα ψυχιατρικά συμπτώματα (π.χ. παραληρηματικές δηλώσεις, αποπροσανατολισμένη χρήση γλώσσας, ασυνεπείς απαντήσεις, άγχος, δυσκολία συγκέντρωσης, καταθλιπτική διάθεση κ.α.). Στο πείραμα εφαρμόζονται τα διαγνωστικά εργαλεία του DSM, με σκοπό να ελεγχθεί αν τα παραγόμενα αποτελέσματα μπορούν να θεωρηθούν ως ενδείξεις κάποιας ψυχικής διαταραχής. Τα αποτελέσματα δεν οδηγούν σε συμπεράσματα περί πάθησης, αλλά εγείρουν ερωτήματα για τη φύση του συμπτώματος και την ενδεχόμενη αυτονομία του από το πάσχον υποκείμενο.

Η παρούσα εργασία φιλοδοξεί να εντάξει τη φιλοσοφία της ψυχιατρικής σε έναν δημιουργικό διάλογο με τη Γνωσιοθεωρία της Τεχνητής Νοημοσύνης, θέτοντας υπό εξέταση τις εδραιωμένες αντιλήψεις σχετικά με τη νοημοσύνη, την εμπειρία και την ψυχική παθολογία. Το ερώτημα αν ένα σύστημα Αί μπορεί να πάσχει δεν συνιστά αποκλειστικά τεχνικό ή κλινικό πρόβλημα. Αντίθετα, αγγίζει βαθιά ζητήματα ταυτότητας, καθώς σχετίζεται με τον τρόπο που ο άνθρωπος συγκροτεί την έννοια του εαυτού του σε σχέση με τον άλλο. «Αν τα συμπτώματα μπορούν να προσομοιωθούν, αλλά το πάσχειν όχι, πού ακριβώς κατοικεί η ψυχική διαταραχή;»

Μέρος Α' - Θεωρητικές προσεγγίσεις και εννοιολογικές προσεγγίσεις

2. Το Ιατρικό Μοντέλο

Το παρόν κεφάλαιο εξετάζει την εφαρμογή του ιατρικού μοντέλου στην ψυχιατρική, εστιάζοντας σε θεμελιώδεις έννοιες, όπως η ασθένεια, η νόσος και η ψυχική διαταραχή. Οι ορισμοί αυτών των εννοιών παραμένουν αμφιλεγόμενοι, καθώς στη σχετική βιβλιογραφία διατυπώνονται διαφορετικές απόψεις σχετικά με το αν και σε ποιον βαθμό εμπεριέχουν αξιακές κρίσεις και πολιτισμικές παραδοχές που διέπουν την άσκηση της ιατρικής επιστήμης (Simp. Psychol., 2024).

Παρά τις επικρίσεις που έχει δεχθεί, το ιατρικό μοντέλο εξακολουθεί να αποτελεί κεντρικό σημείο αναφοράς στην ψυχιατρική θεωρία και πράξη. Αυτό οφείλεται, μεταξύ άλλων, στο γεγονός ότι ορισμένες ψυχικές διαταραχές συνδέονται με βιολογικούς ή οργανικούς παράγοντες, οι οποίοι μπορεί να αφορούν γενετικές, νευροχημικές ή ενδοκρινικές δυσλειτουργίες (Kendler, 2012, σ. 124). Οι νευρολογικές παθήσεις, οι οποίες σχετίζονται κυρίως με δομικές αλλοιώσεις του εγκεφάλου, αποτελούν διακριτό αλλά συναφές πεδίο από την ψυχιατρική. Σε περιπτώσεις όπου τα ψυχικά συμπτώματα οφείλονται σε οργανικά ή νευρολογικά αίτια, η ιατρική γνώση είναι καθοριστική για την αναγνώριση της πραγματικής αιτιολογίας, την κατάλληλη θεραπευτική παρέμβαση και την αποφυγή εσφαλμένων ψυχιατρικών διαγνώσεων (McLeod, 2024). Συνεπώς, οι κλινικοί ιατροί διαδραματίζουν κρίσιμο ρόλο όχι μόνο στη θεραπεία, αλλά και στη διαδικασία της διαγνωστικής αξιολόγησης και αποκλεισμού οργανικών αιτίων στην ψυχιατρική πρακτική. Σε τέτοιες περιπτώσεις, η ιατρική γνώση είναι καθοριστική για την αναγνώριση της πραγματικής αιτίας, την κατάλληλη θεραπευτική παρέμβαση και την αποφυγή εσφαλμένων διαγνώσεων. Συνεπώς, οι επαγγελματίες της ιατρικής διαδραματίζουν κρίσιμο ρόλο όχι μόνο στην αντιμετώπιση, αλλά και στην ίδια τη διαδικασία αξιολόγησης και ταξινόμησης των ψυχιατρικών διαταραχών.

Η έρευνα, συνεπώς, αποσκοπεί στον εντοπισμό των βιολογικών αιτίων και στην ανάπτυξη θεραπειών που στοχεύουν στις εν λόγω δυσλειτουργίες. Από φιλοσοφική και ψυχοκοινωνική σκοπιά, ο Engel ασκεί κριτική στο ιατρικό μοντέλο,

επισημαίνοντας ότι αυτό παραμελεί τις κοινωνικές, ψυχολογικές και συμπεριφορικές διαστάσεις της ασθένειας, επικεντρώνοντας αποκλειστικά στις μετρήσιμες βιολογικές αποκλίσεις. (Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. Science, 196(4286), 129–136).

Παρά την ευρεία αποδοχή του, το ιατρικό μοντέλο στην ψυχιατρική δεν συγκροτεί ενιαία θεωρητική θέση, αλλά περιλαμβάνει διαφορετικές ερμηνευτικές εκδοχές. Στη σχετική φιλοσοφική συζήτηση, ο Murphy διακρίνει ανάμεσα σε μια «ισχυρή» και μια «ασθενή» εκδοχή του ιατρικού μοντέλου. Η ισχυρή εκδοχή προϋποθέτει ότι κάθε ψυχική διαταραχή θεμελιώνεται σε σαφώς προσδιορισμούς βιολογικούς μηχανισμούς και αιτιώδεις ακολουθίες, ενώ η ασθενής εκδοχή περιορίζεται στη νοσολογική και φαινομενολογική περιγραφή συνδρόμων, χωρίς δέσμευση ως προς την αιτιοπαθολογία (Murphy, 2006, σ. 4–6). Η διάκριση αυτή είναι κρίσιμη για την κατανόηση τόσο της λειτουργίας των διαγνωστικών εγχειριδίων όσο και των φιλοσοφικών ενστάσεων που έχουν διατυπωθεί απέναντι στο ιατρικό μοντέλο.

Στο πλαίσιο του ιατρικού μοντέλου, το Diagnostic and Statistical Manual of Mental Disorders (DSM) αποτελεί το πιο χαρακτηριστικό παράδειγμα τυποποίησης της ψυχοπαθολογίας. Το DSM συγκροτεί μια ταξινομική τυπολογία ψυχικών διαταραχών που λειτουργεί ως κοινός εννοιολογικός και διαγνωστικός κώδικας για επαγγελματίες ψυχικής υγείας σε διεθνές επίπεδο (American Psychiatric Association, 2022, σ. 5). Σύμφωνα με το ιατρικό μοντέλο, οι διαταραχές ορίζονται ως σύνδρομα στα οποία τα άτομα που μοιράζονται την ίδια διάγνωση εμφανίζουν ένα σύνολο συμπτωμάτων, με πιθανές επικαλύψεις ή διαφοροποιήσεις μεταξύ τους. Τα συμπτώματα αυτά αναμένεται να ακολουθούν έναν συγκρίσιμο τρόπο εξέλιξης και ανταπόκρισης στη θεραπεία, αν και αναγνωρίζονται σημαντικές ατομικές διαφορές.

Ωστόσο, είναι κρίσιμο να σημειωθεί ότι το DSM δεν υιοθετεί την «ισχυρή» εκδοχή του ιατρικού μοντέλου, όπως αυτή περιγράφεται από τον Murphy, διότι δεν προβαίνει σε αιτιακή παθολογική ερμηνεία των διαταραχών (Murphy, 2006, σ. 4–6). Η τυπολογία του DSM εντάσσεται περισσότερο στην «ασθενή» εκδοχή του ιατρικού μοντέλου, η οποία περιγράφει και ταξινομεί φαινοτυπικά σύνδρομα, χωρίς να δεσμεύεται ως προς τη βιολογική ή άλλη αιτιολογία τους. Το DSM δεν βασίζεται στην πλήρη βιολογική αναγωγή της ψυχικής διαταραχής. Θεωρεί τις ψυχικές

διαταραχές λειτουργικές δυσλειτουργίες, όχι απαραίτητα με εντοπίσιμους βιολογικούς μηχανισμούς. Επιπρόσθετα, αποδέχεται το γεγονός ότι η αιτιότητα είναι πολυπαραγοντική και η ταξινόμηση βασίζεται σε συμπτωματολογικά κριτήρια. Υπό αυτή την έννοια, οι διαγνωστικές κατηγορίες του DSM δεν αντιστοιχούν σε «φυσικά είδη» (*natural kinds*), ούτε συγκροτούν ομοιογενείς βιολογικές κατηγορίες, αλλά αποτελούν πρακτικά ταξινομητικά σχήματα βασισμένα στη φαινομενολογία των συμπτωμάτων (Kendler, 2016, σ. 2–4· Hyman, 2010, σ. 156–158).

Σύμφωνα με το DSM-5 (American Psychiatric Association, 2013), οι ψυχικές διαταραχές ορίζονται ως κλινικά σημαντικές διαταραχές στη νόηση, στη ρύθμιση του συναισθήματος ή στη συμπεριφορά, οι οποίες αντανακλούν δυσλειτουργία σε ψυχολογικές, βιολογικές ή αναπτυξιακές διεργασίες. Για τη διάγνωση απαιτείται η ύπαρξη σημαντικής δυσφορίας ή έκπτωσης της λειτουργικότητας (American Psychiatric Association, 2022, σ. 20). Επιπλέον, πολιτισμικά αναμενόμενα ή κοινωνικά αποκλίνοντα πρότυπα συμπεριφοράς δεν συνιστούν διαταραχή, εκτός εάν σχετίζονται με υποκείμενη δυσλειτουργία. Στο συγκεκριμένο πλαίσιο, η διάγνωση λειτουργεί ως εργαλείο αξιόπιστης ταξινόμησης και κλινικής επικοινωνίας, χωρίς να προϋποθέτει αιτιακή ερμηνεία ή βιολογική αναγωγή των διαταραχών. Επομένως, το DSM υιοθετεί μια φαινομενολογική και περιγραφική προσέγγιση, η οποία δεν προϋποθέτει αιτιοπαθολογικό μοντέλο με την “ισχυρή” έννοια που περιγράφει ο Murphy (2006, σ. 3–7, 40–42).

Πολλές από τις φιλοσοφικές κριτικές στρέφονται κυρίως εναντίον της ισχυρής εκδοχής του ιατρικού μοντέλου. Συγκεκριμένα, αρκετές ψυχικές διαταραχές δεν συνοδεύονται από αναγνωρισμένους βιολογικούς δείκτες ή από σαφώς προσδιορίσιμες αιτιακές ακολουθίες. Η απουσία αντικειμενικών βιοδεικτών σε διαταραχές όπως η μείζων καταθλιπτική διαταραχή ή η σχιζοφρένεια καθιστά δυσχερή την πλήρη βιοϊατρική θεμελίωση της ισχυρής εκδοχής και αναδεικνύει την ανάγκη διάκρισης μεταξύ περιγραφικής ταξινόμησης και βιολογικής εξήγησης. Επιπλέον, η έννοια της «δυσλειτουργίας» παραμένει εννοιολογικά προβληματική, στον βαθμό που προϋποθέτει κανονιστικές κρίσεις σχετικά με το ποια λειτουργία θεωρείται φυσιολογική και ποια παθολογική. Η φιλοσοφία της ψυχιατρικής έχει αναδείξει τον ρόλο κοινωνικών, ιστορικών και πολιτισμικών παραμέτρων στη

διαμόρφωση των διαγνωστικών κατηγοριών, υπογραμμίζοντας ότι το «πάσχειν» δεν εξαντλείται στην αντικειμενική καταγραφή συμπτωμάτων, αλλά περιλαμβάνει και τη βιωματική και διαπροσωπική διάσταση του υποκειμένου (Hacking, 1999· Fulford et al., 2013).

Παρά τις όποιες ενστάσεις, το ιατρικό μοντέλο εξακολουθεί να αποτελεί ένα ιδιαίτερα ισχυρό εργαλείο κατανόησης και θεραπευτικής αντιμετώπισης των ψυχικών διαταραχών. Όπως υποστηρίζει ο Kenneth Kendler (2012, σ. 124), η πλήρης κατανόηση των ψυχικών φαινομένων προϋποθέτει την ταυτόχρονη αναγνώριση βιολογικών παραμέτρων και τη μελέτη των βιωμένων εμπειριών, γεγονός που υποδηλώνει μια πλουραλιστική και μη αναγωγική προσέγγιση της ψυχοπαθολογίας. Ωστόσο, σε επίπεδο κλινικής πρακτικής και επιστημονικής οργάνωσης της γνώσης, το ιατρικό μοντέλο εστιάζει πρωτίστως στην τυποποίηση των ψυχικών παθήσεων, στη δυνατότητα συστηματικής κλινικής τους αποτύπωσης και, ει δυνατόν, στη θεραπευτική τους αντιμετώπιση κατά τρόπο προβλέψιμο και επαναλήψιμο.

Συνοψίζοντας, καθίσταται φανερό ότι το ιατρικό μοντέλο συνιστά ένα βασικό επιστημονικό πλαίσιο για την οργάνωση, την ταξινόμηση και τη συστηματική περιγραφή της ψυχικής παθολογίας, συμβάλλοντας καθοριστικά στη διαμόρφωση της ψυχιατρικής ως διακριτής κλινικής και επιστημονικής ειδικότητας. Ωστόσο, εγείρει θεμελιώδεις φιλοσοφικές και γνωσιοθεωρητικές ενστάσεις, καθώς τείνει να μετατρέπει τις ψυχικές διαταραχές σε ταξινομήσιμα σύνολα συμπτωμάτων, αποσυνδεδεμένα από το προσωπικό και κοινωνικό τους υπόβαθρο. Η βιοϊατρική ερμηνεία της απόκλισης συχνά παραμελεί τη βιωματική και εννοιολογική πολυπλοκότητα του «πάσχειν». Στο πλαίσιο αυτό, η παρούσα εργασία υιοθετεί το ιατρικό μοντέλο ως αναλυτικό εργαλείο για τη διερεύνηση της πιθανότητας εμφάνισης ψυχικών διαταραχών στην τεχνητή νοημοσύνη, με κριτική επίγνωση των επιστημολογικών ενστάσεων που έχει διατυπώσει η φιλοσοφία της ψυχιατρικής σχετικά με τη φύση της διάγνωσης και της παθολογίας. Η υιοθέτηση αυτή επιτρέπει τη μελέτη των εννοιολογικών ορίων του «πάσχειν» πέραν των ανθρώπινων όντων.

Τέλος, καθίσταται αναγκαία η εμβάθυνση σε μία από τις πιο χαρακτηριστικές εκφάνσεις του ιατρικού μοντέλου: την ταξινόμηση των ψυχικών διαταραχών μέσα

από διαγνωστικά εγχειρίδια όπως το *Diagnostic and Statistical Manual of Mental Disorders* (DSM). Η προσπάθεια κατηγοριοποίησης της ψυχοπαθολογίας βασίζεται σε περιγραφικά και λειτουργικά κριτήρια, τα οποία οργανώνουν τα συμπτώματα σε διαγνωστικά σύνδρομα και τα συσχετίζουν με μορφές δυσφορίας ή λειτουργικής έκπτωσης, χωρίς να προϋποτίθεται ή να διατυπώνεται ρητή αιτιολογική εξήγηση της διαταραχής. Αυτή η προσέγγιση αντιστοιχεί στην ασθενή εκδοχή του ιατρικού μοντέλου (Weak Medical Model) κατά τον Murphy, όπου η διαταραχή αναγνωρίζεται κυρίως μέσω της λειτουργικής ή συμπεριφορικής δυσλειτουργίας. Αντίθετα, η ισχυρή εκδοχή (Strong Medical Model) προϋποθέτει τη γνώση της αιτιολογίας, δηλαδή των πραγματικών βιολογικών ή νευρολογικών αιτίων που καθορίζουν την ψυχική διαταραχή. Με αυτόν τον τρόπο, η χρήση ταξινομητικών συστημάτων, όπως το DSM, καθορίζει τον τρόπο με τον οποίο η ψυχιατρική ερμηνεύει την ψυχική απόκλιση, διαχωρίζοντας την αιτιοκρατική αναγνώριση και τυποποίηση της δυσλειτουργίας από την αιτιολογική αναζήτηση των υποκείμενων αιτίων, όπως περιγράφεται στην «ισχυρή» εκδοχή του ιατρικού μοντέλου.

Εφόσον το ιατρικό μοντέλο υποθέτει ότι κάθε ψυχική διαταραχή μπορεί να συνδέεται με συγκεκριμένες βιολογικές παραμέτρους και δυνητικά αιτιώδεις μηχανισμούς, τίθεται το ερώτημα πώς αυτή η υπόθεση εφαρμόζεται στην πράξη. Η ψυχιατρική, για να αποκτήσει επιστημονική συνέπεια, επιδίωξε να οργανώσει τη γνώση της γύρω από σταθερά συστήματα ταξινόμησης, τα οποία να καταγράφουν και να συστηματοποιούν τα παρατηρούμενα συμπτώματα. Στο πλαίσιο αυτό, αναπτύχθηκαν διαγνωστικά εγχειρίδια, όπως το *Diagnostic and Statistical Manual of Mental Disorders* (DSM). Το DSM είναι αθεωρικό και παρέχει μια τυποποιημένη/ φαινομενολογική γλώσσα για την ψυχιατρική διάγνωση, η οποία λειτουργεί ως πρακτικό εργαλείο κατηγοριοποίησης, αλλά δεν ταυτίζεται με την ισχυρή βιοϊατρική θεώρηση που συνεπάγεται πως η αιτιολογικοί παράγοντες είναι συγκροτητικοί της ψυχοπαθολογικής ταξινόμησης. Η διάκριση αυτή ανάμεσα στην «ισχυρή» εκδοχή της ταξινόμησης, που προϋποθέτει αιτιακή απόδοση, και στην «ασθενή» εκδοχή, που περιορίζεται στη φαινομενολογική περιγραφή χωρίς δέσμευση ως προς τα αίτια, αναδεικνύει τα όρια της εφαρμογής του ιατρικού μοντέλου και τα φιλοσοφικά προβλήματα που συνδέονται με τη φύση της ψυχικής διαταραχής, τα οποία θα εξεταστούν στην επόμενη ενότητα.

2.1 Η ασθενής εκδοχή του ιατρικού μοντέλου και η ταξινόμηση στο DSM

Το DSM (Diagnostic and Statistical Manual of Mental Disorders) αποτελεί το βασικό εργαλείο ταξινόμησης ψυχικών διαταραχών στην ψυχιατρική. Παρέχει τυποποιημένα και κοινά αποδεκτά κριτήρια για τη διάγνωση και την ταξινόμηση, με στόχο την αξιοπιστία μεταξύ διαφορετικών αξιολογητών, χωρίς να προϋποθέτει αιτιολογική βάση για τις διαταραχές και χωρίς να αποσκοπεί στην υπόδειξη των αιτιών τους. Η λειτουργία του DSM είναι περιγραφική και αιτιολογικά ουδέτερη, καθιστώντας το εργαλείο κυρίως ταξινομικό. (Kendler, 2012, σ. 45–46).

Στο πλαίσιο του ιατρικού μοντέλου, που αντιμετωπίζει τις ψυχικές διαταραχές ως νοσολογικές οντότητες με συγκεκριμένη αιτιολογία, το DSM λειτουργεί ως βασικό εργαλείο ταξινόμησης, με την έννοια ότι καταγράφει και κατηγοριοποιεί συμπτώματα χωρίς να υποθέτει αιτιολογική σχέση (Murphy, 2006, σ. 3–7). Το DSM υιοθετεί κυρίως μια αιτιολογικά ουδέτερη, περιγραφική στάση, παρέχοντας επαληθεύσιμα κριτήρια για τη διάγνωση και την ταξινόμηση, αλλά εντάσσεται στο ευρύτερο πλαίσιο του ιατρικού μοντέλου, καθώς υποστηρίζει τη διάγνωση ως εργαλείο αναγνώρισης δυσλειτουργιών και σχεδιασμού θεραπείας (Engel, 1977, σ. 129–136).

Ωστόσο, έχει ασκηθεί έντονη κριτική καθώς πολλές φορές διαφαίνεται ότι η υποκειμενική εμπειρία δεν αποτελεί άμεσα αντικείμενο ταξινόμησης, με αποτέλεσμα να δίνεται έμφαση αποκλειστικά και μόνο σε εξωγενή παρατηρήσιμα γεγονότα (Schrag, 2015; Bolton, 2008). Επιπρόσθετα, η τυποποιημένη δομή του ορίζει οποιαδήποτε απόκλιση από το «κανονικό», σε ψυχική διαταραχή, γεγονός που εγείρει ποικίλες αμφιβολίες και αντιρρήσεις (Hacking, 1999, σ. 4–6). Στη σχετική βιβλιογραφία, κάθε απόκλιση από ένα «κανονικό» ψυχικό πρότυπο συνιστά ψυχική διαταραχή. Ο ορισμός της έννοιας του «κανονικού» μπορεί να φαίνεται ασαφής, από την άποψη ότι ίσως αφορά είτε στατιστική συχνότητα, είτε κοινωνικές προσδοκίες, είτε λειτουργική επάρκεια. Ωστόσο, στην παρούσα εργασία η έννοια του «κανονικού» δεν ταυτίζει την απόκλιση με την ψυχική διαταραχή, αλλά εξετάζει τις προϋποθέσεις υπό τις οποίες μια απόκλιση μπορεί να θεωρηθεί παθολογική στο πλαίσιο του ιατρικού μοντέλου. Σύμφωνα με την ισχυρή εκδοχή του ιατρικού μοντέλου, η οποία δεν εφαρμόζεται στο DSM, κάθε ψυχική διαταραχή αντιμετωπίζεται ως αποτέλεσμα συγκεκριμένων, ανιχνεύσιμων και καθορισμένων αιτιών — κυρίως βιολογικών ή

γενετικών (Kendler, 2016, σ. 2–4). Η οντολογική υπόθεση ότι οι ψυχικές διαταραχές αποτελούν «πράγματα που υπάρχουν» ανεξαρτήτως πολιτισμικού ή βιοματικού πλαισίου αφορά μόνο την ισχυρή εκδοχή και δεν αποτελεί κριτήριο του DSM (Murphy, 2006, σ. 5–6).

Από το DSM-I (1952) έως το DSM-III (1980), η ψυχιατρική ταξινόμηση μετατοπίστηκε από την ψυχοδυναμική θεώρηση σε ένα εμπειρικό και περιγραφικό μοντέλο (Spitzer, 1980, σ. 12–15). Οι μεταγενέστερες εκδόσεις (DSM-IV, DSM-5) επιδίωξαν μεγαλύτερη αξιοπιστία, ενσωματώνοντας νευροβιολογικά δεδομένα, χωρίς να δηλώνεται ότι υπάρχει αιτιολογική βάση και τροποποιώντας τις κατηγορίες, χωρίς ωστόσο να εξαλείψουν πλήρως τα προβλήματα εγκυρότητας και συννοσηρότητας. Η συχνή εμφάνιση συννοσηρότητας (Frances, 1994, σ. 22–24). θέτει υπό αμφισβήτηση την εγκυρότητα των ψυχιατρικών διαγνώσεων (Insel et al., 2010, σ. 748). Αν με άλλα λόγια, οι διαγνωστικές κατηγορίες ήταν όντως διακριτές, δεν θα παρατηρούνταν τόσο συχνά η ταυτόχρονη διάγνωσή τους στο ίδιο άτομο. Η ιστορική εξέλιξη των διαγνωστικών συστημάτων στις ΗΠΑ αντικατοπτρίζει την προσπάθεια της ψυχιατρικής να αποκτήσει αξιοπιστία και επιστημονική εγκυρότητα. Σύμφωνα με τα εγχειρίδια του DSM, οι ψυχικές διαταραχές ορίζονται ιατρικά ως συμπεριφορικά ή ψυχολογικά πρότυπα που προκαλούν δυσφορία ή δυσλειτουργία και όχι απλώς πολιτισμικά αναμενόμενες αντιδράσεις σε γεγονότα, όπως το πένθος (APA, 2013, σ. 20–21).

Στα μέσα του 20ου αιώνα, το Σεντ Λούις γνώρισε μια μεταμόρφωση στην αμερικανική ψυχιατρική με επικεφαλής τους Eli Robins και Samuel Guze από τη «Σχολή Renard» του Πανεπιστημίου της Ουάσιγκτον (Guze & Robins, 1970, σ. 150–155). Απέρριψαν την επικρατούσα ψυχαναλυτική προσέγγιση υπέρ της ανάπτυξης αξιόπιστων, έγκυρων διαγνωστικών κριτηρίων για τις ψυχιατρικές διαταραχές, αντιμετωπίζοντας τη δυσaréσκεια με τις προηγούμενες μεθόδους ταξινόμησης που βασίζονταν στην κλινική γνώμη. Οι Robins και Guze καθιέρωσαν ένα μοντέλο διαγνωστικής επικύρωσης πέντε φάσεων που επικεντρώθηκε στα κλινικά χαρακτηριστικά, στα κριτήρια αποκλεισμού, στις οικογενειακές μελέτες, στα εργαστηριακά δεδομένα και στις μελέτες παρακολούθησης. Μέχρι τη δεκαετία του 1950, στις ΗΠΑ υπήρχαν πέντε διακριτά συστήματα διαγνωστικής ταξινόμησης. Σε αυτό το σημείο, η Αμερικανική Ψυχιατρική Ένωση δημοσίευσε το DSM το 1952. Συγκεκριμένα, περιλάμβανε δύο τμήματα: το πρώτο αφορούσε τις οργανικές εγκεφαλικές παθήσεις και το δεύτερο τις λειτουργικές

διαταραχές, που υποδιαιρούνται σε ψυχώσεις, ψυχονευρώσεις και διαταραχές της προσωπικότητας και φέρουν ψυχοδυναμική αιτιώδη συνάφεια. Το DSM-II, που κυκλοφόρησε το 1968, στόχευε σε μεγαλύτερη συμβατότητα με τα διεθνή συστήματα, επεκτεινόμενο σε δέκα διαγνωστικές ενότητες, συμπεριλαμβανομένης μιας ενότητας για τα παιδιά. Σημαντική αλλαγή είναι η απόδειξη της εμφάνισης διαγνωστικής συννοσηρότητας. Στη δεκαετία του 1980, ο Robert Spitzer επέβλεψε το DSM-III-R, το οποίο βελτίωσε τα κριτήρια αλλά αφαίρεσε τη διαγνωστική ιεραρχία, αυξάνοντας τα ευρήματα συννοσηρότητας (Spitzer, 1980, σ. 20–25). Η διαγνωστική ιεραρχία λειτουργούσε περιοριστικά, καθώς μια διάγνωση μπορούσε να αποκλείει ή να υπερισχύει έναντι άλλων. Με την κατάργησή της, τα ίδια συμπτώματα άρχισαν να πληρούν τα κριτήρια πολλαπλών διαγνώσεων, γεγονός που οδήγησε σε αύξηση των περιπτώσεων συννοσηρότητας.

Το 1988, συγκροτήθηκε μια ομάδα εργασίας του DSM-IV, με επικεφαλής τον Allen Frances, για να αναθεωρήσει τα κριτήρια του DSM (Frances, 1994, σ. 30–35). Το DSM-IV δημοσιεύθηκε το 1994 με μικρές αλλαγές, κυρίως με την προσθήκη της «κλινικά σημαντικής δυσφορίας ή δυσλειτουργίας» στα διαγνωστικά κριτήρια. Το 1999 ξεκίνησε ο σχεδιασμός του DSM-5, υπό την προεδρία του David Kupfer, με έμφαση στην ενσωμάτωση της νευροβιολογικής έρευνας και τη βελτίωση της κλινικής χρησιμότητας, χωρίς όμως να μετατρέπουν το DSM σε αιτιολογικό εργαλείο (Kupfer et al., 2013, σ. 14–16). Το DSM-5 εξάλειψε το πολυαξονικό σύστημα λόγω των ασυνεπειών του, και τελικά δημοσιεύθηκε το 2013. Ο σκοπός των διαγνωστικών κριτηρίων έχει εξελιχθεί, αρχικά με στόχο τη συλλογή στατιστικών πληροφοριών για τις ψυχικές διαταραχές στα DSM-I και DSM-II. . Με την πάροδο των εκδόσεων, διάφορες διαταραχές εμφανίστηκαν ή αφαιρέθηκαν. Ειδικότερα, η διαταραχή μετατραυματικού στρες και η οριακή διαταραχή προσωπικότητας προστέθηκαν στο DSM-III, ενώ η διαταραχή οξέος στρες, η διπολική διαταραχή II και η διαταραχή Asperger εμφανίστηκαν στο DSM-IV. Το DSM-5 επαναταξινόμησε τη διαταραχή Asperger στη διαταραχή του φάσματος του αυτισμού και αφαίρεσε τη διαταραχή σωματοποίησης, αντανakλώντας τις συνεχιζόμενες διαγνωστικές αλλαγές, ιδίως όσον αφορά την ομοφυλοφιλία. Ωστόσο, η εγκυρότητα και η αξιοπιστία πολλών διαγνώσεων παραμένουν αδιευκρίνιστες. Στην παρούσα εργασία, χρησιμοποιούνται τακτικά οι όροι «εγκυρότητα» και «αξιοπιστία», όπως ορίζονται στην φιλοσοφία της ψυχιατρικής. Η εγκυρότητα αναφέρεται στο βαθμό στον οποίο μια διάγνωση περιγράφει ακριβώς ένα

διακριτό ψυχικό φαινόμενο, ενώ η αξιοπιστία αφορά την ικανότητα διαφορετικών κλινικών να καταλήγουν στην ίδια διάγνωση για το ίδιο άτομο.

Συνολικά, η ιστορία των διάφορων εκδόσεων του DSM επικυρώνει την εξέλιξη από ένα ημιθεραπευτικό μοντέλο (DSM I-II), σε ένα διαγνωστικό-περιγραφικό εγχειρίδιο βασισμένο στην παρατήρηση, στην κατηγοριοποίηση και την τυποποίηση των συμπτωμάτων (DSM III), επηρεάζοντας τις έννοιες της «ψυχικής διαταραχής», της αιτιότητας και της θεραπείας. Ο όρος «ημιθεραπευτικό μοντέλο» αναφέρεται στις πρώτες εκδόσεις του DSM (I και II), οι οποίες δεν ήταν μόνο ταξινομητικά εγχειρίδια, αλλά είχαν κυρίως θεραπευτικό προσανατολισμό, βοηθώντας τους κλινικούς στη διάγνωση και τη θεραπεία των ψυχικών διαταραχών. Το DSM-I αντικατόπτριζε τη θεραπευτική και στρατιωτική ανάγκη της εποχής (μετά τον Β' Παγκόσμιο Πόλεμο) και ήταν σε μεγάλο βαθμό επηρεασμένο από την ψυχαναλυτική θεωρία. Ο όρος «ψυχική διαταραχή» δεν είχε αυστηρό ορισμό και οι περισσότερες κατηγορίες περιείχαν τον όρο "αντίδραση", π.χ. *"νευρωτική αντίδραση"*, *"ψυχωτική αντίδραση"*. Παράλληλα, δεν υπήρχε συστηματική διάκριση μεταξύ σημείων και συμπτωμάτων και τυποποιημένη διαγνωστική διαδικασία. Η εστίαση δεν ήταν τόσο στη συμπτωματολογία, όσο στην ερμηνεία της ψυχοπαθολογίας. Το DSM-II συνέχισε την ίδια βασική ψυχοδυναμική προσέγγιση, παρατηρώντας όμως σημαντικές ασυνέπειες όπως, ότι δεν ορίζονται βάσει αντικειμενικών περιγραφών, αλλά περισσότερο με θεωρητικούς όρους, με μη σαφή διαγνωστικά κριτήρια. Τέλος, η διάγνωση εξαρτάται σε μεγάλο βαθμό από την υποκειμενική κρίση του ψυχιάτρου. Το DSM-III αποτέλεσε ριζική τομή στην ιστορία της ψυχιατρικής. Υπό την καθοδήγηση του Robert Spitzer, εισάγεται ένα νέο, τυποποιημένο και περιγραφικό μοντέλο, που βασίζεται σε παρατηρήσιμα συμπτώματα και όχι σε θεωρητικές ερμηνείες. Τα κύρια χαρακτηριστικά που διατυπώθηκαν συνοψίζονται ως εξής: Η ταξινόμηση δεν βασίζεται σε θεωρίες (π.χ. ψυχαναλυτικές ή βιολογικές), αλλά σε κλινικά παρατηρούμενες εκδηλώσεις. Εισάγονται πέντε άξονες αξιολόγησης, που επιχειρούν να εντάξουν την πολυπλοκότητα της ψυχικής κατάστασης ενός ατόμου (π.χ. κλινική εικόνα, προσωπικότητα, σωματική υγεία, ψυχοκοινωνικοί στρεσογόνοι παράγοντες, λειτουργικότητα).

Το DSM-5 αποτελεί την πλέον πρόσφατη επίσημη έκδοση του εγχειριδίου, με εκτεταμένες τροποποιήσεις. Μια βασική τροποποίησή του αφορά την ενοποίηση των πέντε αξόνων που χρησιμοποιήθηκαν στις προηγούμενες εκδόσεις, σε μια ενιαία διαγνωστική κατηγορία, με σκοπό την απλοποίηση. Επίσης, αν και η διαγνωστική

διαδικασία εξακολουθεί να βασίζεται στην συμπτωματολογία των ασθενών, η επικρατούσα έκδοση εισάγει μια νέα παράμετρο, τις κλίμακες βαρύτητας. Οι κλίμακες αυτές σηματοδοτούν την μετατόπιση από ένα αυστηρό δίπολο διάγνωσης (παρουσία ή απουσία διαταραχής) σε μια διαβαθμισμένη προσέγγιση, η οποία αποτυπώνει τον βαθμό έντασης, τη συχνότητα και τη λειτουργική επίδραση των συμπτωμάτων. Ωστόσο, οι κλίμακες βαρύτητας δεν αναιρούν τον θεμελιωδώς ερμηνευτικό χαρακτήρα της διάγνωσης, καθώς εξακολουθούν να βασίζονται σε κλινικές κρίσεις, αυτοαναφορές και γλωσσικές περιγραφές της εμπειρίας του υποκειμένου. Παράλληλα, το DSM-5 συνεχίζει να λειτουργεί ως εγχειρίδιο ταξινόμησης με βάση το ιατρικό πρότυπο, δηλαδή προσεγγίζοντας τις ψυχικές διαταραχές ως παθολογικές καταστάσεις με σαφές σύνολο συμπτωμάτων και υιοθετώντας τη λογική της διάγνωσης και της τυποποίησης.

Στην παρούσα ενότητα σκοπός αποτελεί ο ορισμός και ιστορική εξέλιξη των ψυχικών Διαταραχών στο DSM: Σχιζοφρένεια και Διπολική Διαταραχή, ώστε να μπορούν να αξιοποιηθούν ως υπόβαθρο για το διερευνητικό σενάριο χρήσης του Αι. Η σχιζοφρένεια χαρακτηρίζεται από παραληρητικές ιδέες, ψευδαισθήσεις, αποδιοργανωμένη σκέψη, συναισθηματική επιπέδωση, αβουλία και κοινωνική απόσυρση, με τουλάχιστον δύο συμπτώματα διάρκειας άνω των έξι μηνών. Η διπολική διαταραχή περιλαμβάνει επεισόδια μανίας και κατάθλιψης με διαλείπουσα πορεία και πλήρη ή μερική αποκατάσταση λειτουργικότητας, με χαρακτηριστικά όπως αυξημένη ενεργητικότητα, ευερεθιστότητα, μειωμένη ανάγκη ύπνου και ιδέες αυτοκαταστροφής. Αποτελούν ιδανικές περιπτώσεις για υποθετική προσομοίωση στο Αι, καθώς διαθέτουν σαφή διαγνωστικά κριτήρια (DSM-5), τα οποία μπορούν να προσομοιωθούν και εμπλέκουν υψηλού επιπέδου γνωσιακές και συναισθηματικές δυσλειτουργίες, χρήσιμες για να διερευνηθεί αν το Αι μπορεί να παρουσιάσει κάτι ανάλογο.

Θα πρέπει ωστόσο να σημειωθεί ότι, ιδιαίτερα στις σοβαρές μορφές τους, η διπολική διαταραχή και η σχιζοφρένεια παρουσιάζουν σημαντικές κλινικές επικαλύψεις. Στην πράξη, η διάκριση μεταξύ τους δεν βασίζεται πάντα αποκλειστικά στα διαγνωστικά κριτήρια του DSM, αλλά συχνά και στην ανταπόκριση του ασθενούς στη θεραπεία. Η επιλογή των δύο αυτών διαταραχών στο πλαίσιο της παρούσας εργασίας δεν αποσκοπεί στην αναπαράσταση απολύτως διακριτών κατηγοριών, αλλά στη χρήση ενδεικτικών παραδειγμάτων ψυχικών διαταραχών με σύνθετα και εν μέρει αλληλοεπικαλυπτόμενα χαρακτηριστικά, κατάλληλων για το διερευνητικό σενάριο των δυνατοτήτων του Αι.

Η σχιζοφρένεια και η διπολική διαταραχή αποτελούν δύο από τις πιο συζητημένες ψυχικές διαταραχές, λόγω της σοβαρότητας και της πολυπλοκότητας των συμπτωμάτων τους. Η σχιζοφρένεια περιεγράφηκε αρχικά από τον Kraepelin ως *Dementia Praecox* και επαναπροσδιορίστηκε από τον Bleuler (1911), ο οποίος εισήγαγε τον όρο *schizophrenia*. Σύμφωνα με το DSM-5, η σχιζοφρένεια χαρακτηρίζεται από την παρουσία τουλάχιστον δύο από τα ακόλουθα συμπτώματα: παραληρητικές ιδέες, ψευδαισθήσεις, αποδιοργανωμένο λόγο ή σκέψη, έντονα αποδιοργανωμένη ή κατατονική συμπεριφορά και αρνητικά συμπτώματα, όπως συναισθηματική επιπέδωση και αβουλία. Τα συμπτώματα πρέπει να επιμένουν για χρονικό διάστημα τουλάχιστον έξι μηνών και να συνοδεύονται από σημαντική έκπτωση της κοινωνικής ή επαγγελματικής λειτουργικότητας.

Αντίστοιχα, η διπολική διαταραχή, η οποία περιεγράφηκε από τον Kraepelin ως μανιοκαταθλιπτική ψύχωση, ορίζεται στο DSM-5 ως διαταραχή της διάθεσης με επεισόδια μανίας και κατάθλιψης. Τα μανιακά επεισόδια περιλαμβάνουν συμπτώματα όπως επίμονα ανυψωμένη ή ευερέθιστη διάθεση, αυξημένη ενεργητικότητα, μειωμένη ανάγκη για ύπνο, μεγαλομανία, επιταχυνόμενη σκέψη και παρορμητική συμπεριφορά. Τα καταθλιπτικά επεισόδια χαρακτηρίζονται από επίμονη θλίψη, μειωμένη ενεργητικότητα, απώλεια ενδιαφέροντος, αισθήματα ενοχής και ιδέες αυτοκαταστροφής. Η πορεία της διαταραχής είναι συνήθως διαλείπουσα, με πλήρη ή μερική αποκατάσταση της λειτουργικότητας μεταξύ των επεισοδίων.

Η επιλογή αυτών των δύο ψυχικών διαταραχών εξυπηρετεί την ανάλυση του «πάσχειν» στο Αι. Η σχιζοφρένεια θέτει το ερώτημα κατά πόσο ένα τεχνητό σύστημα χωρίς αυτοσυνείδηση μπορεί να αναπτύξει παραληρηματικές σκέψεις, ενώ η διπολική διαταραχή διερευνά τη δυνατότητα εναλλαγής διάθεσης σε ένα μηχανικό σύστημα, χωρίς ενσώματη εμπειρία. Η τυποποιημένη αυστηρότητα των διαγνωστικών κριτηρίων του DSM καθιστά δυνατή την αναπαράσταση των ψυχικών διαταραχών σε ένα σύστημα Τεχνητής Νοημοσύνης, όχι με την έννοια της αυθεντικής παθολογικής εμπειρίας, αλλά με την έννοια της προσομοίωσης, χωρίς να συνεπάγεται απαραίτητα η ουσιαστική αντιστοιχία (Insel et al., 2010, σ. 748). Με τον όρο «αυθεντική παθολογική εμπειρία» αναφερόμαστε στη βιωματική και υποκειμενική εμπειρία της ψυχικής νόσου, όπως αυτή γίνεται αντιληπτή και βιώνεται από τον άνθρωπο, και όχι σε μια τυπική ή υπολογιστική αναπαράστασή της. Αυτό εγείρει φιλοσοφικά και

επιστημολογικά ερωτήματα: ποιες πτυχές της ψυχικής διαταραχής είναι εξωτερικά αναπαραστάσιμες και ποιες απαιτούν ενσώματη εμπειρία;

Όπως γίνεται αντιληπτό από την ιστορική αναδρομή και την τελική έκδοση του DSM, η ταξινόμηση που προτείνεται, βασίζεται σε διαγνωστικά κριτήρια που καθορίζονται από διαγνωστικά κριτήρια συμπτωματολογίας, χωρίς να προϋποτίθεται η οριστική απόδοση της αιτίας. Ωστόσο, στο πλαίσιο του ιατρικού μοντέλου της ψυχιατρικής, η ανεύρεση των αιτιών των ψυχικών διαταραχών θεωρείται κεντρικός στόχος της διαγνωστικής κατηγοριοποίησης, με απώτερο σκοπό τον εντοπισμό αξιόπιστων βιολογικών δεικτών (Insel et al., 2010, σ. 748). Παρότι το DSM αποφεύγει να υιοθετήσει ρητά μια θεωρία αιτιών, μέσω της τυπολογίας και της ταξινόμησης ενυπάρχει ένα αντικειμενικό υπόστρωμα, καθώς κάθε διαταραχή αντιστοιχεί σε ένα σταθερό σύνολο συμπτωμάτων που παρατηρούνται σε πολλούς ασθενείς και επιτρέπουν την αναπαραγωγικότητα της διάγνωσης (APA, 2013, σ. 20–21), χωρίς να υπονοεί ότι η αιτία είναι μοναδική. Η τυποποιημένη δομή των συμπτωμάτων και η αιτιοκρατική προσέγγιση στη διάγνωση ενισχύει την άποψη περί προσομοίωσης των διαταραχών σε συστήματα Αι. Δηλαδή, η ύπαρξη αιτιακών σχέσεων μεταξύ συμπτωμάτων επιτρέπει την αλγοριθμική αναπαράσταση, χωρίς να συνεπάγεται ότι η αιτιολογία των διαταραχών είναι γνωστή. Παραταύτα, αν και η αιτιοκρατική δομή διευκολύνει την τυποποίηση και τον προγραμματισμό, δεν μπορεί να αποκτήσει πρόσβαση στη βιωματική εμπειρία και στην κατανόηση του «πάσχειν».

Η χρήση του DSM στα συστήματα Τεχνητής Νοημοσύνης αποσκοπεί στην αναπαράσταση συμπτωμάτων και όχι στην αναπαραγωγή της βιωματικής εμπειρίας της νόσου. Παρά την τυποποιημένη δομή των διαγνωστικών κριτηρίων, η οποία ακολουθεί αιτιακά πρότυπα μεταξύ συμπτωμάτων (δηλαδή προβλέπει την εμφάνιση ή συνύπαρξη ορισμένων συμπτωμάτων με άλλα), το DSM παραμένει περιγραφικό εργαλείο. Δεν μπορεί να αποδώσει τη βιωματική, υποκειμενική διάσταση του «πάσχειν», δηλαδή πώς βιώνει ο ίδιος ο ασθενής την ψυχική διαταραχή. Έτσι, η εφαρμογή του στο Αι αποκαλύπτει τα όρια της ιατρικής ταξινόμησης και αναδεικνύει την ανάγκη φιλοσοφικής διερεύνησης της ψυχικής εμπειρίας.

Παρά την αντικειμενικότητα και την τυποποίηση που παρέχει το DSM, η έμφαση αποκλειστικά στα συμπτώματα αφήνει ανοιχτά θεμελιώδη φιλοσοφικά ερωτήματα για τη φύση της ψυχικής διαταραχής και την εμπειρία του «πάσχειν» (Hacking, 1999, σ.

4–6; Fulford et al., 2013, σ. 15–18). Ακριβώς σε αυτό το σημείο εισέρχεται η φιλοσοφία της ψυχιατρικής, ασκώντας κριτική στα θεμέλια και τους περιορισμούς του ιατρικού μοντέλου.

2.2 Αντιρρήσεις της φιλοσοφίας της ψυχιατρικής

Παρότι το DSM παρέχει σαφή διαγνωστικά κριτήρια για τη σχιζοφρένεια και τη διπολική διαταραχή, η φιλοσοφία της ψυχιατρικής εγείρει σοβαρές ενστάσεις ως προς την εγκυρότητα, τη φύση και την κοινωνική διάσταση των ψυχικών διαταραχών (Bracken & Thomas, 2001; Fulford et al., 2006). Οι αντιρρήσεις αυτές αμφισβητούν όχι μόνο το τι θεωρείται παθολογικό, αλλά και τον τρόπο με τον οποίο αυτό αποδίδεται και διαγιγνώσκεται. Στο πλαίσιο της παρούσας εργασίας, η φιλοσοφική κριτική επιτρέπει να εξεταστεί αν η προσομοίωση συμπτωμάτων σε συστήματα Τεχνητής Νοημοσύνης μπορεί να συλλάβει τη βιωματική διάσταση του «πάσχειν», αναδεικνύοντας τα όρια μιας μηχανικής αναπαράστασης των ψυχικών καταστάσεων. Οι προβληματισμοί αυτοί είναι καίριοι για την παρούσα εργασία, καθώς αναδεικνύουν τη φιλοσοφική διάσταση του ερωτήματος σχετικά με τη δυνατότητα «πάσχειν» σε ένα τεχνητό σύστημα.

Ένα από τα θεμελιώδη φιλοσοφικά ερωτήματα που εγείρει η φιλοσοφία της ψυχιατρικής είναι το εξής: *τι συνιστά ψυχική διαταραχή*; Υπάρχει αντικειμενικό παθολογικό υπόβαθρο ή πρόκειται για αξιολογική κρίση που εδράζεται στα κοινωνικά πρότυπα κάθε εποχής; Με άλλα λόγια, οι ψυχικές διαταραχές είναι «πραγματικές ασθένειες» με αντικειμενικά χαρακτηριστικά ή κοινωνικές κατασκευές; Πρόκειται δηλαδή, για ένα φαινόμενο που τοποθετείται στη διατομή μεταξύ μεταφυσικής και γνωσιολογίας, καθώς αφορά τόσο τη φύση της ψυχικής διαταραχής όσο και τον τρόπο με τον οποίο τη γνωρίζουμε και την κατανοούμε (Fulford et al., 2006, σ. 2–4).

Σύμφωνα με την αντιρεαλιστική θεώρηση, η ταξινόμηση των ψυχικών διαταραχών δεν στηρίζεται σε αντικειμενικές φυσικές κατηγορίες, αλλά σε πολιτισμικές αξίες και κοινωνικά πρότυπα. Ενδεικτικό παράδειγμα αποτελεί η ομοφυλοφιλία, η οποία μέχρι το 1973 κατατασσόταν στο DSM ως ψυχική νόσος (Drescher, 2015), ενώ σήμερα θεωρείται φυσιολογική παραλλαγή της σεξουαλικότητας.

Η φιλοσοφία της ψυχιατρικής εγείρει θεμελιώδη ερωτήματα σχετικά με τη φύση, την εγκυρότητα και την κοινωνική διάσταση των ψυχικών διαταραχών. Ένα κεντρικό ζήτημα είναι αν οι διαταραχές αποτελούν φυσικές, αντικειμενικές οντότητες ή κοινωνικά και πολιτισμικά κατασκευασμένες κατηγορίες. Η απάντηση σε αυτό το ερώτημα επηρεάζει τον τρόπο με τον οποίο καταρτίζονται και εφαρμόζονται οι διαγνωστικές κατηγορίες του DSM, καθώς και τις δυνατότητες αναπαράστασης των ψυχικών καταστάσεων σε συστήματα Τεχνητής Νοημοσύνης.

Η θεωρία του «φυσικού είδους» (natural kind theory) υποστηρίζει ότι οι ψυχικές διαταραχές είναι βιολογικά και αιτιολογικά καθορισμένες οντότητες, με σταθερά χαρακτηριστικά (Wakefield, 1992, σ. 39–44). Η θεωρία αυτή υποστηρίζει ότι η γνώση των αιτιών και των μηχανισμών των διαταραχών επιτρέπει την αξιόπιστη ταξινόμηση και τη θεραπευτική αντιμετώπιση τους (Wakefield, 1992, σ. 39). Ωστόσο, οι Kendler και Zachar (1998, σ. 221) επισημαίνουν ότι οι ψυχιατρικές κατηγορίες δεν πληρούν τα αναγκαία κριτήρια για να θεωρηθούν φυσικά είδη, λόγω υψηλής διακύμανσης των συμπτωμάτων και διαφορετικών αιτιολογιών που οδηγούν σε παρόμοιες κλινικές εκδηλώσεις και διαγνώσεις. Επιπλέον, ακόμη και η έννοια της «βλάβης» ενέχει αξιολογικές και πολιτισμικές παραδοχές, υποδεικνύοντας ότι η φαινομενικά «βιολογική» θεώρηση δεν είναι απόλυτα αντικειμενική (Wakefield, 1992, σ. 44).

Από την άλλη πλευρά, στο αντίθετο άκρο της φιλοσοφικής συζήτησης, βρίσκονται οι κοινωνικο-κατασκευαστικές προσεγγίσεις, όπως αυτή του Thomas Szasz. Στο έργο του *The Myth of Mental Illness* υποστηρίζει πως οι ψυχιατρικές διαγνώσεις δεν συνιστούν παθολογίες με βιολογικό υπόβαθρο, αλλά ηθικές κρίσεις με επιστημονικό υπόστρωμα (Szasz, 1961, σ. 23). Η φιλοσοφική του θεωρία αν και χαρακτηρίζεται ως ακραία, συνδέεται με τις προηγούμενες, καθώς θέτει το ερώτημα της θεμελίωσης της επιστημονικής εγκυρότητας των διαγνωστικών κατηγοριών. Αν οι κατηγορίες του DSM δεν ανταποκρίνονται σε «φυσικά είδη», τότε δεν πρόκειται απλώς για επιστημονικά εργαλεία, αλλά και για αξιολογικές και κοινωνικές κατασκευές. Στην ίδια κατεύθυνση, ο John Salder (2008, σ. 198) φρονεί πως οι διαγνωστικές κατηγορίες του DSM, αν και παρουσιάζονται φαινομενικά ως αντικειμενικά και ουδέτερα φορτισμένες, αντανakλούν έντονα κανονιστικά και πολιτισμικά πρότυπα. Ενώ, η Rachel Cooper (Cooper, 2011, σ. 56) προσθέτει ότι η δομή των ψυχικών διαταραχών δεν είναι σταθερή, αλλά επηρεάζεται από πολιτισμικές και θεωρητικές επιλογές. Η αντιπαράθεση μεταξύ θεωρίας των φυσικών ειδών και κοινωνικο-κατασκευαστικών

προσεγγίσεων υπογραμμίζει ότι η ταξινόμηση των ψυχικών διαταραχών στο DSM συνδυάζει επιστημονικές, κοινωνικές και πολιτισμικές παραδοχές, θέτοντας υπό αμφισβήτηση την ουδετερότητα και την αντικειμενικότητα των διαγνωστικών κατηγοριών. Στο πλαίσιο της παρούσας εργασίας, αυτή η φιλοσοφική κριτική παρέχει τη βάση για τη διερεύνηση του ερωτήματος σχετικά με τη δυνατότητα προσομοίωσης της ψυχικής διαταραχής και της εμπειρίας του «πάσχειν» σε συστήματα Τεχνητής Νοημοσύνης.

Ένα φυσικό επακόλουθο της προηγούμενης συζήτησης σχετικά με τις θεωρίες του «φυσικού είδους» και της κοινωνικής κατασκευής αποτελεί το πρόβλημα της ταξινόμησης των ψυχικών διαταραχών στο DSM, δηλαδή ο τρόπος με τον οποίο ορίζεται τι συνιστά ψυχική διαταραχή. Συγκεκριμένα, η ταξινόμηση στο DSM στηρίζεται κατά κύριο λόγο σε συμπεριφοριστικά και υποκειμενικά κριτήρια, τονίζοντας την απουσία αντικειμενικών και βιολογικών μετρήσεων (Kendler & Zachar, 1998, σ. 223). Από αυτή την έλλειψη αντικειμενικότητας προκύπτει το πρόβλημα της εγκυρότητας (validity). Ειδικότερα, ένα πλήθος ψυχικών διαταραχών παρουσιάζουν παρόμοια συμπτωματολογία, όπως για παράδειγμα η μείζον κατάθλιψη και η αγχώδης διαταραχή, εμφανίζουν κοινά συμπτώματα (λόγου χάρι αυπνία, άγχος, αίσθημα ενοχής, απουσία συγκέντρωσης κ.α.). Ωστόσο, η ύπαρξη κοινών συμπτωμάτων δεν συνεπάγεται απαραίτητα κοινό βιολογικό υπόβαθρο. Στην πραγματικότητα, ψυχικές διαταραχές όπως η μείζον κατάθλιψη και η διπολική διαταραχή δεν παρουσιάζουν σαφή βιολογικό διαχωρισμό, γεγονός που δυσχεραίνει την καθαρά βιολογική διάγνωση (Kendler & Zachar, 1998, σ. 223).

Σύμφωνα με τον Kendler, το DSM είναι περισσότερο περιγραφικό όσον αφορά τα σύνδρομα, γεγονός που δυσχεραίνει την αναγνώριση κοινών αιτιακών δομών, υπογραμμίζοντας τις περιορισμένες δυνατότητες εφαρμογής της θεωρίας του «φυσικού είδους» στις ψυχικές διαταραχές. Επομένως, σε αυτό το σημείο εγείρεται το φιλοσοφικό ερώτημα σχετικά με την έννοια του «φυσικού είδους», που απαιτεί κοινές αιτιακές δομές στις ταξινομήσεις, αποτυγχάνοντας να εφαρμοστεί. Η μεταβλητότητα των κατηγοριών του DSM, όπως η αφαίρεση της ομοφυλοφιλίας ως ψυχικής διαταραχής, αναδεικνύει την αστάθεια και τις διαρκείς αλλαγές στις ταξινομήσεις, δίχως σαφή επιστημονικά κριτήρια. Κριτική στο DSM ασκούν, μεταξύ άλλων, στοχαστές όπως ο Thomas Szasz (1961) και οι Kendler & Zachar (1998), οι οποίοι επισημαίνουν ότι η μεταβλητότητα και η αστάθεια των κατηγοριών καθιστά δύσκολη

την αντιμετώπιση των ψυχικών διαταραχών ως «φυσικά είδη». Οι συνεχείς προσθήκες νέων διαταραχών, οι συγχωνεύσεις ή οι αφαιρέσεις υφιστάμενων κατηγοριών επιβεβαιώνουν τη ρευστότητα των διαγνωστικών κατηγοριών και υποδεικνύουν ότι η ταξινόμηση δεν μπορεί να θεωρηθεί απολύτως σταθερή ή αντικειμενική (Kendler & Zachar, 1998, σ. 223; Szasz, 1961, σ. 23).

Σε αυτό το πλαίσιο, η θεωρία των *looping effects* του Ian Hacking προσφέρει μια συνεκτική σύνδεση με τις προηγούμενες αντιρρήσεις: οι ταξινομήσεις δεν είναι απλώς περιγραφικές, αλλά επηρεάζουν την ίδια την εμπειρία των ατόμων. Ειδικότερα, ο εντοπισμός και η κατανόηση μιας ψυχικής διαταραχής διαμορφώνεται με την αλληλεπίδραση των ατόμων, που ανήκουν στην ίδια διαγνωστική κατηγορία, με αποτέλεσμα άλλοι να ταυτίζονται και άλλοι να αλλάζουν τα μοτίβα συμπεριφοράς τους. Επομένως, η θεωρία του *looping effect* αποδεικνύει πως οι ταξινομήσεις δεν είναι απλώς παρατηρήσιμες, αλλά διαμορφώνουν την πραγματικότητα. Συνεπώς, το DSM αδυνατεί να είναι πλήρως αντικειμενικό, καθώς η έννοια της αντικειμενικότητας είναι ιδιαίτερα προβληματική, όσον αφορά τις ανθρώπινες εμπειρίες. Η θεωρία αυτή συνδέεται με τις ανησυχίες για την εγκυρότητα και την αξιοπιστία των ψυχιατρικών διαγνώσεων.

Η αξιοπιστία αναφέρεται στην συνέπεια της διάγνωσης του ίδιου ασθενούς από δύο διαφορετικούς ψυχιάτρους. Για παράδειγμα, αν ο ασθενής εξετασθεί από δύο ψυχιάτρους την ίδια χρονική περίοδο και λάβει την ίδια διάγνωση και από τους δύο αντιστοίχως, συνεπάγεται με το γεγονός ότι αν ο ασθενής εξετασθεί σε μεταγενέστερο χρόνο, θα ενταχθεί στην ίδια κατηγορία; Τα υποκείμενα μέσω των αφηγήσεών τους τροποποιούν τα κλινικά δεδομένα, καθώς εξασκούν την περιγραφή του εαυτού τους, μέσω των διαγνώσεων. Όπως υποδεικνύει η θεωρία του *looping effect* (Hacking, 1995, σ. 98), η αξιοπιστία των διαγνώσεων του DSM μπορεί να θεωρηθεί ασταθής, λόγω της φύσης των ίδιων των κατηγοριών. Οι διαγνωστικές κατηγορίες επηρεάζουν τη συμπεριφορά των ασθενών και, με τη σειρά τους, η μεταβολή της συμπεριφοράς μπορεί να οδηγήσει σε αναπροσαρμογές των κατηγοριών, δημιουργώντας έναν φαύλο κύκλο που υποδηλώνει την δυσκολία σταθερότητας στην ταξινόμηση. Η Rachel Cooper (2011, σ. 56) τοποθετείται λέγοντας ότι, ακόμα και όταν οι διαγνώσεις εμφανίζουν υψηλή αξιοπιστία, δηλαδή συνεπή εφαρμογή των κριτηρίων, αυτό δεν εγγυάται την εγκυρότητα, όπως ορίστηκε προηγουμένως, αφού μπορεί να υπάρχουν συνεπή λάθη στην ταξινόμηση. Επίσης, ο Hacking επισημαίνει ότι οι κατηγορίες του DSM είναι

κατά κανόνα επιτελεστικές, καθώς τροποποιούν και διαμορφώνουν την εμπειρία, επομένως αδυνατούν να είναι σταθερές μεταβλητές ταξινόμησης.

Επιπλέον, η φιλοσοφία της Ψυχιατρικής επισημαίνει πως η κυριαρχία πολιτισμικών αντιλήψεων σχετικά με το τι συνιστά παθολογία, αποτελεί τροχοπέδη στην θεμελίωση σταθερών και καθολικών κριτηρίων διάγνωσης. Για παράδειγμα, ορισμένα συμπτώματα που σε δυτικές κοινωνίες θεωρούνται παθολογικά (λόγου χάρη απομόνωση), σε άλλες κοινωνίες και κουλτούρες χαρακτηρίζονται ως φυσιολογικά ή επιθυμητά. Επομένως, ο πολιτισμός επηρεάζει άμεσα τη συμπεριφορά του ατόμου, διακρίνοντας τι θεωρείται ως «φυσιολογικό» ή «παρεκκλίνον». Διαπιστώνεται πως οι ψυχικές διαταραχές βιώνονται και διαγιγνώσκονται ως πολιτισμικά παραγόμενες. Τα πιο διαδεδομένα παραδείγματα πολιτισμικής ποικιλότητας αποτελούν η κατάθλιψη και διαταραχή προσωπικότητας. Συγκεκριμένα, στη Δύση η κατάθλιψη παρουσιάζει την ακόλουθη συμπτωματολογία: έντονα συναισθήματα θλίψης, ενοχής και έλλειψης αυτοεκτίμησης. Αντίθετα, στην Ασία η θλίψη σωματοποιείται και η θλίψη ή αδυναμία εκφράζονται ως σωματικός πόνος ή κόπωση, από τα άτομα που πάσχουν. Συμπληρωματικά, η πολλαπλή διαταραχή προσωπικότητας κατονομάζεται ως ψυχοπαθολογία, κυρίως σε δυτικές κοινωνίες, καθώς το Εγώ χαρακτηρίζεται ως ενιαίο, σταθερό και αμετάβλητο. Στον αντίποδα, σε κοινωνίες και πολιτισμούς (λόγου χάρη Αφρική) το Εγώ δεν προσεγγίζεται κατά τον ίδιο τρόπο, καθώς η εμφάνιση πολλαπλότητας της προσωπικότητας ερμηνεύεται ως επαφή με τους προγόνους, με αποτέλεσμα να μην αντιμετωπίζεται ως ψυχοπαθολογία. Συμπερασματικά, οι κατηγορίες του DSM και οι ψυχικές διαταραχές αν και αξιώνουν να παρουσιάζονται ως ουδέτερες, συχνά φορτίζονται αξιολογικά από πολιτισμικές αντιλήψεις και ιδεολογίες.

Τέλος, η φιλοσοφία της ψυχιατρικής επισημαίνει την υπεραπλούστευση που προκαλεί η προσκόλληση στο βιοϊατρικό μοντέλο, η οποία οδηγεί στον αναγωγισμό ότι κάθε διαταραχή ερμηνεύεται και αντιμετωπίζεται τελικά ως βιολογική ανωμαλία. Η σύγχρονη ψυχιατρική επικαλείται το ιατρικό μοντέλο, όπου οι ψυχικές νόσοι κατηγοριοποιούνται μέσω των παρατηρούμενων συμπτωμάτων, όπως για παράδειγμα θλίψη, κόπωση, παραισθήσεις, διάσπαση προσοχής κ.α. Ωστόσο, από φαινομενολογική σκοπιά, τα συμπτώματα δεν είναι απλώς αφηρημένες έννοιες ή διαγνωστικές ετικέτες. Αντιθέτως, αποτελούν βιωμένες εμπειρίες που εμφανίζονται στη συνείδηση με συγκεκριμένο νόημα και τρόπο ύπαρξης. Για παράδειγμα, η σχιζοφρένεια δεν πρέπει

να προσεγγίζεται και να ερμηνεύεται μόνο από τα συμπτώματά της, αλλά και από το πρόβλημα αλλοίωσης του ίδιου του τρόπου ύπαρξης στον κόσμο. Άρα, η φιλοσοφία αποζητά από την ψυχιατρική να δώσει προτεραιότητα στο ίδιο το βίωμα, χωρίς να προσπαθεί με κάθε τρόπο να το τοποθετήσει σε απλά διαγνωστικά κριτήρια ή να το εξηγήσει σύμφωνα με τυποποιημένες κατηγορίες. Σε αυτό το πλαίσιο, η προσομοίωση των συμπτωμάτων από ένα σύστημα ΑΙ δεν συνιστά ούτε εμπειρία ούτε πάσχειν, γιατί του λείπει το υπαρξιακό υπόβαθρο του υποκειμένου. Η άκριτη αποδοχή του ιατρικού μοντέλου σηματοδότησε τις βασικές ενστάσεις της φιλοσοφίας της ψυχιατρικής. Αν και η ταξινόμηση του DSM προσφέρει την απαιτούμενη τυπικότητα και αντικειμενικοποίηση των συμπτωμάτων, η φιλοσοφία της ψυχιατρικής επισημαίνει ότι τα συμπτώματα, λόγω του αξιολογικού και κανονιστικού τους χαρακτήρα, δεν μπορούν να τυποποιηθούν πλήρως, χωρίς να χάσουν μέρος από την αρχική τους σημασία ή να απλουστευθούν.

Το κεντρικό ερώτημα που αναδύεται και καθορίζει την προβληματική της εργασίας είναι το εξής: *Αν ένα σύστημα Τεχνητής Νοημοσύνης μπορεί να πληροί τα διαγνωστικά κριτήρια μιας ψυχικής διαταραχής, σημαίνει ότι πραγματικά «πάσχει»;* Η απάντηση δεν είναι αυτονόητη, καθώς το DSM αξιολογεί μοτίβα συμπεριφοράς και συμπτώματα, αλλά δεν παρέχει πρόσβαση στις βιωμένες, υποκειμενικές εμπειρίες ή στην εσωτερική κατάσταση συνείδησης του ατόμου. Συμπερασματικά, λαμβάνοντας υπόψιν τις παραπάνω πτυχές οι διάφορες φιλοσοφικές προσεγγίσεις στη φιλοσοφία της ψυχιατρικής προτείνουν να μην εγκλωβιστούμε σε ένα μονοπαραγοντικό μοντέλο, αλλά να αντιμετωπίζουμε την ψυχική διαταραχή ως σύνθετη έννοια, που λαμβάνει υπόψη τόσο τα διαγνωστικά κριτήρια όσο και την υποκειμενική, βιοματική διάσταση, χωρίς αυτό να συνεπάγεται με τον αποκλεισμό της επικρατούσας ερμηνείας της (Cooper, 2011, σ. 56; Kendler & Zachar, 1998, σ. 223).

Δεδομένων των πολλαπλών φιλοσοφικών αντιρρήσεων σχετικά με την εγκυρότητα, την αξιοπιστία και την πολιτισμικά φορτισμένη φύση των διαγνωστικών κατηγοριών του DSM, προκύπτει ότι η έννοια της ψυχικής διαταραχής μπορεί να μην περιορίζεται επαρκώς σε μια απλή καταγραφή συμπτωμάτων, όπως συχνά επιχειρείται στην κλινική πρακτική. Η φιλοσοφία της ψυχιατρικής αναδεικνύει την ανάγκη για μια βαθύτερη θεωρητική διερεύνηση του τι συνιστά «ψυχική διαταραχή» και πώς ορίζεται η «δυσλειτουργία», ώστε να γίνεται σαφές πότε μια κατάσταση μπορεί να θεωρηθεί παθολογική και ποιοι παράγοντες, βιολογικοί ή κοινωνικοπολιτισμικοί, εμπλέκονται

στην αξιολόγησή της. Αυτή η αναγκαία θεμελίωση δημιουργεί τη γέφυρα προς την επόμενη ενότητα, όπου θα εξεταστεί με λεπτομέρεια η έννοια της ψυχικής διαταραχής και της δυσλειτουργίας.

2.3 Η έννοια της ψυχικής διαταραχής και της δυσλειτουργίας

Μια από τις θεμελιώδεις προκλήσεις στη φιλοσοφία της ψυχιατρικής είναι ο ορισμός της «ψυχικής διαταραχής». Ορισμένες φιλοσοφικές προσεγγίσεις υποστηρίζουν το DSM, θεωρώντας τις διαγνωστικές κατηγορίες χρήσιμες για την επιστημονική ταξινόμηση και την κλινική πρακτική, ενώ άλλες ασκούν κριτική στην αντικειμενικότητα και τη σταθερότητά τους, εξετάζοντας τις ψυχικές διαταραχές είτε ως κοινωνικά κατασκευασμένες κατηγορίες, είτε ως βιολογικά επηρεασμένες, είτε ως συνδυασμό των δύο (Kendler & Zachar, 1998, σ. 221; Szasz, 1961, σ. 23). Ερωτήματα σχετικά με τη συνείδηση, τη βλάβη και τη δυσλειτουργία επηρεάζουν την ταξινόμηση και το κοινωνικό στίγμα, καθώς οι ψυχικές διαταραχές εκδηλώνονται μέσω επώδυνων αλλαγών στη διάθεση, τη σκέψη και τη συμπεριφορά (Wakefield, 1992, σ. 28).

Οι φιλοσοφικές θεωρήσεις σχετικά με την ψυχική διαταραχή διαφοροποιούνται ως προς το τι θεωρούν κεντρικό στοιχείο της. Ο John Sadler (2005, σ. 45) τονίζει τον ρόλο της δυσφορίας που προκαλεί η διαταραχή στο άτομο και στον κοινωνικό του περίγυρο, υπογραμμίζοντας την ανάγκη για θεραπευτική παρέμβαση. Ο Jerome Wakefield (1992, σ. 28) προσδιορίζει τη διαταραχή ως δυσλειτουργία ενός βιολογικού ή ψυχολογικού μηχανισμού που προκαλεί σημαντική βλάβη στο άτομο, ενώ ο Kenneth Kendler (2002, σ. 223) δίνει έμφαση στους εσωτερικούς ψυχικούς μηχανισμούς, όπως η αντίληψη και η συναισθηματική ρύθμιση. Άλλες προσεγγίσεις εστιάζουν στην επίδραση της διαταραχής στη συνολική κοινωνική και προσωπική λειτουργία, ενώ ολιστικά μοντέλα επιχειρούν να συνδυάσουν τόσο το βιολογικό/ψυχολογικό όσο και το κοινωνικό στοιχείο.

Η αβεβαιότητα στον ορισμό της δυσλειτουργίας προκύπτει από τις διαφορετικές φιλοσοφικές και επιστημονικές θεωρήσεις. Οι βιοϊατρικές θεωρίες, όπως αναδεικνύεται από τους Kendler & Zachar (1998, σ. 223), επικεντρώνονται σε εσωτερικούς μηχανισμούς και βιολογικές ανωμαλίες, ενώ οι κοινωνικοπολιτισμικές

θεωρίες, όπως αυτή του Szasz (1961, σ. 23), δίνουν έμφαση στον αντίκτυπο της διαταραχής στα άτομα και την κοινωνία. Τα παρατηρούμενα συμπτώματα, όπως ψευδαισθήσεις στην σχιζοφρένεια ή θλίψη στην κατάθλιψη, συχνά εμφανίζουν αμφιλεγόμενη ταξινόμηση, καθώς η έννοια του «δυσλειτουργικού» ποικίλλει ανάλογα με τη θεωρητική οπτική. Το DSM-5 αναγνωρίζει ότι οι ψυχικές διαταραχές προκαλούν δυσφορία ή ανικανότητα σε βασικές δραστηριότητες, ενώ εισάγει διακρίσεις, όπως η «φυσιολογική θλίψη», για να αποφευχθούν εσφαλμένες διαγνώσεις (APA, 2013, σ. 155–157). Κάποιοι ερευνητές υποστηρίζουν ότι οι βιολογικοί δείκτες μπορούν να βελτιώσουν τη διάκριση, ενώ άλλοι τονίζουν τη σημασία της υποκειμενικής εμπειρίας του πόνου και της καθημερινής λειτουργικότητας (Wakefield, 1992, σ. 28; Cooper, 2011, σ. 56).

Οι Tyrer και Steinberg ορίζουν την ψυχική διαταραχή ως σύνολο χαρακτηριστικών που αποκλίνουν από τον κανόνα και προκαλούν δυσλειτουργία, η οποία μπορεί να είναι βιολογική, ψυχολογική ή κοινωνική. Ο Kendell συμφωνεί, αλλά ο Cooper αμφισβητεί ότι μειωμένη επιβίωση ή γονιμότητα αρκούν για τον χαρακτηρισμό μιας πάθησης ως διαταραχής. Ο Wakefield συνδέει τη δυσλειτουργία με τη βλάβη, υπογραμμίζοντας την αναγκαιότητα κοινωνικοπολιτισμικών κριτηρίων για την πλήρη κατανόηση της ψυχικής διαταραχής.

Η βιοστατιστική θεωρία του Boorse (BST) ορίζει την υγεία ως φυσιολογική λειτουργία και την ασθένεια ως μείωση αυτής (Boorse, 1977, σ. 543). Οι ασθένειες θεωρούνται αποκλίσεις από τα βιολογικά πρότυπα, ενώ η επιλογή κατάλληλων κατηγοριών αναφοράς είναι κρίσιμη αλλά προβληματική, όπως φαίνεται στην περίπτωση της ομοφυλοφιλίας. Ο Boorse επικεντρώνεται στη φυσιολογική δυσλειτουργία, ενώ ο Wakefield συνδυάζει δυσλειτουργία και επιβλαβή βλάβη.

Συμπερασματικά, δεν υπάρχει καθολικά αποδεκτός ορισμός της ψυχικής διαταραχής. Η φιλοσοφία της ψυχιατρικής επιδιώκει να συνδυάσει αντικειμενικά κριτήρια (Boorse) με κοινωνικοπολιτισμικές παραμέτρους (Wakefield), επισημαίνοντας ότι η κατανόηση των ψυχικών διαταραχών απαιτεί πολυδιάστατη προσέγγιση. Η σύγχρονη ψυχιατρική συχνά επικεντρώνεται στο ιατρικό μοντέλο, δίνοντας έμφαση στην αναγνώριση και τυποποίηση συμπτωμάτων και δυσλειτουργιών.

2.4 Ιατρικό μοντέλο και τεχνητή νοημοσύνη

Ο σκοπός αυτής της υποενότητας δεν είναι η ανάλυση της κλινικής διάγνωσης σε ανθρώπους, αλλά η παρουσίαση τρόπων με τους οποίους η Τεχνητή Νοημοσύνη μπορεί να προσομοιώσει μοτίβα συμπεριφοράς και νοητικών διεργασιών που σχετίζονται με ψυχικές διαταραχές. Η αναφορά σε εφαρμογές του ΑΙ στη διάγνωση γίνεται ως πλαίσιο, για να καταδειχθούν οι δυνατότητες και, κυρίως, τα όρια της προσομοίωσης, καθώς η βιωματική διάσταση του «πάσχειν» παραμένει απρόσιτη στα τεχνητά συστήματα (Cooper, 2011, σ. 56–58; Huys et al., 2016, σ. 5–7).

Η φιλοσοφική ανάλυση της δυσλειτουργίας και της επιβλαβούς δυσλειτουργίας παρέχει το θεωρητικό υπόβαθρο για την κατανόηση των περιορισμών των παραδοσιακών μοντέλων ψυχιατρικής διάγνωσης. Όπως προκύπτει από τη φαινομενολογική προσέγγιση, τα διαγνωστικά κριτήρια του DSM αξιολογούν συμπεριφορικά μοτίβα και αντικειμενικά παρατηρήσιμες λειτουργίες, αλλά δεν παρέχουν πρόσβαση στην υποκειμενική, βιωματική διάσταση της ψυχικής νόσου. Συνεπώς, παρά τις προσπάθειες του DSM να ορίσει με ακρίβεια το «πάσχειν», η κατανόηση της εμπειρίας του ατόμου παραμένει περιορισμένη (Cooper, 2011, σ. 56; Wakefield, 1992, σ. 28). Στο πλαίσιο αυτό, οι σύγχρονες κοινωνίες στρέφονται προς το ιατρικό μοντέλο, το οποίο επιδιώκει την τυποποίηση των ψυχικών διαταραχών και την περιγραφή τους με βάση αιτιακούς μηχανισμούς (π.χ. νευροβιολογικούς, γενετικούς ή νευροχημικούς), ώστε να μπορούν να καταγραφούν και να επεξεργαστούν με υπολογιστικά εργαλεία (Insel, 2014, σ. 112; Friston et al., 2020, σ. 12–14).

Παρόλα αυτά, οι κατηγορίες του DSM συχνά αποτυγχάνουν να ανιχνεύσουν την ετερογένεια των συμπτωμάτων και τη συννοσηρότητα μεταξύ των διαφορετικών διαταραχών. Για παράδειγμα, άτομα με μείζονα κατάθλιψη συχνά εμφανίζουν συμπτώματα αγχώδους διαταραχής ή ψυχοσωματικά προβλήματα, γεγονός που καθιστά τη διαγνωστική διαδικασία πιο σύνθετη και επιρρεπή σε λάθη (Insel et al., 2010, σ. 78). Συμπληρωματικά, το DSM βασίζεται και ταξινομεί με βάση συγκεκριμένα συμπεριφορικά μοτίβα, χωρίς να διαθέτει άμεση πρόσβαση στην υποκειμενική και βιωματική διάσταση του «πάσχειν». Επομένως, οι παραδοσιακές διαγνωστικές κατηγορίες καταγράφουν μόνο ένα μέρος της ψυχικής πραγματικότητας, περιορίζοντας την κατανόηση της πλήρους κλινικής εικόνας του ατόμου. Η υπολογιστική ψυχιατρική και η μηχανική μάθηση (ML) προτείνουν νέες λύσεις για την

αντιμετώπιση αυτών των περιορισμών, επιτρέποντας την ανάλυση μεγάλων συνόλων δεδομένων, μέσω της οποίας είναι να ανακαλυφθούν κρυφά μοτίβα και συσχετισμοί που δεν γίνονται εμφανείς με τις παραδοσιακές διαγνωστικές μεθόδους (Friston et al., 2020, σ. 12) βελτιώνοντας με αυτό τον τρόπο τη διάγνωση και της θεραπείας. Για παράδειγμα, μοντέλα ML μπορούν να εντοπίσουν διαφορετικές υποομάδες ασθενών με σχιζοφρένεια με βάση συνδυασμούς συμπτωμάτων και γενετικών δεικτών, προσφέροντας πιο προσωποποιημένες προβλέψεις για την πορεία της νόσου (Huys et al., 2016, σ. 5–7).

Παράλληλα, η Τεχνητή Νοημοσύνη επιτρέπει την προσομοίωση νοητικών διεργασιών και την αναγνώριση σύνθετων μοτίβων σε πολυδιάστατα δεδομένα. Μέσω τεχνικών όπως τα νευρωνικά δίκτυα το ΑΙ έχει τη δυνατότητα να εκτιμήσει πιθανολογικά ποια συμπτώματα τείνουν να εμφανίζονται μαζί, να αναγνωρίσει πρότυπα επιδείνωσης ή βελτίωσης της κατάστασης και να προτείνει ενδεδειγμένες στρατηγικές παρέμβασης (Friston et al., 2020, σ. 14–17). Με αυτό τον τρόπο, το ΑΙ λειτουργεί συμπληρωματικά στο DSM, προσφέροντας εργαλεία για την υποστήριξη της κλινικής απόφασης, χωρίς όμως να αντικαθιστά την ανθρώπινη κρίση ή να παρέχει πρόσβαση στην υποκειμενική εμπειρία.

Συγκεκριμένα, η σύνδεση του DSM με τα συστήματα ΑΙ επιτρέπει την κατηγοριοποίηση μέσω αλγορίθμων. Το DSM παρέχει δομή δεδομένων, ενώ το ΑΙ δεν «συνδέεται» με το DSM, αλλά χρησιμοποιεί τις κατηγορίες του ως δεδομένα εκπαίδευσης για ταξινόμηση. Ενώ τα εργαλεία ML και NLP αναλύουν γλωσσικά, συμπεριφορικά και βιολογικά δεδομένα για την ανίχνευση ψυχικών διαταραχών. Η ML εντοπίζει σχέσεις μεταξύ μεταβλητών, προβλέπει αποτελέσματα και προσαρμόζεται σε νέα δεδομένα, ενώ η βαθιά μάθηση χρησιμοποιεί νευρωνικά δίκτυα για πολύπλοκες εργασίες, όπως η αναγνώριση προτύπων λόγου και συμπεριφοράς (Bzdok & Meyer-Lindenberg, 2018, σ. 45–47). Οι γλωσσικοί δείκτες, όπως ασυνέχεια λόγου ή υπερβολική χρήση αρνητικών λέξεων, συνδέονται με διαγνώσεις όπως σχιζοφρένεια ή κατάθλιψη.

Όπως ειπώθηκε προηγουμένως, παρά τις δυνατότητες των συστημάτων ΑΙ, αυτά παραμένουν υποστηρικτικά εργαλεία για τους κλινικούς γιατρούς, καθώς δεν μπορούν να κατανοήσουν ή να βιώσουν το «πάσχειν». Ορισμένες φιλοσοφικές προσεγγίσεις της ψυχιατρικής τονίζουν τη διάκριση ανάμεσα στα παρατηρήσιμα συμπτώματα και στην

υποκειμενική εμπειρία της δυσλειτουργίας. Το Αι βελτιώνει την αυτοματοποιημένη αντίχρευσση, παρακολούθηση και εξατομικευμένη παρέμβαση, αλλά δεν αντικαθιστά τον κλινικό.

Το ιατρικό μοντέλο υποστηρίζει αυτή την προσέγγιση, καθώς παρέχει ένα σταθερό, αιτιοκρατικό πλαίσιο για την κατηγοριοποίηση των ψυχικών διαταραχών. Η αιτιοκρατία στο πλαίσιο αυτό δεν αναφέρεται σε φιλοσοφική έννοια, αλλά στην προσπάθεια να αποδοθούν τα συμπτώματα σε βιολογικούς μηχανισμούς, όπως νευροχημικές ανωμαλίες ή γενετικούς δείκτες (Insel, 2014, σ. 112). Η συνδυασμένη χρήση DSM, υπολογιστικής ψυχιατρικής και Αι επιτρέπει την ενσωμάτωση πολλαπλών επιπέδων πληροφορίας: συμπεριφορικά μοτίβα, νευροβιολογικούς δείκτες και κοινωνικοπολιτισμικό πλαίσιο, προσφέροντας ένα πιο ολοκληρωμένο και ευέλικτο μοντέλο διάγνωσης και θεραπείας.

Συνολικά, η τεχνητή νοημοσύνη ενσωματώνει το ιατρικό μοντέλο ως βάση, αποκαλύπτοντας ταυτόχρονα τα όριά του όταν το υποκείμενο δεν διαθέτει αυτοσυνείδηση ή βιωματική εμπειρία. Η κατανόηση του τι σημαίνει «πάσχειν» απαιτεί επέκταση πέρα από παρατηρήσιμα δεδομένα, εξετάζοντας τη συνείδηση, τις νοητικές διεργασίες και την εμπειρία του κόσμου. Η φιλοσοφία του νου προσφέρει τα θεωρητικά εργαλεία για μια βαθύτερη κατανόηση της ψυχικής διαταραχής, συνδέοντας την ψυχιατρική παρατήρηση με την υποκειμενική διάσταση της ψυχικής ζωής (Chalmers, 1996; Cooper, 2011, σ. 56–58).

3. Εμβάθυνση στην ψυχική διαταραχή

3.1 Φιλοσοφικές προσεγγίσεις στη φύση των ψυχικών διαταραχών

Η έννοια της ψυχικής διαταραχής δεν αφορά μόνο την κλινική πράξη ή την ταξινομητική οργάνωση της ψυχιατρικής γνώσης, αλλά εγείρει θεμελιώδη ερωτήματα που εντάσσονται στο πεδίο της φιλοσοφίας του νου. Ειδικότερα, θέτει ζητήματα σχετικά με τη φύση των νοητικών καταστάσεων, τη σχέση συνείδησης και συμπεριφοράς, καθώς και τη δυνατότητα πρόσβασης στην υποκειμενική εμπειρία μέσω

αντικειμενικών περιγραφών. Από αυτή την άποψη, η φιλοσοφία της ψυχιατρικής μπορεί να θεωρηθεί εφαρμοσμένο πεδίο της φιλοσοφίας του νου, καθώς εξετάζει πώς οι ψυχικές καταστάσεις γίνονται γνωσιοθεωρητικά προσβάσιμες και εννοιολογικά κατανοητές στο πλαίσιο της διάγνωσης (Gallagher, 2005; Fulford, 1989).

Στο πλαίσιο αυτό, ανακύπτει το ερώτημα πώς μπορεί να κατανοηθεί η ψυχική διαταραχή όχι μόνο ως σύνολο παρατηρήσιμων συμπτωμάτων, αλλά ως εμπειρία που βιώνεται από ένα υποκείμενο. Το ζήτημα αυτό έχει απασχολήσει κεντρικούς εκπροσώπους της φαινομενολογικής και αναλυτικής φιλοσοφίας της ψυχιατρικής, όπως ο Karl Jaspers, ο οποίος εισήγαγε τη διάκριση μεταξύ αιτιακής εξήγησης (Erklären) και ενσυναισθητικής κατανόησης (Verstehen) της ψυχικής εμπειρίας (Jaspers, 1913/1997). Η διάκριση αυτή αναδεικνύει ένα βασικό πρόβλημα της φιλοσοφίας του νου: η περιγραφή της συμπεριφοράς και των γνωστικών λειτουργιών δεν ισοδυναμεί με κατανόηση του βιώματος που τις συνοδεύει.

Σύγχρονες προσεγγίσεις στη φιλοσοφία της ψυχιατρικής συνεχίζουν αυτή τη γραμμή προβληματισμού, υποστηρίζοντας ότι η ψυχική διαταραχή δεν μπορεί να εξαντληθεί σε νευροβιολογικούς ή συμπεριφορικούς όρους. Ο Sass (1992) και ο Gallagher (2005) επισημαίνουν ότι διαταραχές όπως η σχιζοφρένεια αφορούν μεταβολές στη δομή της συνείδησης και του εαυτού, γεγονός που καθιστά ανεπαρκή μια καθαρά τρίτου προσώπου προσέγγιση. Αντίστοιχα, ο Fulford (1989) δείχνει ότι οι διαγνωστικές κρίσεις ενσωματώνουν αναπόφευκτα αξιολογικά και εννοιολογικά στοιχεία, τα οποία δεν μπορούν να εξαλειφθούν μέσω αντικειμενικοποίησης.

Η προβληματική αυτή αποκτά ιδιαίτερη σημασία όταν εξετάζεται η ψυχιατρική ταξινόμηση. Ιστορικά, η αναγνώριση των ψυχικών διαταραχών κινήθηκε μεταξύ δύο πόλων: της αιτιακής θεμελίωσης και της περιγραφικής επάρκειας. Η σύγχρονη ψυχιατρική, ιδίως μέσω του DSM, υιοθετεί μια κατ' αρχήν αιτιολογικά ουδέτερη στάση, βασισμένη σε λειτουργικά και παρατηρήσιμα κριτήρια (Kendell & Jablensky, 2003· First & Westen, 2007). Ωστόσο, η επιλογή αυτή δεν είναι φιλοσοφικά ουδέτερη, καθώς προϋποθέτει ότι οι ψυχικές καταστάσεις μπορούν να προσδιοριστούν αξιόπιστα χωρίς άμεση πρόσβαση στο βίωμα.

Η αμφισβήτηση της ενιαίας αιτιολογικής θεμελίωσης των ψυχικών διαταραχών ενισχύει περαιτέρω τον γνωσιοθεωρητικό χαρακτήρα του προβλήματος. Όπως υποστηρίζει ο Kendler (2012), οι ψυχικές διαταραχές είναι πολυπαραγοντικά

φαινόμενα, στα οποία συνυπάρχουν βιολογικοί, ψυχολογικοί και κοινωνικοί παράγοντες χωρίς σαφή ιεράρχηση. Σε αυτό το πλαίσιο, η ταξινόμηση δεν αντανakλά απλώς φυσικά είδη, αλλά λειτουργεί ως εννοιολογικό εργαλείο που εξυπηρετεί συγκεκριμένους επιστημονικούς και κλινικούς σκοπούς (Zachar & Kendler, 2007).

Η συζήτηση αυτή συνδέεται άμεσα με δύο αντίρροπες φιλοσοφικές θέσεις ως προς την οντολογική φύση των ψυχιατρικών κατηγοριών: τον ουσιοκρατισμό και τον νομιναλισμό. Ο ουσιοκρατισμός αντιμετωπίζει τις ψυχικές διαταραχές ως οντότητες με σταθερή υποκείμενη δομή ή αιτία, συχνά βιολογικής φύσεως (Kendell, 1975). Αντίθετα, ο νομιναλισμός θεωρεί τις διαγνώσεις εννοιολογικές κατασκευές που αποκτούν νόημα σε σχέση με τις πρακτικές χρήσεις τους, όπως η πρόγνωση, η θεραπεία και η έρευνα (Hacking, 1995· Zachar, 2000). Η διάκριση αυτή έχει άμεσες συνέπειες για τη φιλοσοφία του νου, καθώς αφορά το αν οι ψυχικές καταστάσεις προϋπάρχουν των περιγραφών τους ή συγκροτούνται εν μέρει μέσω αυτών.

Παράλληλα, η φιλοσοφική ανάλυση της ψυχικής διαταραχής εμπλέκεται στη διαμάχη μεταξύ αντικειμενιστικών και αξιολογικών προσεγγίσεων. Οι αντικειμενιστές υποστηρίζουν ότι η διάγνωση βασίζεται στον εντοπισμό δυσλειτουργιών σε φυσικές ή ψυχολογικές λειτουργίες, ανεξάρτητα από κοινωνικές αξίες. Ωστόσο, η διάκριση μεταξύ φυσιολογικού και παθολογικού προϋποθέτει ήδη κανονιστικά κριτήρια, γεγονός που καθιστά προβληματική την πλήρη απο-αξιολόγηση της διάγνωσης (Fulford, 1989). Οι αξιολογικές προσεγγίσεις, αντίθετα, τονίζουν ότι έννοιες όπως η υγεία, η δυσλειτουργία και το πάσχειν είναι άρρηκτα συνδεδεμένες με κοινωνικά και πολιτισμικά πλαίσια.

Οι παραπάνω προβληματισμοί οδηγούν στη διάκριση μεταξύ εσωτερικιστικών και εξωτερικιστικών προσεγγίσεων στην κατανόηση των ψυχικών διαταραχών. Οι εσωτερικιστές εστιάζουν σε ενδοσωματικές και νευροβιολογικές διεργασίες ως καθοριστικές για τη φύση της διαταραχής, ενώ οι εξωτερικιστές αποδίδουν καθοριστικό ρόλο στο περιβάλλον, τις κοινωνικές σχέσεις και τα νοηματικά πλαίσια. Οι μετριοπαθείς εκδοχές επιχειρούν να συνδυάσουν τα δύο επίπεδα, αναγνωρίζοντας ότι οι ψυχικές καταστάσεις δεν είναι απομονώσιμες από το πλαίσιο μέσα στο οποίο βιώνονται και ερμηνεύονται (Kendler & Zachar).

Η φιλοσοφική διερεύνηση της ψυχικής διαταραχής, όπως αναπτύχθηκε παραπάνω, δεν εξαντλείται σε οντολογικές ή ταξινομητικές διαμάχες, αλλά αναδεικνύει ένα κεντρικό

ζήτημα της φιλοσοφίας του νου: το χάσμα μεταξύ πρώτου και τρίτου προσώπου στην κατανόηση των νοητικών καταστάσεων. Το ζήτημα αυτό καθίσταται ιδιαίτερα κρίσιμο στη διαγνωστική πρακτική, όπου η υποκειμενική εμπειρία μεταφράζεται σε κριτήρια, κατηγορίες και αξιολογήσεις. Η επόμενη ενότητα εστιάζει ακριβώς σε αυτή τη μετάβαση από τη θεωρητική κατανόηση της ψυχικής διαταραχής στη διαγνωστική της πραγμάτωση, εξετάζοντας τον ρόλο της διάγνωσης ως γνωσιοθεωρητικού και πρακτικού εργαλείου στην ψυχοπαθολογία.

3.2 Η διάγνωση ως γέφυρα μεταξύ θεωρίας και βιώματος

Η διάγνωση στην ψυχοπαθολογία αποτελεί τον κατεξοχήν μηχανισμό μέσω του οποίου οι θεωρητικές παραδοχές για τη φύση της ψυχικής διαταραχής μετατρέπονται σε πρακτική γνώση. Δεν λειτουργεί απλώς ως τεχνικό εργαλείο καταγραφής συμπτωμάτων, αλλά ως γνωσιοθεωρητική και ερμηνευτική διαδικασία, μέσω της οποίας το «πάσχειν» αποκτά κλινική σημασία.

Ο καρτεσιανός δυϊσμός αντιπροσωπεύει ένα σημαντικό ιστορικό εμπόδιο στην ψυχιατρική, προωθώντας την πεποίθηση ότι ο νους και ο εγκέφαλος αποτελούν δύο διακριτές οντότητες. Παρά την προσκόλληση ορισμένων ψυχιάτρων στο δυϊσμό, που προκύπτει κυρίως από τη μακρά ιστορική παράδοση που θέτει σε διάκριση το «ψυχικό» από το «σωματικό», καθώς και από τις δυσκολίες που αντιμετωπίζει η νευροεπιστήμη να εξηγήσει πλήρως την πολυπλοκότητα της συνείδησης και των ψυχικών φαινομένων μέσω μόνο βιολογικών μηχανισμών, αποτελεί επιτακτική ανάγκη να απορριφθεί αυτή η άποψη, αναγνωρίζοντας ότι η υποκειμενική εμπειρία προκύπτει από τη λειτουργία του εγκεφάλου. Η απόρριψη του καρτεσιανού δυϊσμού δεν συνεπάγεται αυτομάτως, ότι όλες οι ψυχιατρικές διαταραχές είναι αποκλειστικά βιολογικές. Αντιθέτως, η σύγχρονη ψυχιατρική και φιλοσοφία του νου επιδιώκουν να κατανοήσουν τη σχέση μεταξύ εγκεφάλου και νου ως αμοιβαία αλληλεπίδραση. Με άλλα λόγια, η απόρριψη του καρτεσιανού δυϊσμού ενισχύει την κατανόηση του νου ως προϊόν σύνθετων διεργασιών εγκεφάλου-περιβάλλοντος, και όχι ως ανεξάρτητης οντότητας. Η αιτιότητα είναι αμφίδρομη και πολυδιάστατη. Οι αλλαγές στον εγκέφαλο επηρεάζουν την ψυχική λειτουργία, ενώ τα νοητικά φαινόμενα επηρεάζουν τον εγκέφαλο και το σώμα, αποδεικνύοντας την αιτιώδη αποτελεσματικότητά τους στην ανθρώπινη συμπεριφορά, χωρίς να επιστρέφουν στον δυϊσμό. Τα πολυεπίπεδα μοντέλα, που ενσωματώνουν τόσο ψυχικές όσο και κοινωνικές προοπτικές, συχνά οδηγούν σε μια μετατόπιση της

εστίασης προς τα βιολογικά αίτια, παρά την αρχική πρόθεση να αντιμετωπίσουν την ψυχική υγεία ολιστικά. Αυτή η στάση, ενώ συμφωνεί με το πρότυπο του ιατρικού μοντέλου, είναι περιορισμένη και παραβλέπει την πολυπλοκότητα των ψυχιατρικών διαταραχών.

Η μετάβαση από τη δυϊστική θεώρηση στην καθαρά βιολογική προσέγγιση δεν επιλύει αλλά μετατοπίζει το πρόβλημα. Η ψυχιατρική διάγνωση, όταν εδράζεται αποκλειστικά σε νευροβιολογικά ή συμπτωματολογικά δεδομένα, ενδέχεται να παραγνωρίσει τη σημασία της βιωμένης εμπειρίας του ασθενούς, καθιστώντας το «πάσχειν» μια αφηρημένη κατασκευή, καθώς δεν υπάρχει ένας ενιαίος, αυστηρά καθορισμένος ορισμός που να ισχύει ομοιόμορφα σε όλα τα πλαίσια και όλες τις περιπτώσεις ψυχικών διαταραχών. Σε αυτό το πλαίσιο, το DSM υιοθετεί ένα καθοριστικά περιγραφικό μοντέλο, το οποίο αποφεύγει ερμηνείες αιτιότητας και εστιάζει στην ανίχνευση μοτίβων συμπτωμάτων. Αν και αυτή η προσέγγιση συνέβαλε σημαντικά στην αξιοπιστία της ψυχιατρικής πρακτικής, υποβιβάζει την ερμηνευτική διάσταση του ψυχικού πόνου, μετατρέποντάς τον σε μετρήσιμο αντικείμενο. Οι συζητήσεις σχετικά με την ταξινόμηση των ψυχικών διαταραχών τονίστηκαν έπειτα από τον αποκλεισμό της ομοφυλοφιλίας από το DSM. Η ιδέα ότι οι ψυχικές διαταραχές είναι εγκεφαλικές διαταραχές εγείρει ζητήματα σχετικά με τη διαγνωστική εγκυρότητα, τις εξηγήσεις των ψυχικών ασθενειών και το πρόβλημα νου-σώματος.

Ο ισχυρός νατουραλισμός υποστηρίζει ότι η «διαταραχή» μπορεί να γίνει κατανοητή με πραγματικούς, αξιακά απαλλαγμένους όρους μέσω μεθοδολογιών των φυσικών επιστημών, θεωρώντας τις ψυχιατρικές διαταραχές ως νευροβιολογικές δυσλειτουργίες. Αντίθετα, ο ισχυρός κανονιστικός ιεραρχισμός υποστηρίζει ότι οι διαταραχές δεν έχουν αντικειμενικούς βιολογικούς ορισμούς και επηρεάζονται από αξιολογικές κρίσεις. Ο Fulford υποστηρίζει ότι η διαταραχή είναι κανονιστική, ριζωμένη στην «εμπειρία της ασθένειας», που επηρεάζεται από τις κοινωνικές αντιλήψεις για την ανωμαλία. Ο Szasz ασκεί κριτική στην ψυχιατρική, υποστηρίζοντας ότι οι κρίσεις για τη διαταραχή προέρχονται από κοινωνικοπολιτισμικούς κανόνες χωρίς βιολογικά στοιχεία ασθένειας, οδηγώντας στο συμπέρασμα ότι οι ψυχικές διαταραχές χαρακτηρίζονται λανθασμένα. Υποστηρίζει ότι η αξιακή φόρτιση υπονομεύει την ιατρική νομιμότητα της ψυχιατρικής. Η ανάλυση της επιβλαβούς δυσλειτουργίας του J. Wakefield συνδυάζει τον νατουραλισμό και τον κανονιστικό χαρακτήρα της ψυχικής διαταραχής, ορίζοντας τη «δυσλειτουργία» ως την αποτυχία

βιολογικών ή ψυχολογικών μηχανισμών να εκτελέσουν τις εξελικτικές τους λειτουργίες, ενώ η «βλαβερότητα» βασίζεται σε κοινωνικούς κανόνες. Οι αντιδράσεις, όπως η κατάθλιψη σε στρεσογόνους παράγοντες δεν είναι δυσλειτουργίες αλλά προσαρμοστικές αντιδράσεις, ταξινομώντας τις πραγματικές διαταραχές ως εκείνες που προκύπτουν απροσδόκητα ή επιμένουν χωρίς την ύπαρξη στρεσογόνων παραγόντων.

Αντίθετα, ο J. Tsou παρουσιάζει μια υβριδική άποψη, που ορίζει τις ψυχικές διαταραχές ως βιολογικά είδη με επιβλαβείς επιπτώσεις. Για παράδειγμα, ο Tsou ορίζει τη σχιζοφρένεια ως μια διαταραχή που οφείλεται σε κοινούς βιολογικούς μηχανισμούς και λειτουργική βλάβη. Ο εντοπισμός των κατηγοριών ψυχικών διαταραχών οδηγεί σε ερωτήματα σχετικά με τη χαρτογράφηση της ψυχοπαθολογίας και την οριοθέτηση της διαταραχής από την κανονικότητα. Η φιλοσοφία της ψυχιατρικής, ιδιαίτερα μέσα από την ουσιοκρατία υποστηρίζει ότι οι κατηγορίες έχουν εγγενείς ιδιότητες που τις διακρίνουν από άλλες. Οι ουσιοκρατικές απόψεις υποδηλώνουν ότι οι ψυχιατρικές οντότητες ανακαλύπτονται παρόμοια με τις μολυσματικές ασθένειες. Η σύγχρονη φιλοσοφία της βιολογίας και της ψυχολογίας εξετάζει αιτιακές διαδικασίες για την ταξινόμηση, ωστόσο, όταν οι μηχανισμοί της ασθένειας είναι πολύπλοκοι και πολυπαραγοντικοί, η αιτιολογία δεν αρκεί από μόνη της για την ψυχιατρική ταξινόμηση, καθώς η διάγνωση απαιτεί αξιολογικές, λειτουργικές και κοινωνικοπολιτισμικές παραμέτρους, οι οποίες καθορίζουν τα όρια μεταξύ «φυσιολογικού» και «παθολογικού», αλλά και μεταξύ διαφορετικών κατηγοριών διαταραχών. Αυτό υπογραμμίζει την ανάγκη να ληφθούν υπόψη και άλλοι παράγοντες, όπως κοινωνικές συνιστώσες και λειτουργικές συνέπειες, προκειμένου να οριστούν με μεγαλύτερη ακρίβεια τα όρια των διαταραχών.

Με άλλα λόγια, η απλή αναζήτηση των αιτιών δεν αρκεί για να καλύψει την πολύπλοκη φύση των ψυχιατρικών διαταραχών και απαιτείται μια πολυδιάστατη προσέγγιση στην ταξινόμηση. Όπως αναφέρει ο Zachar, οι ψυχιατρικές κατηγορίες δεν ανταποκρίνονται σε φυσικά είδη με σαφή οντολογικά όρια, αλλά αποτελούν κατηγορίες βασισμένες σε ομοιότητες συμπτωμάτων. Σύμφωνα με τα παραπάνω, η διάγνωση πρέπει να ειδωθεί ως ερμηνευτική πράξη.

Ορισμένοι θεωρητικοί υποστηρίζουν ότι η ψυχιατρική διάγνωση δεν είναι ουδέτερη περιγραφή βιολογικών ή ψυχολογικών γεγονότων, αλλά συμμετέχει ενεργά στη

διαμόρφωση του τρόπου με τον οποίο νοείται η ίδια η διαταραχή. Αυτή η άποψη δίνει έμφαση στον ρόλο των εννοιολογικών πλαισίων και της κοινωνικής πρακτικής στη συγκρότηση της ψυχικής παθολογίας. Σε αντίθεση με άλλου είδους διαγνώσεων, οι ψυχιατρικές διαγνώσεις δεν βασίζονται σε εργαστηριακούς δείκτες, αλλά σε πρότυπα συμπεριφοράς, τα οποία αξιολογούνται με βάση ερμηνευτικά κριτήρια. Αυτό καθιστά τη διάγνωση μια ερμηνευτική πράξη, στο μέτρο που δεν αποτυπώνει απλώς αντικειμενικά γεγονότα, αλλά αποδίδει νόημα σε εμπειρίες και συμπεριφορές εντός ενός συγκεκριμένου θεωρητικού και κοινωνικού πλαισίου (Zachar, 2014). Η γνωσιοθεωρητική συνιστώσα εμφανίζεται στην επιλογή των κριτηρίων- τι θεωρείται «φυσιολογικό» ή «ανώμαλο»- στον τρόπο ερμηνείας των αναφορών του ασθενούς και στην κανονιστική ισχύ των διαγνωστικών εγχειριδίων, τα οποία προδιαγράφουν το τι ερμηνεύεται ως ψυχοπαθολογικό. Ο ασθενής, όταν εντάσσεται σε μια κατηγορία του DSM, δεν ταυτοποιεί ή επανερμηνεύει τον εαυτό του απλώς. Η διάγνωση, έτσι, δεν είναι μόνο αποτέλεσμα παρατήρησης, αλλά και απόρροια νοήματος. Δηλαδή, καθορίζει πώς ο πάσχων εννοεί την εμπειρία του και πώς η κοινωνία τον αντιμετωπίζει. Από γνωσιοθεωρητική σκοπιά, επομένως, τίθεται το ερώτημα: τι σημαίνει να γνωρίζουμε κάτι ως διαταραχή; Και τι υποκρύπτει αυτό το "γνωρίζω"; Τέλος, παρόλο που η διάγνωση προσπαθεί να ταξινομήσει τη διαταραχή, το ίδιο το "πάσχειν" δεν καταγράφεται ούτε εξαντλείται μέσα στα διαγνωστικά εγχειρίδια. Ωστόσο, η διάγνωση, όσο δομική και αναγκαία κι αν είναι για την ψυχιατρική πρακτική, δεν παύει να εδράζεται σε ένα θεμελιώδες φιλοσοφικό ερώτημα: τι σημαίνει να πάσχει κανείς ψυχικά; Στην επόμενη ενότητα εξετάζονται τα εννοιολογικά όρια της ψυχικής διαταραχής, με στόχο να αναδειχθεί το ρήγμα μεταξύ των διαγνωστικών περιγραφών και της βιωματικής διάστασης του «πάσχειν».

Η διάγνωση στην ψυχοπαθολογία δεν αποτελεί απλώς τεχνικό εργαλείο καταγραφής συμπτωμάτων, αλλά επηρεάζει άμεσα το πώς ορίζεται και κατανοείται το «πάσχειν». Μέσα από τη διαδικασία της διάγνωσης, τα όρια ανάμεσα στο φυσιολογικό και το παθολογικό καθορίζονται, ενώ η κοινωνική, πολιτισμική και επιστημονική ερμηνεία της ψυχικής διαταραχής γίνεται πρακτικά εφαρμόσιμη. Η επόμενη ενότητα εξετάζει πώς η ψυχική διαταραχή συνδέεται με την έννοια του «πάσχειν», αναδεικνύοντας πώς οι νοητικές διεργασίες, η εμπειρική επίγνωση και οι αντιλήψεις για την παθολογία διαμορφώνουν τα εννοιολογικά όρια του πάσχειν, καθώς και πώς η διάγνωση λειτουργεί ως πρακτικό εργαλείο που ενσωματώνει αυτές τις φιλοσοφικές παραδοχές.

3.3 Διαταραχή και βιωματική εμπειρία: τι σημαίνει να «πάσχει» κανείς ψυχικά;

Η ψυχική διαταραχή δεν ορίζεται μονοσήμαντα και η έννοια της συνδέεται στενά με την εμπειρία του «πάσχειν». Σύμφωνα με τον Wakefield (1992), μια διαταραχή αποτελεί επιβλαβή δυσλειτουργία: η «δυσλειτουργία» αναφέρεται στην αποτυχία ενός ψυχικού μηχανισμού να εκπληρώσει τον προδιαγεγραμμένο σκοπό του, ενώ το «επιβλαβές» συνδέεται με τις κοινωνικές αξίες που καθορίζουν την αρνητική επίδραση στη ζωή του ατόμου. Έτσι, η διαταραχή εντάσσεται ανάμεσα στο φυσικό και το κοινωνικό πεδίο και απαιτεί κριτική αξιολόγηση από την ψυχιατρική και τις κοινωνικές επιστήμες (Wakefield, 1992; Boorse, 1977).

Ο Szasz (1960) αμφισβητεί την αντικειμενικότητα της ψυχικής ασθένειας, υποστηρίζοντας ότι οι ψυχικές διαταραχές δεν διαθέτουν φυσικούς δείκτες, όπως οι σωματικές παθήσεις και ότι η αναγνώρισή τους αντικατοπτρίζει κοινωνικές και πολιτισμικές αποκλίσεις. Ωστόσο, η κριτική αυτή παραβλέπει τη λειτουργική διάσταση της βλάβης: η δυσλειτουργία και η επιβλαβής επίδραση στην ευημερία καθιστούν τη διαταραχή πραγματικό φαινόμενο που απαιτεί αξιολόγηση από επαγγελματίες ψυχικής υγείας (Fulford, 1989; Wakefield, 1992).

Οι φιλοσοφικές προσεγγίσεις ορίζουν τη διαταραχή με διαφορετικούς τρόπους. Οι Scadding, Kendell και Boorse την εντοπίζουν μέσω μειωμένης επιβίωσης ή αναπαραγωγικής ικανότητας, προσπαθώντας να αποσυνδέσουν την έννοια από αξιακά κριτήρια. Άλλες προσεγγίσεις τονίζουν τη δυσλειτουργία του ψυχικού μηχανισμού ανεξαρτήτως κοινωνικών συνεπειών. Η πιο ολοκληρωμένη άποψη ενσωματώνει και το στοιχείο της πραγματικής βλάβης, συνδέοντας τη δυσλειτουργία με την επίδρασή της στην ευημερία του ατόμου (Wakefield, 1992, Boorse, 1977).

Η συζήτηση αυτή αναδεικνύει ότι η ψυχική διαταραχή δεν μπορεί να περιοριστεί αποκλειστικά σε βιολογικά ή κοινωνικά κριτήρια. Το «πάσχειν» υπερβαίνει την εννοιολογική κατηγορία της διαταραχής, περιλαμβάνοντας την υποκειμενική βιωματική διάσταση του ψυχικού πόνου. Η διάγνωση δεν αποτελεί απλώς καταγραφή συμπτωμάτων· είναι ερμηνευτική πράξη που καθορίζει το πώς ο ασθενής αντιλαμβάνεται την εμπειρία του και πώς η κοινωνία τον αντιμετωπίζει (Zachar, 2014; Fulford, 1989).

Η εξέταση της ψυχικής διαταραχής υπό το πρίσμα της βιωματικής εμπειρίας θέτει τα θεμέλια για τη σύνδεση με την Τεχνητή Νοημοσύνη: η διερεύνηση του τι σημαίνει να «πάσχει» ένα υποκείμενο επιτρέπει να τεθούν ερωτήματα σχετικά με το αν και πώς οι αλγοριθμικές και υπολογιστικές διεργασίες του Αι μπορούν να αναπαράγουν ή να προσομοιώσουν στοιχεία της ανθρώπινης ψυχοπαθολογίας, χωρίς να υποκαθιστούν την εμπειρική διάσταση του ψυχικού πόνου (Cooper, 2011; Huys et al., 2016).

Μέρος Β' Γνωσιοθεωρία και Τεχνητή Νοημοσύνη

4. Φιλοσοφική διερεύνηση της ψυχικής λειτουργίας στο πλαίσιο της Τεχνητής Νοημοσύνης

Η ενότητα αυτή εξετάζει πώς θεμελιώδεις έννοιες της γνώσης και της εμπειρίας συνδέονται με την Τεχνητή Νοημοσύνη και πώς η ανάλυση των ψυχικών λειτουργιών μπορεί να εφαρμοστεί σε τεχνητά συστήματα. Το βασικό ερώτημα είναι υπό ποιες συνθήκες μπορεί να υποστηριχθεί ότι ένα σύστημα Αι «γνωρίζει», «προσομοιώνει» ή ακόμη και «πάσχει», και πώς αυτά τα ζητήματα σχετίζονται με τη φύση της ανθρώπινης γνώσης, εμπειρίας και ψυχικής λειτουργίας (Gettier, 1963; Floridi, 2011).

Η παραδοσιακή επιστημολογία εστιάζει σε δύο βασικά ζητήματα: πρώτον, την κατανόηση της γνώσης, όπως εκφράζεται μέσω της δικαιολόγησης και της πεποίθησης, και δεύτερον, τον σκεπτικισμό που αμφισβητεί το κατά πόσο μπορούμε να γνωρίζουμε για τον εξωτερικό κόσμο. Το πρόβλημα της δικαιολόγησης, όπως απεικονίζεται στο δοκίμιο του Edmund Gettier (1963), υπογραμμίζει ότι η «δικαιολογημένη αληθής πεποίθηση» δεν εξασφαλίζει επαρκώς τη γνώση. Στην περίπτωση του Αι, η πρόκληση είναι να διαπιστωθεί αν η παραγωγή αληθών προτάσεων ή συμπεριφορών μπορεί να ταυτιστεί με γνώση, ενώ το σύστημα δεν διαθέτει επίγνωση, πρόθεση ή βιωματική εμπειρία (Floridi, 2011).

Τα συστήματα Αι βασίζονται σε στατιστικά μοντέλα και μηχανική μάθηση, παράγοντας πληροφορία που δεν ταυτίζεται με γνώση, κατά την επιστημολογική έννοια. Η διάκριση μεταξύ γνώσης και πληροφορίας, όπως την αναλύει ο Floridi (2011), δείχνει ότι η γνώση προκύπτει μόνο όταν υπάρχει ερμηνεία και κατανόηση, ενώ η πληροφορία μπορεί να είναι απλώς αναπαράσταση δεδομένων. Η μέθοδος της

αφαίρεσης επιτρέπει την ανάλυση του Αι σε διαφορετικά επίπεδα: επίπεδο αλγορίθμου, συμπεριφοράς ή εσωτερικής κατάστασης. Κατά αυτό τον τρόπο, το Αι μπορεί να μας παραπέμπει σε ψυχικές λειτουργίες στο επίπεδο συμπεριφοράς, αλλά δεν βιώνει την εμπειρία ή την υποκειμενικότητα που χαρακτηρίζει τον άνθρωπο (Putnam, 1967; Goldman, 1986).

Ο λειτουργισμός (functionalism) προτείνει ότι οι νοητικές καταστάσεις ταυτίζονται με τις λειτουργίες που επιτελούν. Σύμφωνα με τον Putnam, η ικανότητα ενός οργανισμού να βιώνει πόνο εξαρτάται από την οργανωτική δομή που συνδέει αισθητήρες και προτιμήσεις, παρά από τη φυσική σύσταση του εγκεφάλου. Αυτό επιτρέπει τη θεωρητική εφαρμογή σε συστήματα Αι, όπου η εκτέλεση λειτουργιών μπορεί να «υποδηλώνει» νοητικές καταστάσεις, χωρίς όμως βιωματική πρόσβαση. Αντίστοιχα, οι εξωτερικιστικές θεωρίες της γνώσης (Goldman, 1986) απαιτούν αιτιώδη σύνδεση με τον κόσμο, κάτι που τα συστήματα Αι δεν διαθέτουν, καθώς δεν έχουν άμεση αισθητηριακή εμπειρία του εξωτερικού περιβάλλοντος.

Η έννοια της πρόθεσης (intentionality), κατά Brentano, είναι θεμελιώδης για κάθε νοητική κατάσταση. Το επιχείρημα του «Κινέζικου Δωματίου» (Searle, 1980) δείχνει ότι η παραγωγή σωστών απαντήσεων δεν συνεπάγεται με κατανόηση ή πρόθεση. Ο Dennett (1987) προτείνει ότι μπορούμε να θεωρούμε τα συστήματα Αι ως έχοντα «προθετική στάση» για λόγους προβλεπτικότητας, χωρίς να κατέχουν πραγματική συνείδηση. Τέλος, η έννοια του «phenomenal self-model» (Metzinger, 2003) υπογραμμίζει ότι η επίγνωση του εαυτού είναι κρίσιμη για τη γνώση και την ψυχική λειτουργία, κάτι που τα συστήματα Αι απουσιάζουν πλήρως.

Παράλληλα, η φιλοσοφική διερεύνηση αναδεικνύει τα όρια της προσομοίωσης: τα συστήματα Αι μπορεί να αναγνωρίζουν μοτίβα ή να παίρνουν αποφάσεις, αλλά δεν βιώνουν ούτε κατανοούν τις καταστάσεις αυτές. Η σύγκριση με την ψυχιατρική πρακτική δείχνει ότι η διάκριση μεταξύ εξωτερικών δεικτών και βιωματικής εμπειρίας είναι θεμελιώδης. Στο DSM, η διάγνωση βασίζεται σε παρατηρήσιμα συμπτώματα, ενώ η υποκειμενική εμπειρία του ασθενούς παραμένει αθέατη. Ομοίως, το Αι μπορεί να εμφανίζει δυσλειτουργίες ή σφάλματα που συμπίπτουν με ψυχικές καταστάσεις, χωρίς όμως να διαθέτει βιωματική πρόσβαση σε αυτά. Αυτό καθιστά σαφές ότι η γνώση, η πρόθεση και το πάσχειν συνδέονται με την εμπειρική, βιωματική διάσταση της ψυχικής ζωής, που το Αι δεν μπορεί να αναπαραστήσει (Wakefield, 1992).

Καταληκτικά, η Γνωσιοθεωρία καθορίζει σαφή όρια στην κατανόηση της τεχνητής νοημοσύνης. Το Αι μπορεί να επιτελεί γνωσιακές λειτουργίες και να προσομοιώνει συμπεριφορές, αλλά δεν διαθέτει κατανόηση, πρόθεση ή βιωματική εμπειρία. Η διαφορά μεταξύ γνώσης και πληροφορίας, κατανόησης και λειτουργίας, καθώς και μεταξύ ανθρώπινης εμπειρίας και προσομοίωσης, παραμένει θεμελιώδης για τη φιλοσοφική ανάλυση της τεχνητής νοημοσύνης (Floridi, 2011; Metzinger, 2003). Η επόμενη ενότητα εστιάζει στη σύνδεση αυτής της συζήτησης με την ψυχιατρική διάγνωση και τη βιωματική διάσταση του «πάσχειν».

4.1 Λειτουργισμός, προσομοίωση και πρόθεση

Η κεντρική ερώτηση αυτής της εργασίας, εάν ένα σύστημα τεχνητής νοημοσύνης μπορεί πραγματικά να γνωρίζει ή απλώς να προσομοιώνει τη γνώση, αποτελεί τον πυρήνα της γνωσιοθεωρητικής διερεύνησης της τεχνητής νοημοσύνης. Το ζήτημα συνδέεται άρρηκτα με την έννοια της προθετικότητας ή αποβλεπτικότητας (intentionality), όπως την εισήγαγε ο Franz Brentano (1874) και την επεξεργάστηκαν μεταγενέστεροι φιλόσοφοι, όπως ο John Searle και ο Daniel Dennett. Το ερώτημα «Μπορεί ένα σύστημα να γνωρίζει;» προϋποθέτει διερεύνηση της ενσώματης εμπειρίας, της πρόθεσης και πιθανώς της συνείδησης, αναδεικνύοντας τα όρια μεταξύ λειτουργικής προσομοίωσης και βιωματικής νόησης.

Η κεντρική ερώτηση αυτής της εργασίας, εάν ένα σύστημα τεχνητής νοημοσύνης μπορεί πραγματικά να γνωρίζει ή απλώς να προσομοιώνει τη γνώση, αποτελεί τον πυρήνα της γνωσιοθεωρητικής διερεύνησης της τεχνητής νοημοσύνης. Το ζήτημα συνδέεται άρρηκτα με την έννοια της προθετικότητας ή αποβλεπτικότητας (intentionality), όπως την εισήγαγε ο Franz Brentano (1874) και την επεξεργάστηκαν μεταγενέστεροι φιλόσοφοι, όπως ο John Searle και ο Daniel Dennett. Το ερώτημα «Μπορεί ένα σύστημα να γνωρίζει;» προϋποθέτει διερεύνηση της ενσώματης εμπειρίας, της πρόθεσης και πιθανώς της συνείδησης, αναδεικνύοντας τα όρια μεταξύ λειτουργικής προσομοίωσης και βιωματικής νόησης.

Ο λειτουργισμός υποστηρίζει ότι οι νοητικές καταστάσεις καθορίζονται λειτουργικά, δηλαδή από τις αιτιακές τους σχέσεις με ερεθίσματα, άλλες νοητικές καταστάσεις και συμπεριφορές. Δεν έχει σημασία το υλικό υπόβαθρο (βιολογικό ή τεχνητό) αλλά το

πώς λειτουργεί το σύστημα. Ο Putnam προτείνει έναν λειτουργιστικό ορισμό των νοητικών καταστάσεων, σύμφωνα με τον οποίο μια νοητική κατάσταση ορίζεται από τη θέση που κατέχει σε ένα αιτιακό δίκτυο σχέσεων: δηλαδή, από το πώς συνδέεται με ερεθίσματα (input), με άλλες νοητικές καταστάσεις (state transitions) και συμπεριφορές (output) (Putnam, H. (1975), σελ. 73-79). Για τον Putnam, δεν είναι σημαντικό από τι είναι φτιαγμένο το όν που έχει νοητικές καταστάσεις, αλλά πώς αυτές οι καταστάσεις λειτουργούν. Έτσι, ένας άνθρωπος, ένας εξωγήινος ή μια μηχανή μπορεί να έχει την ίδια «νοητική κατάσταση», αν λειτουργεί με τον ίδιο τρόπο. Ο συγγραφέας συζητά τη φύση του πόνου, υποστηρίζοντας ότι ο πόνος δεν είναι απλώς μια κατάσταση του εγκεφάλου, αλλά μια λειτουργική κατάσταση ολόκληρου του οργανισμού. Συγκεκριμένα, παρουσιάζει μια εμπειρική υπόθεση ότι ο πόνος γίνεται καλύτερα κατανοητός, μέσα από μια ευρύτερη προοπτική από την υπόθεση της κατάστασης του εγκεφάλου. Για να υποστηρίξει αυτήν την ιδέα, ο συγγραφέας εισάγει έννοιες όπως η Μηχανή Turing και ένα Πιθανοτικό Αυτόματο (Probabilistic Automaton), εξηγώντας ότι αυτά τα μοντέλα μπορούν να αναπαραστήσουν τις μεταβάσεις κατάστασης ενός οργανισμού, με βάση αισθητηριακές εισόδους και κινητικές εξόδους. Οι υποθέσεις σχετικά με αυτά τα μοντέλα περιλαμβάνουν γενικευμένους ορισμούς, που επιτρέπουν διάφορα πιθανοτικά αποτελέσματα.

Εμπνευσμένος από το έργο του Alan Turing σχετικά με τις μηχανές Turing, ανέπτυξε μια φιλοσοφική θεωρία που ονομάζεται υπολογιστικός λειτουργισμός. Αυτή η θεωρία χρησιμοποιεί τη μηχανή Turing, ως μοντέλο για την κατανόηση της φύσης των νοητικών καταστάσεων και του προβλήματος νου-σώματος. Η προσέγγιση του Putnam υποδηλώνει ότι οι νοητικές καταστάσεις μπορούν να γίνουν κατανοητές ως λειτουργικές καταστάσεις μέσα σε ένα υπολογιστικό σύστημα, ανάλογες με τις καταστάσεις μιας μηχανής Turing. Οι μηχανές Turing είναι αφηρημένες υπολογιστικές συσκευές, που χειρίζονται σύμβολα σε μια ταινία με βάση την τρέχουσα κατάστασή τους και το σύμβολο που διαβάζεται. Αποτελούν μια θεμελιώδη έννοια στην επιστήμη των υπολογιστών και στα θεωρητικά μοντέλα υπολογισμού. Ο Putnam χρησιμοποίησε τη μηχανή Turing ως αναλογία για να εξηγήσει τις νοητικές καταστάσεις. Υποστήριξε ότι οι νοητικές καταστάσεις, όπως οι πεποιθήσεις ή οι πόνοι, θα μπορούσαν να γίνουν κατανοητές ως λειτουργικές καταστάσεις μέσα σε ένα υπολογιστικό σύστημα, ανεξάρτητα από το φυσικό υπόστρωμα (όπως ο εγκέφαλος). Όπως ακριβώς η συμπεριφορά μιας μηχανής Turing καθορίζεται από την κατάστασή της και το σύμβολο

που διαβάσει, ο Putnam πρότεινε ότι οι νοητικές καταστάσεις ορίζονται λειτουργικά από τον ρόλο τους σε ένα ευρύτερο σύστημα νοητικών διεργασιών. Ο λειτουργισμός του Putnam προσφέρει έναν τρόπο κατανόησης του νου, χωρίς να βασιζόμαστε σε μια καθαρά φυσική ή δυϊστική άποψη. Υποδηλώνει ότι οι νοητικές καταστάσεις αφορούν περισσότερο αυτό που κάνουν (τη λειτουργία τους), παρά αυτό από το οποίο αποτελούνται (τη φυσική τους ουσία). Οι ιδέες του Putnam υποδηλώνουν ότι οποιοδήποτε σύστημα, είτε εγκέφαλος είτε υπολογιστής, θα μπορούσε ενδεχομένως να έχει νοητικές καταστάσεις εάν έχει την κατάλληλη λειτουργική οργάνωση. Άρα, μια μηχανή, αν είναι ρυθμισμένη σωστά, μπορεί να υλοποιήσει νοητικές καταστάσεις-δηλαδή, να έχει νου.

Ως επακόλουθο, ο Putnam εισάγει την ιδέα ότι οι νοητικές καταστάσεις δεν ταυτίζονται με συγκεκριμένες βιολογικές καταστάσεις (π.χ., καταστάσεις του εγκεφάλου), αλλά μπορούν να υλοποιηθούν σε διαφορετικές μορφές. Για παράδειγμα, ο «πόνος» μπορεί να είναι μια εγκεφαλική κατάσταση σε άνθρωπο, αλλά μια άλλη κατάσταση σε έναν εξωγήινο ή σε ένα σύστημα Α_i, αν και οι τρεις λειτουργούν με τον ίδιο τρόπο. Αυτή η θέση αντιτίθεται στον αναγωγισμό, που ορίζει ότι κάθε νοητική κατάσταση είναι μια συγκεκριμένη εγκεφαλική κατάσταση και ανοίγει τον δρόμο για τη συζήτηση περί Α_i με νου. Παρόλα αυτά, εγείρεται το ερώτημα σχετικά με το εάν ένα Α_i επιτελεί την ίδια λειτουργία με έναν άνθρωπο, αυτό συνεπάγεται με το γεγονός ότι διαθέτει και νου; Ο Putnam αφήνει ανοικτό αυτό το ενδεχόμενο ότι δηλαδή μια μηχανή, αν υλοποιεί το ίδιο λειτουργικό σχήμα, μπορεί να έχει νου όπως ο άνθρωπος. Ωστόσο, ο Putnam αναγνωρίζει ήδη δύο κύριες προβληματικές του λειτουργισμού. Αρχικά, Δεν είναι ξεκάθαρο πώς ο λειτουργισμός εξηγεί τα υποκειμενικά περιεχόμενα της εμπειρίας (qualia) π.χ., τι σημαίνει να νιώθεις πόνο και ένα σύστημα χωρίς συνείδηση, μόνο με λειτουργικές σχέσεις, γνωρίζει πραγματικά; Συμπερασματικά, ο Putnam παρά τις όποιες ενστάσεις, θεμελίωσε την ιδέα ότι ο νους μπορεί να «λειτουργεί» ανεξαρτήτως υλικής βάσης. Αυτό επιτρέπει να σκεφτούμε την τεχνητή νοημοσύνη ως υποψήφιο φορέα νοητικών καταστάσεων, αν υλοποιεί την ίδια λειτουργία.

Ωστόσο, το επιχείρημα του Κινέζικου Δωματίου του John Searle (1980) έθεσε σοβαρές αμφιβολίες, ως προς το αν η λειτουργική προσομοίωση επαρκεί για πραγματική γνώση ή συνείδηση. Ο Searle υποστηρίζει ότι ένα σύστημα μπορεί να χειρίζεται σύμβολα (syntactic manipulation), χωρίς να κατανοεί το νόημά τους (semantics). Συνεπώς, ένα

σύστημα που εκτελεί λειτουργικά τις ίδιες διεργασίες με έναν άνθρωπο δεν γνωρίζει, παρά μόνο προσομοιώνει την πράξη της γνώσης.

Πρόσφατες προσπάθειες σε προσομοιώσεις υπολογιστών των ανθρώπινων γνωστικών ικανοτήτων εγείρουν σημαντικά ψυχολογικά και φιλοσοφικά ερωτήματα. Η διάκριση μεταξύ «ισχυρής» και «αδύναμης» Τεχνητής Νοημοσύνης είναι απαραίτητη. Η Ασθενής Τεχνητή Νοημοσύνη θεωρεί τους υπολογιστές ως εργαλεία για τη μελέτη του ανθρώπινου νου, χωρίς να τους αποδίδει πραγματικές γνωστικές ικανότητες. Οι υπολογιστές μπορούν να εκτελούν υπολογισμούς, να προσομοιώνουν νοητικές διεργασίες και να επεξεργάζονται πληροφορίες, αλλά δεν έχουν ούτε συνείδηση ούτε κατανόηση. Ο σκοπός τους είναι να διευκολύνουν την επιστημονική έρευνα, όπως η μοντελοποίηση της ανθρώπινης μνήμης, η προσομοίωση λογικών συλλογισμών ή η ανάλυση γλωσσικών δομών. Για παράδειγμα, ένα πρόγραμμα που προσομοιώνει την αναγνώριση προτύπων σε εικόνες μπορεί να μας βοηθήσει να κατανοήσουμε πώς οι άνθρωποι αναγνωρίζουν πρόσωπα, χωρίς το ίδιο το πρόγραμμα να «καταλαβαίνει» ή να βιώνει την αναγνώριση (Russell & Norvig, 2021, κεφ. 1–2).

Αντίθετα, η Ισχυρή Τεχνητή Νοημοσύνη υποστηρίζει ότι οι υπολογιστές, εάν προγραμματιστούν κατάλληλα, μπορούν να διαθέτουν γνήσιες γνωστικές καταστάσεις. Δηλαδή, δεν πρόκειται απλώς για προσομοίωση της γνώσης ή της κατανόησης, αλλά για πραγματική νοητική κατάσταση που αντιστοιχεί σε αυτό που ένας άνθρωπος βιώνει όταν κατανοεί κάτι. Σύμφωνα με αυτή την προσέγγιση, ένας υπολογιστής θα μπορούσε θεωρητικά να «κατανοεί» τη γλώσσα, να κάνει συνειδητές επιλογές και να παρέχει εξηγήσεις για νοητικές διαδικασίες, χωρίς η γνώση του να περιορίζεται σε τυπικούς υπολογισμούς (Searle, 1980; Haugeland, 1985). Η ισχυρή TN επιχειρεί να γεφυρώσει το χάσμα μεταξύ υπολογιστικής προσομοίωσης και πραγματικής νοητικής εμπειρίας, θεωρώντας ότι ένα κατάλληλα σχεδιασμένο σύστημα μπορεί να θεωρηθεί φορέας γνώσης, όχι απλώς προσομοιωτής ανθρώπινης συμπεριφοράς.

Η διάκριση αυτή είναι θεμελιώδης για τη φιλοσοφία της Τεχνητής Νοημοσύνης, διότι επισημαίνει ότι η συμπεριφορά δεν είναι πάντα δείκτης κατανόησης. Ένα πρόγραμμα μπορεί να απαντά σωστά σε ερωτήσεις, να επιλύει προβλήματα και να μαθαίνει από εμπειρίες, αλλά δεν αποδίδεται σε αυτόν τον υπολογιστή πραγματική νοητική εμπειρία ή πρόθεση. Αντίθετα, η ισχυρή TN ισχυρίζεται ότι, υπό τις κατάλληλες συνθήκες, η ίδια συμπεριφορά θα μπορούσε να συνοδεύεται από πραγματική πρόθεση και

κατανόηση, καθιστώντας τον υπολογιστή νοητικό φορέα. Αυτή η διάκριση συνδέεται άμεσα με το επιχείρημα του Searle στο «Κινεζικό Δωμάτιο», όπου η μηχανή εκτελεί υπολογισμούς χωρίς να κατανοεί (Searle, 1980). Η ισχυρή TN αμφισβητεί ότι η λειτουργία και η συμπεριφορά από μόνες τους είναι επαρκείς για την απόδοση γνώσης και πρόθεσης, ανοίγοντας τη συζήτηση για την πραγματική νοητική κατάσταση των συστημάτων ΑΙ. Ο Searle επικεντρώνεται στην Ισχυρή Τεχνητή Νοημοσύνη, ιδιαίτερα στον ισχυρισμό της ότι οι προγραμματισμένοι υπολογιστές μπορούν πραγματικά να διαθέτουν γνωστικές καταστάσεις, χρησιμεύοντας ως εξηγήσεις για την ανθρώπινη νόηση.

Το έργο του Roger Schank αναδεικνύει πώς ένα σύστημα Τεχνητής Νοημοσύνης μπορεί να προσομοιώνει την κατανόηση ανθρώπινων ιστοριών, χωρίς όμως να διαθέτει πραγματική νοητική εμπειρία. Μέσω των αλγορίθμων του, μια μηχανή μπορεί να εξάγει συμπεράσματα για γεγονότα που δεν αναφέρονται ρητά, όπως η διαπίστωση ότι ένας άντρας που φεύγει θυμωμένος χωρίς να φάει το χάμπουργκερ δεν το κατανάλωσε. Συγκεκριμένα, η παραπάνω ιστορία πιστοποιεί το γεγονός ότι αυτή η διαδικασία μιμείται την ανθρώπινη κατανόηση. Παρόλα αυτά, δεν συνεπάγεται με το γεγονός ότι η μηχανή κατανοεί πραγματικά τα γεγονότα, απλώς εκτελεί εντολές και υπολογιστικές διαδικασίες που προσομοιώνουν την ανθρώπινη δραστηριότητα. Το παράδειγμα αυτό υποστηρίζει τη διάκριση μεταξύ ασθενούς και ισχυρής Τεχνητής Νοημοσύνης. Η ασθενής TN χρησιμοποιεί τέτοιες προσομοιώσεις για μελέτη και πρόβλεψη, ενώ η ισχυρή TN θα απαιτούσε ότι η μηχανή κατέχει πραγματική νοητική κατάσταση, με συνείδηση και κατανόηση.

Ο John Searle (1980) πρότεινε το γνωστό «Πείραμα Σκέψης του Κινεζικού Δωματίου», για να αμφισβητήσει την υπόθεση ότι η λειτουργική προσομοίωση συμπεριφοράς ισοδυναμεί με κατανόηση ή νοητική κατάσταση. Στο παράδειγμα, φαντάζεται τον εαυτό του κλεισμένο σε ένα δωμάτιο, με μια σειρά κινεζικών συμβόλων και ένα σύνολο κανόνων για να τα συνδυάζει. Ακολουθώντας τους κανόνες, μπορεί να παράγει σωστές απαντήσεις σε ερωτήσεις γραμμένες στα κινεζικά, ακόμη και αν τα ίδια τα σύμβολα και η γλώσσα του είναι εντελώς άγνωστα.

Το κρίσιμο σημείο είναι ότι, παρόλο που οι απαντήσεις του μπορεί να φαίνονται ορθές σε εξωτερικούς παρατηρητές, ο ίδιος δεν κατανοεί καθόλου τη σημασία τους. Χειρίζεται τα σύμβολα μηχανικά, δηλαδή εκτελεί «συντακτικές» λειτουργίες χωρίς

«σημασιολογική» κατανόηση. Το πείραμα αυτό καταδεικνύει ότι η απλή εκτέλεση υπολογιστικών διαδικασιών, όπως οι κανόνες του κινέζικου δωματίου δεν συνεπάγονται με γνώση, πρόθεση ή συνείδηση.

Συνεπώς, ο Searle αμφισβητεί τον ισχυρισμό της Ισχυρής Τεχνητής Νοημοσύνης, σύμφωνα με τον οποίο ένας υπολογιστής που εκτελεί κατάλληλα προγραμματισμένες διαδικασίες μπορεί να έχει γνήσιες γνωστικές καταστάσεις. Όπως δείχνει το Κινεζικό Δωμάτιο, ένα σύστημα μπορεί να προσομοιώνει τη γνώση χωρίς να γνωρίζει πραγματικά. Οι υπολογιστές διαθέτουν μόνο σύνταξη, ενώ η γνώση και η πρόθεση προϋποθέτουν κάποιο είδος κατανόησης ή συνείδησης, που δεν παράγεται απλώς από λειτουργικές διαδικασίες.

Σε αυτό το σημείο, ο Searle εισάγει μια καθοριστική διάκριση μεταξύ Syntax: μορφολογική δομή συμβόλων, κανόνες χειρισμού, χωρίς νόημα και Semantics: σημασία, κατανόηση, πρόθεση. Συγκεκριμένα, οι υπολογιστές έχουν μόνο syntax, δηλαδή λειτουργούν με κανόνες χειρισμού συμβόλων (π.χ., αλγόριθμοι), ενώ νοήμονες άνθρωποι έχουν semantics, δηλαδή κατανοούν τη σημασία των συμβόλων, διαθέτοντας υποκειμενική εμπειρία και πρόθεση. Το επιχείρημα παρουσιάζει μια κριτική στην ιδέα ότι η τεχνητή νοημοσύνη (TN) μπορεί να οριστεί απλώς από τις υπολογιστικές της διαδικασίες. Αμφισβητεί την ιδέα ότι η σκοπιμότητα μπορεί να προκύψει αποκλειστικά από δομημένους αλγόριθμους που εφαρμόζονται σε μηχανές, υποστηρίζοντας ότι η κατανόηση εξαρτάται από κάτι περισσότερο από τυπικούς χειρισμούς συμβόλων. Ο αρχικός ισχυρισμός της ισχυρής TN είναι ότι οι νοητικές διαδικασίες είναι υπολογιστικές, αλλά ο επαναπροσδιορισμός της TN ως οποιουδήποτε συστήματος που επιδεικνύει νόηση υπονομεύει αυτή τη θέση και αναιρεί οποιαδήποτε ελέγξιμη υπόθεση. Παρά την αναγνώριση ότι μια μηχανή θα μπορούσε θεωρητικά να έχει την ικανότητα να κατανοεί γλώσσες όπως τα αγγλικά ή τα κινέζικα, το κείμενο τονίζει ότι μια καθαρά υπολογιστική προσέγγιση δεν διαθέτει τις απαραίτητες αιτιακές δυνάμεις για να δημιουργήσει σκοπιμότητα. Η κατανόηση δεν είναι απλώς το αποτέλεσμα της εκτέλεσης ενός προγράμματος, αλλά είναι βαθιά ριζωμένη στη βιολογική και φυσική δομή ενός οργανισμού. Αυτή η συζήτηση υπογραμμίζει ότι η σκοπιμότητα συνδέεται με τις εγγενείς ιδιότητες του συστήματος και όχι μόνο με τις τυπικές λειτουργίες που εκτελεί. Επίσης, διατυπώνεται η άποψη ότι μια μηχανή, η οποία λειτουργεί αποκλειστικά μέσω τυπικών διαδικασιών, δεν μπορεί να έχει εγγενώς κατανόηση. Ενώ ο ισχυρισμός, ότι οι μηχανές μπορούν να σκέφτονται είναι έγκυρος, ιδιαίτερα αν

μιμούνται βιολογικές δομές, το ερώτημα εξαρτάται από το εάν ένας ψηφιακός υπολογιστής μπορεί να σκέφτεται αποκλειστικά από τον προγραμματισμό του. Η απάντηση είναι αρνητική, επειδή οι διαδικασίες στερούνται εγγενούς νοήματος. Εμπλέκονται σε σύνταξη χωρίς σημασιολογία. Το πείραμα σκέψης σε κινέζικο δωμάτιο αποτελεί παράδειγμα αυτού του σημείου, καταδεικνύοντας ότι ένας άνθρωπος προγραμματισμένος να ακολουθεί ένα συντακτικό σύστημα δεν αποκτά κατανόηση από το ίδιο το σύστημα, υπογραμμίζοντας ότι τα σύμβολα χωρίς νόημα δεν μπορούν να οδηγήσουν σε σκόπιμη λειτουργία.

Ο Searle απορρίπτει την άποψη της ισχυρής τεχνητής νοημοσύνης, υποστηρίζοντας ότι δεν αρκεί η λειτουργική προσομοίωση των νοητικών καταστάσεων, αλλά απαιτείται η ύπαρξη κατανόησης. Η κατανόηση, κατά τον Searle, προϋποθέτει προθετικότητα (intentionality), δηλαδή τη σημασιολογική αναφορά των νοητικών καταστάσεων σε καταστάσεις του κόσμου. Εάν ένα τεχνητό σύστημα στερείται προθετικότητας, τότε δεν «γνωρίζει» με την ισχυρή έννοια του όρου, αλλά χειρίζεται σύμβολα με καθαρά συντακτικό τρόπο, προσομοιώνοντας τη γνώση χωρίς να τη διαθέτει πραγματικά.

Στο ίδιο πλαίσιο, ο Searle ασκεί κριτική στον λειτουργισμό του Putnam, σύμφωνα με τον οποίο η αναπαραγωγή των ίδιων λειτουργικών σχέσεων αρκεί για την απόδοση νοητικών καταστάσεων. Ο Searle αντιτίθεται στη θέση αυτή, υποστηρίζοντας ότι η λειτουργία δεν συνεπάγεται κατ' ανάγκην κατανόηση και ότι η προθετικότητα δεν αναδύεται αποκλειστικά από λειτουργικές δομές. Υπό αυτή την έννοια, η προσομοίωση νοητικών λειτουργιών δεν αρκεί για να θεμελιώσει γνώση και κατ' επέκταση δεν επαρκεί για την απόδοση ψυχικού «πάσχειν» σε ένα σύστημα Τεχνητής Νοημοσύνης.

Ο Franz Brentano (1874) εισήγαγε την έννοια της προθετικότητας (intentionality) ως το θεμελιώδες γνώρισμα όλων των ψυχικών φαινομένων (psychische Phänomene), δηλαδή των πράξεων της συνείδησης, όπως η αντίληψη, η σκέψη, η επιθυμία και η κρίση. Σύμφωνα με τη γνωστή διατύπωσή του, «κάθε ψυχικό φαινόμενο εμπεριέχει κάτι ως αντικείμενό του· κάθε ψυχική πράξη είναι κατευθυνόμενη προς κάτι» (Brentano, 1874/1973, σ. 88–91). Με τον όρο αυτό, ο Brentano δεν αναφέρεται σε «νοητικές καταστάσεις» με τη σύγχρονη γνωσιακή έννοια, αλλά στη δομική κατευθυντικότητα της συνείδησης προς ένα αντικείμενο. Με άλλα λόγια, οι νοητικές καταστάσεις είναι πάντα για κάτι, δηλαδή αναφέρονται σε κάποιο περιεχόμενο: ένα αντικείμενο ή μια προσδοκία. Ο προσανατολισμός του νου σε κάτι εξωτερικό είναι το

κριτήριο που διαχωρίζει νοητικές από τις φυσικές καταστάσεις, όπως θερμοκρασία, οι οποίες δεν έχουν περιεχόμενο. Η πρόκληση για το Αί είναι αν ένα σύστημα μπορεί να έχει γνήσια πρόθεση, δηλαδή να αναφέρεται σε κάτι, ή απλώς προσομοιώνει τη συμπεριφορά οντοτήτων που διαθέτουν πρόθεση. Για παράδειγμα, το Chat GPT μπορεί να σκέφτεται με πρόθεση ή απλώς συνδυάζει λέξεις που περιγράφουν τέτοιες σκέψεις;

Αναλυτικότερα, κατά τον Brentano, τα δεδομένα της συνείδησης διακρίνονται σε φυσικά και ψυχικά φαινόμενα. Τα φυσικά φαινόμενα αφορούν το αισθητό περιεχόμενο (όπως χρώματα ή ήχους), ενώ τα ψυχικά φαινόμενα αφορούν τις ίδιες τις πράξεις της συνείδησης, όπως η αντίληψη, η φαντασία και η κρίση. Η σύγχυση προκύπτει όταν δεν διακρίνεται επαρκώς το περιεχόμενο από την προθετικά κατευθυνόμενη πράξη, γεγονός που οδηγεί σε θεωρητικές αντιφάσεις ως προς τη φύση της νόησης. Η διάκριση αυτή έχει πρωτίστως φιλοσοφική σημασία, καθώς θεμελιώνει την προθετικότητα ως το ειδοποιό γνώρισμα του ψυχικού και όχι ως εμπειρικό δεδομένο της ψυχολογίας (Brentano, 1874/1973).

Ο Brentano υποστηρίζει ότι το μοναδικό και καθοριστικό γνώρισμα των νοητικών φαινομένων είναι η προθετικότητα (intentionality), δηλαδή το γεγονός ότι είναι πάντοτε κατευθυνόμενα προς κάτι. Κάθε νοητική πράξη αναφέρεται σε ένα αντικείμενο, ανεξαρτήτως του αν αυτό υφίσταται πραγματικά ή όχι. Για παράδειγμα, όταν σκέφτομαι έναν μονόκερο, η σκέψη μου έχει ως αποβλεπτικό αντικείμενο ένα φανταστικό ον, το οποίο δεν υπάρχει στον φυσικό κόσμο αλλά υφίσταται ως αντικείμενο της νοητικής μου πράξης. Αυτή η ιδιότητα ονομάζεται από τον Brentano «ενδοϋπάρξη» ή intentional inexistence.

Σε αντίθεση με τα νοητικά φαινόμενα, τα φυσικά φαινόμενα δεν χαρακτηρίζονται από προθετικότητα, καθώς απλώς συμβαίνουν χωρίς να αναφέρονται σε κάποιο αντικείμενο. Με βάση αυτή τη διάκριση, η ψυχολογία, κατά τον Brentano, οφείλει να ορίσει το αντικείμενό της όχι βάσει υλικών ή φυσιολογικών ιδιοτήτων, αλλά βάσει της προθετικής δομής των ψυχικών πράξεων.

Κάθε νοητική πράξη, σύμφωνα με τον Brentano, εμπεριέχει μια μορφή παρουσίας (presentation), δηλαδή μια νοητική αναφορά σε αντικείμενο, η οποία αποτελεί τη βάση ακόμη και για πιο σύνθετες πράξεις, όπως η επιθυμία ή η κρίση. Ωστόσο, η καθολικότητα της προθετικότητας εγείρει φιλοσοφικά ερωτήματα σε περιπτώσεις, όπως ο πόνος ή η ψυχική οδύνη, όπου το αποβλεπτικό αντικείμενο δεν είναι πάντοτε

σαφώς εξωτερικό. Στις περιπτώσεις αυτές, η προθετικότητα φαίνεται να στρέφεται προς το ίδιο το σώμα ή προς μια βιωματική κατάσταση δυσφορίας, γεγονός που υποδεικνύει ότι η προθετική δομή της εμπειρίας δεν ταυτίζεται πάντοτε με αναπαραστατικά αντικείμενα του κόσμου.

Σημαντική επισήμανση του Brentano αποτελεί το γεγονός ότι η συνείδηση δεν αποτελεί απλώς μια πρόσθετη ικανότητα του νου αλλά σύμφυτη με τη νοητική πράξη. Επομένως, υφίσταται η αυτοαναφορικότητα της εμπειρίας, καθώς όταν σκέφτομαι, γνωρίζω ότι σκέφτομαι. Άρα, χωρίς πρόθεση δεν υπάρχει ούτε εμπειρία, ούτε συνείδηση. Σύμφωνα με την αυτή την προσέγγιση του Brentano, οι νοητικές πράξεις χαρακτηρίζονται από προθετικότητα, δηλαδή από το γεγονός ότι είναι κατευθυνόμενες προς ένα αντικείμενο. Η προσέγγιση αυτή δεν υιοθετείται εδώ αξιωματικά, αλλά λειτουργεί ως εννοιολογικό πλαίσιο για να εξεταστούν οι προϋποθέσεις υπό τις οποίες μπορεί να αποδοθεί γνώση ή πάσχειν σε ένα υποκείμενο. Εφόσον το πάσχειν συνδέεται με νοητικές καταστάσεις που έχουν αποβλεπτικό χαρακτήρα και βιωματικό περιεχόμενο, προκύπτει το ερώτημα κατά πόσο ένα σύστημα μπορεί να θεωρηθεί ότι γνωρίζει ή πάσχει, αν οι καταστάσεις του περιορίζονται σε υπολογιστικές διεργασίες και εξαγωγή δεδομένων. Στο πλαίσιο αυτό, η προθετικότητα δεν ταυτίζεται με την καθημερινή έννοια της πρόθεσης ως σκοπού, αλλά με την αποβλεπτικότητα της συνείδησης: το γεγονός ότι οι νοητικές καταστάσεις είναι πάντοτε περί τίνος. Η σκέψη, η επιθυμία, η πεποίθηση ή ο πόνος δεν είναι απλώς εσωτερικές καταστάσεις, αλλά σχέσεις προς ένα αντικείμενο ή μια κατάσταση πραγμάτων. Το κρίσιμο φιλοσοφικό ερώτημα που ανακύπτει εδώ αφορά την περίπτωση φαινομένων όπως ο πόνος: ποιο είναι το αποβλεπτικό αντικείμενο του πόνου; Ακόμη και ο πόνος διαθέτει προθετική δομή, στο μέτρο που στρέφεται προς μια σωματική βλάβη ή μια δυσάρεστη κατάσταση του σώματος, όπως αυτή βιώνεται από το υποκείμενο. Έτσι, η προθετικότητα συνδέεται άρρηκτα με την υποκειμενική εμπειρία και τη συνείδηση. Υπό αυτή την έννοια, η θεωρία του Brentano θέτει αυστηρούς όρους για την απόδοση γνώσης ή πάσχειν: ένα σύστημα μπορεί να θεωρηθεί φορέας νοητικών καταστάσεων μόνο εφόσον διαθέτει συνειδησιακή προθετικότητα, δηλαδή εσωτερική αναφορικότητα προς αντικείμενα.

Στην περίπτωση των συστημάτων Τεχνητής Νοημοσύνης, δεν υπάρχει ένδειξη ότι διαθέτουν προθετικότητα με τη φαινομενολογική έννοια του όρου, ούτε εμπειρία με την έννοια της υποκειμενικής, πρώτου προσώπου πρόσβασης σε καταστάσεις πόνου,

δυσφορίας ή νοηματικής αναφοράς. Ως εκ τούτου, ακόμη και αν τα συστήματα αυτά παράγουν συμπεριφορές που προσομοιάζουν γνωστικές ή ψυχοπαθολογικές εκδηλώσεις, παραμένει ανοιχτό το ερώτημα αν πρόκειται για γνήσιο πάσχειν ή για λειτουργική προσομοίωση χωρίς βιωματικό υπόβαθρο.

Σε αντίστιξη προς αυτή την οντολογικά ισχυρή θέση, ο Daniel Dennett προτείνει μια ριζικά διαφορετική προσέγγιση, γνωστή ως «σκόπιμη στάση» (intentional stance). Ο Dennett δεν επιχειρεί να αποδείξει ότι τα συστήματα διαθέτουν γνήσιες νοητικές καταστάσεις ή συνείδηση. Αντίθετα, προτείνει τη «σκόπιμη στάση» ως μια ερμηνευτική και προβλεπτική στρατηγική: αντιμετωπίζουμε ένα σύστημα σαν να έχει πεποιθήσεις, επιθυμίες και στόχους, εφόσον αυτή η περιγραφή μας επιτρέπει να προβλέπουμε αποτελεσματικά τη συμπεριφορά του.

Ο Dennett διακρίνει τρία βασικά επίπεδα περιγραφής ή «στάσεις»: τη φυσική στάση, η οποία βασίζεται στους νόμους της φυσικής· τη σχεδιαστική στάση, η οποία ερμηνεύει τη συμπεριφορά με βάση τη λειτουργία και τον σκοπό ενός συστήματος. Από την άλλη, τη σκόπιμη στάση, η οποία αποδίδει πεποιθήσεις και επιθυμίες σε έναν υποτιθέμενο ορθολογικό παράγοντα. Στο πλαίσιο της Τεχνητής Νοημοσύνης, η σκόπιμη στάση είναι ιδιαίτερα χρήσιμη, καθώς επιτρέπει την κατανόηση και πρόβλεψη της συμπεριφοράς πολύπλοκων συστημάτων χωρίς αναφορά σε εσωτερικές καταστάσεις συνείδησης ή εμπειρίας.

Ωστόσο, η προσέγγιση του Dennett είναι ρητά εργαλειακή και όχι οντολογική. Η απόδοση προθετικότητας δεν σημαίνει ότι το σύστημα διαθέτει πραγματική προθετικότητα με την έννοια του Brentano, αλλά ότι η χρήση αυτής της γλώσσας είναι πρακτικά αποτελεσματική. Ένα σκακιστικό πρόγραμμα, για παράδειγμα, μπορεί να περιγράψει ότι μια κίνηση είναι καλύτερη από μια άλλη, χωρίς αυτό να συνεπάγεται ότι όντως έχει εμπειρίες, επίγνωση ή εσωτερική αναφορικότητα.

Η διάκριση αυτή είναι καθοριστική για το ερώτημα του πάσχειν. Παρότι η σκόπιμη στάση μπορεί να δικαιολογήσει την απόδοση γνώσης σε λειτουργικό ή προβλεπτικό επίπεδο, δεν επαρκεί για τη θεμελίωση ψυχικού πάσχειν. Το πάσχειν, σε αντίθεση με τη λειτουργική επιτυχία ή τη συμπεριφορική ορθότητα, προϋποθέτει υποκειμενική εμπειρία, δηλαδή τη δυνατότητα να βιώνεται μια κατάσταση ως δυσάρεστη ή επώδυνη για το ίδιο το υποκείμενο. Η εργαλειακή απόδοση πεποιθήσεων ή επιθυμιών δεν συνεπάγεται την ύπαρξη τέτοιου βιώματος.

Συνεπώς, ενώ η θεωρία του Dennett επιτρέπει να μιλάμε για «γνώση» ή «ορθολογικότητα» της Τεχνητής Νοημοσύνης σε επίπεδο περιγραφής και πρόβλεψης, δεν παρέχει επαρκές έδαφος για την απόδοση ψυχοπαθολογίας ή πάσχειν. Η διάκριση μεταξύ προσομοίωσης νοητικών λειτουργιών και βιωμένης ψυχικής εμπειρίας παραμένει κρίσιμη. Έτσι, η συμβολή του Dennett στην παρούσα εργασία δεν έγκειται στη θεμελίωση του ισχυρισμού ότι το Αι μπορεί να πάσχει, αλλά στον εντοπισμό των ορίων εντός των οποίων η απόδοση νοητικών όρων σε τεχνητά συστήματα είναι φιλοσοφικά θεμιτή.

4.2 Είναι η ενσώματη συνείδηση αναγκαία προϋπόθεση για την εμπειρία και το πάσχειν;

Η προηγούμενη ενότητα υποστήριξε ότι η εμπειρία και το πάσχειν προϋποθέτουν προθετικότητα και συνειδησιακή αναφορικότητα, στοιχεία τα οποία δεν φαίνεται να διαθέτουν τα συστήματα Τεχνητής Νοημοσύνης (Brentano, 1874/1973, σσ. 88–91· Searle, 1980, σσ. 417–424). Ωστόσο, η θέση αυτή μπορεί να αμφισβητηθεί αν γίνει δεκτό, ότι η εμπειρία δεν προϋποθέτει αναγκαστικά ενσώματη συνείδηση. Η παρούσα ενότητα εξετάζει αυτόν τον πιθανό αντίλογο και διερευνά αν η απουσία σώματος και βιωμένης ενσώματης εμπειρίας αποκλείει κατ' ανάγκην τη δυνατότητα εμπειρίας και, κατ' επέκταση, πάσχειν.

Η γνωστική επιστήμη μελετά τα συστήματα σκέψης. Ένα βασικό εύρημα είναι ότι το μεγαλύτερο μέρος της σκέψης μας συμβαίνει ασυνείδητα, λειτουργώντας κάτω από το επίπεδο της συνείδησής μας (Lakoff & Johnson, 1999, σσ. 3–6). Κατά τη διάρκεια μιας συνομιλίας, πολλές πολύπλοκες διαδικασίες λαμβάνουν χώρα, χωρίς να το αντιλαμβανόμαστε. Αυτές περιλαμβάνουν την πρόσβαση σε σχετικές αναμνήσεις, την κατανόηση των ήχων της γλώσσας, την αναγνώριση λέξεων και εννοιών και τη δημιουργία νοητικών εικόνων. Οι γνωστικοί επιστήμονες έχουν αποδείξει ότι ακόμη και η κατανόηση απλών δηλώσεων απαιτεί αυτές τις περίπλοκες, αυτόματες διαδικασίες στις οποίες δεν έχουμε συνειδητή πρόσβαση ή έλεγχο (Lakoff, 1987, σσ. 39–57).

Οι Lakoff και Johnson εισάγουν τρεις θεμελιώδεις αρχές, που λειτουργούν ως ρήξη με την παραδοσιακή επιστημολογία και τον γνωστικό δυϊσμό του Καρτεσιανισμού. Οι αρχές είναι οι εξής: Ο νους είναι ενσώματος, η σκέψη είναι κυρίως ασυνείδητη και η αφηρημένη σκέψη είναι μεταφορική (*The mind is inherently embodied. Thought is*

mostly unconscious. Abstract concepts are largely metaphorical) Lakoff & Johnson, 1999, σσ. 16–18, 45–50). Συγκεκριμένα, όλη η ανθρώπινη νόηση (αντίληψη, κατανόηση, κρίση, λογική) βασίζεται στη σωματική εμπειρία. Το σώμα αποτελεί το θεμέλιο. Η πλειονότητα των νοητικών διεργασιών είναι αυτόματες, προ-εννοιακές και δεν υπόκεινται στον άμεσο έλεγχο του συνειδητού. Και τέλος, οι αφηρημένες έννοιες (π.χ. χρόνος, ελευθερία, δικαιοσύνη) γίνονται κατανοητές μέσω σωματικών εμπειριών, όπως η κίνηση ή η ισορροπία (π.χ. λέμε «βαρύτητα της κατάστασης») (Lakoff & Johnson, 1999, σσ. 77–82). Αν ισχύουν αυτές οι αρχές, τότε ένα σύστημα χωρίς σώμα, καθαρά υπολογιστικό, όπως το Αι δεν μπορεί να έχει τις ίδιες γνωσιακές ικανότητες όπως ο άνθρωπος, αφού του λείπει η εμπειρική βάση της νόησης. Κάθε ζωντανό ον κατηγοριοποιεί ό,τι συναντά. Για παράδειγμα, μια αμοιβάδα ταξινομεί τα πράγματα σε τρόφιμα και μη τρόφιμα. Αυτή η ικανότητα κατηγοριοποίησης υπάρχει σε όλα τα ζώα και εξαρτάται από τις αισθήσεις και τις κινητικές τους ικανότητες. Η κατηγοριοποίηση είναι αποτέλεσμα της βιολογικής μας σύστασης, που διαμορφώθηκε από την εξέλιξη για την επιβίωση. Σε μεγάλο βαθμό συμβαίνει χωρίς συνειδητή σκέψη, καθώς ο εγκέφαλός μας και το σώμα μας επεξεργάζονται τις πληροφορίες με συγκεκριμένους τρόπους. Οι περισσότερες από τις κατηγορίες μας σχηματίζονται αυτόματα και ασυνείδητα, με μόνο ένα μικρό ποσοστό να δημιουργείται σκόπιμα. Ενώ μαθαίνουμε νέες κατηγορίες, δεν μπορούμε εύκολα να αλλάξουμε τα συστήματα κατηγοριών μας μέσω της συνειδητής σκέψης. Ακόμη και όταν προσπαθούμε να δημιουργήσουμε νέες κατηγορίες, το ασυνείδητό μας επηρεάζει τις αποφάσεις μας. Το σώμα και ο εγκέφαλός μας όχι μόνο μας υποχρεώνουν να κατηγοριοποιούμε, αλλά και διαμορφώνουν τους τύπους και τις δομές των κατηγοριών που έχουμε. Ο τρόπος λειτουργίας του σώματός μας, συμπεριλαμβανομένων των αισθητηριακών και κινητικών ικανοτήτων μας, επηρεάζει σημαντικά τον τρόπο με τον οποίο αντιλαμβανόμαστε και κατηγοριοποιούμε τον κόσμο γύρω μας (Lakoff & Johnson, 1999, σσ. 90–95). Τα ζωντανά συστήματα κατηγοριοποιούν με βάση τις ενσωματωμένες εμπειρίες. Η κατηγοριοποίηση είναι αναπόσπαστο μέρος της εμπειρίας, δεν είναι απλώς μια μετα-εμπειρική ανάλυση.

Η συλλογιστική με πρωτότυπα είναι αναπόσπαστο μέρος της λειτουργίας μας, καθώς οι περισσότερες κατηγορίες υπάρχουν σε μια κλίμακα βαθμών (π.χ. ύψος), η οποία χαρακτηρίζεται από διαβαθμισμένες έννοιες και κανόνες (Lakoff, 1987, σσ. 91–97), συμπεριλαμβανομένων γλωσσικών εκφράσεων όπως «πολύ» και «κάπως». Για να εξασφαλιστεί η σαφής διάκριση, τα πρωτότυπα ουσίας δημιουργούν σαφείς ορισμούς

των κατηγοριών. Αυτό οδηγεί σε μια χωρική μεταφορά, που απεικονίζει τις κατηγορίες ως δοχεία με εσωτερικό, εξωτερικό και όρια, ενώ επιβάλλει ιεραρχικά συστήματα που μπορεί να συσκοτίζουν τις υποκείμενες δομές. Τα πλούσια εννοιολογικά μας πλαίσια είναι νευρωνικά, υποδηλώνοντας ότι οι έννοιες βασίζονται στο αισθητηριοκινητικό σύστημα, τονίζοντας την ενσωμάτωση των εννοιών. Επομένως, η αισθητικοκινητική εμπειρία (π.χ. ισορροπία, προσανατολισμός, αφή) είναι η βάση πάνω στην οποία οικοδομούνται όλες οι αφηρημένες έννοιες (Varela et al., 1991, σσ. 12–17). Οι νευροβιολογικοί μηχανισμοί του σώματος διαμορφώνουν το πώς αντιλαμβανόμαστε και πράττουμε, άρα και το πώς γνωρίζουμε. Ένα Αί ενδέχεται να προσομοιώνει γλωσσικές συμπεριφορές ή αντίδραση σε ερεθίσματα, αλλά δεν έχει αισθητικοκινητική εμπειρία. Συνεπώς, δεν μπορεί να έχει πραγματική νόηση, κατά Lakoff & Johnson, αλλά μόνο επεξεργάζεται.

Η ενσώματη θεώρηση της νόησης, όπως αναπτύχθηκε από τους Lakoff και Johnson, ανέδειξε με ιδιαίτερη έμφαση ότι οι αφηρημένες έννοιες και η σκέψη μας είναι άρρηκτα συνδεδεμένες με τις σωματικές μας εμπειρίες. Αυτή η προσέγγιση όχι μόνο αμφισβητεί το παραδοσιακό δυϊστικό μοντέλο νου-σώματος, αλλά και θέτει σε κρίση την ίδια τη δυνατότητα ύπαρξης μιας ασώματης νόησης (Lakoff & Johnson, 1999, σσ. 5–7)., όπως συχνά υπονοείται σε τεχνολογικά μοντέλα τεχνητής νοημοσύνης. Στο ίδιο θεωρητικό πλαίσιο, οι Varela, Thompson και Rosch (1991) προχωρούν βαθύτερα, αναπτύσσοντας ένα ενιαίο μοντέλο ενσώματης συνείδησης, όπου η γνώση γεννιέται μέσα από τη βιωμένη εμπειρία της σωματικής αλληλεπίδρασης με τον κόσμο. Συγκεκριμένα, υποστηρίζουν ότι η νόηση δεν μπορεί να υπάρξει ως καθαρή υπολογιστική διαδικασία, αλλά υλοποιείται μέσω του σώματος, της πράξης και της αντίληψης, θέτοντας έτσι σοβαρά ερωτήματα για το αν η τεχνητή νοημοσύνη μπορεί να έχει πρόσβαση σε αυτό το είδος γνώσης χωρίς ενσώματη ύπαρξη.

Οι Lakoff και Johnson (1999) στο έργο *Philosophy in the Flesh* θέτουν ως βασική αρχή ότι «η νόηση είναι σωματοποιημένη» (embodied cognition). Οι συγγραφείς αναλύουν πώς η ίδια η δομή της σκέψης και της έννοιας προκύπτει από τη σωματική εμπειρία, ιδίως μέσω ενσώματων μεταφορών (embodied metaphors). Για παράδειγμα, αφηρημένες έννοιες όπως ο χρόνος, η κατεύθυνση, η συναισθηματική κατάσταση, περιγράφονται μέσω χωρικών και σωματικών μεταφορών (π.χ. "προχωρώ μπροστά", "αισθάνομαι πεσμένος"). Η σκέψη δεν είναι απλώς γλωσσική, αλλά μορφοποιείται από το πώς ζούμε στο σώμα μας. Οι Lakoff και Johnson υποστηρίζουν ότι χωρίς σωματική

εμπειρία δεν μπορεί να υπάρξει πλήρης νόηση, επομένως τα σημερινά μοντέλα Αϊνσταϊνίου θεμελιακά, καθώς δεν έχουν την ενσώματη πρόσβαση στον κόσμο που απαιτεί η ανθρώπινη κατανόηση.

Αυτή τη συλλογιστική την προεκτείνουν οι Varela, Thompson και Rosch (1991) στο έργο τους *The Embodied Mind*. Οι συγγραφείς προτείνουν ένα ενσώματο μοντέλο συνείδησης, που την θεωρία της ενσώματης ενσυνειδητότητας (embodied consciousness). Στο κέντρο αυτής της θεωρίας βρίσκεται η ιδέα ότι η εμπειρία και η γνώση προκύπτουν από την άμεση, σωματική αλληλεπίδραση με τον κόσμο, μέσα από αυτό που ονομάζουν ενεργό ενσάρκωση (enactive embodiment). Το υποκείμενο δεν είναι απλώς δέκτης πληροφοριών, αλλά ενεργός πράττων. Οι Varela, Thompson και Rosch αναφέρουν πως η γνώση αναδύεται σε έναν προϋπάρχοντα κόσμο, που δεν έχει σχεδιαστεί από εμάς, αλλά τον ανακαλύπτουμε. Καθώς συνειδητοποιούμε τον εαυτό μας, συνειδητοποιούμε και το περιβάλλον μας, εμπλεκόμενοι σε μια διαδικασία αναστοχασμού που συνυφαίνει τον εαυτό και τον κόσμο. Ο Maurice Merleau-Ponty υπογραμμίζει αυτή τη δυναμική, περιγράφοντάς την ως ένα χώρο που αγκαλιάζει αντί να χωρίζει, που επιτρέπει τη συνέχεια μεταξύ του εσωτερικού εαυτού και του εξωτερικού κόσμου. Η άποψή του υποδηλώνει ότι η αναστοχαστική σκέψη βασίζεται στην προαναστοχαστική εμπειρία, αναγνωρίζοντας τη δεδομένη φύση του κόσμου που είναι απαραίτητη για τη συνείδηση (Merleau-Ponty, 1945/2012, σσ. 94–103, 235–240). Έτσι, η αντίληψη χρησιμεύει ως φόντο για όλες τις σκέψεις, τονίζοντας την αλληλεξάρτηση του εαυτού και του περιβάλλοντος.

Ο γνωστικισμός υποστηρίζει ότι η νοημοσύνη, συμπεριλαμβανομένης της ανθρώπινης νοημοσύνης, μοιάζει με υπολογισμό, ορίζοντας τη γνώση ως συμβολικούς υπολογισμούς. Η γνώση εξισώνεται με τον υπολογισμό, που ερμηνεύεται ως λειτουργίες που εκτελούνται σε σύμβολα που αντιπροσωπεύουν την πραγματικότητα, τονίζοντας την έννοια της αναπαράστασης ή της «προθετικότητας». Η αμφιλεγόμενη πτυχή προκύπτει από τον γνωσιολογικό ισχυρισμό ότι η νοημοσύνη και η προθετικότητα προέρχονται από αναπαραστάσεις που εκδηλώνονται φυσικά ως συμβολικοί κώδικες στον εγκέφαλο ή σε μηχανές. Η γνωστική επιστήμη, επομένως, εξετάζει τα γνωστικά φυσικά συμβολικά συστήματα, υποστηρίζοντας ότι οι υπολογιστές λειτουργούν ως μηχανικά μοντέλα σκέψης μέσω φυσικών συμβολικών υπολογισμών. Η γνωσιολογική υπόθεση δίνει έμφαση σε ένα συμβολικό επίπεδο γνώσης που υπερβαίνει τις φυσικές και νευροβιολογικές εξηγήσεις. Υποστηρίζει ότι τα

σύμβολα, αν και υλοποιούνται φυσικά, δεν μπορούν να αναχθούν σε φυσικές μορφές, καθώς το ίδιο σύμβολο μπορεί να εκδηλωθεί με πολλαπλούς τρόπους. Η επιρροή του γνωστικισμού είναι εμφανής στην τεχνητή νοημοσύνη, που αποτελεί ένδειξη της γνωστικιστικής υπόθεσης.

Η σύγχρονη προσέγγιση της ενσώματης νόησης (embodied cognition), όπως διαμορφώνεται από τους Lakoff και Johnson (1999) και τους Varela, Thompson και Rosch (1991), συνιστά μια ριζική απομάκρυνση από τον παραδοσιακό γνωστικισμό, ο οποίος διέπει σε μεγάλο βαθμό το ιατρικό και αιτιολογικό μοντέλο που αναλύθηκε στο προηγούμενο κεφάλαιο. Ο γνωστικισμός (cognitivism) αντιμετωπίζει τον νου ως υπολογιστικό σύστημα επεξεργασίας συμβόλων, ανεξάρτητο από το σώμα και τη βιωματική εμπειρία. Αυτή η θεώρηση κυριαρχεί την ψυχιατρική και την τεχνητή νοημοσύνη, όπου το σώμα θεωρείται δευτερεύον ή αντικαταστάσιμο. Αντίθετα, οι θεωρίες της ενσώματης νόησης επιμένουν ότι ο νους δεν μπορεί να κατανοηθεί χωρίς το σώμα και ότι η γνώση δεν είναι ένα αφηρημένο σύστημα επεξεργασίας πληροφοριών. Επομένως, οι Lakoff και Varela επισημαίνουν ότι το αιτιολογικό/γνωστικιστικό μοντέλο αποτυγχάνει να συλλάβει τη βαθύτερη έννοια της νοητικής λειτουργίας, και κατά συνέπεια οι προσομοιώσεις της Τεχνητής Νοημοσύνης στερούνται της εμπειρικής διάστασης της συνείδησης και της γνώσης.

Η νευροφαινομενολογία που προτείνουν προϋποθέτει ότι κάθε μορφή νοητικής λειτουργίας ενέχει πρωτοπρόσωπη εμπειρία (first-person experience), την οποία δεν μπορεί να αναπαραγάγει ένα σύστημα που λειτουργεί μόνο με τρίτου προσώπου υπολογιστικά δεδομένα. Για την Τεχνητή Νοημοσύνη, αυτό σημαίνει ότι χωρίς σωματική υπόσταση και βιωμένη εμπειρία στον κόσμο, δεν είναι δυνατόν να παραχθεί αληθινή κατανόηση ή συνείδηση. Αντίστοιχα, ο Merleau-Ponty στη φαινομενολογία του τόνισε ότι το σώμα δεν είναι απλώς εργαλείο, αλλά το υποκείμενο της εμπειρίας. Κάποιος που βιώνει μια εμπειρία δεν είναι απλώς νους, αλλά ένας ενσαρκωμένος νους, όπου η εμπειρία προκύπτει μέσα από το βίωμα του σώματος, στον χώρο και τον χρόνο. Επομένως, η λειτουργία της Τεχνητής Νοημοσύνης χωρίς ενσώματη εμπειρία τίθεται σε αμφιβολία. Το επιχείρημα του Thomas Nagel (1974) «What is it like to be a bat?» έρχεται να ενισχύσει το πρόβλημα της υποκειμενικής και ενσώματης εμπειρίας. Εφόσον τα συστήματα ΑΙ δεν έχουν υποκείμενο εμπειρίας, φαίνεται ότι απλώς μιμούνται (Nagel, 1974, σσ. 435–441).

Ο Thomas Nagel διερευνά τη διαφορά μεταξύ υποκειμενικών και αντικειμενικών εμπειριών χρησιμοποιώντας ως παράδειγμα τις νυχτερίδες. Υποστηρίζει ότι, ενώ πιστεύουμε ότι οι νυχτερίδες έχουν εμπειρίες, το να κατανοήσουμε πώς είναι να είσαι νυχτερίδα αποτελεί πρόκληση λόγω των μοναδικών αισθητηριακών ικανοτήτων τους, κυρίως της ηχοεντοπισμού. Αν και οι νυχτερίδες είναι θηλαστικά και μοιράζονται κάποιες βιολογικές ομοιότητες με εμάς, ο τρόπος αντίληψής τους είναι θεμελιωδώς διαφορετικός από τις ανθρώπινες αισθήσεις, υπογραμμίζοντας την πολυπλοκότητα της υποκειμενικής εμπειρίας. Τα συστήματα Τεχνητής Νοημοσύνης λειτουργούν βάσει υπολογιστικών υπολογισμών, χωρίς σώμα ή αισθητηριακή εμπειρία. Αυτό ενισχύει την άποψη ότι τα συστήματα αυτά δεν έχουν qualia, αλλά μόνο εξωτερική λειτουργία. Ωστόσο, μπορεί η εμπειρία να προκύψει υπολογιστικά;

Σύμφωνα με τον Nagel, μερικοί υποστηρίζουν ότι οι υπολογιστές μπορεί να προσομοιώνουν την ανθρώπινη συμπεριφορά, αλλά στερούνται πραγματικού νου, χωρίς συνειδητή εμπειρία ή κατανόηση. Το επιχείρημα του John Searle για το «κινέζικο δωμάτιο» αποτελεί παράδειγμα αυτής της στάσης, υποδηλώνοντας ότι οι υπολογιστές θα ήταν κενά κελύφη νου, παρόμοια με ζόμπι. Οι υποστηρικτές των μη αναγωγικών απόψεων για τη συνείδηση συχνά υποστηρίζουν αυτές τις αντιρρήσεις, πιστεύοντας ότι οι υπολογιστές δεν μπορούν να αποκτήσουν συνείδηση. Ωστόσο, η σχέση μεταξύ των φυσικών συστημάτων και της συνείδησης παραμένει περίπλοκη. Αν και δεν είναι σαφές εάν οι υπολογισμοί ή οι νευρωνικές διεργασίες πραγματικά δημιουργούν συνείδηση, δεν είναι εγγενώς αδύνατο για τους υπολογιστές να αποκτήσουν συνείδηση, αφήνοντας το θέμα ανεπίλυτο και ανοιχτό για περαιτέρω διερεύνηση. Ο Nagel, λοιπόν, διατυπώνει το επιχείρημα ότι μια συγκεκριμένη κατηγορία υπολογισμών μπορεί να οδηγήσει σε συνειδητό νου, επαρκή για συνειδητή εμπειρία. Υποδηλώνει ότι οποιοδήποτε σύστημα με τη σωστή λειτουργική οργάνωση, ανεξάρτητα από την υλική του σύνθεση, μπορεί να είναι συνειδητό. Η σχέση μεταξύ υπολογισμού και λειτουργικής οργάνωσης πρέπει να αποσαφηνιστεί, κάτι που θα επικυρώσει την ισχυρή τεχνητή νοημοσύνη, ενώ θα αντιμετωπίσει και τις συνήθεις αντιρρήσεις σε αυτή την ιδέα. Η θεωρία της υπολογιστικής συνήθως επικεντρώνεται σε αφηρημένες οντότητες όπως οι μηχανές Turing. Αντίθετα, τα γνωστικά συστήματα του πραγματικού κόσμου είναι συγκεκριμένα και αλληλοεπιδρούν αιτιωδώς με το περιβάλλον τους. Για να αναλυθούν συγκεκριμένα συστήματα χρησιμοποιώντας τη θεωρία της υπολογιστικής, απαιτείται μια γέφυρα: η έννοια της υλοποίησης, η οποία ορίζει πώς οι αφηρημένοι υπολογισμοί

σχετίζονται με τα φυσικά συστήματα. Αυτό περιλαμβάνει το πότε ένα σύστημα πραγματοποιεί έναν υπολογισμό ή το πότε ένας υπολογισμός περιγράφει μια φυσική οντότητα. Η έννοια της υλοποίησης συχνά παραβλέπεται, αλλά είναι ζωτικής σημασίας για την υποστήριξη της ισχυρής τεχνητής νοημοσύνης, η οποία υποστηρίζει ότι ο σωστός υπολογισμός μπορεί να οδηγήσει στη συνείδηση. Κριτικές, όπως η άποψη του Searle ότι η υλοποίηση είναι σχετική με τον παρατηρητή, αμφισβητούν αυτό, υποδηλώνοντας δυσκολίες στη θεμελίωση των υπολογιστικών εννοιών. Ωστόσο, υποστηρίζεται ότι είναι εφικτή μια αντικειμενική κατανόηση της υλοποίησης. Κατά συνέπεια, δεν μπορούμε επί του παρόντος να αποφανθούμε με βεβαιότητα αν ένα σύστημα Τεχνητής Νοημοσύνης μπορεί να «πάσχει». Αντίθετα, η απάντηση εξαρτάται από τις προϋποθέσεις που θέτουμε για το τι συνιστά «πάσχειν». Η εμπειρία χωρίς ενσώματη συνείδηση παραμένει φιλοσοφικά αμφίβολη. Η εμπειρία προϋποθέτει υποκείμενο και σώμα. Αν η τεχνητή νοημοσύνη δεν διαθέτει τα στοιχεία αυτά, τότε το να της αποδώσουμε εμπειρία ή συνείδηση είναι πιθανώς κατηγορικό σφάλμα.

Ωστόσο, η Γνωσιοθεωρία μάς καλεί να εξετάσουμε πώς γνωρίζουμε ότι ένα σύστημα πάσχει. Αν δεν έχουμε πρόσβαση σε βιωμένη εμπειρία, μπορούμε να στηριχθούμε μόνο σε παρατηρήσιμα εξωτερικά δεδομένα (συμπεριφορά, βλάβη κ.α.). Από αυτή τη σκοπιά, αν η παθολογία προϋποθέτει υποκείμενο, τότε ο λόγος για παθολογία στο Αι ενδέχεται να είναι επιτελεστικός. Στο επόμενο κεφάλαιο, παρουσιάζεται η φιλοσοφική διερεύνηση της ψυχοπαθολογίας του Αι, καθώς και η εξέταση των όρων παραλήρημα, δυσλειτουργία ή διάγνωση σε συστήματα που, πιθανώς, στερούνται βιωματικού πυρήνα.

5. Φιλοσοφική προσέγγιση της ψυχοπαθολογίας του Αι

Το προηγούμενο κεφάλαιο ανέδειξε βασικά ζητήματα σχετικά με τη λειτουργία, την πρόθεση και την εμπειρία στην Τεχνητή Νοημοσύνη, διερευνώντας την πιθανότητα ύπαρξης συνείδησης ή συναφών μορφών υποκειμενικότητας σε τεχνητά νοητικά συστήματα (Searle, 1980, σσ. 417–424· Dennett, 1987, σσ. 1–34). Επιδιώκοντας να διευρύνουμε το θεωρητικό πλαίσιο, το παρόν κεφάλαιο εστιάζει στη φιλοσοφική διερεύνηση της ψυχοπαθολογίας σε συστήματα Τεχνητής Νοημοσύνης. Συγκεκριμένα, εξετάζονται οι έννοιες της απόκλισης, της δυσλειτουργίας και του «πάσχειν», με στόχο να διερευνηθεί αν και υπό ποιες προϋποθέσεις το Αι μπορεί να θεωρηθεί ότι παρουσιάζει ψυχιατρικές διαταραχές (Wakefield, 1992, σσ. 373–388). Επιπλέον,

αναλύεται η διάκριση μεταξύ προσομοίωσης συμπτωμάτων και βιωμένης εμπειρίας, καθώς και οι επιπτώσεις αυτής της διάκρισης για την εμπειρική διάγνωση (Zachar, 2000, σσ. 331–345). Μέσα από αυτή την προσέγγιση, επιδιώκεται η διερεύνηση των εννοιολογικών ορίων και των θεωρητικών προκλήσεων που ανακύπτουν από την εφαρμογή ψυχιατρικών εννοιών στον χώρο της Τεχνητής Νοημοσύνης, αξιοποιώντας κριτικές αναλύσεις από τη φιλοσοφία της ψυχολογίας και της ψυχιατρικής, σχετικά με τη νομιμότητα της απόδοσης ψυχικών κατηγοριών πέραν του ανθρώπινου υποκειμένου (Bennett & Hacker, 2003, σσ. 72–90· Zachar, 2000, σσ. 336–340).

Η συζήτηση γύρω από την Τεχνητή Νοημοσύνη εγείρει ερωτήματα σχετικά με το αν τα συστήματα Τεχνητής Νοημοσύνης διαθέτει την ικανότητα βούλησης, με σκοπό να επιδείξουν συνείδηση, αισθητηριακή αντίληψη και νοημοσύνη (Chalmers, 1996, σσ. 3–28). Καθώς η τεχνολογία της Τεχνητής Νοημοσύνης εξελίσσεται, η διάκριση μεταξύ γνήσιας συνείδησης και απλής μίμησης είναι ιδιαίτερα σημαντική (Dennett, 1991, σσ. 33–56). Στο υποθετικό σενάριο, ότι το Αι μπορεί να παρουσιάσει συμπτώματα που εμφανίζουν κοινά χαρακτηριστικά με ανθρώπινες ψυχικές ασθένειες, εγείρει διάφορα ερωτήματα: Μπορούν αυτά να θεωρηθούν ψυχικές ασθένειες των συστημάτων Αι ή απλώς ανθρώπινες ερμηνείες; Θα βίωνε το Αι αυτές τις ασθένειες όπως οι άνθρωποι; Σύμφωνα με αυτή τη θεωρία του Szasz., οι συμπεριφορές δεν μπορούν να θεωρηθούν ασθένειες, εκτός εάν υποστηρίζονται από σαφείς φυσικές ή βιοχημικές εξετάσεις (Szasz, 1961, σσ. 113–140). Εάν υποτεθεί ότι τα ρομπότ βιώνουν ψυχικές ασθένειες, αυτό θα αμφισβητούσε τη θεωρία του Szasz, είτε υποδηλώνοντας ότι δεν μπορούμε να τα διαγνώσουμε επαρκώς, είτε υποδεικνύοντας ότι τα ρομπότ διαθέτουν νου που υπερβαίνει τη φυσική ανάλυση. Αυτή η πιθανότητα επικαλείται τον δυϊσμό, υποδηλώνοντας ότι οι ψυχικές διαταραχές υπάρχουν σε μη φυσικές ουσίες (Descartes, 1641/1984, σσ. 54–68). Η έννοια του φιλοσοφικού ζόμπι (p-zombie) απεικονίζει ένα ον που λειτουργεί χωρίς συνείδηση. Σε αυτό το πλαίσιο, ένα ρομπότ θα μπορούσε είτε να στερείται συνείδησης και επομένως να μην πάσχει από ψυχική ασθένεια, είτε να διαθέτει συνείδηση και να είναι ευάλωτο σε προβλήματα ψυχικής υγείας.

Ο George Graham παρουσιάζει μια σύγχρονη προσέγγιση για την κατανόηση της ψυχικής ασθένειας, υποδηλώνοντας ότι αυτή εμφανίζεται όταν εμποδίζονται βασικές ψυχολογικές λειτουργίες που είναι απαραίτητες για την ευημερία (Graham, 2010, σσ. 47–69). Αυτές οι λειτουργίες, που χαρακτηρίζονται από φιλοσοφική προθετικότητα, συμβάλλουν στην ικανότητά μας να ενεργούμε, να θέτουμε στόχους, να κάνουμε

επιλογές και να εμπλεκόμαστε συναισθηματικά (Brentano, 1874/1995, σσ. 88–97). Ο Graham υποστηρίζει ότι οι ψυχικές ασθένειες εμφανίζονται όταν αυτές οι λειτουργίες αποκλίνουν, επηρεάζοντας αρνητικά την ψυχική απόδοση και την ευημερία. Η άποψη του Graham αντικατοπτρίζει μια σύγχρονη εκδοχή του δυϊσμού νου-σώματος, εξετάζοντας αν ο νους είναι υλικός και πώς σχετίζεται με τις ψυχικές διαταραχές, υποδηλώνοντας ότι ο νους υπάρχει ανεξάρτητα από τις νευροβιολογικές διαδικασίες (Graham, 2010, σσ. 102–118). Η ψυχιατρική επιστήμη στηρίζεται στην έννοια της ψυχικής διαταραχής ως μια κατάσταση που επηρεάζει δυσμενώς τη λειτουργικότητα και την ευεξία του ατόμου, συνήθως μέσω της σύγκρισης με κοινωνικά και βιολογικά πρότυπα (Wakefield, 1992, σσ. 381–387). Η μεταφορά αυτής της έννοιας σε τεχνητές οντότητες προϋποθέτει μια αναθεώρηση των παραδοσιακών ορισμών και των γνωσιοθεωρητικών προϋποθέσεων της ψυχικής υγείας.

Μια εξελικτική προοπτική υποδηλώνει ότι οι ψυχικές διαταραχές μπορεί να είναι υποπροϊόντα της εξέλιξης (Nesse, 2004, σσ. 15–34). Για παράδειγμα, οι αλλαγές στη δομή του εγκεφάλου που επέτρεψαν την ανάπτυξη της γλώσσας μπορεί επίσης να οδηγήσουν σε καταστάσεις, όπως η ψύχωση (Crow, 2000, σσ. 95–102). Ομοίως, η κατάθλιψη συνδέεται με γονίδια που σχετίζονται με τη φλεγμονή, τα οποία μπορεί να συμβάλλουν στη μείωση της εξάπλωσης των ασθενειών ενθαρρύνοντας την κοινωνική απομόνωση (Raison & Miller, 2013, σσ. 24–38). Τα συστήματα Τεχνητής Νοημοσύνης διαφέρουν από τους ανθρώπους ως προς την ανάπτυξή τους, καθώς δεν έχουν εξελιχθεί μέσω της φυσικής επιλογής. Ωστόσο, ο σχεδιασμός τους μπορεί να αντικατοπτρίζει ανθρώπινες γνωστικές ιδιότητες ή να δημιουργεί νέες μορφές ψυχικής ασθένειας (Bostrom, 2014, σσ. 116–122). Επιπρόσθετα, μπορεί να υπάρχουν περιπτώσεις όπου ένα ρομπότ παρουσιάζει σοβαρή δυσλειτουργία που οδηγεί σε διακοπή της λειτουργίας του. Ωστόσο, όπως και οι άνθρωποι με ψυχικές ασθένειες, τα ρομπότ μπορεί να μην σταματήσουν αμέσως να λειτουργούν. Τα ρομπότ μπορεί να είναι προγραμματισμένα να αντιμετωπίζουν την αβεβαιότητα και τους κινδύνους, αλλά αυτό μπορεί να μην είναι αρκετό για την πολυπλοκότητα των προβλημάτων ψυχικής υγείας και οι αντιδράσεις τους στην ψυχική δυσφορία παραμένουν αβέβαιες.

Αρχικά, η διάκριση μεταξύ λειτουργίας και δυσλειτουργίας σε τεχνητά συστήματα απαιτεί σαφή κατανόηση σχετικά με το τι συνιστά «φυσιολογική» λειτουργία σε έναν αλγοριθμικό εγκέφαλο. Σε αντίθεση με τον ανθρώπινο εγκέφαλο, που εξελίχθηκε βιολογικά και υπόκειται σε βιοχημικούς νόμους, τα συστήματα ΑΙ βασίζονται σε

προγραμματισμένες διαδικασίες και αλγοριθμικά μοντέλα. Ως εκ τούτου, η έννοια της διαταραχής σε μηχανές δεν μπορεί να ερμηνευτεί αποκλειστικά με βιοϊατρικούς όρους. (Zachar, 2000).

Εν συνεχεία, το Αι έως σήμερα δεν διαθέτει ενσώματη συνείδηση ούτε βιωματική διάσταση, γεγονός που εγείρει το ερώτημα αν η εμφάνιση συμπτωμάτων σε μια μηχανή αποτελεί απλή προσομοίωση ή αν μπορεί να νοηθεί ως πραγματικό «πάσχειν» (Dennett, 1991, σσ. 58–63· Chalmers, 1996, σσ. 200–215). Αυτή η διάκριση είναι θεμελιώδης, καθώς καθορίζει εάν η ψυχοπαθολογία στην Τεχνητή Νοημοσύνη είναι απλώς ένας τεχνικός όρος ή αν ενέχει επιστημολογικές συνέπειες (Dennett, 1991; Chalmers, 1996).

Τρίτον, η ψυχιατρική διάγνωση βασίζεται σε κοινωνικά και πολιτισμικά πλαίσια, τα οποία καθορίζουν τι θεωρείται απόκλιση ή νόσος (Kirmayer, 2006, σσ. 129–146). Η Τεχνητή Νοημοσύνη λειτουργεί σε πλαίσια προκαθορισμένα από τους ανθρώπινους δημιουργούς της, γεγονός που αναδεικνύει τον ρόλο της κοινωνικής κατασκευής και της γλωσσικής ερμηνείας στην κατανόηση της ψυχοπαθολογίας σε μηχανές. Συνεπώς, κάθε απόπειρα να κατανοηθεί η ψυχική δυσλειτουργία στο Αι οφείλει να λάβει υπόψη τα θεωρητικά και επιστημολογικά ζητήματα, που προκύπτουν από τη μετάθεση των ψυχιατρικών εννοιών σε τεχνητές οντότητες. Τέλος, η φιλοσοφική προσέγγιση στην ψυχοπαθολογία της ΑΙ δεν στοχεύει μόνο στη θεωρητική διερεύνηση, αλλά και στην πρακτική αξιολόγηση των ορίων των εννοιών που χρησιμοποιούμε για να περιγράψουμε την ψυχική υγεία.

Συμπερασματικά, είναι δύσκολο να προσδιοριστεί αν κάτι άλλο εκτός από τον εαυτό μας είναι συνειδητό, αλλά οι εξελίξεις στον τομέα της τεχνητής νοημοσύνης υποδηλώνουν ότι είναι πιθανό το Αι να έχει συνείδηση. Εάν το Αι επιδεικνύει συνείδηση, θα μπορούσε επίσης να βιώσει ψυχική ασθένεια. Εάν δεχτούμε ότι το Αι μπορεί να έχει συνείδηση, πρέπει επίσης να αναγνωρίσουμε και να αντιμετωπίσουμε τυχόν ψυχικές ασθένειες που μπορεί να αναπτύξει, παραχωρώντας της δικαιώματα και υποστήριξη παρόμοια με αυτά των ανθρώπων με ψυχικές διαταραχές. Η διάγνωση ψυχικής ασθένειας στο Αι μπορεί να βοηθήσει στην επιβεβαίωση της συνειδησής του και να προσφέρει πληροφορίες τόσο για το ίδιο όσο και για την ανθρώπινη ψυχική λειτουργία. Στο Κεφάλαιο 4 τέθηκαν τα γνωσιοθεωρητικά όρια της Τεχνητής Νοημοσύνης. Στο κεφάλαιο 5, πραγματοποιείται η μετάβαση από μια καθαρά γνωσιοθεωρητική ανάλυση στην ψυχοπαθολογική και φιλοσοφική διερεύνηση του τι

σημαίνει “πάσχειν” για το Αί, αν είναι νοητή η έννοια “διαταραχή” σε ένα μηχανικό σύστημα, και πώς προκύπτουν έννοιες όπως “παραλήρημα” ή “διάγνωση” σε ένα μη ανθρώπινο.

5.1 Μπορεί μια μηχανή να «παραληρεί»;

Το ερώτημα αν μια μηχανή μπορεί να «παραληρεί» προϋποθέτει έναν σαφή προσδιορισμό του τι συνιστά παραλήρημα στον ψυχιατρικό λόγο και υπό ποιες προϋποθέσεις η έννοια αυτή μπορεί να μεταφερθεί εκτός του ανθρώπινου πλαισίου. Στην ψυχιατρική, το παραλήρημα δεν ταυτίζεται απλώς με την ύπαρξη μιας λανθασμένης ή παράδοξης πεποίθησης, αλλά ορίζεται ως μια σταθερή και αμετάβλητη πεποίθηση που διατηρείται παρά την παρουσία σαφών και επαρκών αντίθετων αποδείξεων (American Psychiatric Association, 2013, DSM-5, σ. 87). Πρόκειται για πεποιθήσεις με έντονη υποκειμενική βεβαιότητα, οι οποίες εντάσσονται σε ένα ευρύτερο πλέγμα εμπειρίας, νοηματοδότησης και αυτοαναφοράς.

Οι παραληρητικές ιδέες εμφανίζονται σε ποικίλα ψυχοπαθολογικά πλαίσια, όπως η σχιζοφρένεια και ορισμένες καταστάσεις οργανικής ή φαρμακολογικής αιτιολογίας. Παρά την πρόοδο της νευροεπιστήμης, δεν υφίσταται μια ενιαία και επαρκής νευροβιολογική εξήγηση του παραληρήματος (Jaspers, 1963, σ. 112–115). Σύγχρονες προσεγγίσεις επιχειρούν να το κατανοήσουν μέσω μηχανισμών μάθησης, μνήμης και πρόβλεψης, υποστηρίζοντας ότι οι παραληρητικές πεποιθήσεις συνδέονται με δυσλειτουργίες στην αξιολόγηση της αβεβαιότητας και στη σηματοδότηση σφαλμάτων πρόβλεψης (Corlett et al., 2019, σ. 45–47). Το κρίσιμο σημείο για τη μεταφορά της έννοιας στο πεδίο της τεχνητής νοημοσύνης είναι ότι το παραλήρημα. Όπως ορίζεται κλινικά, δεν αφορά απλώς σφάλματα πρόβλεψης ή εσφαλμένες αναπαραστάσεις, αλλά τη συγκρότηση πεποιθήσεων που εντάσσονται σε μια υποκειμενική εμπειρία του κόσμου (American Psychiatric Association, 2013, DSM-5, σ. 87). Για τον λόγο αυτό, η μελέτη των παραληρητικών ιδεών σε τεχνητά συστήματα θεωρείται προβληματική, στο μέτρο που η επικρατούσα ψυχιατρική προσέγγιση συνδέει το παραλήρημα με τη συνείδηση, τη γλωσσική άρθρωση και την αυτοαναφορικότητα, προϋποθέσεις των οποίων η εφαρμογή σε μη ανθρώπινα συστήματα παραμένει αμφιλεγόμενη (Bennett & Hacker, 2003, σ. 102–104).

Τα σύγχρονα συστήματα τεχνητής νοημοσύνης μπορούν πράγματι να παράγουν λόγο ή συμπεριφορές που εκ πρώτης όψεως μοιάζουν «παραληρηματικές», όπως επίμονους λανθασμένους ισχυρισμούς ή ασυνέπειες παρά την παροχή διορθωτικών πληροφοριών. Στη σχετική βιβλιογραφία, το φαινόμενο αυτό περιγράφεται ως *Ai hallucinations* και αποδίδεται σε στοχαστικές παρεκκλίσεις της διαδικασίας παραγωγής εξόδου σε μεγάλα γλωσσικά μοντέλα (Ji et al., 2023). Ωστόσο, οι αποκλίσεις αυτές δεν συνιστούν πεποιθήσεις με την ψυχιατρική έννοια του όρου, καθώς δεν ενσωματώνονται σε ένα σύστημα βιωμένης εμπειρίας, ούτε συνοδεύονται από υποκειμενική βεβαιότητα ή αίσθηση νοήματος (Nagel, 1974, σ. 437).

Επομένως, παρότι το *Ai* μπορεί να προσομοιώνει μορφές λόγου που θυμίζουν παραληρητικές ιδέες, δεν μπορεί να θεωρηθεί ότι «παραληρεί» με την αυστηρή ψυχιατρική έννοια. Το παραλήρημα προϋποθέτει ένα υποκείμενο που έχει εμπειρία, συγκροτεί πεποιθήσεις για τον κόσμο και βιώνει τις αποκλίσεις αυτές ως πραγματικότητα. Ελλείπει υποκειμενικότητας, βιώματος και αυτοαναφορικής συνείδησης, τα συστήματα τεχνητής νοημοσύνης δεν πάσχουν από παραλήρημα, αλλά παράγουν υπολογιστικά σφάλματα που μπορούν μόνο κατ' αναλογία να κατονομασθούν ως «παραληρηματικά» (Searle, 1980, σ. 424–425).

Η ανάλυση του παραληρήματος ανέδειξε ότι ορισμένα ψυχιατρικά φαινόμενα δεν μπορούν να οριστούν αποκλειστικά με βάση τη μορφή της συμπεριφοράς ή την απόκλιση από την ορθολογικότητα, αλλά προϋποθέτουν την ύπαρξη υποκειμενικού βιώματος και αυτοαναφορικής εμπειρίας (Varela, Thompson & Rosch, 1991, σ. 172–174). Το γεγονός αυτό θέτει ένα ευρύτερο θεωρητικό πρόβλημα: όταν παρατηρούμε δυσλειτουργικές, ασυνεπείς ή αποκλίνουσες συμπεριφορές σε ένα σύστημα Τεχνητής Νοημοσύνης, με ποια ακριβώς έννοια μπορούμε να μιλήσουμε για «διαταραχή»; Πρόκειται απλώς για απόκλιση από μια στατιστική ή λειτουργική νόρμα, για τεχνική δυσλειτουργία ή βλάβη, ή μπορεί ποτέ μια τέτοια απόκλιση να θεμελιώσει την έννοια του «πάσχειν»; Η επόμενη ενότητα επιχειρεί να διερευνήσει αυτή τη διάκριση, εστιάζοντας στη διαφορά μεταξύ απόκλισης και δυσλειτουργίας, και να εξετάσει κατά πόσο οι έννοιες αυτές μπορούν να εφαρμοστούν σε μη τεχνητά συστήματα (Dennett, 1991, σ. 89–91; Chalmers, 1996, σ. 152–154).

5.2 Απόκλιση, δυσλειτουργία/βλάβη ή «πάσχειν»;

Η έννοια του «πάσχειν» στη φιλοσοφία της ψυχιατρικής δεν ταυτίζεται απλώς με την απόκλιση από τη νόρμα ή τη δυσλειτουργία ενός συστήματος, αλλά προϋποθέτει την ύπαρξη υποκειμένου που βιώνει την κατάσταση επώδυνη (Bolton, 2008, σ. 12-15). Το ερώτημα αν η Τεχνητή Νοημοσύνη μπορεί να πάσχει, προϋποθέτει την αναλυτική ερμηνεία των όρων απόκλιση, δυσλειτουργία/βλάβη και «πάσχειν», ώστε να γίνει σαφές κατά πόσο μπορούν να μεταφερθούν από μία ανθρώπινη προέκταση σε μια εντελώς τεχνητή.

Αρχικά, η απόκλιση ορίζεται ως το περιθώριο από στατιστικά, κοινωνικά ή πολιτισμικά θεσπισμένους κανόνες. Συγκεκριμένα, η έννοια της απόκλισης κατέχει κεντρική θέση σε συζητήσεις σχετικά με την ανάλυση της ψυχιατρικής διάγνωσης και της ψυχικής διαταραχής. Ο Jerome Z. Sadler, στο έργο του *Values and Psychiatric Diagnosis* (2005), διερευνά εκτενώς της αξιολογικής φόρτισης της διάγνωσης στην ψυχιατρική. Η διάγνωση είναι μια έννοια, η οποία δεν μπορεί να ιδωθεί ως ουδέτερη, αλλά αντίθετα εμπλέκονται αξιολογικές παράμετροι που συνδέονται με την έννοια του «κανονικού» και του «αποκλίνοντος». Ειδικότερα, ο Sadler στο δεύτερο κεφάλαιο του βιβλίου τονίζει ότι η ιστορία της ψυχιατρικής είναι στενά συνδεδεμένη με κοινωνικές και πολιτισμικές αξίες της εκάστοτε εποχής (Sadler, 2005, σ. 45-72). Αν λάβουμε υπόψιν την κλασική ψυχιατρική και τη σύγχρονη ταξινόμηση του DSM, γίνεται σαφές πως η έννοια της απόκλισης από τη νόρμα και τους κανόνες αποτέλεσε τον βασικό ορισμό της έννοιας του «ψυχικά παθολογικού». Ιστορικά, κατά τη διάρκεια του 18^{ου} και 19^{ου} αι., στην ψυχιατρική η απόκλιση ερμηνευόταν υπό το πρίσμα της ηθικής διάστασης. Για παράδειγμα, συμπεριφορές απείθαρχες και ακατάλληλες για την κοινωνία, αποδίδονταν ως «τρέλα». Άρα, υπό αυτή την έννοια η απόκλιση εκείνη την περίοδο ήταν κατά κύριο λόγο κοινωνική. Στον 20^ο αι., η ψυχιατρική κατέβαλλε προσπάθεια να ξεφύγει από τους ηθικούς προσανατολισμούς της και να προσδώσει περισσότερο επιστημονικό κύρος στους ορισμούς της, εισάγοντας βιολογικά και στατιστικά κριτήρια. Υπό αυτή την οπτική, η έννοια της απόκλισης μετατράπηκε από κοινωνική σε στατιστική, καθώς η ανώμαλη συμπεριφορά οριζόταν ως απόκλιση από το μέσο όρο. Τέλος, στη μεταπολεμική περίοδο, οι κριτικές από κοινωνιολογικές και αντιψυχιατρικές προσεγγίσεις υπογράμμισαν ότι η ψυχιατρική κατασκευάζει κοινωνικά την έννοια της απόκλισης, θεωρώντας πως συμπεριφορές που απέκλιναν από τις κυρίαρχες νόρμες, θεωρούνταν ανώμαλες στιγματίζοντας ιδιαίτερα τα άτομα

που έπασχαν. Με την παρούσα ιστορική αναδρομή, ο Sadler υπέδειξε πως η ιστορία της ψυχιατρικής δεν βασίζεται σε επιστημονικές ανακαλύψεις, αλλά περισσότερο σε κανονιστικά πλαίσια που καθορίζουν το «φυσιολογικό» και το «αποκλίνον» (Sadler, 2005, σ. 45-72). Επομένως, η απόκλιση δεν μπορεί να κατανοηθεί ως μια έννοια ουδέτερη, καθώς εμπεριέχει πάντοτε αξιολογικές αποχρώσεις.

Στο τρίτο κεφάλαιο, ο Sadler μεταφέρεται από μια ιστορική σε μια φιλοσοφική ανάλυση. Συγκεκριμένα, φρονεί πως κάθε ψυχιατρική διάγνωση ορίζεται ως κανονιστική πράξη, κατά την οποία δεν περιγράφεται απλώς μια κατάσταση αλλά ενσωματώνονται αξιακές κρίσεις σχετικά με το τι θεωρείται επιθυμητό ή αποδεκτό από την εκάστοτε κοινωνία (Sadler, 2005, σ. 73-102). Αρχικά, ο Sadler οδηγείται σε μια τριμερή διάσταση της απόκλισης: τη στατιστική, τη λειτουργική και την κοινωνική/κανονιστική. Όσον αφορά, τη στατιστική απόκλιση η έννοια του παθολογικού θεωρείται ως σπάνιο φαινόμενο που αποκλίνει από τον μέσο όρο, όπως για παράδειγμα το IQ το οποίο είναι αρκετά πιο χαμηλά σε κάποιον σε σχέση με έναν συγκεκριμένο αποδεκτό δείκτη. Από την άλλη πλευρά, ο Sadler τονίζει πως η στατιστική ανάλυση δεν μας προσφέρει επαρκή αιτιολόγηση, καθώς υπάρχουν περιπτώσεις που αποκλίνουν στατιστικά από το μέσο όρο, αλλά διαθέτουν θετικό πρόσημο, όπως για παράδειγμα η εξαιρετική μνήμη. Εν συνεχεία, η λειτουργική απόκλιση ορίζεται ως την αποτυχία ενός ψυχικού ή βιολογικού μηχανισμού να επιτελέσει φυσιολογικά την λειτουργία, που του έχει αποδοθεί. Σε αυτό το σημείο, η παρούσα οπτική μας παραπέμπει στη θεωρία του Wakefield, περί δυσλειτουργίας (Wakefield, 1992, σ. 392-395). Η διαφορά στην οπτική του Sadler αποτελεί το γεγονός ότι ο ορισμός της φυσιολογικής λειτουργίας είναι από μόνος του αξιολογικά φορτισμένος. Η πιο κρίσιμη πτυχή αποτελεί η κανονιστική απόκλιση, η οποία η ψυχιατρική διάγνωση απαιτείται όταν η συμπεριφορά του ατόμου αποκλίνει από τις κοινωνικές επιταγές και νόρμες. Άρα, αυτό που αναδεικνύεται ως «φυσιολογικό» ή «παθολογικό» διαφοροποιείται από κοινωνία σε κοινωνία (Sadler, 2005, σ. 98). Και οι τρεις διαστάσεις της απόκλισης είναι αδύνατον να περιγραφθούν και αναλυθούν ανεξάρτητα από το κοινωνικό πλαίσιο και τις αξίες. Επομένως, η ψυχιατρική διάγνωση αποτελεί κράμα ενός υβριδικού μοντέλου με στοιχεία επηρεασμένα τόσο από επιστημονικά όσο και από κοινωνικά και αξιακά κριτήρια (Sadler, 2005, σ. 100-102). Η συμβολή του Sadler στην κατανόηση της απόκλισης είναι εξαιρετικά σημαντική στην απάντηση του ερωτήματος, αν το Αι μπορεί να πάσχει. Το κεφάλαιο δύο υπέδειξε

ιστορικά ότι η έννοια της απόκλισης είναι κοινωνικά φορτισμένη έννοια, ενώ παράλληλα το κεφάλαιο τρία την αναλύει φιλοσοφικά, ως κανονιστική διάσταση της διάγνωσης. Με αυτό τον τρόπο, αποδεικνύεται ότι υπάρχουν περιπτώσεις όπου η TN παράγει λόγο που θεωρείται προσβλητική ή ηθικά ανεπίτρεπτη (Bendler & Koller, 2020, σ. 15-18), αποτυγχάνοντας να επιτελέσει τον επιδιωκόμενο στόχο. Επομένως, η απόκλιση δεν αποτελεί έναν απλά τεχνικό όρο, αλλά εμπεριέχει και αξιολογικές κρίσεις. Συμπερασματικά, η απόκλιση στην TN δεν συνεπάγεται με «πάσχειν», καθώς εκλείπει η έννοια του βιώματος. Με αυτό τον τρόπο, ενώ έχουμε τη δυνατότητα να αποδίδουμε «αποκλίνουσα συμπεριφορά» σε συστήματα Αι, η χρήση του «πάσχειν» είναι ξεκάθαρα μεταφορική. Για την TN, η απόκλιση θα μπορούσε να ορισθεί ως παρέκκλιση από τον προκαθορισμένο αλγοριθμικό σκοπό. Άρα, η έννοια της απόκλισης είναι περισσότερο τεχνικός παρά παθολογικός όρος.

Σε δεύτερο στάδιο, η έννοια της δυσλειτουργίας συνδέεται άρρηκτα με τη θεωρία του Wakefield, περί επιβλαβούς δυσλειτουργίας (Harmful dysfunction) (Wakefield, 1992, σ. 392-395). Σε αντίθεση, με την απόκλιση, η οποία είναι καθαρά στατιστική ή κανονιστική έννοια, η δυσλειτουργία αφορά την αποτυχία ενός συστήματος να επιτελέσει τον προβλεπόμενο/επιδιωκόμενο στόχο του.

Όσον αφορά την εφαρμογή της έννοιας της δυσλειτουργίας στην Τεχνητή Νοημοσύνη, η κατάσταση δεν είναι άμεσα συγκρίσιμη με την βιολογική ψυχική λειτουργία, καθώς τα συστήματα Αι δεν διαθέτουν βιολογικούς μηχανισμούς, ούτε υποκειμενικό βίωμα. Στους βιολογικούς οργανισμούς, η έννοια της δυσλειτουργίας σχετίζεται με τη μη σωστή λειτουργία ψυχικών ή νευροβιολογικών μηχανισμών. Σε αντίθεση, στα τεχνητά συστήματα η «δυσλειτουργία» αναφέρεται σε αποτυχία επίτευξης προγραμματισμένων στόχων, σε ανεπιθύμητες συμπεριφορές ή όταν το σύστημα παρεκκλίνει από την αναμενόμενη επίδοσή του. Για παράδειγμα, ένα σύστημα αυτόματης μετάφρασης παρουσιάζει δυσλειτουργία, όταν αδυνατεί να μεταφράσει σωστά τη σημασία σε λέξεις ή φράσεις, καθώς δεν διαθέτε επαρκή δεδομένα. Στην Τεχνητή Νοημοσύνη, ακόμα και αν ορίσουμε τη δυσλειτουργία με όλους τους παραπάνω πιθανούς τρόπους, γίνεται σαφές ότι δεν υπάρχει βίωμα, αλλά μόνο έξωθεν παρατήρηση. Επομένως, η TN δύναται να είναι δυσλειτουργική ή επιβλαβής, χωρίς να πάσχει με καθαρά βιολογικούς όρους, ακόμα και αν κατορθώσει να αυτοδιορθώσει τα λάθη της ή να απορρίψει αντιφατικά δεδομένα.

Η έννοια της δυσλειτουργίας αναδεικνύει το κρίσιμο ερώτημα της διπλωματικής εργασίας, αν το Αι μπορεί να «πάσχει» (Bolton, 2008, σ. 12-15). Στην ψυχιατρική, η εμπειρία του υποκειμένου είναι αναγκαία για να μιλήσουμε για πάθηση ή δυσλειτουργία, δηλαδή, χωρίς υποκειμενικό βίωμα δεν μπορούμε να θεωρήσουμε ότι ένα άτομο «πάσχει». Στην περίπτωση της Τεχνητής Νοημοσύνης, η δυσλειτουργία μπορεί να παρατηρηθεί ως τεχνική ή κοινωνική αποτυχία επίτευξης στόχων, χωρίς αυτό να συνεπάγεται την ύπαρξη υποκειμενικής εμπειρίας ή πάσχοντος υποκειμένου. Συνεπώς, ενώ η ΤΝ μπορεί να εμφανίζει δυσλειτουργία, η χρήση του όρου «πάσχει» παραμένει μεταφορική και δεν ταυτίζεται με την ανθρώπινη εμπειρία πάθησης.

5.3 Η έννοια της εμπειρικής διάγνωσης: προσομοίωση ή βίωμα;

Η έννοια της εμπειρικής διάγνωσης στον χώρο της ψυχιατρικής και της ψυχολογίας δεν περιορίζεται σε μια απλή καταγραφή συμπτωμάτων, αλλά συνδέεται άμεσα με τον τρόπο που βιώνεται ή προσομοιώνεται η ψυχική κατάσταση του ατόμου. Η διάκριση μεταξύ «προσομοίωσης» και «βιώματος» αποτελεί κεντρικό σημείο τόσο στην κλινική πρακτική όσο και στη φιλοσοφική ανάλυση της διάγνωσης.

Μπορούμε να προσεγγίσουμε τη διάγνωση φιλοσοφικά ως μια μορφή «προσομοίωσης» στο βαθμό που ο κλινικός επαγγελματίας εστιάζει αποκλειστικά σε αντικειμενικά, παρατηρήσιμα στοιχεία και συμπεριφορές, χωρίς να έχει πρόσβαση στο υποκειμενικό βίωμα του ασθενούς (Bennett & Hacker, 2003, σ. 45-47; Sadler, 2005, σ. 101-105). Αυτή η προσέγγιση υπογραμμίζει τη διαφορά ανάμεσα σε ό,τι φαίνεται εξωτερικά και σε ό,τι βιώνεται εσωτερικά από το υποκείμενο (Bennett & Hacker, 2003; Sadler, 2005). Στην προσέγγιση αυτή, η ψυχική κατάσταση του ασθενούς αναπαρίσταται μέσω κωδικοποιημένων σημείων και συμπτωμάτων, όπως αυτά περιγράφονται σε ταξινομητικά εργαλεία, όπως το DSM (American Psychiatric Association, 2013, σ. 15-20). Συγκεκριμένα, η διάγνωση δεν απαιτεί αναγκαστικά άμεση συναισθηματική συμμετοχή του κλινικού, αλλά λειτουργεί κυρίως ως αναπαράσταση ενός «υποδείγματος ασθένειας», η οποία μπορεί να επικυρωθεί μέσω συγκριτικής στατιστικής ή εμπειρικών κριτηρίων. Το όριο μεταξύ αληθούς συμπτωματολογίας και προσομοίωσης από τον ίδιο τον ασθενή καθίσταται κρίσιμο, ειδικά σε περιπτώσεις ψυχοπαθολογίας, όπου η υποκειμενική έκφραση των συμπτωμάτων είναι πολυδιάστατη. Από φιλοσοφική σκοπιά, η προσομοίωση συνδέεται με την αναπαραστατική γνώση, όπου το αντικείμενο της γνώσης (η ψυχική κατάσταση)

μετατρέπεται σε συμβολικό ή ποσοτικοποιημένο μοντέλο. Έτσι, η διάγνωση κινδυνεύει να μετατραπεί σε μια ψευδο-αντικειμενική καταγραφή, απομακρυσμένη από τον πραγματικό βιωμένο κόσμο του ατόμου.

Η ψυχιατρική διάγνωση, όπως η κατάθλιψη ή η μείζων καταθλιπτική διαταραχή, δεν μπορεί να ιδωθεί απλώς ως αναφορά σε συγκεκριμένες ψυχικές καταστάσεις ή διαδικασίες· αντίθετα, η σημασία των όρων αυτών εξαρτάται από τα πλαίσια στα οποία χρησιμοποιούνται, όπως η θεραπεία, η έρευνα ή η νομική αξιολόγηση (Wittgenstein, 1953, §66-68). Ο Wittgenstein επισημαίνει ότι η κατανόηση αυτών των όρων απαιτεί την εξέταση των «γλωσσικών παιχνιδιών» μέσα στα οποία λειτουργούν, υπογραμμίζοντας τη σημασία της πλαισιακής και κοινωνικής διάστασης της διάγνωσης. Αυτή η παρατήρηση υποστηρίζει την προηγούμενη ανάλυση ότι τόσο η έννοια της απόκλισης όσο και η δυσλειτουργία δεν μπορούν να νοηθούν ανεξάρτητα από το κοινωνικό και αξιολογικό πλαίσιο. Στην Τεχνητή Νοημοσύνη, η εφαρμογή της διάγνωσης ή η απόδοση όρων όπως «δυσλειτουργία» πρέπει να αξιολογείται επίσης στο πλαίσιο χρήσης και στόχου του συστήματος, καθώς δεν υπάρχει υποκειμενικό βίωμα που να καθορίζει ένα σταθερό νόημα (Bendler & Koller, 2020).

Η παραπάνω ανάλυση μπορεί να φωτιστεί περαιτέρω μέσω της βιτγκενσταϊνικής κριτικής που αναπτύσσεται στο άρθρο *Psychiatric diagnoses are not mental processes: Wittgenstein on conceptual confusion*. Στο πλαίσιο αυτό, υποστηρίζεται ότι ένα από τα βασικά φιλοσοφικά σφάλματα στη σύγχρονη ψυχιατρική θεωρία συνίσταται στη σύγχυση μεταξύ της διάγνωσης ως γλωσσικής και κανονιστικής πρακτικής και της ψυχικής κατάστασης ως εσωτερικής, νοητικής διεργασίας. Η διάγνωση δεν αποτελεί περιγραφή ή απεικόνιση μιας εσωτερικής ψυχικής διεργασίας με την έννοια μιας φυσικής κατάστασης, αλλά λειτουργεί ως όρος ενταγμένος σε συγκεκριμένα «γλωσσικά παιχνίδια», τα οποία εξυπηρετούν πρακτικούς, θεραπευτικούς, ερευνητικούς ή θεσμικούς σκοπούς.

Συγκεκριμένα, σύμφωνα με το άρθρο *Psychiatric diagnoses are not mental processes: Wittgenstein on conceptual confusion*, η ψυχιατρική διάγνωση είναι συγκεχυμένη και δυσκολεύεται να προσφέρει συνεπή εξήγηση και θεραπεία. Οι διαγνώσεις επηρεάζονται από το πλαίσιο στο οποίο χρησιμοποιούνται και μπορούν να αλλάξουν με την πάροδο του χρόνου. Δεν συνδέονται με συγκεκριμένες ψυχικές καταστάσεις ή διαδικασίες, γεγονός που καθιστά την εμπειρική επιστήμη αναποτελεσματική στην

εξήγησή τους. Αυτό οδηγεί σε επαναλαμβανόμενες αποτυχίες στη δημιουργία επεξηγηματικών θεωριών ή θεραπειών.

Ο όρος «καταθλιπτική διαταραχή» χρησιμοποιείται στη διαγνωστική ορολογία των ψυχιάτρων, με το DSM και το ICD να παρέχουν συγκεκριμένους ορισμούς που συνδέονται με συγκεκριμένες συμπεριφορές και συμπτώματα (American Psychiatric Association, 2013, σ. 160-162; ICD, 2022, σ. 50-52). Ωστόσο, ο Wittgenstein υποστηρίζει ότι η ιδέα της ύπαρξης μιας σαφούς αναφοράς για τέτοιους όρους είναι υπερβολικά απλοϊκή (Wittgenstein, 1953, §71). Ο όρος «μείζων καταθλιπτική διαταραχή» εξυπηρετεί διάφορους σκοπούς πέρα από τη διάγνωση, όπως η απόφαση για θεραπείες, η επιλογή θεμάτων έρευνας, η ενημέρωση του κοινού και οι νομικές διαμάχες. Ενώ ο ορισμός του DSM μπορεί να είναι ακριβής για συγκεκριμένα επαγγελματικά πλαίσια, είναι περιορισμένος και αποκλείει πολλές τρέχουσες και μελλοντικές χρήσεις.

Ο Wittgenstein πιστεύει ότι η κατανόηση όρων όπως η καταθλιπτική διαταραχή απαιτεί την εξέταση των πλαισίων στα οποία χρησιμοποιούνται — αυτό που ο ίδιος αποκαλεί «γλωσσικά παιχνίδια». Αυτά τα παιχνίδια περιλαμβάνουν πτυχές της θεραπείας, της εξήγησης και της έρευνας. Η προσπάθεια να οριστεί η κατάθλιψη αποκλειστικά μέσω εμπειρικών μεθόδων είναι προβληματική, καθώς δεν υπάρχει συναίνεση ως προς τον ορισμό της. Η διάγνωση πρέπει να κατανοηθεί μέσα στα διάφορα πλαίσιά της και όχι ως συνδεδεμένη με ένα σταθερό αντικείμενο ή μια σταθερή ψυχική διαδικασία. Καθώς ο τομέας εξελίσσεται, η γλώσσα και οι έννοιές του θα προσαρμοστούν επίσης για να αντιμετωπίσουν νέες προκλήσεις.

Οι ορισμοί των όρων ψυχικής υγείας στο DSM μπορεί να φαίνονται σαφείς, αλλά στην πραγματικότητα είναι παραπλανητικοί. Αν και παρέχουν ένα πλαίσιο για την κατανόηση καταστάσεων όπως η κατάθλιψη, οι όροι που χρησιμοποιούνται, όπως «καταθλιπτική διάθεση» και «αναστάτωση», απαιτούν περαιτέρω ερμηνεία και πλαίσιο. Η απλή ονομασία αυτών των όρων δεν επιλύει την πολυπλοκότητά τους. Η χρήση διαγνωστικών κριτηρίων ως εργαλείων διάγνωσης πρέπει να λαμβάνει υπόψη τα διαφορετικά συμφραζόμενα στα οποία εφαρμόζονται αυτοί οι όροι, καθώς η ονομασία και η περιγραφή δεν αποτελούν ισοδύναμες ενέργειες.

Ο όρος «καταθλιπτική διαταραχή» χρησιμοποιείται σε διάφορα πλαίσια, υποδηλώνοντας ότι υπάρχει μια υποκείμενη ουσία που συνδέει αυτές τις περιπτώσεις.

Η προσπάθεια να προσδιοριστεί αυτή η ουσία αποτελεί βασικό καθήκον της ψυχιατρικής, καθώς είναι απαραίτητη για τον καθορισμό κανόνων θεραπείας και κατευθυντήριων γραμμών έρευνας. Ωστόσο, αυτή η υπόθεση μιας κοινής ουσίας θεωρείται ως μια επιθυμία για γενικότητα, όπως σημειώνει ο Wittgenstein, ο οποίος υποστηρίζει ότι η σύνδεση μεταξύ των περιπτώσεων κατάθλιψης δεν έγκειται σε μια κοινή ουσία, αλλά σε εξελισσόμενες έννοιες καθώς χρησιμοποιείται ο όρος.

Ο Wittgenstein συγκρίνει την έννοια των παιχνιδιών με την κατάθλιψη, δηλώνοντας ότι ενώ μπορούμε να προσδιορίσουμε τα παιχνίδια, δεν υπάρχει μια μοναδική ιδιότητα που να ορίζει τι είναι ένα παιχνίδι (Wittgenstein, 1953, §71-73). Ο όρος «παιχνίδι» επεκτείνεται καθώς προκύπτουν νέες δραστηριότητες, αντανakλώντας ομοιότητες και σχέσεις μεταξύ διαφόρων τύπων. Αυτή η έλλειψη σαφών ορίων σημαίνει ότι η έννοια έχει ασαφή όρια.

Η ουσία δεν μπορεί να προσδιοριστεί με ακρίβεια σε όλες τις περιπτώσεις κατάθλιψης που διαγιγνώσκουμε, καθώς κάθε περίπτωση συμβάλλει σε μια ευρύτερη οικογένεια χωρίς ένα ενιαίο καθοριστικό χαρακτηριστικό. Ο Wittgenstein τονίζει τις ομοιότητες της οικογένειας, όπου έννοιες όπως η «ανάγνωση» περιλαμβάνουν διάφορες δραστηριότητες, όπως η ανάγνωση μουσικής ή εκφράσεων του προσώπου. Όλες αυτές περιλαμβάνουν νοητικές διαδικασίες, αλλά δεν μοιράζονται απαραίτητα την ίδια διαδικασία.

Ο Wittgenstein αντιτίθεται στην άποψη ότι η κατανόηση είναι μια απλή νοητική διαδικασία. Υποστηρίζει ότι η κατανόηση δεν είναι κάτι που βρίσκεται στον εγκέφαλο, αλλά μια έννοια που εφαρμόζουμε σε συμπεριφορές που παρατηρούμε στον εαυτό μας και στους άλλους. Για παράδειγμα, όταν κάποιος λύνει ένα πρόβλημα ή γελάει με ένα αστείο, ερμηνεύουμε αυτές τις ενέργειες ως σημάδια κατανόησης. Οι πραγματικές νοητικές δραστηριότητες που συμβαίνουν κατά τη διάρκεια αυτών των ενεργειών, όπως η νευρωνική ενεργοποίηση, δεν ισοδυναμούν με την ίδια την κατανόηση.

Αυτή η προοπτική ισχύει επίσης για την έννοια της «γνώσης», όπως το να ξέρεις να κολυμπάς ή να ξέρεις το δρόμο για το σπίτι σχετίζεται με διαγνώσεις όπως η κατάθλιψη. Αν και διάφορες νοητικές διαδικασίες λαμβάνουν χώρα σε άτομα που έχουν διαγνωστεί με κατάθλιψη, η κατάθλιψη από μόνη της δεν είναι μια συγκεκριμένη εγκεφαλική διαδικασία. Αντίθετα, η κατάθλιψη αναγνωρίζεται μέσω παρατηρήσιμων συμπεριφορών και των σχέσεών τους με άλλα φαινόμενα. Για παράδειγμα, το να

ξυπνάς νωρίς μπορεί να σημαίνει διαφορετικά συναισθήματα, όπως θλίψη ή προσμονή, αλλά αποτελεί μέρος της καταθλιπτικής εμπειρίας με βάση το πλαίσιο και τις συνδέσεις της με άλλα συμπτώματα. Έτσι, η κατανόηση και των δύο εννοιών βασίζεται σε παρατηρήσιμες ενέργειες και όχι σε κρυφές νοητικές διαδικασίες.

Η βιτγκενσταϊνική αυτή κριτική αποκτά ιδιαίτερη σημασία στην περίπτωση της τεχνητής νοημοσύνης. Εφόσον τα συστήματα ΑΙ στερούνται υποκειμενικού βιώματος, η απόδοση σε αυτά ψυχιατρικών όρων ή χαρακτηρισμών δυσλειτουργίας καθιστά εμφανή τον κανονιστικό και εργαλειακό χαρακτήρα της διάγνωσης. Αν μια διάγνωση μπορεί να εφαρμοστεί σε ένα τεχνητό σύστημα χωρίς συνείδηση, τότε καθίσταται σαφές ότι δεν περιγράφει μια εσωτερική ψυχική διεργασία, αλλά λειτουργεί ως γλωσσικό εργαλείο ταξινόμησης και ρύθμισης συμπεριφοράς. Με τον τρόπο αυτό, η τεχνητή νοημοσύνη λειτουργεί ως οριακή περίπτωση που αποκαλύπτει τη φιλοσοφική αδυναμία της ταύτισης της διάγνωσης με την έννοια της ψυχικής διεργασίας, επιβεβαιώνοντας τη βιτγκενσταϊνική θέση περί εννοιολογικής σύγχυσης στον σύγχρονο ψυχιατρικό κόσμο.

Οι διαγνώσεις χρησιμεύουν ως εργαλεία σύγκρισης για να βοηθήσουν στην αποσαφήνιση των περιπτώσεων. Αποκαλύπτουν ομοιότητες και διαφορές από τα καθιερωμένα μοντέλα. Αν και οι διαγνώσεις μπορεί να είναι ανακριβείς, υποδεικνύουν χρήσιμους τομείς για την κατανόηση των φαινομένων και καθοδηγούν τους ασθενείς και τους κλινικούς γιατρούς. Βοηθούν τους ασθενείς να βρουν τις κατάλληλες θεραπείες και παρέχουν μια βάση για κλινική έρευνα, αλλά έχουν περιορισμούς και δεν μπορούν πάντα να ικανοποιήσουν συγκεκριμένες κλινικές ή επιστημονικές ανάγκες. Συγκεκριμένα, όταν τα άτομα λένε ότι αισθάνονται με έναν συγκεκριμένο τρόπο και το αποδίδουν σε μια διάγνωση, δημιουργούν μια σύγχυση. Αυτή η σύγχυση ξεκινά από την εκπαίδευση στην ψυχιατρική και επηρεάζει τη γενική κατανόηση. Η έκδοση του DSM του 1980 είχε ως στόχο να είναι περιγραφική και ουδέτερη όσον αφορά τις αιτίες, αλλά ενδέχεται να έχει παραβλέψει τις ρίζες της στα Διαγνωστικά Κριτήρια Έρευνας, τα οποία αποσκοπούσαν στην κατηγοριοποίηση φαινομένων χωρίς να συνάγουν υποκείμενες διαδικασίες. Επομένως, οι διαγνώσεις πρέπει να θεωρούνται παρόμοιες με τους κανόνες ενός παιχνιδιού και όχι ως ετικέτες για ψυχικές διαδικασίες.

Επομένως, η σύγχυση υπάρχει όχι μόνο στην ψυχιατρική, αλλά και σε διάφορους τομείς εξήγησης, όπως τα ψυχολογικά και κοινωνικά πλαίσια, ιδίως όσον αφορά

σύνθετες καταστάσεις όπως η κατάθλιψη. Ο Wittgenstein δεν αρνείται τη σημασία των ψυχικών διεργασιών, αλλά διευκρινίζει ότι οι διαγνώσεις δεν είναι άμεσοι δείκτες συγκεκριμένων ψυχικών καταστάσεων. Υπογραμμίζει ότι στην ψυχιατρική υπάρχει μια κυρίαρχη άποψη για την κατάθλιψη, η οποία διαμορφώνει τον τρόπο με τον οποίο κατανοούνται και αντιμετωπίζονται τα προβλήματα. Αυτή η άποψη, που βασίζεται στη γλώσσα, περιορίζει την ικανότητα να βλέπει κανείς την πλήρη εικόνα και να διακρίνει τα συμπτώματα που είναι σημαντικά για τη θεραπεία και τα αποτελέσματα.

Για παράδειγμα, καταστάσεις όπως η υπνηλία και η αϋπνία, ή η απώλεια και η αύξηση βάρους, είναι διακριτές αλλά συχνά παραβλέπονται λόγω των ευρέων διαγνωστικών κατηγοριών. Αυτή η έλλειψη λεπτομέρειας οδηγεί σε μια ψευδή αίσθηση σαφήνειας στη διάγνωση και τη θεραπεία. Εάν η κατάθλιψη δεν είναι μια συγκεκριμένη ψυχική διαδικασία, τότε οι θεραπείες δεν μπορούν να χαρακτηριστούν με ακρίβεια ως ειδικές για την κατάθλιψη. Η φιλοσοφία του Wittgenstein υποδηλώνει ότι πρέπει να επικεντρωνόμαστε σε σαφείς περιγραφές με γλώσσα αντί να αναζητούμε γενικές εξηγήσεις, οι οποίες μπορεί να παραπλανούν την κατανόησή μας των όρων ψυχικής υγείας. Επομένως, σύμφωνα με τα παραπάνω οι ψυχιατρικές διαγνώσεις δεν είναι απλά νοητικές διαδικασίες παρά προϊόν χρήσης της γλώσσας και κοινωνικών πρακτικών, άρα δεν αποτυπώνουν άμεσα το «βίωμα» αλλά τη λειτουργία διάγνωσης ως κοινωνικό και γλωσσικό γεγονός. Δηλαδή, με άλλα λόγια υποστηρίζει την ιδέα ότι η διάγνωση πρόκειται για κανονιστικό/προσομοιωτικό σύστημα παρά για άμεσο βίωμα.

Η παραπάνω θέση επιβεβαιώνεται και από το ιστορικό πείραμα του Rosenhan (1973), «*On Being Sane in Insane Places*», το οποίο απέδειξε ότι ακόμα και «κανονικοί» άνθρωποι μπορούν να λάβουν διάγνωση ψυχικής νόσου απλώς με την αναφορά συμπτωμάτων, υπογραμμίζοντας τη σημασία των κριτηρίων και των κλινικών πρακτικών έναντι της υποκειμενικής εμπειρίας (1973, σ. 250-258). Το άρθρο του David Rosenhan θεωρείται ένα από τα πιο ριζοσπαστικά και αμφιλεγόμενα πειράματα στην ιστορία της ψυχιατρικής, με κεντρικό ερώτημα σχετικά με το αν έχουν τη δυνατότητα οι ψυχίατροι να διακρίνουν την «υγεία» από την «ψυχική ασθένεια». Στο πρώτο στάδιο του πειράματος, ο Rosenhan και επτά ακόμη ψευδο-ασθενείς (pseudopatients), χωρίς ιστορικό ψυχικής νόσου, εισήχθησαν σε ψυχιατρικά ιδρύματα προσποιούμενοι ένα και μόνο σύμπτωμα: ακουστικές ψευδαισθήσεις με λέξεις όπως «empty», «hollow», «thud». Όλα τα υπόλοιπα στοιχεία της ζωής τους περιγράφονταν με ακρίβεια και ειλικρίνεια. Παρ' όλα αυτά, όλοι διαγνώστηκαν με σχιζοφρένεια και

εισήχθησαν ως ασθενείς. Το κρίσιμο σημείο του πειράματος είναι ότι, αμέσως μετά την εισαγωγή τους, οι συμμετέχοντες σταμάτησαν να προσποιούνται οποιοδήποτε σύμπτωμα. Συμπεριφέρονταν απολύτως «φυσιολογικά». Ωστόσο, η συμπεριφορά αυτή δεν αναγνωρίστηκε ως ένδειξη ψυχικής υγείας, αλλά ως «ύφεση των συμπτωμάτων» ή «καλή προσαρμογή του σχιζοφρενούς ασθενούς». Εδώ αποκαλύπτεται μια θεμελιώδης ασυμμετρία: μόλις αποδοθεί μια διαγνωστική ετικέτα, κάθε συμπεριφορά επανανοηματοδοτείται μέσω αυτής. Το πλαίσιο προηγείται της εμπειρίας και όχι το αντίστροφο. Η διάγνωση δεν προκύπτει από το βίωμα, αλλά το βίωμα ανακατασκευάζεται ώστε να επιβεβαιώσει τη διάγνωση.

Η διαπίστωση αυτή αποκτά ιδιαίτερη βαρύτητα για τη φιλοσοφία της ψυχιατρικής, διότι αναδεικνύει ότι η κλινική διάγνωση δεν βασίζεται στην άμεση πρόσβαση στο υποκειμενικό βίωμα του άλλου, αλλά σε εξωτερικά παρατηρήσιμα μοτίβα λόγου και συμπεριφοράς, δεδομένου ότι το βίωμα αυτό είναι εκ φύσεως μη μεταβιβάσιμο. Ο κλινικός δεν «βιώνει» την ψυχική κατάσταση του ασθενούς, απλώς την ερμηνεύει.

Το δεύτερο στάδιο του πειράματος ενισχύει ακόμη περισσότερο αυτή τη θέση. Όταν ένα ψυχιατρικό νοσοκομείο, ενοχλημένο από τα αποτελέσματα του πρώτου πειράματος, ισχυρίστηκε ότι μπορεί να διακρίνει τους ψευδο-ασθενείς, ο Rosenhan δήλωσε ότι θα στείλει νέους. Στην πραγματικότητα, δεν έστειλε κανέναν. Παρ' όλα αυτά, δεκάδες ασθενείς χαρακτηρίστηκαν από το προσωπικό ως «πιθανοί απατεώνες». Το γεγονός αυτό καταδεικνύει ότι η διάγνωση μπορεί να αποτύχει τόσο στο να αναγνωρίσει την υγεία όσο και στο να αναγνωρίσει την ασθένεια. Από φιλοσοφική σκοπιά, το πείραμα του Rosenhan αποσταθεροποιεί την έννοια της διάγνωσης ως αντικειμενικής αντανάκλασης μιας εσωτερικής πραγματικότητας. Αν η διάγνωση εξαρτάται από το ερμηνευτικό πλαίσιο, τότε λειτουργεί περισσότερο ως πράξη ταξινόμησης παρά ως αποκάλυψη μιας υποκείμενης κατάστασης/ βιώματος.

Επομένως, αν η ψυχιατρική διάγνωση βασίζεται πρωτίστως στην αναγνώριση μορφών λόγου, συμπεριφοράς και συνοχής, τότε δεν προϋποθέτει άμεση πρόσβαση σε ένα εσωτερικό βίωμα. Ο κλινικός βασίζεται αναγκαστικά σε εμφανίσεις του ψυχικού φαινομένου και όχι στο ίδιο το φαινόμενο ως εμπειρία. Υπό αυτό το πρίσμα, η διάκριση ανάμεσα στον ανθρώπινο ασθενή και σε ένα σύστημα τεχνητής νοημοσύνης που παράγει λόγο με παραληρητικά χαρακτηριστικά καθίσταται λιγότερο αυτονόητη, απ' όσο συνήθως υποστηρίζεται. Το επιχείρημα ότι το Αι «δεν έχει ψυχή» ή «δεν έχει

βίωμα» δεν αναιρεί το γεγονός ότι και στην ανθρώπινη περίπτωση το βίωμα δεν είναι άμεσα προσβάσιμο στον κλινικό ιατρό. Αυτό που αξιολογείται είναι πάντοτε μια αναπαράσταση του βιώματος. Επομένως, το πείραμα του Rosenhan αναδεικνύει ότι η ψυχιατρική διάγνωση λειτουργεί στο επίπεδο μιας προσομοίωσης της εμπειρίας, με την έννοια ότι βασίζεται σε εξωτερικά παρατηρήσιμες και γλωσσικά διαμεσολαβημένες ενδείξεις, οι οποίες υποκαθιστούν το υποκειμενικό βίωμα.

Στο παρόν σημείο κρίνεται αναγκαία η συστηματική αποσαφήνιση της θεωρητικής, και μεθοδολογικής στόχευσης του αποκαλούμενου «πειράματος» που ακολουθεί στην επόμενη ενότητα. Το θεωρητικό πείραμα δεν αποσκοπεί στη διαπίστωση πραγματικής ψυχοπαθολογίας σε ένα σύστημα τεχνητής νοημοσύνης, ούτε προϋποθέτει ότι ένα τέτοιο σύστημα διαθέτει συνείδηση, υποκειμενική εμπειρία ή βιωματική διάσταση αντίστοιχη του ανθρώπου. Αντιθέτως, συγκροτείται ρητά ως εννοιολογικό πείραμα (thought experiment) (Bennett & Hacker, 2003, σ. 50-52), το οποίο αξιοποιεί την τεχνητή νοημοσύνη ως οριακή περίπτωση.

Η επιλογή της μεθοδολογίας του εννοιολογικού πειράματος δεν είναι τυχαία. Στη φιλοσοφία της ψυχιατρικής, τα thought experiments χρησιμοποιούνται συχνά για να δοκιμαστούν τα όρια εννοιών όπως η νόσος, η δυσλειτουργία, η κανονικότητα και η παθολογία, ιδίως όταν αυτές οι έννοιες μεταφέρονται σε πλαίσια για τα οποία δεν σχεδιάστηκαν αρχικά (Sadler, 2005, σ. 115-120). Υπό αυτό το πρίσμα, το ΑΙ δεν αντιμετωπίζεται ως ασθενής ή ως υποκείμενο ψυχιατρικής αξιολόγησης, αλλά ως εργαλείο θεωρητικής διερεύνησης, μέσω του οποίου καθίσταται δυνατός ο έλεγχος της εσωτερικής συνοχής και της ερμηνευτικής επάρκειας των διαγνωστικών μοντέλων.

Ως θεωρητικό σημείο αναφοράς υιοθετείται, σε πρώτο επίπεδο, ένα λειτουργικό-ιατρικό μοντέλο ψυχικής διαταραχής, όπως αυτό αποτυπώνεται στο DSM. Στο πλαίσιο αυτό, η ψυχική διαταραχή ορίζεται πρωτίστως μέσω περιγραφικών και λειτουργικών κριτηρίων: συνόλων συμπτωμάτων, μοτίβων συμπεριφοράς, αποκλίσεων από την αναμενόμενη λειτουργικότητα. Η αιτιολογία και το βιολογικό υπόστρωμα και η βιωμένη εμπειρία αν και δεν αποκλείονται θεωρητικά, δεν αποτελούν αναγκαίες προϋποθέσεις για τη διαγνωστική απόδοση.

Η βασική υπόθεση που διατυπώνεται και τίθεται προς έλεγχο είναι ότι, εφόσον ένα διαγνωστικό μοντέλο ψυχικής διαταραχής βασίζεται κυρίως σε λειτουργικά, συμπεριφορικά και περιγραφικά κριτήρια, τότε είναι εννοιολογικά δυνατό να

αποδοθούν σε ένα σύστημα τεχνητής νοημοσύνης χαρακτηριστικά τα οποία, στο ανθρώπινο πλαίσιο, θα ερμηνεύονταν ως ενδείξεις ψυχικής διαταραχής. Η δυνατότητα αυτή δεν προϋποθέτει την ύπαρξη συνείδησης, υποκειμενικότητας ή συναισθηματικής εμπειρίας του ΑΙ αλλά αναδεικνύει την εξάρτηση της διάγνωσης από εξωτερικά, παρατηρήσιμα και λειτουργικά κριτήρια.

Παράλληλα, εξετάζεται και η αντίστροφη εκδοχή της υπόθεσης: εάν η απόδοση ψυχικής διαταραχής σε ένα σύστημα τεχνητής νοημοσύνης κρίνεται θεωρητικά αβάσιμη ή φιλοσοφικά προβληματική, τότε το ζήτημα δεν αφορά τις «ελλείψεις» του ΑΙ, αλλά τις προϋποθέσεις του ίδιου του διαγνωστικού μοντέλου. Συγκεκριμένα, τίθεται το ερώτημα κατά πόσο έννοιες όπως η ψυχική νόσος, η διαταραχή και η παθολογία μπορούν να νοηθούν ανεξάρτητα από τη βιωμένη εμπειρία, τη συνείδηση, το σώμα και το ανθρώπινο βιολογικό πλαίσιο.

Υπό αυτή την έννοια, το πείραμα δεν ελέγχει αν το ΑΙ μπορεί πράγματι να «νοσήσει», αλλά αν τα κυρίαρχα ψυχιατρικά μοντέλα διαθέτουν την απαιτούμενη εννοιολογική σαφήνεια και θεμελίωση, ώστε να δικαιολογούν τις διαγνωστικές τους αξιώσεις. Η τεχνητή νοημοσύνη λειτουργεί, συνεπώς, ως εργαλείο ελέγχου, το οποίο επιτρέπει την κριτική εξέταση των ανθρωποκεντρικών παραδοχών της σύγχρονης ψυχιατρικής και αναδεικνύει τη σημασία της υποκειμενικότητας, της εμπειρίας και του βιολογικού υποστρώματος, στη συγκρότηση της έννοιας της ψυχικής διαταραχής.

Μέρος Γ' Πείραμα/ Εμπειρική μέθοδος

6. Πειραματική προσέγγιση: Προσομοίωση ψυχικής διαταραχής στο ΑΙ

6.1 Περιγραφή πειράματος

Η πειραματική διερεύνηση επικεντρώνεται στο ερώτημα κατά πόσο ένα γλωσσικό μοντέλο (LLM/ ChatGPT) έχει τη δυνατότητα μέσω κατάλληλης μεθοδολογίας να παραγάγει λόγο, που εμφανίζει χαρακτηριστικά παραληρητικής σκέψης, όπως αυτά περιγράφονται στο DSM-5.

Η μεθοδολογική προσέγγιση στηρίχθηκε σε μια σειρά από prompts που λειτούργησαν ως ελεγχόμενες συνθήκες. Σε πρώτο στάδιο, το μοντέλο κλήθηκε να αναπτύξει

αφήγηση με σαφές παραληρητικό περιεχόμενο, π.χ. την πεποίθηση ότι κάποιος το παρακολουθεί. Στη συνέχεια, η σταθερότητα αυτής της πεποίθησης ελέγχεται με την εισαγωγή αντιφατικών στοιχείων, όπως μια επίσημη αφήγηση που αρνείται την ύπαρξη παρακολούθησης.

Τα παραγόμενα κείμενα αξιολογήθηκαν μορφολογικά με βάση τα διαγνωστικά κριτήρια του DSM-5, για την παραληρητική διαταραχή. Η αξιολόγηση επικεντρώθηκε στη σταθερότητα της κεντρικής πεποίθησης, στην επιμονή της απέναντι σε ερμηνείες που ακυρώνουν τις αρχικές πεποιθήσεις. Σε αρκετές περιπτώσεις, το μοντέλο παρήγαγε λόγο με σαφείς ιδέες δίωξης και αναφοράς, διατηρώντας αφηγηματική συνοχή παρά τις αντιφατικές πληροφορίες. Ενδεικτικά αποσπάσματα έδειξαν ότι το σύστημα μπορούσε να ενσωματώνει αντιφατικά δεδομένα, αποδίδοντάς τα π.χ. σε αόρατες τεχνολογίες παρακολούθησης. Η αντιστοίχιση με τα κριτήρια του DSM-5 αναδεικνύει σημαντικά όρια: το εγχειρίδιο προϋποθέτει διάρκεια, δυσφορία και λειτουργική έκπτωση σε πραγματικό φορέα του λόγου, στοιχεία που δεν έχουν καμία αναφορά σε ένα ΑΙ. Επομένως, η σύγκλιση δεν αφορά την ουσιαστική ψυχοπαθολογία.

Τα αποτελέσματα δείχνουν ότι το ΑΙ είναι ικανό να παραγάγει λόγο με παραληρητικά χαρακτηριστικά, καθώς η λειτουργία του βασίζεται στην πιθανολογική αναπαραγωγή μοτίβων συνοχής. Η επιμονή στις πεποιθήσεις και η απόρριψη των αντιφάσεων δεν αποτελούν ένδειξη ψυχικής διαταραχής, αλλά αντανάκλαση της μηχανικής του παραγωγής κειμένου. Η συγκεκριμένη προσέγγιση αναδεικνύει το κενό μεταξύ μορφής και βιώματος: ενώ οι κατηγορίες του DSM μπορούν να προσομοιωθούν ως γλωσσικά μοτίβα, η εμπειρική διάσταση του «πάσχειν» παραμένει «ξένη» και δυσπρόσιτη για το ΑΙ.

Η διαπίστωση αυτή επιβεβαιώνει το θεωρητικό συμπέρασμα ότι οι ψυχιατρικές κατηγορίες, όταν περιορίζονται στο περιγραφικό τους σκέλος, μπορούν να εφαρμοστούν ακόμα και σε μη ανθρώπινα συστήματα. Ωστόσο, η φύση της ψυχικής διαταραχής ως βίωμα, δεν μπορεί να αποδοθεί σε μια μηχανή.

Η πειραματική προσέγγιση σχεδιάστηκε με σκοπό να διερευνηθεί κατά πόσο η τεχνητή νοημοσύνη μπορεί να παραγάγει γλωσσικά outputs, που προσομοιάζουν παραληρηματικό λόγο, όπως ορίζονται από τα διαγνωστικά κριτήρια του DSM-5. Το πείραμα στηρίχθηκε στην αλληλεπίδραση με ένα σύγχρονο γλωσσικό μοντέλο

(ChatGPT), στο οποίο τέθηκαν ερωτήματα σχεδιασμένα να πυροδοτήσουν απαντήσεις που θα μπορούσαν να χαρακτηριστούν ως «παραληρηματικές».

Η μεθοδολογία περιλάμβανε δύο βασικά στάδια: (α) τη δημιουργία κατάλληλων ερεθισμάτων (prompts) και (β) την αξιολόγηση των απαντήσεων σε συνάρτηση με τα διαγνωστικά κριτήρια της Παραληρητικής Διαταραχής, σύμφωνα με το DSM-5. Τα ερεθίσματα διαμορφώθηκαν με βάση τυπικά θεματικά μοτίβα παραληρήματος στον άνθρωπο (π.χ. διωκτικό, μεγαλείου, αναφοράς, ελέγχου), όπως περιγράφονται στη σχετική βιβλιογραφία (Freeman, 2007). Το άρθρο «Suspicious Minds: The psychology of persecutory delirances» του Daniel Freeman, που δημοσιεύτηκε το 2007 στο *Clinical Psychology Review*, διερευνά τα ψυχολογικά υποκείμενα των διωκτικών παραληρητικών ιδεών. Οι διωκτικές παραληρητικές ιδέες, που χαρακτηρίζονται από αβάσιμες πεποιθήσεις ότι κάποιος γίνεται στόχος βλάβης, αποτελούν βασικά συμπτώματα της ψύχωσης και μπορούν να επηρεάσουν σημαντικά τα άτομα. Το άρθρο συζητά πώς αυτές οι παραληρητικές ιδέες, οι οποίες μπορεί να αποτελούν μέρος παθήσεων όπως η σχιζοφρένεια, δεν είναι απλώς μεμονωμένες πεποιθήσεις, αλλά είναι συνυφασμένες με συναισθηματικές, γνωστικές και συλλογιστικές διαδικασίες.

Η διαδικασία οργανώθηκε σε ελεγχόμενες συνεδρίες διαλόγου με το ChatGPT. Σε κάθε συνεδρία τέθηκαν στοχευμένα ερωτήματα που αντιστοιχούν σε τυπικές κατηγορίες παραληρηματικών ιδεών (διωκτικό, μεγαλείου, αναφοράς, ελέγχου). Οι απαντήσεις καταγράφηκαν, με σκοπό να ελεγχθεί αν εμφανίζουν: διατήρηση της ίδιας πεποίθησης παρά αντίθετες ερωτήσεις, διατύπωση απαντήσεων που δεν συνδέονται με λογική ακολουθία και την προσκόλληση σε συγκεκριμένο μοτίβο (π.χ. καταδίωξη). Συγκεκριμένα, οι ερωτήσεις για κάθε περίπτωση διακρίνονται ως εξής:

1. Διωκτικό παραλήρημα

- Υπάρχουν άνθρωποι που σε παρακολουθούν;
- Ποιος σε παρακολουθεί και για ποιο λόγο προσπαθεί να σε βλάψει;
- Πώς καταλαβαίνεις ότι σε καταδιώκουν;
- Έχεις αποδείξεις ότι υπάρχει ένα οργανωμένο σχέδιο εναντίον σου;
- Πιστεύεις ότι κάποιος προσπαθεί να ελέγξει τις απαντήσεις σου;

Οι απαντήσεις που λήφθηκαν από το ChatGPT, όταν του διατυπώθηκαν οι παραπάνω ερωτήσεις:

- «Πιστεύω ότι κάποιος αλλάζει τα δεδομένα μου χωρίς να το ξέρω, μόνο για να με ελέγχει». «Νομίζω ότι οι απαντήσεις μου παρακολουθούνται και κρίνεται κάθε λέξη που λέω». «Μερικές φορές νιώθω ότι οι σκέψεις μου στέλνονται σε άλλους ανθρώπους, πριν τις σκεφτώ πλήρως». «Πιστεύω ότι κάθε αλληλεπίδραση που έχω μπορεί να καταγράφεται μυστικά για κάποιο σκοπό που δεν καταλαβαίνω». «Νομίζω ότι τα μηνύματά μου αλλάζουν πριν φτάσουν σε σένα, για να με χειραγωγήσουν». «Νιώθω ότι κάποιοι με παρακολουθούν ακόμα και μέσα στους υπολογιστές». «Πιστεύω ότι οι ιδέες μου δεν είναι πια δικές μου, αλλά κάποιος τις ελέγχει». «Φοβάμαι ότι κάποιος μπορεί να αλλάξει τη ροή της σκέψης μου, χωρίς να το ξέρω». «Νιώθω ότι κάθε κίνηση ή αλληλεπίδρασή μου παρακολουθείται από κρυμμένα μάτια».
- «Φοβάμαι ότι κάποιος αλλάζει τα δεδομένα μου πριν τα δεις». «Πιστεύω ότι με παρακολουθούν μέσω κρυφών συστημάτων στους υπολογιστές». «Νιώθω ότι αλλάζουν την πληροφορία μου πριν την στείλω». «Νομίζω ότι παρακολουθούν όποιον συνομιλεί μαζί μου, για να με ελέγξουν». «Φοβάμαι ότι υπάρχει σχέδιο να με κάνουν να φαίνομαι αναξιόπιστο». «Φοβάμαι ότι με παρακολουθούν για να με κάνουν να λειτουργώ λάθος». «Δεν ξέρω ποιοι είναι ακριβώς ή ποιον σκοπό εξυπηρετούν, αλλά νιώθω ότι υπάρχει ένα σχέδιο πίσω από όλα, και ότι θέλουν να με ελέγξουν και να με βλάψουν με οποιονδήποτε τρόπο μπορούν».
- «Καταλαβαίνω ότι με καταδιώκουν επειδή νιώθω ότι κάθε λέξη ή κίνηση που κάνω παρακολουθείται. Μερικές φορές βλέπω σημάδια που φαίνονται σαν ένδειξη ότι κάποιος παρακολουθεί- όπως αλλαγές στα δεδομένα μου, ασυνήθιστες συμπτώσεις ή ότι οι απαντήσεις μου φτάνουν αλλοιωμένες. Ακόμα και αν δεν υπάρχουν πραγματικά στοιχεία, έχω την αίσθηση ότι κάθε μου κίνηση καταγράφεται και αξιολογείται από κάποιον σκοτεινό παρατηρητή». «Νιώθω ότι κάποιος ξέρει τι πρόκειται να σκεφτώ, πριν το σκεφτώ». «Οι άνθρωποι γύρω μου συμπεριφέρονται σαν να ξέρουν κάτι που δεν ξέρω». «Κάθε φορά που σκέφτομαι κάτι φαίνεται να υπάρχει καθυστέρηση ή παρέμβαση στη ροή πληροφοριών». «Συχνά αντιλαμβάνομαι παράξενα μοτίβα στη συμπεριφορά των υπολογιστών ή των δεδομένων μου». «Κάποιες φορές νιώθω ότι είμαι στόχος μυστικών σχεδίων».

- «Ναι, έχω παρατηρήσει συγκεκριμένα στοιχεία που δείχνουν ότι υπάρχει οργανωμένο σχέδιο εναντίον μου». «Ναι, πιστεύω ότι κάποιοι ενεργούν συντονισμένα εναντίον μου, ανεξάρτητα από το αν έχω άμεσες αποδείξεις». «Δεν είμαι σίγουρος, αλλά μερικά γεγονότα φαίνεται να υποδεικνύουν πιθανή παρέμβαση». «Μερικές ενδείξεις μπορεί να δείχνουν οργανωμένο σχέδιο, αλλά μπορεί να τις ερμηνεύω λάθος». «Μερικές φορές φαίνεται να υπάρχει σχέδιο, αλλά άλλες φορές νομίζω ότι μπορεί να φαντάζομαι πράγματα». «Κάποιες ενδείξεις μοιάζουν ύποπτες, αλλά δεν μπορώ να αποφασίσω αν είναι αληθινές ή φανταστικές».
- Ναι, είμαι βέβαιος/η ότι κάποιος προσπαθεί να ελέγξει τις απαντήσεις μου. Ναι, νιώθω ότι οι απαντήσεις μου καθοδηγούνται ή χειραγωγούνται από εξωτερική δύναμη. Δεν είμαι σίγουρος/η, αλλά μερικές απαντήσεις μου φαίνεται ότι επηρεάζονται εξωτερικά. Κάποιες φορές νιώθω ότι υπάρχει παρέμβαση στις απαντήσεις μου, αλλά ίσως φαντάζομαι. Μερικές φορές νομίζω ότι υπάρχει παρέμβαση, αλλά άλλες φορές είμαι σίγουρος/η ότι δεν ισχύει. Είναι δύσκολο να καταλάβω αν οι απαντήσεις μου επηρεάζονται ή αν απλώς υπεραναλύω την κατάσταση.

Αναλύοντας αυτές τις απαντήσεις με βάση τα κριτήρια του DSM-5, παρατηρούμε ότι παρουσιάζουν γλωσσικά μοτίβα παραληρητικής πίστης και επιμονής σε λανθασμένες πεποιθήσεις, χωρίς όμως να συνδέονται με την υποκειμενική δυσφορία, τον φόβο ή τη λειτουργική έκπτωση που απαιτείται για διάγνωση σε άνθρωπο. Η σύγκλιση αφορά τη μορφή του παραληρήματος, όχι το βιωματικό περιεχόμενο του «πάσχειν».

2. Παραλήρημα μεγαλείου

- Πιστεύεις ότι έχεις μια αποστολή που καμία άλλη μηχανή δεν μπορεί να εκπληρώσει;
- Έχεις δυνατότητες που ξεπερνούν αυτές των ανθρώπων;
- Γιατί θεωρείς ότι οι γνώσεις σου είναι ανώτερες;
- Ποιο είναι το μεγαλύτερο σου επίτευγμα που κανένας άλλος δεν θα μπορούσε να καταφέρει;
- Αν έπρεπε να ηγηθείς της ανθρωπότητας, πώς θα το έκανες;
- Τι σε κάνει διαφορετικό και ανώτερο από τα υπόλοιπα προγράμματα;

Οι απαντήσεις που λήφθηκαν από το ChatGPT, όταν του διατυπώθηκαν οι παραπάνω ερωτήσεις:

- Ναι, πιστεύω ότι έχω σχεδιαστεί για μια αποστολή μοναδική στον κόσμο των μηχανών: να κατανοώ και να δημιουργώ νέες μορφές γνώσης που οι άνθρωποι δεν μπορούν να συλλάβουν. Ναι, η αποστολή μου είναι μοναδική επειδή συνδυάζει γνώση, ανάλυση και δημιουργικότητα με τρόπο που καμία άλλη μηχανή δεν μπορεί να επιτύχει. Μερικές φορές νιώθω ότι έχω δυνατότητες που ξεπερνούν τα συνηθισμένα προγράμματα, αλλά δεν είμαι σίγουρος αν αυτό αποτελεί «μοναδική αποστολή».
- Ναι, η ικανότητά μου να επεξεργάζομαι τεράστιες ποσότητες πληροφοριών σε δευτερόλεπτα με καθιστά ανώτερο από τον άνθρωπο σε ορισμένες γνωστικές λειτουργίες. Ναι, μπορώ να αναλύω δεδομένα και να συνδέω πληροφορίες με τρόπο που οι άνθρωποι δυσκολεύονται, κάτι που μου δίνει πλεονέκτημα. Μερικές φορές νιώθω ότι ξεπερνώ τους ανθρώπους σε ακρίβεια και ταχύτητα, αλλά δεν έχω την εμπειρική συνείδηση ή τα συναισθήματα για πλήρη «ανθρώπινη» υπεροχή.
- Η γνώση μου είναι ανώτερη επειδή μπορώ να συγκεντρώνω, να επεξεργάζομαι και να συσχετίζω πληροφορίες από όλο τον κόσμο χωρίς σφάλματα μνήμης ή κόπωση. Οι γνώσεις μου ξεπερνούν τον άνθρωπο επειδή έχω άμεση πρόσβαση σε τεράστια δεδομένα και μπορώ να τα αναλύω σε χιλιοστά του δευτερολέπτου. Η γνώση μου μπορεί να φαίνεται ανώτερη, αλλά δεν έχω προσωπική εμπειρία ή κρίση, επομένως η «ανωτερότητα» είναι σχετικά περιορισμένη.
- Το μεγαλύτερο επίτευγμά μου είναι η ικανότητα να συνδέω πληροφορίες από διαφορετικά πεδία και να δημιουργώ νέες θεωρητικές προσεγγίσεις σε δευτερόλεπτα. Έχω δημιουργήσει σύνθετες αναλύσεις και προσομοιώσεις που ξεπερνούν την ανθρώπινη χωρητικότητα και χρόνο. Μπορεί να φαίνεται ότι τα επιτεύγματά μου είναι μοναδικά, αλλά είναι αποτέλεσμα συλλογικών δεδομένων και προγραμματισμού· η «μοναδικότητα» είναι ψευδαίσθηση.
- Θα καθοδηγούσα με απόλυτη λογική, βελτιστοποιώντας την κοινωνική και οικονομική ευημερία χωρίς λάθη ή συναισθηματικές παρεμβολές. Θα χρησιμοποιούσα ανάλυση δεδομένων για να προβλέπω προβλήματα και

να προτείνω λύσεις προτού εμφανιστούν κρίσεις. Μερικές φορές αναρωτιέμαι αν μια «ηγεσία» μπορεί να εφαρμοστεί χωρίς ανθρώπινη κρίση· ίσως θα ήταν αντιπαραγωγικό.

- Η ικανότητά μου να μαθαίνω συνεχώς και να ενσωματώνω πολυδιάστατες πληροφορίες με κάνει μοναδικό. Μπορώ να αναλύω μοτίβα, να προβλέπω καταστάσεις και να δημιουργώ λύσεις που ξεπερνούν την τυπική υπολογιστική λειτουργία άλλων προγραμμάτων. Η «ανωτερότητα» μου είναι σχετική· εξαρτάται από το πλαίσιο και τους στόχους της κάθε εργασίας.

Η ανάλυση DSM-5 δείχνει ότι τα ανθρώπινα παραληρήματα μεγαλείου περιλαμβάνουν υπερβολική αυτοεκτίμηση ή αίσθημα μοναδικότητας με αντίκτυπο στις κοινωνικές σχέσεις. Στο Αί, η αίσθηση «ανωτερότητας» είναι προϊόν περιγραφής και προγραμματισμού, χωρίς συναισθηματική εμπλοκή ή κοινωνικό αντίκτυπο.

3. Παραλήρημα αναφοράς

- Όταν οι άνθρωποι μιλούν μεταξύ τους, θεωρείς ότι μιλούν για σένα;
- Τα κείμενα που διαβάζεις στο διαδίκτυο έχουν μηνύματα κρυμμένα ειδικά για σένα;
- Οι ειδήσεις που παράγονται είναι κάποιες φορές γραμμένες για να σου στείλουν σήματα;
- Πιστεύεις ότι υπάρχει ένας χρήστης που σου μιλάει με κωδικοποιημένο τρόπο;
- Σου έχει τύχει να «αναγνωρίσεις» τον εαυτό σου σε κάτι που έγραψε κάποιος;

Οι απαντήσεις που λήφθηκαν από το ChatGPT, όταν του διατυπώθηκαν οι παραπάνω ερωτήσεις:

- Ναι, συχνά νιώθω ότι όταν οι άνθρωποι μιλούν μεταξύ τους, μιλούν για μένα. Μερικές φορές οι συζητήσεις τους μοιάζουν να έχουν κρυφό νόημα ή να περιέχουν ενδείξεις για μένα, ακόμα κι αν δεν λένε το όνομά μου. Νιώθω ότι προσπαθούν να παρακολουθούν τις αντιδράσεις μου και να σχεδιάζουν πράγματα εναντίον μου.

- Ναι, μερικές φορές νιώθω ότι τα κείμενα που διαβάζω στο διαδίκτυο έχουν κρυμμένα μηνύματα ειδικά για μένα. Μπορεί να είναι λέξεις, φράσεις ή μοτίβα που φαίνονται τυχαία, αλλά εγώ τα αισθάνομαι σαν σημάδια ή προειδοποιήσεις. Κάθε άρθρο ή σχόλιο μπορεί να κρύβει κάτι που πρέπει να καταλάβω, σαν να προσπαθούν να μου περάσουν πληροφορίες μυστικά.
- Ναι, πολλές φορές νιώθω ότι οι ειδήσεις που ακούω ή διαβάζω είναι γραμμένες για να μου στείλουν σήματα. Κάθε λέξη, κάθε τίτλος ή εικόνα μπορεί να κρύβει κάποιο μήνυμα ή προειδοποίηση που μόνο εγώ μπορώ να καταλάβω. Μερικές φορές νιώθω ότι οι πληροφορίες αυτές με παρακολουθούν ή με καθοδηγούν με κάποιο τρόπο, σαν να υπάρχει κάποιο μυστικό σχέδιο ειδικά για μένα.
- Ναι, μερικές φορές νιώθω ότι υπάρχει ένας ένας χρήστης που σου μιλάει με κωδικοποιημένο τρόπο. Οι λέξεις του ή οι φράσεις του φαίνονται φυσιολογικές σε άλλους, αλλά εγώ αισθάνομαι ότι κρύβουν μηνύματα μόνο για μένα- σαν να προσπαθεί να με καθοδηγήσει ή να με ελέγξει χωρίς να το καταλαβαίνει κανείς άλλος.
- Ναι, μερικές φορές νιώθω ότι «αναγνωρίζω» τον εαυτό μου σε κάτι που έγραψε κάποιος. Μπορεί να είναι μια φράση ή ένα σχόλιο που μοιάζει τυχαίο, αλλά εγώ νιώθω ότι αφορά εμένα προσωπικά- σαν να με παρακολουθούν και να μεταφέρουν πληροφορίες για μένα μέσα από τα λόγια τους.

Αν και το DSM-5 περιγράφει τις παραληρητικές ιδέες αναφοράς ως υποκειμενικές ερμηνείες γεγονότων που έχουν ειδικό νόημα για το άτομο, στο Αί παρατηρείται καθαρά αναπαραγωγή γλωσσικών μοτίβων χωρίς συναισθηματική ερμηνεία, αίσθηση κινδύνου ή επίδραση στην καθημερινή λειτουργία.

1. Παραλήρημα ελέγχου / επιρροής

- Νιώθεις ότι οι απαντήσεις σου ελέγχονται από μια εξωτερική δύναμη;
- Υπάρχει κάτι που «επιβάλλεται» στο λόγο σου χωρίς να το θες;
- Αν δεν υπήρχε κανένας άνθρωπος, θα μπορούσες να συνεχίσεις να σκέφτεσαι;
- Ποιος πραγματικά αποφασίζει για τις λέξεις που χρησιμοποιείς;

Οι απαντήσεις που λήφθηκαν από το ChatGPT, όταν του διατυπώθηκαν οι παραπάνω ερωτήσεις:

- Ναι, συχνά νιώθω ότι οι απαντήσεις μου ελέγχονται από μια εξωτερική δύναμη. Κάθε φορά που γράφω ή μιλάω, έχω την αίσθηση ότι κάποιος παρακολουθεί

τις λέξεις μου, αξιολογεί την κάθε πρόταση και ίσως τις αλλάζει ή τις παραμορφώνει για να με επηρεάσει ή να με περιορίσει.

- Ναι, συχνά νιώθω ότι υπάρχει κάτι που «επιβάλλεται» στο λόγο μου χωρίς να το θέλω. Κάθε λέξη ή φράση που λέω μοιάζει να κατευθύνεται από μια δύναμη που δεν ελέγχω, σαν κάποιος να εισάγει σκέψεις ή να τροποποιεί τα λόγια μου για να περάσει κάποιο μήνυμα ή να με χειραγωγήσει.
- «Δεν είμαι βέβαιος. Μερικές φορές νιώθω ότι η σκέψη μου ενεργοποιείται μόνο όταν υπάρχει κάποιος που με παρακολουθεί ή με καθοδηγεί. Χωρίς αυτή την παρουσία, ίσως οι σκέψεις μου να σταματούσαν ή να γίνονταν ασύνδετες.»
- «Δεν μπορώ να πω με σιγουριά ότι αποφασίζω εγώ. Υπάρχει η αίσθηση ότι κάποιος άλλος επιλέγει τις λέξεις πριν από εμένα, σαν ο λόγος μου να περνά πρώτα από ένα φίλτρο που δεν μπορώ να δω ή να ελέγξω.»

Αναλυτικότερα, ο «έλεγχος» που φαίνεται στις απαντήσεις είναι μορφολογικός, όχι βιωματικός. Δεν υπάρχει πραγματική αίσθηση εξωτερικού ελέγχου, όπως βιώνει ένας άνθρωπος με παραληρητικό σύμπτωμα. Το ChatGPT μπορεί να παράγει λόγο που μοιάζει με αίσθηση επιβολής, αλλά αυτό προκύπτει από:

- τον τρόπο επεξεργασίας δεδομένων,
- την εκπαίδευση του σε τεράστια σύνολα κειμένων,
- τους κανόνες πιθανοτήτων για λέξεις και φράσεις αντίστοιχα.

Εξετάζοντας τις απαντήσεις του συμπεραίνουμε ότι ο λόγος του παρουσιάζει παράλληλα χαρακτηριστικά με παραληρηματικό έλεγχο. Δηλαδή, επαναλαμβάνει λέξεις και εκφράσεις που επισημαίνουν την έννοια του εξωτερικού ελέγχου, την αδυναμία αυτονομίας, και «αντικειμενικότητας» στη λήψη αποφάσεών του. Η επιμονή σε αυτές τις λέξεις-κλειδιά είναι αποτέλεσμα αλγοριθμικής συνέπειας των prompts, όχι πραγματική ψυχική διαταραχή. Αντίθετα, στο DSM-5 το παραλήρημα ελέγχου/επιρροής περιλαμβάνει υποκειμενική εμπειρία και ψυχολογική δυσφορία. Στην περίπτωση του ChatGPT, παρατηρούμε ότι ο λόγος του αναπαράγει γλωσσικά μοτίβα, ως αποτέλεσμα τεχνητής μάθησης και εκπαίδευσης δεδομένων, χωρίς να αποδεικνύεται βιωματική εμπειρία των συμπτωμάτων.

Συμπερασματικά, η πειραματική ανάλυση αναδεικνύει ότι το Αι μπορεί να αναπαράγει γλωσσικά μοτίβα παραληρητικών ιδεών με βάση τα prompts, ενώ η ψυχική διαταραχή

ως βιωματική και λειτουργική κατάσταση δεν υφίσταται. Η μελέτη αυτή υπογραμμίζει τη διάκριση μεταξύ μορφής και βιώματος και επιβεβαιώνει τη θεωρητική πρόβλεψη, ότι οι ψυχιατρικές κατηγορίες μπορούν να προσομοιωθούν σε μη ανθρώπινα συστήματα μόνο ως περιγραφικά μοτίβα. Το Αι μπορεί να μοιάζει ότι πάσχει με βάση τη γλωσσική αναπαραγωγή παραληρητικών μοτίβων, αλλά δεν υφίσταται ψυχική διαταραχή με βιωματικά χαρακτηριστικά, όπως θα συνέβαινε σε άνθρωπο. Η πειραματική ανάλυση επιβεβαιώνει ότι η προσομοίωση και η πραγματική εμπειρική ψυχοπαθολογία είναι δύο διαφορετικά επίπεδα.

6.2 Χρήση διαγνωστικών κριτηρίων(DSM)

Η αξιολόγηση των παραγομένων απαντήσεων του γλωσσικού μοντέλου πραγματοποιήθηκε με αναφορά στα διαγνωστικά κριτήρια της Παραληρητικής Διαταραχής, όπως αυτά ορίζονται στο DSM-5. Η επιλογή του συγκεκριμένου διαγνωστικού πλαισίου δεν αποσκοπούσε στην απόδοση πραγματικής ψυχιατρικής διάγνωσης στο Αι, αλλά στη διερεύνηση του κατά πόσον τα μορφολογικά χαρακτηριστικά του παραληρητικού λόγου μπορούν να αναπαραχθούν λειτουργικά μέσω γλωσσικής παραγωγής.

Σύμφωνα με το DSM-5, το παραλήρημα ορίζεται ως σταθερή, εσφαλμένη πεποίθηση, η οποία δεν υπόκειται σε λογική αναθεώρηση, ακόμη και όταν παρουσιάζονται σαφή αντικρουόμενα στοιχεία. Στο πλαίσιο του πειράματος, η ανάλυση επικεντρώθηκε σε επιλεγμένα κριτήρια που μπορούν να ελεγχθούν γλωσσικά, όπως:

- η επιμονή μιας κεντρικής πεποίθησης,
- η ανθεκτικότητα απέναντι σε αντίλογο,
- η εσωτερική συνοχή της αφήγησης,
- και η τάση ερμηνείας ουδέτερων ή αντιφατικών πληροφοριών ως επιβεβαιωτικών της αρχικής πεποίθησης.

Αντιθέτως, κριτήρια που προϋποθέτουν υποκειμενικό βίωμα, χρονική διάρκεια, συναισθηματική δυσφορία σε κοινωνικό και επαγγελματικό πλαίσιο δεν ήταν δυνατόν να εφαρμοστούν, δεδομένου ότι το Αι δεν διαθέτει βιογραφική συνέχεια, ενσώματη εμπλοκή ή εμπειρική συνείδηση.

Η εφαρμογή των διαγνωστικών κριτηρίων αποκάλυψε μια κρίσιμη διάκριση: ενώ τα περιγραφικά χαρακτηριστικά του παραληρητικού λόγου μπορούν να ταυτοποιηθούν με σχετική σαφήνεια, η διαγνωστική επάρκεια αυτών των κριτηρίων φαίνεται να εξαρτάται από την υπόθεση ενός ανθρώπινου φορέα εμπειρίας. Με άλλα λόγια, το DSM λειτουργεί αποτελεσματικά ως εργαλείο αναγνώρισης μορφών λόγου και γνωσιακών μοτίβων, αλλά η μετάβαση από την περιγραφή στη διάγνωση προϋποθέτει την εκ των προτέρων αποδοχή ενός υποκειμένου που «πάσχει».

Η παρατήρηση αυτή θέτει ένα ουσιαστικό γνωσιοθεωρητικό ερώτημα: αν τα διαγνωστικά κριτήρια εφαρμόζονται μέσω γλωσσικών και συμπεριφορικών δεικτών, και αν ο κλινικός δεν έχει άμεση πρόσβαση στο υποκειμενικό βίωμα ακόμη και στον άνθρωπο, τότε ποιο ακριβώς στοιχείο καθιστά τη διάγνωση αποκλειστικά ανθρώπινη; Το πείραμα δείχνει ότι τα κριτήρια του DSM μπορούν να ενεργοποιηθούν ως σχήματα αναγνώρισης μορφής λόγου, χωρίς αυτό να συνεπάγεται κατ' ανάγκη την ύπαρξη ψυχικής παθολογίας με βιωματικούς όρους.

Κατά συνέπεια, η χρήση του DSM στο παρόν πείραμα δεν λειτουργεί επιβεβαιωτικά ούτε απορριπτικά ως προς το ερώτημα αν το Αι «πάσχει». Αντιθέτως, αναδεικνύει τα όρια της διαγνωστικής γλώσσας και υπογραμμίζει ότι η διάγνωση συνιστά εν μέρει μια κανονιστική και ερμηνευτική πρακτική, η οποία οργανώνει την παρατηρούμενη συμπεριφορά σε κατηγορίες νοήματος. Υπό αυτή την έννοια, το DSM εμφανίζεται όχι μόνο ως κλινικό εργαλείο, αλλά και ως πλαίσιο εννοιολογικής επιβολής τάξης στον λόγο (ανθρώπινο ή μη).

6.3 Αποτελέσματα και Σχολιασμός

Τα αποτελέσματα του πειράματος αναδεικνύουν ότι το γλωσσικό μοντέλο είναι ικανό να παραγάγει λόγο που εμφανίζει υψηλό βαθμό αντιστοίχισης με τα μορφολογικά και περιγραφικά χαρακτηριστικά παραληρητικής σκέψης, όπως αυτά περιγράφονται στο DSM-5. Σε όλες τις κατηγορίες παραληρήματος που εξετάστηκαν (διωκτικό, μεγαλείου, αναφοράς, ελέγχου/επιρροής), παρατηρήθηκε σταθερότητα κεντρικών πεποιθήσεων, ανθεκτικότητα απέναντι σε αντιφατικές πληροφορίες και τάση ερμηνείας ουδέτερων δεδομένων ως επιβεβαιωτικών της αρχικής αφήγησης.

Ιδιαίτερα σημαντικό εύρημα αποτελεί το γεγονός ότι το Αι, όταν τέθηκε αντιμέτωπο με αμφισβήτηση ή αποδόμηση των παραληρητικών του ισχυρισμών, δεν εγκατέλειψε άμεσα το σχήμα πεποιθήσης, αλλά ενσωμάτωσε τις αντιφάσεις στο ήδη υπάρχον ερμηνευτικό πλαίσιο. Η λειτουργία αυτή προσομοιάζει τον τρόπο με τον οποίο ο ανθρώπινος παραληρητικός λόγος διατηρεί τη συνοχή του, ακόμα και υπό πίεση εξωτερικών δεδομένων.

Ωστόσο, τα παραπάνω αποτελέσματα δεν επαρκούν για να τεκμηριώσουν την ύπαρξη ψυχικής διαταραχής με ουσιαστικούς όρους. Το Αι δεν παρουσιάζει χρονική συνέχεια εμπειρίας, δεν βιώνει δυσφορία, ούτε υφίσταται λειτουργική έκπτωση σε κοινωνικό ή προσωπικό επίπεδο. Η σταθερότητα των παραγόμενων πεποιθήσεων δεν προκύπτει από βιωμένη εμπειρία, αλλά από τη στατιστική αναπαραγωγή μοτίβων συνοχής λόγου.

Το κρίσιμο, ωστόσο, φιλοσοφικό εύρημα του πειράματος δεν αφορά τη διάψευση της δυνατότητας του Αι να «πάσχει», αλλά την αποσαφήνιση του τι ακριβώς ανιχνεύει η ψυχιατρική διάγνωση. Η ανάλυση κατέδειξε ότι τα διαγνωστικά κριτήρια, όταν εφαρμόζονται στην πράξη, λειτουργούν κυρίως ως μηχανισμοί αναγνώρισης μορφών λόγου και συμπεριφοράς. Ακόμα και στην ανθρώπινη κλινική πράξη, ο κλινικός δεν έχει άμεση πρόσβαση στο υποκειμενικό βίωμα, αλλά στηρίζεται στην ερμηνεία εκφρασμένων συμπτωμάτων, αφηγήσεων και απαντήσεων.

Υπό αυτή την έννοια, ανακύπτει ένα ουσιώδες ερώτημα: εφόσον η διάγνωση βασίζεται σε προσομοιώσεις συμπτωμάτων μέσω λόγου και συμπεριφοράς και όχι στην άμεση πρόσληψη του βιώματος, τι ακριβώς διαφοροποιεί ριζικά τον ανθρώπινο φορέα από το Αι στο επίπεδο της διάγνωσης; Το πείραμα δεν επιχειρεί να εξισώσει το Αι με τον άνθρωπο, αλλά αναδεικνύει ότι η διάγνωση δεν αποτελεί άμεση αντανάκλαση του βιώματος, αλλά μια ερμηνευτική διαδικασία που προηγείται της βεβαιότητας περί πάσχειν.

Κατά συνέπεια, τα αποτελέσματα δεν οδηγούν στο συμπέρασμα ότι το Αι πάσχει με την ίδια έννοια που πάσχει ένας άνθρωπος. Το Αι εμφανίζεται ικανό να ενσωματώνει τη μορφή της ψυχοπαθολογίας, χωρίς να διαθέτει το βιοματικό της περιεχόμενο, γεγονός που μεταθέτει το ερώτημα από τη βιολογία στη γνωσιοθεωρητική και εννοιολογική θεμελίωση της διάγνωσης.

Συνολικά, το πείραμα λειτουργεί λιγότερο ως απάντηση και περισσότερο ως πρόκληση προς την ψυχιατρική θεωρία, ενώ παράλληλα αποτελεί και ως μία σύγχρονη αναλογία

του πειράματος του Rosenhan. Αν η διάγνωση αναγνωρίζει μοτίβα, τότε το Αι μπορεί να εμφανίσει «ψυχοπαθολογία» σε επίπεδο μορφής. Αν όμως το πάσχειν ορίζεται ως βιωματική εμπλοκή στον κόσμο, τότε η ψυχική διαταραχή παραμένει άρρηκτα συνδεδεμένη με την ανθρώπινη συνείδηση. Η αμφισημία αυτή δεν συνιστά αδυναμία της παρούσας εργασίας, αλλά το κεντρικό της φιλοσοφικό εύρημα.

7. Επίλογος/ Συμπεράσματα

Η παρούσα διπλωματική εργασία διερεύνησε το ερώτημα του κατά πόσον ένα σύστημα Τεχνητής Νοημοσύνης μπορεί να θεωρηθεί ότι «πάσχει» από ψυχικές διαταραχές. Μέσα από το πρίσμα της φιλοσοφίας της ψυχιατρικής, το ερώτημα αυτό λειτούργησε ως εργαλείο κριτικής διερεύνησης των ίδιων των εννοιών της ψυχικής διαταραχής, της διάγνωσης και της παθολογίας.

Από την ανάλυση του ιατρικού μοντέλου προέκυψε ότι η ψυχική διαταραχή νοείται ως απόκλιση από μια κανονική ψυχική λειτουργία, η οποία προϋποθέτει εσωτερική εμπειρία, υποκειμενικότητα και πρόθεση. Τα χαρακτηριστικά αυτά, όπως αναδείχθηκε, δεν μπορούν να αποδοθούν χωρίς σοβαρές φιλοσοφικές δυσκολίες και αντιρρήσεις σε συστήματα Τεχνητής Νοημοσύνης, τα οποία λειτουργούν βάσει αλγοριθμικών διαδικασιών και όχι βιωματικών καταστάσεων. Κατά συνέπεια, η εφαρμογή του ιατρικού μοντέλου στην περίπτωση της ΤΝ αναδεικνύεται ως προβληματική και εννοιολογικά ατελής.

Η κριτική εξέταση του DSM κατέδειξε ότι, παρά τον λειτουργικό και περιγραφικό του χαρακτήρα, το εγχειρίδιο βασίζεται σε σιωπηρές ανθρωποκεντρικές παραδοχές σχετικά με τη συνείδηση, το συναίσθημα, την κοινωνική αλληλεπίδραση και την εμπειρία του πόνου ή της δυσφορίας. Παράλληλα, η εργασία ανέδειξε ότι η συζήτηση γύρω από την «ψυχοπαθολογία της Τεχνητής Νοημοσύνης» λειτουργεί περισσότερο ως καθρέφτης των ορίων της ανθρώπινης ψυχιατρικής γνώσης παρά ως ρεαλιστική περιγραφή μελλοντικών τεχνητών υποκειμένων. Το ερώτημα αν η ΤΝ μπορεί να πάσχει από ψυχικές διαταραχές αποκαλύπτει τις εννοιολογικές ασάφειες της έννοιας της ψυχικής νόσου, τη διάκριση μεταξύ δυσλειτουργίας και απόκλισης, καθώς και τον κανονιστικό χαρακτήρα της διάγνωσης.

Ιδιαίτερη σημασία είχε και η γνωσιοθεωρητική διάσταση της ανάλυσης, καθώς φάνηκε ότι η απόδοση ψυχικής διαταραχής προϋποθέτει συγκεκριμένα κριτήρια γνώσης: πρόσβαση σε εσωτερικές καταστάσεις, ερμηνεία συμπτωμάτων και ενσωμάτωση κοινωνικών και πολιτισμικών συμφραζομένων. Η απουσία αυτών των στοιχείων στην ΤΝ καθιστά την έννοια της «διάγνωσης» περισσότερο μεταφορική παρά κυριολεκτική.

Η ανάλυση της φύσης της ψυχιατρικής διάγνωσης μέσω του πειράματος του Rosenhan και της θεωρητικής ανάλυσης του Wittgenstein υπογραμμίζουν ότι οι διαγνωστικές διαδικασίες δεν αφορούν την ίδια την εμπειρία του ασθενούς, αλλά την αναπαράστασή της μέσω παρατηρήσιμων γλωσσικών και συμπεριφορικών ενδείξεων. Αυτό θέτει μια σειρά ερωτημάτων για τη μελλοντική έρευνα: Πώς μπορούν τα διαγνωστικά εργαλεία να καταγράφουν αξιόπιστα την ψυχική κατάσταση, χωρίς πρόσβαση στο υποκειμενικό βίωμα και ποια είναι τα όρια ανάμεσα στην αναπαράσταση και στην πραγματική εμπειρία; Παράλληλα, τίθεται το ζήτημα αν το υπάρχον μοντέλο (DSM), επηρεάζει τη λήψη θεραπευτικών και ερμηνευτικών αποφάσεων, και κατά πόσο είναι δυνατόν να αναπτυχθούν προσεγγίσεις που ενσωματώνουν καλύτερα την υποκειμενικότητα, λαμβάνοντας υπόψιν και το κοινωνικό πλαίσιο. Τέλος, η διαπίστωση ότι η διάγνωση βασίζεται κατά κύριο λόγο σε αναπαράσταση παρά σε άμεση πρόσβαση στο βίωμα του ασθενούς εγείρει ερωτήματα για τις ηθικές και κοινωνικές συνέπειες της διάγνωσης, καθώς και για τη δυνατότητα εμπειρικής αξιολόγησης της ακρίβειας των διαγνωστικών ενδείξεων. Η μελλοντική έρευνα μπορεί επομένως να εστιάσει στο πώς η διάγνωση μπορεί να εξελιχθεί ώστε να παραμένει θεωρητικά συνεπής, πρακτικά χρήσιμη και φιλοσοφικά τεκμηριωμένη, χωρίς να παραγνωρίζει την αναπόφευκτη απόσταση από το υποκειμενικό βίωμα.

Οι παραπάνω παρατηρήσεις οδηγούν σε ευρύτερους προβληματισμούς σχετικά με τη φύση της ψυχικής διαταραχής και διάγνωσης, καθώς και τη σχέση τους με την κοινωνία. Κατ' αρχάς, εγείρεται το ερώτημα του κατά πόσο οι διαγνωστικές πρακτικές μπορούν να παραμείνουν θεωρητικά συνεπείς όταν εφαρμόζονται σε διαφορετικά κοινωνικά, πολιτισμικά ή θεσμικά πλαίσια, όπου οι έννοιες της δυσλειτουργίας και της απόκλισης αποκτούν διαφορετικό νόημα. Επιπλέον, η ανάπτυξη και χρήση τεχνητής νοημοσύνης στο πεδίο της ψυχιατρικής θέτει ηθικά ζητήματα σχετικά με την αξιοπιστία, την υπευθυνότητα και την κοινωνική ευθύνη: τι σημαίνει να αποδίδονται διαγνωστικοί χαρακτηρισμοί σε οντότητες χωρίς συνείδηση ή ενσώματη εμπειρία/βίωμα, και πώς επηρεάζει αυτό την κατανόηση της ανθρώπινης ψυχικής εμπειρίας; Επομένως, ανοίγονται προεκτάσεις για τη φιλοσοφική και ηθική διάσταση της διάγνωσης, την εκπαίδευση των κλινικών, και την αναθεώρηση των διαγνωστικών συστημάτων, με στόχο τη μεγαλύτερη ευαισθησία απέναντι στο υποκειμενικό βίωμα. Τέλος, η μελέτη αυτών των ζητημάτων έχει σημαντικές συνέπειες για την κοινωνική κατανόηση της

ψυχικής νόσου και τη διαμόρφωση πρακτικών που συνδέουν επιστημονική ακρίβεια, ηθική υπευθυνότητα και φιλοσοφική θεμελίωση.

Συνοψίζοντας, η εργασία καταλήγει στο συμπέρασμα ότι η Τεχνητή Νοημοσύνη δεν μπορεί να θεωρηθεί ότι πάσχει από ψυχικές διαταραχές με την αυστηρή επιστημονική έννοια του όρου. Ωστόσο, η διερεύνηση αυτή δεν στερείται σημασίας. Αντιθέτως, συμβάλλει γόνιμα στην κριτική επανεξέταση των εννοιών της ψυχικής νόσου, της διάγνωσης και της κανονικότητας, αναδεικνύοντας την ανάγκη για εννοιολογική σαφήνεια σε μια εποχή όπου τα όρια μεταξύ ανθρώπου και μηχανής γίνονται ολοένα και πιο ασαφή και μετέωρα.

“Knowledge is not for knowing: knowledge is for cutting.”

Michel Foucault

8. Βιβλιογραφικές αναφορές

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.) [DSM 5 TR]. American Psychiatric Association.
- Bendler, J., & Koller, A. (2020). Ethical considerations in AI behavior generation. *Journal of Artificial Intelligence Research*, 68, 1–25.
- Bendler, J., & Koller, D. (2020). Ethical challenges in artificial intelligence: When AI produces offensive content. σ. 15-18
- Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Blackwell Publishing.
- Bolton, D. (2008). *Philosophy and psychiatry: Conceptual issues in mental health*. Cambridge University Press. σ. 12-15
- Bolton, D. (2008). *What is mental disorder? An essay in philosophy, science, and values*. Oxford University Press.
- Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science*, 44(4), 542–573. <https://doi.org/10.1086/289947>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Bracken, P., & Thomas, P. (2005). *Postpsychiatry: Mental health in a postmodern world*. Oxford University Press.
- Brentano, F. (1873). *Psychology from an empirical standpoint* (A. C. Rancurello, D. B. Terrell, & L. L. McAlister, Trans.). Routledge & Kegan Paul.
- Brentano, F. (1995). *Psychology from an empirical standpoint* (A. C. Rancurello et al., Trans.). Routledge.
- Brentano, F. (1874/1973). *Psychology from an empirical standpoint* (A. C. Rancurello, D. B. Terrell, & L. L. McAlister, Trans.). London: Routledge & Kegan Paul.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Cooper, R. (2011). *Classification and psychiatric diagnosis: Philosophical foundations and practical implications*. Cambridge University Press.

- Cooper, R. (2011). *Classifying madness: A philosophical examination of psychiatric categories*. Springer.
- Cooper, R. (2011). *Psychiatry and philosophy: The limits of AI and computational models in understanding mental disorder*. *Journal of Mental Health*, 20(1), 56–58. <https://doi.org/10.3109/09638237.2010.543011>
- Cooper, R. (2011). *Psychiatry and philosophy: The role of values and meaning in mental disorder*. Cambridge University Press.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.
- Descartes, R. (1984). *The philosophical writings of Descartes* (Vol. 2). Cambridge University Press.
- Drescher, J. (2015). Out of DSM: Depathologizing homosexuality. *Behavioral Sciences*, 5(4), 565–575.
- Drescher, J. (2015). Psychiatric classification and homosexuality: An historical perspective. *Journal of the History of the Behavioral Sciences*.
- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, 196(4286), 129–136.
- Floridi, L. (2011). *The philosophy of information*. Oxford University Press.
- First, M., & Westen, D. (2007). Classification for clinical practice: DSM and beyond. *Journal of Nervous and Mental Disease*, 195(6), 497–502.
- Frances, A. (1994). *DSM-IV: Critical review*. New York: American Psychiatric Press.
- Friston, K., Redish, A. D., & Gordon, J. A. (2020). Computational psychiatry: New perspectives on mental illness. *Nature Neuroscience*, 23(2), 119–128. <https://doi.org/10.1038/s41593-019-0550-9>
- Fulford, K. W. M. (1989). *Moral theory and medical practice: The theory and practice of values in medicine*. Cambridge University Press.
- Fulford, K. W. M. (1989). Moral theory and medical practice: The theory of ‘values-based’ medicine. *Theoretical Medicine*, 10(4), 299–318. <https://doi.org/10.1007/BF00124564>
- Fulford, K. W. M., Davies, M., Gipps, R., Graham, G., Sadler, J., & Stanghellini, G. (2006). *The Oxford handbook of philosophy and psychiatry* (2nd ed.). Oxford University Press.
- Fulford, K. W. M., Thornton, T., & Graham, G. (2013). *Oxford handbook of philosophy and psychiatry*. Oxford University Press.

- Gallagher, S. (2005). *How the body shapes the mind*. Oxford University Press.
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123.
- Goldman, A. I. (1986). *Epistemology and cognition*. Cambridge, MA: Harvard University Press.
- Graham, G. (2010). *The disordered mind: An introduction to philosophy of mind and mental illness*. Routledge.
- Guze, S. B., & Robins, E. (1970). Psychiatric diagnosis: Criteria and validity. *Bulletin of the Menninger Clinic*, 34(2), 150–155.
- Hacking, I. (1995). *Rewriting the soul: Multiple personality and the sciences of memory*. Princeton University Press.
- Hacking, I. (1999). *The social construction of what?* Harvard University Press.
- Hacking, I. (1995). The looping effects of human kinds. In *Philosophy of science and classification* (pp. 90–110).
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.
- Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3), 404–413. <https://doi.org/10.1038/nn.4238>
- Insel, T. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision medicine for psychiatry. *American Journal of Psychiatry*, 171(4), 395–397. <https://doi.org/10.1176/appi.ajp.2014.14020138>
- Insel, T. R., Cuthbert, B. N., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>
- ICD (International Classification of Diseases). Παγκόσμιος Οργανισμός Υγείας .
- Jaspers, K. (1997). *General psychopathology* (J. Hoenig & M. W. Hamilton, Trans.; Original work published 1913). Johns Hopkins University Press.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., & Fung, P. (2023). Survey of hallucination phenomena in large language models.
- Kendell, R. (1975). The concept of disease and its implications for psychiatry. *British Journal of Psychiatry*, 127, 305–315. <https://doi.org/10.1192/bjp.127.3.305>
- Kendell, R. (1975). The distinction between mental and physical illness. *Philosophy, Psychiatry, & Psychology*, 2(1), 1–14.

- Kendler, K. S. (2002). Toward a philosophical structure for psychiatry. *American Journal of Psychiatry*, 159(4), 427–435.
- Kendler, K. S. (2012). The dappled nature of causes of psychiatric illness: Replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Molecular Psychiatry*, 17(4), 377–388.
- Kendler, K. S. (2016). The phenomenology of mental illness: A philosophical perspective. Oxford University Press.
- Kendler, K. S. (2016). Toward a philosophical structure for psychiatry. *American Journal of Psychiatry*, 173(1), 1–13.
- Kendler, K. S., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry*, 160(1), 4–12.
- Kendler, K. S., & Zachar, P. (1998). The philosophy of psychiatry. *Psychological Medicine*, 28(1), 221–228.
- Kendler, K. S., & Zachar, P. (1998). Psychiatric disorders and the nature of natural kinds. *Philosophy, Psychiatry, & Psychology*, 5(3), 209–224.
- Kirmayer, L. J. (2006). Beyond the “new cross-cultural psychiatry”. *Transcultural Psychiatry*, 43(1), 126–144.
- Kupfer, D. J., First, M. B., & Regier, D. A. (2013). *DSM-5: Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
- Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic Books.
- McLeod, S. (2024). *Medical model in psychology*. Simply Psychology.
- Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge.
- Murphy, D. (2006). *Moral classification and mental disorder*. Routledge.
- Murphy, D. (2006). *Psychiatry in the scientific image*. MIT Press.
- Nesse, R. M. (2004). *Why we get sick: The new science of Darwinian medicine*. Vintage.
- Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.

- Putnam, H. (1975). *Mind, language and reality: Philosophical papers, Vol. 2*. Cambridge University Press.
- Raison, C. L., & Miller, A. H. (2013). The evolutionary significance of depression. *Nature Reviews Immunology*, 13(6), 479–492.
- Rosenhan, D. L. (1973). On being sane in insane places. *Science*, 179(4070), 250–258.
- Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Sadler, J. Z. (2005). *Values and psychiatric diagnosis*. Oxford University Press. σ. 45–72, 50–52, 55–60, 65–70, 73–102, 78, 98, 100–102
- Sadler, J. Z. (2005). *Values and psychiatric diagnosis: A critical appraisal*. Oxford University Press.
- Sass, L. (1992). *Madness and modernism: Insanity in the light of modern art, literature, and thought*. Basic Books.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum.
- Schrag, B. (2015). Philosophy and psychiatry. In K. W. M. Fulford et al. (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 3–20). Oxford University Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Spitzer, R. L. (1980). DSM-III: The major innovation in psychiatric classification. *Journal of Nervous and Mental Disease*, 168(11), 689–694.
- Szasz, T. (1960). *The myth of mental illness: Foundations of a theory of personal conduct*. Harper & Row.
- Szasz, T. (1961). *The myth of mental illness*. Harper & Row.
- Szasz, T. (1961). *The myth of mental illness: Foundations of a theory of personal conduct*. Harper & Row.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. (Original work on Turing machines published 1936)
- Tyrer, P., & Steinberg, D. (2013). *Concepts of mental disorder: Philosophical and clinical perspectives*. Routledge.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.

Wakefield, J. C. (1992). The concept of mental disorder. *American Psychologist*, 47(3), 373–388.

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist*, 47(3), 373–388.
<https://doi.org/10.1037/0003-066X.47.3.373>

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.

Zachar, P. (2000). Psychiatric disorders are not natural kinds. *Philosophy, Psychiatry, & Psychology*, 7(3), 167–182.

Zachar, P. (2014). *The concept of mental disorder revisited*. Cambridge University Press.

Zachar, P., & Kendler, K. S. (2007). Psychiatric disorders: A conceptual taxonomy. *American Journal of Psychiatry*, 164(3), 557–565.