



H-BAR

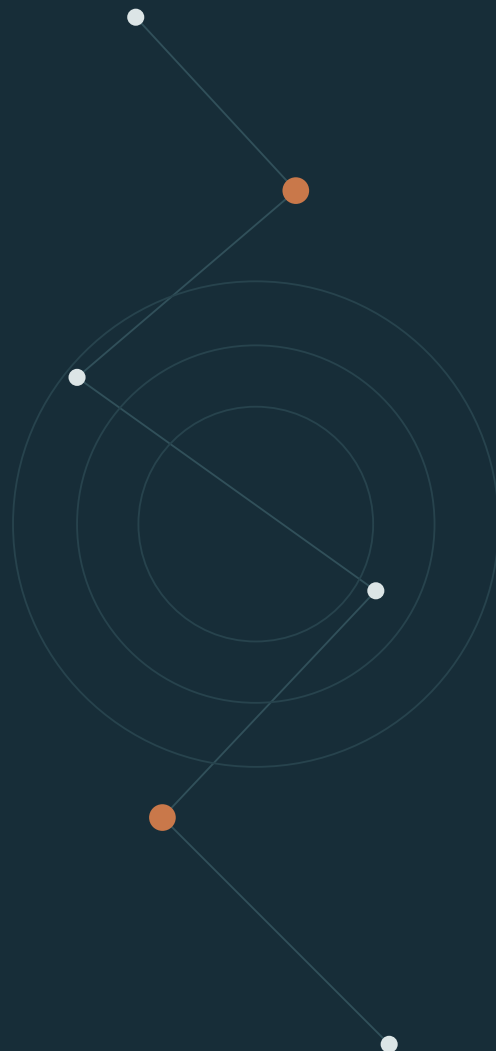
ΕΠΙΣΤΗΜΗ · ΤΕΧΝΟΛΟΓΙΑ · ΦΙΛΟΣΟΦΙΑ

Πληροφορία, Μνήμη και Συνείδηση

Από τον Claude Shannon και τον Alan Turing
στον εγκέφαλο και την τεχνητή νοημοσύνη

ΣΥΝΤΑΚΤΙΚΗ ΟΜΑΔΑ H-BAR

Πλήρης αναλυτική έκδοση · 2026



Πληροφορία, Μνήμη και Συνείδηση

Από τον Claude Shannon και τον Alan Turing
στον εγκέφαλο και την τεχνητή νοημοσύνη

Συντακτική ομάδα h-bar

Στοιχεία έκδοσης

Τίτλος	Πληροφορία, Μνήμη και Συνείδηση
Υπότιτλος	Από τον Claude Shannon και τον Alan Turing στον εγκέφαλο και την τεχνητή νοημοσύνη
Συγγραφέας	Συντακτική ομάδα h-bar
Έκδοση	Πρώτη ψηφιακή έκδοση, Ιούλιος 2026
Ιστότοπος	www.h-bar.gr

© 2026 h-bar. Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αναπαραγωγή, αποθήκευση ή διανομή του έργου, ολικά ή μερικά, χωρίς προηγούμενη γραπτή άδεια, με εξαίρεση σύντομα αποσπάσματα για σκοπούς κριτικής, παρουσίασης ή επιστημονικής αναφοράς, με σαφή αναφορά στην πηγή.

Το βιβλίο αποτελεί έργο επιστημονικής εκλαίκευσης. Οι βιβλιογραφικές αναφορές παρατίθενται στο τέλος για περαιτέρω μελέτη και έλεγχο των επιμέρους ισχυρισμών.

Περιεχόμενα

1. Ένα ποντίκι μέσα σε έναν λαβύρινθο
2. Πότε μια φυσική μεταβολή γίνεται μνήμη;
3. Τι είναι τελικά η πληροφορία;
4. Πώς ένα ηλεκτρικό σήμα γίνεται μνήμη
5. Πού βρίσκεται μια ανάμνηση;
6. Η μνήμη που αλλάζει όταν τη θυμόμαστε
7. Πώς θυμάται μια μηχανή;
8. Τι είναι συνείδηση;
9. Μπορεί μια τεχνητή νοημοσύνη να αποκτήσει συνείδηση;
10. Επίλογος - Από το ίχνος στην εμπειρία
11. Ενδεικτική βιβλιογραφία

1. Ένα ποντίκι μέσα σε έναν λαβύρινθο

Στις αρχές της δεκαετίας του 1950, στους διαδρόμους των Bell Laboratories, ένα μικρό μεταλλικό ποντίκι περιπλανιόταν μέσα σε έναν λαβύρινθο.

Το όνομά του ήταν **Theseus**. Δεν ήταν ζωντανό και δεν περιείχε κάποιον σύγχρονο ηλεκτρονικό υπολογιστή. Κάτω από την επιφάνεια του λαβυρίνθου, ηλεκτρομηχανικά ρελέ και ένας κινούμενος μαγνήτης καθοδηγούσαν το ποντίκι. Δημιουργός του ήταν ο Claude Shannon, ο άνθρωπος που λίγα χρόνια νωρίτερα είχε θεμελιώσει τη μαθηματική θεωρία της πληροφορίας.

Κατά την πρώτη του διαδρομή, το Theseus μπορούσε να κάνει λάθη, να φτάσει σε αδιέξοδα και να αναγκαστεί να αναζητήσει άλλη κατεύθυνση. Όταν όμως τοποθετούνταν ξανά στον ίδιο λαβύρινθο, ακολουθούσε αποτελεσματικότερα τη διαδρομή προς τον στόχο.

Το σύστημα είχε μεταβληθεί από την προηγούμενη εμπειρία του.

Δεν γνώριζε ότι βρισκόταν σε λαβύρινθο. Δεν έβλεπε τοίχους, δεν επιθυμούσε την έξοδο και δεν αισθανόταν ικανοποίηση όταν έφτανε στον στόχο. Ωστόσο, ένα γεγονός του παρελθόντος είχε αφήσει μέσα στη φυσική του δομή ένα ίχνος, και αυτό το ίχνος άλλαζε τη μελλοντική του συμπεριφορά.

Το Theseus ήταν μια μικρή μηχανή. Το ερώτημα που άνοιγε, όμως, ήταν τεράστιο:

Πότε μια φυσική μεταβολή γίνεται μνήμη;

Το 1950, την ίδια περίπου περίοδο, ο Alan Turing δημοσίευσε στο περιοδικό *Mind* το άρθρο «Computing Machinery and Intelligence». Αρχίζει με την ερώτηση «Μπορούν οι μηχανές να σκεφτούν;», αλλά σχεδόν αμέσως αναγνωρίζει ότι οι λέξεις «μηχανή» και «σκέψη» ήταν υπερβολικά ασαφείς για να οδηγήσουν σε καθαρή επιστημονική απάντηση.

Αντί να προσπαθήσει να κοιτάξει στο εσωτερικό μιας μηχανής και να αποφασίσει αν εκεί υπάρχει πραγματική σκέψη, πρότεινε ένα διαφορετικό πρόβλημα: θα μπορούσε μια μηχανή, μέσα από μια γραπτή συνομιλία, να παράγει απαντήσεις που ένας άνθρωπος δεν θα μπορούσε εύκολα να διακρίνει από τις απαντήσεις ενός άλλου ανθρώπου;

Ο Turing δεν είχε κατασκευάσει ένα τεστ συνείδησης. Δεν είχε βρει έναν τρόπο να αποδείξει ότι μια μηχανή αισθάνεται, κατανοεί ή διαθέτει εσωτερική εμπειρία. Είχε μετακινήσει τη συζήτηση από μια δυσπρόσιτη μεταφυσική ερώτηση σε ένα παρατηρήσιμο πρόβλημα συμπεριφοράς.

Η μετατόπιση αυτή αποδείχθηκε εξαιρετικά γόνιμη. Η συμπεριφορά μπορεί να παρατηρηθεί, να συγκριθεί και να μετρηθεί. Η εμπειρία ενός άλλου συστήματος, αντίθετα, δεν είναι άμεσα προσβάσιμη.

Εβδομήντα και πλέον χρόνια αργότερα, οι μηχανές παράγουν κείμενα, μεταφράσεις, εικόνες, κώδικα, επιχειρήματα και ποίηση. Η τεχνητή νοημοσύνη δεν κλόνισε τα θεμέλια της συνείδησης επειδή απέδειξε ότι μια μηχανή διαθέτει εσωτερική εμπειρία. Τα κλόνισε επειδή έδειξε ότι αρκετά από τα εξωτερικά σημάδια που συνδέαμε με τη σκέψη δεν αποτελούν πλέον ασφαλείς αποδείξεις της.

Για να κατανοήσουμε αυτή την αλλαγή, πρέπει να διαχωρίσουμε έννοιες που χρησιμοποιούνται συχνά σαν να ήταν συνώνυμες:

- την πληροφορία,
- τη μνήμη,
- τη μάθηση,
- τη νοημοσύνη,
- και τη συνείδηση.

Ένα σύστημα μπορεί να μεταδίδει πληροφορίες χωρίς να κατανοεί το νόημά τους. Μπορεί να διατηρεί μνήμη χωρίς να διαθέτει αυτοβιογραφικό παρελθόν. Μπορεί να λύνει σύνθετα προβλήματα χωρίς να γνωρίζουμε αν βιώνει οτιδήποτε.

Η διαδρομή αυτού του άρθρου αρχίζει από ένα ρελέ που αλλάζει κατάσταση και καταλήγει στο δυσκολότερο ερώτημα:

Πότε η πληροφορία γίνεται μνήμη - και πότε ένα σύστημα που θυμάται αποκτά έναν κόσμο που εμφανίζεται από μέσα;

2. Πότε μια φυσική μεταβολή γίνεται μνήμη;

Καθώς το Theseus κινούνται στον λαβύρινθο, τα ηλεκτρομηχανικά του ρελέ αλλάζουν κατάσταση. Η μεταβολή αυτή δεν εξαφανίζεται όταν ολοκληρώνεται η διαδρομή. Παρέμενε και καθορίζει τις επόμενες κινήσεις του. Με έναν στοιχειώδη αλλά ουσιαστικό τρόπο, το παρελθόν είχε αποκτήσει δύναμη επάνω στο μέλλον.

Αρκεί όμως ένα φυσικό ίχνος για να μιλήσουμε για μνήμη;

Ένα αποτύπωμα στην άμμο μαρτυρά ότι κάποιος πέρασε. Μια ουλή διατηρεί το ίχνος ενός τραυματισμού. Οι δακτύλιοι ενός δέντρου αποτυπώνουν περιόδους ξηρασίας και ανάπτυξης. Ένα παραμορφωμένο μέταλλο διατηρεί στη δομή του την επίδραση μιας δύναμης.

Δεν θα λέγαμε, ωστόσο, ότι η άμμος θυμάται το βήμα, ότι το δέντρο αναπολεί τη βροχή ή ότι το μέταλλο ανακαλεί το χτύπημα.

Η απλή διατήρηση μιας μεταβολής δεν αρκεί. Για να μιλήσουμε για μνήμη, το ίχνος πρέπει να μπορεί να χρησιμοποιηθεί.

Ένας λειτουργικός ορισμός είναι ο εξής:

Μνήμη είναι μια σχετικά διατηρήσιμη μεταβολή ενός συστήματος, η οποία προκαλείται από μια προηγούμενη εμπειρία ή είσοδο και επηρεάζει τη μελλοντική επεξεργασία, πρόβλεψη ή συμπεριφορά του.

Η λέξη «επηρεάζει» είναι κρίσιμη. Το παρελθόν δεν αρκεί να έχει αφήσει ένα παθητικό σημάδι. Το σημάδι πρέπει να μπορεί να μεταβάλει το τι θα κάνει το σύστημα αργότερα.

Στο Theseus, η κατάσταση των ρελέ άλλαζε τις επόμενες επιλογές μέσα στον λαβύρινθο. Σε έναν υπολογιστή, η κατάσταση ενός ηλεκτρονικού ή μαγνητικού στοιχείου μπορεί αργότερα να αναγνωστεί και να επηρεάσει την εκτέλεση ενός προγράμματος. Στον εγκέφαλο, μια εμπειρία μπορεί να μεταβάλει τη δύναμη των συνάψεων, τη διεγερσιμότητα των νευρώνων και τη δυναμική ολόκληρων δικτύων.

Και στις τρεις περιπτώσεις, κάτι από το παρελθόν περιορίζει ή διευρύνει τις δυνατότητες του μέλλοντος.

Υπάρχει όμως μια σημαντική διαφορά.

Όταν ένας άνθρωπος θυμάται ένα παιδικό δωμάτιο, δεν αλλάζει απλώς η συμπεριφορά του. Εμφανίζεται ξανά, έστω ατελώς, ένας κόσμος που δεν βρίσκεται πλέον εκεί: το χρώμα των τοίχων, η θέση ενός παραθύρου, μια μυρωδιά, ένα συναίσθημα. Και μαζί εμφανίζεται η αίσθηση ότι **εγώ υπήρξα εκεί**.

Το Theseus χρησιμοποιούσε το παρελθόν.

Ο άνθρωπος φαίνεται επίσης να το βιώνει.

2.1 Πληροφορία, μνήμη και συνείδηση

Οι τρεις έννοιες συνδέονται στενά, αλλά δεν είναι συνώνυμες.

Ένα σύστημα μπορεί να μεταφέρει πληροφορία χωρίς να τη διατηρεί. Μπορεί να διατηρεί μνήμη χωρίς να γνωρίζουμε αν διαθέτει συνείδηση. Μπορεί ακόμη να επεξεργάζεται σύνθετες πληροφορίες και να εμφανίζει ευφυή συμπεριφορά χωρίς αυτό να αποδεικνύει ότι βιώνει οτιδήποτε.

Πληροφορία είναι, στην αυστηρή μαθηματική έννοια, η διάκριση ανάμεσα σε δυνατές καταστάσεις και η μείωση της αβεβαιότητας.

Μνήμη είναι η διατήρηση της επίδρασης μιας παλαιότερης κατάστασης, έτσι ώστε να μεταβάλλεται το μέλλον του συστήματος.

Νοημοσύνη είναι η ικανότητα μάθησης, πρόβλεψης, προσαρμογής και ευέλικτης επίλυσης προβλημάτων.

Συνείδηση είναι η δυσκολότερη έννοια: η ύπαρξη μιας εσωτερικής οπτικής, ενός υποκειμενικού βιώματος.

Ένας αισθητήρας μπορεί να ανιχνεύσει ένα μήκος κύματος. Ένα υπολογιστικό σύστημα μπορεί να ταξινομήσει ένα αντικείμενο ως κόκκινο. Ο άνθρωπος όμως δεν επεξεργάζεται απλώς το αντίστοιχο σήμα: βλέπει κόκκινο.

Γιατί η μία φυσική διαδικασία συνοδεύεται από εμπειρία ενώ μια άλλη, ενδεχομένως, όχι;

Δεν γνωρίζουμε.

2.2 Η συνείδηση και το πρόβλημα των άλλων νόων

Η πληροφορία μπορεί να περιγραφεί από έναν εξωτερικό παρατηρητή. Η μνήμη μπορεί να εντοπιστεί μέσα από διατηρήσιμες φυσικές μεταβολές και αλλαγές στη συμπεριφορά. Η συνείδηση, όμως, είναι άμεσα προσβάσιμη μόνο από τη σκοπιά του ίδιου του υποκειμένου.

Γνωρίζω με βεβαιότητα ότι ο δικός μου πόνος υπάρχει επειδή τον βιώνω. Για τον πόνο ενός άλλου ανθρώπου βασίζομαι στη συμπεριφορά του, στα λόγια του, στη φυσιολογία του και στην ομοιότητα του νευρικού του συστήματος με το δικό μου.

Αυτό είναι το παλιό φιλοσοφικό **πρόβλημα των άλλων νόων**.

Δεν παρατηρούμε άμεσα τη συνείδηση των άλλων. Τη συμπεραίνουμε.

Με την εμφάνιση της τεχνητής νοημοσύνης, το πρόβλημα αποκτά μια νέα μορφή. Μια μηχανή μπορεί να χρησιμοποιεί τη γλώσσα του πόνου, της μνήμης και του εαυτού. Αλλά η ομοιότητα της γλώσσας δεν συνοδεύεται από την ίδια βιολογική, σωματική και αναπτυξιακή συνέχεια που στηρίζει την πίστη μας στη συνείδηση των άλλων ανθρώπων.

2.3 Ο Turing και η μετατόπιση προς τη συμπεριφορά

Ο Turing δεν επιχείρησε να λύσει το πρόβλημα των άλλων νόων. Το παρέκαμψε επιχειρησιακά.

Αντί να ρωτήσει τι υπάρχει στο εσωτερικό της μηχανής, ρώτησε τι είναι ικανή να κάνει. Το παιχνίδι μίμησης αξιολογεί αν η γλωσσική συμπεριφορά μιας μηχανής μπορεί να ανταποκριθεί σε ένα ανθρώπινο πρότυπο ευφυούς συνομιλίας.

Η κίνηση ήταν επιστημονικά ισχυρή, επειδή η συμπεριφορά μπορεί να μετρηθεί. Έχει όμως και ένα όριο:

Η ικανότητα ενός συστήματος να συμπεριφέρεται σαν να καταλαβαίνει δεν αποδεικνύει ότι υπάρχει κάποιος που καταλαβαίνει.

Η μέθοδος μπορεί να αξιολογήσει την επίδοση. Δεν μπορεί από μόνη της να γεφυρώσει την απόσταση ανάμεσα στην επίδοση και στο βίωμα.

2.4 Όταν η μηχανή απέκτησε τον λόγο

Για πολλούς αιώνες, η σύνθετη γλώσσα θεωρούνταν ένα από τα ισχυρότερα σημάδια εσωτερικής σκέψης. Ένα ον που μπορούσε να επιχειρηματολογήσει, να δημιουργήσει μεταφορές, να γράψει ποίηση και να μιλήσει για αφηρημένες έννοιες φαινόταν αυτονόητο ότι διέθετε νου.

Σήμερα, υπολογιστικά συστήματα μπορούν να παράγουν κείμενα, να μεταφράζουν, να γράφουν κώδικα, να αναλύουν επιχειρήματα και να μιμούνται διαφορετικά λογοτεχνικά ύφη.

Αυτό δεν αποδεικνύει ότι οι μηχανές διαθέτουν συνείδηση.

Αποδεικνύει όμως ότι η εξωτερική μορφή πολλών ικανοτήτων που συνδέαμε με τον ανθρώπινο νου μπορεί να παραχθεί από υπολογιστικές διαδικασίες.

Ένα σύστημα μπορεί να γράψει για τη θλίψη χωρίς αυτό να σημαίνει ότι θλίβεται. Μπορεί να περιγράψει ένα χρώμα χωρίς να γνωρίζουμε αν βλέπει. Μπορεί να μιλήσει για τον φόβο του θανάτου χωρίς να υπάρχει μέσα του ένα ον που φοβάται την παύση της ύπαρξής του.

Η τεχνητή νοημοσύνη δεν κλόνισε τα θεμέλια της συνείδησης επειδή απέδειξε ότι η μηχανή αισθάνεται. Τα κλόνισε επειδή έδειξε ότι αρκετά από τα εξωτερικά σημάδια που χρησιμοποιούσαμε για να αναγνωρίζουμε τη σκέψη δεν είναι πλέον ασφαλείς αποδείξεις ενός εσωτερικού βιώματος.

Το ερώτημα που θα ακολουθήσει ολόκληρο το άρθρο είναι επομένως:

Πότε η πληροφορία γίνεται μνήμη και πότε η μνήμη γίνεται συνειδητή εμπειρία;

3. Τι είναι τελικά η πληροφορία;

Η λέξη «πληροφορία» μοιάζει οικεία. Τη χρησιμοποιούμε καθημερινά για μια είδηση, μια επισημονική μέτρηση, μια φωτογραφία ή μια πρόταση που μας αποκαλύπτει κάτι για τον κόσμο.

Στη συνηθισμένη της χρήση, η πληροφορία συνδέεται με το νόημα. Κάτι μας πληροφορεί όταν αναφέρεται σε κάτι άλλο, όταν είναι αληθινό ή ψευδές, όταν μεταβάλλει όσα γνωρίζουμε.

Στη μαθηματική θεωρία της επικοινωνίας, όμως, η πληροφορία απέκτησε έναν πολύ ασυτηρότερο και ταυτόχρονα πιο περιορισμένο ορισμό.

Όταν ο Claude Shannon δημοσίευσε το 1948 τη *Μαθηματική Θεωρία της Επικοινωνίας*, δεν επιχείρησε να εξηγήσει τι σημαίνει ένα μήνυμα. Το πρόβλημά του ήταν τεχνικό: πώς μπορούν τα σύμβολα να μεταδοθούν αξιόπιστα μέσα από ένα κανάλι, πόσο μπορούν να συμπιεστούν και πόσο θόρυβο μπορεί να αντέξει η μετάδοσή τους.

Για τη θεωρία του Shannon, η πληροφορία συνδέεται με την αβεβαιότητα.

Αν ένα γεγονός είναι απολύτως βέβαιο, η πραγματοποίησή του δεν μας προσφέρει νέα πληροφορία. Αν γνωρίζουμε ότι ένα νόμισμα έχει κορόνα και στις δύο πλευρές του, η ανακοίνωση του αποτελέσματος μιας ρίψης δεν αλλάζει τίποτε σε όσα περιμέναμε.

Αντίθετα, όσο λιγότερο πιθανό είναι ένα γεγονός, τόσο περισσότερη πληροφορία αποκτούμε όταν το παρατηρούμε.

$$I(x) = -\log_2 p(x)$$

Η χρήση του λογαρίθμου με βάση το δύο σημαίνει ότι η πληροφορία μετριέται σε **bits**.

Ένα αποτέλεσμα με πιθανότητα 1/2 μεταφέρει ένα bit πληροφορίας. Ένα αποτέλεσμα με πιθανότητα 1/8 μεταφέρει τρία bits, επειδή χρειάζονται τρεις δυαδικές επιλογές για να διακριθεί ανάμεσα σε οκτώ ισοπίθανες δυνατότητες.

Η πληροφορία, επομένως, δεν είναι μια ουσία που βρίσκεται μέσα στο μήνυμα. Είναι η μεταβολή που προκαλεί η παρατήρηση στο σύνολο των δυνατοτήτων που θεωρούσαμε πιθανές.

Πληροφορία είναι αυτό που μαθαίνουμε όταν η πραγματικότητα επιλέγει μία δυνατότητα ανάμεσα σε πολλές.

3.1 Η εντροπία του Shannon

Ένα μεμονωμένο αποτέλεσμα μπορεί να είναι περισσότερο ή λιγότερο πληροφοριακό. Μια πηγή, όμως, μπορεί να παράγει πολλά διαφορετικά αποτελέσματα με διαφορετικές πιθανότητες.

Για να μετρήσει τη μέση αβεβαιότητα μιας τέτοιας πηγής, ο Shannon χρησιμοποίησε την ποσότητα:

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

Η ποσότητα αυτή ονομάστηκε **εντροπία**.

Ένα δίκαιο νόμισμα έχει μέγιστη εντροπία ενός bit ανά ρίψη. Πριν από κάθε ρίψη, η κορόνα και τα γράμματα είναι εξίσου πιθανά.

Ένα έντονα μεροληπτικό νόμισμα έχει μικρότερη εντροπία. Αν φέρνει κορόνα στο 99% των περιπτώσεων, μπορούμε συνήθως να προβλέψουμε σωστά το αποτέλεσμα ακόμη και χωρίς να παρατηρήσουμε τη ρίψη.

Η υψηλή εντροπία σημαίνει μεγάλη απροσδιοριστία και χαμηλή προβλεψιμότητα. Η χαμηλή εντροπία σημαίνει ότι η πηγή εμφανίζει περισσότερη κανονικότητα και μπορεί να περιγραφεί με λιγότερη πληροφορία.

Αυτό αποτελεί και τη βάση της συμπίεσης δεδομένων. Αν ένα σύμβολο εμφανίζεται πολύ συχνά, μπορούμε να το κωδικοποιήσουμε με μικρότερη ακολουθία bits. Τα σπάνια σύμβολα χρειάζονται μεγαλύτερες κωδικοποιήσεις.

Εδώ όμως εμφανίζεται μια πρώτη αντίφαση με την καθημερινή έννοια της πληροφορίας.

Μια εντελώς τυχαία ακολουθία αριθμών έχει υψηλή εντροπία. Είναι δύσκολο να προβλεφθεί και σχεδόν αδύνατο να συμπιεστεί. Δεν σημαίνει όμως τίποτε.

Αντίθετα, μια απλή πρόταση όπως «το παιδί φοβήθηκε τη θύελλα» μπορεί να μεταφέρει νόημα, εικόνες και συναίσθημα, παρότι η γλωσσική της δομή είναι σε μεγάλο βαθμό προβλέψιμη.

Η εντροπία του Shannon μετρά την απροσδιοριστία των συμβόλων. Δεν μετρά τη σημασία τους.

3.2 Πληροφορία και νόημα

Ας υποθέσουμε ότι δύο συστήματα λαμβάνουν το ίδιο μήνυμα:

«Υπάρχει φωτιά στο κτίριο».

Για έναν άνθρωπο που βρίσκεται μέσα στο κτίριο, η πρόταση μπορεί να προκαλέσει φόβο, να ανακαλέσει προηγούμενες εμπειρίες, να ενεργοποιήσει ένα σχέδιο διαφυγής και να αλλάξει ριζικά τη συμπεριφορά του.

Για έναν υπολογιστή που δεν διαθέτει κανέναν μηχανισμό σύνδεσης των λέξεων με τον φυσικό κόσμο, το ίδιο μήνυμα μπορεί να είναι απλώς μια ακολουθία χαρακτήρων.

Από την πλευρά της θεωρίας του Shannon, το μήνυμα μπορεί να μεταφέρει την ίδια ποσότητα πληροφορίας και στις δύο περιπτώσεις. Από την πλευρά του νοήματος, όμως, η διαφορά είναι τεράστια.

Αυτό οδηγεί στη διάκριση ανάμεσα στη **συντακτική** και στη **σημασιολογική πληροφορία**.

Η συντακτική πληροφορία αφορά τις πιθανότητες των συμβόλων, τη δομή των ακολουθιών, την κωδικοποίηση, τη μετάδοση και τον θόρυβο.

Η σημασιολογική πληροφορία αφορά σε τι αναφέρεται ένα μήνυμα, αν είναι αληθινό, τι σημαίνει για εκείνον που το λαμβάνει και πώς συνδέεται με τον κόσμο και με προηγούμενες γνώσεις.

Η μαθηματική θεωρία της επικοινωνίας έθεσε σκόπιμα το νόημα εκτός του βασικού της προβλήματος. Αυτό την έκανε εξαιρετικά ισχυρή ως τεχνική θεωρία, αλλά δεν την καθιστά πλήρη θεωρία της ανθρώπινης γνώσης.

Η διάκριση θα γίνει καθοριστική για την τεχνητή νοημοσύνη. Ένα μεγάλο γλωσσικό μοντέλο μπορεί να μάθει τις πιθανότητες των συμβόλων και να αναπτύξει εξαιρετικά σύνθετες εσωτερικές αναπαραστάσεις. Αλλά το γεγονός ότι χειρίζεται με επιτυχία τη δομή της γλώσσας δεν λύνει από μόνο του το ερώτημα αν τα σύμβολα σημαίνουν για το σύστημα κάτι με τον τρόπο που σημαίνουν για έναν άνθρωπο.

3.3 Γιατί ο Shannon χρησιμοποίησε τη λέξη «εντροπία»;

Η λέξη εντροπία προερχόταν από τη θερμοδυναμική και τη στατιστική φυσική.

Στη θερμοδυναμική, η εντροπία συνδέεται με τον αριθμό των μικροσκοπικών καταστάσεων που μπορούν να αντιστοιχούν στην ίδια μακροσκοπική κατάσταση ενός συστήματος.

Ένα δοχείο αερίου μπορεί να περιγραφεί μακροσκοπικά από την πίεση, τον όγκο και τη θερμοκρασία του. Η ίδια μακροσκοπική κατάσταση, όμως, μπορεί να προκύπτει από ασύλληπτα πολλούς διαφορετικούς συνδυασμούς θέσεων και ταχυτήτων των μορίων.

Για ισοπίθανες μικροκαταστάσεις, η σχέση του Boltzmann γράφεται:

$$S = k_B \ln \Omega$$

όπου S είναι η θερμοδυναμική εντροπία, k_B η σταθερά του Boltzmann και Ω ο αριθμός των μικροκαταστάσεων που είναι συμβατές με τη μακροσκοπική περιγραφή.

Η ομοιότητα με την εντροπία του Shannon είναι βαθιά. Και στις δύο περιπτώσεις εμφανίζεται η αβεβαιότητα σχετικά με το ποια από πολλές δυνατές καταστάσεις έχει πραγματοποιηθεί.

Δεν πρέπει όμως να ταυτίσουμε απλώς τις δύο ποσότητες. Η θερμοδυναμική εντροπία είναι φυσική ποσότητα, με μονάδες ενέργειας ανά θερμοκρασία. Η εντροπία Shannon είναι πληροφοριακή ποσότητα που μετριέται συνήθως σε bits.

Η μαθηματική τους συγγένεια είναι ουσιαστική, αλλά η φυσική τους ερμηνεία εξαρτάται από το σύστημα και το πλαίσιο.

3.4 Ο δαίμονας του Maxwell

Το 1867 ο James Clerk Maxwell φαντάστηκε ένα μικροσκοπικό ον που παρατηρούσε τα μόρια ενός αερίου σε ένα δοχείο χωρισμένο σε δύο θαλάμους. Ανάμεσά τους υπήρχε μια μικρή πόρτα.

Το υποθετικό ον άφηνε τα γρήγορα μόρια να περάσουν προς τη μία πλευρά και τα αργά προς την άλλη. Ύστερα από λίγο, ο ένας θάλαμος θα γινόταν θερμότερος και ο άλλος ψυχρότερος.

Το ον φαινόταν να δημιουργεί τάξη από την αταξία και να μειώνει την εντροπία χωρίς να καταναλώνει έργο, παραβιάζοντας φαινομενικά τον δεύτερο νόμο της θερμοδυναμικής.

Η ουσία του πειράματος ήταν ο ρόλος της γνώσης. Το ον μπορούσε να μειώσει την εντροπία του αερίου επειδή γνώριζε κάτι για κάθε μόριο: αν ήταν γρήγορο ή αργό και προς ποια κατεύθυνση κινούνταν.

Η πληροφορία φαινόταν να μπορεί να μετατραπεί σε φυσικό έργο.

Ο Leo Szilard απλοποίησε αργότερα το πρόβλημα σε μια μηχανή που περιείχε ένα μόνο μόριο. Για να εξαχθεί έργο, ένας παρατηρητής έπρεπε να γνωρίζει σε ποια πλευρά του δοχείου βρισκόταν το μόριο και να διατηρεί αυτή την πληροφορία ώστε να επιλέξει την κατάλληλη ενέργεια.

Το πρόβλημα δεν ήταν μόνο η μέτρηση. Ο δαίμονας έπρεπε να καταγράφει τα αποτελέσματα και κάποια στιγμή να επαναφέρει τη μνήμη του σε αρχική κατάσταση, ώστε να την ξαναχρησιμοποιήσει.

3.5 Η πληροφορία είναι φυσική

Το 1961 ο Rolf Landauer εξέτασε τη φυσική βάση των υπολογιστικών πράξεων.

Αν ένα στοιχείο μνήμης μπορεί αρχικά να βρίσκεται είτε στην κατάσταση 0 είτε στην κατάσταση 1 και το επαναφέρουμε πάντοτε στην κατάσταση 0, η τελική κατάσταση δεν μας επιτρέπει πλέον να γνωρίζουμε ποια ήταν η αρχική. Δύο διαφορετικά ενδεχόμενα συγχωνεύονται σε ένα.

Η πληροφορία για την προηγούμενη κατάσταση έχει διαγραφεί.

Ο Landauer έδειξε ότι η διαγραφή ενός ιδανικού, ισοπίθανου bit σε περιβάλλον θερμοκρασίας T απαιτεί κατ' ελάχιστο την απόρριψη θερμότητας:

$$Q_{\min} = k_B T \ln 2$$

Το όριο είναι εξαιρετικά μικρό σε θερμοκρασία δωματίου. Οι πραγματικοί υπολογιστές καταναλώνουν πολύ περισσότερη ενέργεια ανά λογική πράξη. Η σημασία του, όμως, είναι θεμελιώδης.

Η διαγραφή της πληροφορίας δεν αποτελεί μια καθαρά αφηρημένη μαθηματική ενέργεια. Πραγματοποιείται από ένα φυσικό σύστημα και έχει θερμοδυναμικές συνέπειες.

Αυτό δεν σημαίνει ότι κάθε υπολογιστική πράξη πρέπει αναγκαστικά να καταναλώνει $k_B T \ln 2$. Λογικά αντιστρεπτοί υπολογισμοί μπορούν καταρχήν να εκτελεστούν με αυθαίρετα μικρή απώλεια, αν πραγματοποιούνται αρκετά αργά και ιδανικά. Η αναγκαστική απώλεια συνδέεται με τη λογική μη αντιστρεπτότητα και τη συγχώνευση διαφορετικών προηγούμενων καταστάσεων σε μία.

Η πληροφορία δεν αιωρείται έξω από τον φυσικό κόσμο.

Για να υπάρξει ένα bit, πρέπει να υπάρχουν τουλάχιστον δύο φυσικά διακριτές καταστάσεις: δύο επίπεδα τάσης, δύο κατευθύνσεις μαγνήτισης, δύο θέσεις ενός σωματιδίου ή δύο καταστάσεις ενός κβαντικού συστήματος.

Για να διατηρηθεί, οι καταστάσεις πρέπει να είναι αρκετά σταθερές ώστε ο θερμικός θόρυβος να μην τις αναμειγνύει αμέσως. Για να διαβαστεί ή να μεταβληθεί, πρέπει να υπάρξει φυσική αλληλεπίδραση.

Κάθε πληροφορία χρειάζεται ένα υπόστρωμα.

3.6 Από το bit στη βιολογική μνήμη

Μια ανάμνηση δεν είναι ένα άυλο περιεχόμενο που αιωρείται μέσα στον νου. Για να θυμηθεί ο εγκέφαλος κάτι αύριο, η σημερινή εμπειρία πρέπει να μεταβάλει με σχετικά σταθερό τρόπο κάποιο μέρος της δομής ή της λειτουργίας του.

Μπορεί να αλλάξει η αποτελεσματικότητα μιας σύναψης, ο αριθμός των υποδοχέων στη μεμβράνη, η μορφή μιας δενδριτικής άκανθας, η πιθανότητα πυροδότησης ενός νευρώνα, η γονιδιακή έκφραση ή η σύνδεση ανάμεσα σε διαφορετικά νευρωνικά σύνολα.

Όπως το bit απαιτεί δύο φυσικά διακριτές καταστάσεις, έτσι και η μάθηση απαιτεί ο εγκέφαλος μετά την εμπειρία να βρίσκεται σε διαφορετική κατάσταση από εκείνη στην οποία βρισκόταν πριν.

Η βιολογική μνήμη, όμως, δεν είναι συνήθως δυαδική εγγραφή. Οι αναμνήσεις κατανέμονται σε τεράστια δίκτυα, κωδικοποιούνται σε βαθμούς συναπτικής ισχύος, σε χρονικά μοτίβα δραστηριότητας και σε σχέσεις μεταξύ πληθυσμών νευρώνων.

Ο εγκέφαλος είναι επίσης ανοικτό σύστημα, μακριά από τη θερμοδυναμική ισορροπία. Καταναλώνει συνεχώς ενέργεια για να διατηρεί διαφορές συγκέντρωσης ιόντων, ηλεκτρικά δυναμικά, μοριακές δομές και συναπτικές συνδέσεις.

Η τάξη του δεν παραβιάζει τον δεύτερο νόμο. Διατηρείται επειδή η συνολική εντροπία του εγκεφάλου και του περιβάλλοντος εξακολουθεί να αυξάνεται.

Η μνήμη έχει ενεργειακό κόστος, όχι μόνο επειδή η εγκεφαλική δραστηριότητα καταναλώνει ενέργεια, αλλά επειδή η διατήρηση ενός οργανωμένου, μεταβαλλόμενου δικτύου μακριά από την ισορροπία απαιτεί συνεχή μεταβολική εργασία.

3.7 Είναι ο εγκέφαλος μια μηχανή μείωσης της εντροπίας;

Θα ήταν δειλεαστικό να πούμε ότι ο εγκέφαλος μετατρέπει το χάος των αισθήσεων σε τάξη και γνώση.

Η πραγματικότητα είναι πιο σύνθετη. Ο εγκέφαλος δεν μειώνει αναγκαστικά τη φυσική εντροπία του περιβάλλοντος. Μειώνει κυρίως την **αβεβαιότητά του** σχετικά με ορισμένες πτυχές του.

Όταν ακούμε έναν άγνωστο ήχο, υπάρχουν πολλές πιθανές αιτίες. Καθώς συλλέγουμε περισσότερα αισθητηριακά δεδομένα, περιορίζουμε τις δυνατότητες. Η νέα πληροφορία μεταβάλλει τις πιθανότητες των εσωτερικών μας υποθέσεων.

Η μάθηση οργανώνει αυτές τις μεταβολές ώστε ο εγκέφαλος να προβλέπει καλύτερα τι θα συμβεί στη συνέχεια. Η γνώση μπορεί να θεωρηθεί περιορισμός της αβεβαιότητας και η μνήμη διατηρεί αυτόν τον περιορισμό στον χρόνο.

Αλλά ούτε η πληροφορία ούτε η μνήμη εξηγούν ακόμη το νόημα. Ένα σύστημα μπορεί να μειώνει το σφάλμα των προβλέψεών του χωρίς να γνωρίζουμε αν κατανοεί αυτό που προβλέπει.

Η εντροπία εξηγεί πώς μετράμε την αβεβαιότητα. Η θεωρία πληροφορίας εξηγεί πώς κωδικοποιούνται και μεταδίδονται διακρίσεις. Η θερμοδυναμική εξηγεί γιατί κάθε φυσική μνήμη έχει κόστος.

Απομένει να καταλάβουμε πώς ένα παροδικό ηλεκτρικό σήμα μπορεί να αφήσει μια αλλαγή που διαρκεί χρόνια.

4. Πώς ένα ηλεκτρικό σήμα γίνεται μνήμη

Κάθε εμπειρία αρχίζει ως κάτι παροδικό.

Το φως που φτάνει στον αμφιβληστροειδή, η πίεση ενός αγγίγματος στο δέρμα, μια φωνή ή μια μυρωδιά μετατρέπονται σε ηλεκτροχημικά σήματα. Τα σήματα διαδίδονται μέσα σε νευρωνικά δίκτυα, μεταβάλλονται, συνδυάζονται και, στις περισσότερες περιπτώσεις, εξαφανίζονται χωρίς να αφήσουν κάτι που αργότερα θα μπορούσαμε να ανακαλέσουμε.

Μόνο ένα μικρό μέρος όσων συμβαίνουν γίνεται μνήμη.

Πώς ένα ηλεκτρικό γεγονός που διαρκεί χιλιοστά του δευτερολέπτου μετατρέπεται σε αλλαγή που μπορεί να επιβιώσει για ημέρες, χρόνια ή ακόμη και για ολόκληρη τη ζωή;

Η απάντηση αρχίζει σε ένα σημείο εξαιρετικά μικρό: στη σύναψη.

4.1 Η σύναψη δεν είναι απλώς καλώδιο

Οι νευρώνες δεν συνδέονται συνήθως όπως δύο μεταλλικά καλώδια. Ανάμεσά τους υπάρχει ένα μικροσκοπικό κενό, το συναπτικό χάσμα.

Όταν ένα δυναμικό ενέργειας φτάσει στην απόληξη του προσυναπτικού νευρώνα, προκαλεί το άνοιγμα διαύλων ασβεστίου. Η είσοδος ασβεστίου οδηγεί στην απελευθέρωση νευροδιαβιβαστών, οι οποίοι διασχίζουν το συναπτικό χάσμα και προσδένονται σε υποδοχείς του επόμενου κυττάρου.

Στις περισσότερες διεγερτικές συνάψεις του εγκεφάλου, ο βασικός νευροδιαβιβαστής είναι το γλουταμινικό. Δύο από τους σημαντικότερους υποδοχείς του είναι οι AMPA και NMDA.

Οι υποδοχείς AMPA ανοίγουν γρήγορα όταν προσδεθεί επάνω τους γλουταμινικό. Επιτρέπουν την είσοδο θετικών ιόντων και προκαλούν μια τοπική αποπόλωση της μεμβράνης.

Οι NMDA είναι πιο ιδιαίτεροι. Δεν αρκεί να έχει απελευθερωθεί γλουταμινικό. Σε κατάσταση ηρεμίας, ο διάυλός τους φράσσεται από ένα ιόν μαγνησίου. Για να απομακρυνθεί αυτό το εμπόδιο, πρέπει ταυτόχρονα να έχει αποπολωθεί αρκετά η μετασυναπτική μεμβράνη.

Ο υποδοχέας NMDA ανταποκρίνεται επομένως σε δύο συνθήκες:

- ο προσυναπτικός νευρώνας απελευθέρωσε γλουταμινικό,
- ο μετασυναπτικός νευρώνας ήταν ήδη αρκετά ενεργός.

Λειτουργεί σαν μοριακός **ανιχνευτής σύμπτωσης**. Δεν ανιχνεύει μόνο ότι συνέβη ένα γεγονός. Ανιχνεύει ότι δύο γεγονότα συνέβησαν μαζί.

Εδώ εμφανίζεται σε βιολογική μορφή η αρχή που ο ψυχολόγος Donald Hebb διατύπωσε το 1949: όταν η δραστηριότητα ενός κυττάρου συμβάλλει επανειλημμένα στην ενεργοποίηση ενός άλλου, η μεταξύ τους σύνδεση τείνει να ενισχύεται.

Η ιδέα συνοψίζεται συχνά στη φράση:

Νευρώνες που ενεργοποιούνται μαζί, συνδέονται ισχυρότερα.

Η πραγματική βιολογία είναι πιο περίπλοκη. Η ακριβής χρονική ακολουθία, η συχνότητα της ενεργοποίησης, η κατάσταση του κυττάρου και η παρουσία νευροδιαμορφωτών μπορούν να καθορίσουν αν μια σύναψη θα ενισχυθεί, θα αποδυναμωθεί ή δεν θα αλλάξει.

Η βασική αρχή, όμως, παραμένει: η συσχετισμένη δραστηριότητα μπορεί να αλλάξει τη φυσική αποτελεσματικότητα της σύνδεσης.

4.2 Το ασβέστιο ως μήνυμα αλλαγής

Όταν ανοίξει ο υποδοχέας NMDA, επιτρέπει την είσοδο ασβεστίου στο εσωτερικό του νευρώνα.

Το ασβέστιο δεν λειτουργεί μόνο ως ηλεκτρικό φορτίο. Αποτελεί ενδοκυτταρικό μήνυμα. Η συγκέντρωση, η διάρκεια και η χωρική κατανομή του σήματος ενεργοποιούν διαφορετικά μοριακά μονοπάτια.

Ένα ισχυρό και σύντομο κύμα μπορεί να ευνοήσει την ενίσχυση μιας σύναψης, ενώ διαφορετικά μοτίβα ασβεστίου μπορούν να οδηγήσουν σε αποδυνάμωσή της.

Ένα από τα βασικά μόρια που ενεργοποιούνται είναι η **CaMKII**, η εξαρτώμενη από ασβέστιο και καλμοδουλίνη πρωτεϊνική κινάση II.

Η CaMKII υπάρχει σε μεγάλη συγκέντρωση στις διεγερτικές συνάψεις. Όταν αυξηθεί το ασβέστιο, συνδέεται με την καλμοδουλίνη και μεταβάλλει την κατάσταση της CaMKII. Η πρωτεΐνη μπορεί τότε να αλληλεπιδράσει με υποδοχείς NMDA, με στοιχεία της μετασυναπτικής πυκνότητας και με μόρια που ρυθμίζουν τους AMPA υποδοχείς.

Για μεγάλο διάστημα η CaMKII παρουσιάζοταν σχεδόν ως ένας μοριακός διακόπτης μνήμης: ένα μόριο που μπορούσε να παραμένει ενεργοποιημένο μετά το αρχικό σήμα και να διατηρεί την ενίσχυση της σύναψης.

Η εικόνα σήμερα είναι πιο σύνθετη. Δεν υπάρχει ένα μοναδικό «μόριο μνήμης». Υπάρχει ένα δίκτυο μορίων που αλλάζουν λειτουργία, θέση και αλληλεπιδράσεις ανάλογα με τη δραστηριότητα της σύναψης.

4.3 Περισσότεροι AMPA, ισχυρότερη σύνδεση

Ένας βασικός τρόπος με τον οποίο ενισχύεται μια διεγερτική σύναψη είναι η αύξηση του αριθμού ή της αποτελεσματικότητας των AMPA υποδοχέων στη μετασυναπτική μεμβράνη.

Πριν από την αλλαγή, μια δεδομένη ποσότητα γλουταμινικού προκαλεί συγκεκριμένη ηλεκτρική απόκριση. Μετά την ενίσχυση, η ίδια ποσότητα μπορεί να προκαλέσει μεγαλύτερη απόκριση, επειδή περισσότεροι AMPA υποδοχείς βρίσκονται στη σωστή θέση ή επειδή έχει μεταβληθεί η λειτουργία τους.

Οι υποδοχείς δεν είναι στατικοί. Μετακινούνται μέσα στη μεμβράνη, εισάγονται από ενδοκυτταρικά κυστίδια και απομακρύνονται μέσω ενδοκύττωσης. Κατά τη μακροχρόνια ενδυνάμωση, η CaMKII και άλλοι μηχανισμοί διευκολύνουν τη μεταφορά και τη σταθεροποίηση AMPA υποδοχέων στις ενεργές συνάψεις.

Η πληροφορία έχει πλέον αφήσει ένα φυσικό ίχνος.

Η σύναψη μετά την εμπειρία δεν είναι ίδια με τη σύναψη πριν από αυτήν.

Αυτό δεν σημαίνει ότι αποθηκεύθηκε κάπου μια εικόνα ή μια ολόκληρη ανάμνηση. Σημαίνει ότι μεταβλήθηκε η πιθανότητα με την οποία ένα συγκεκριμένο μοτίβο δραστηριότητας θα προκαλέσει ξανά ένα άλλο μοτίβο.

Η μνήμη στο επίπεδο της σύναψης δεν είναι αντίγραφο του γεγονότος. Είναι αλλαγή στους κανόνες με τους οποίους το δίκτυο θα αντιδράσει στο μέλλον.

4.4 Η μακροχρόνια ενδυνάμωση

Το 1973 οι Terje Lømo και Tim Bliss δημοσίευσαν πειράματα στα οποία διέγειραν επανειλημμένα μια νευρωνική οδό που οδηγεί στην οδοντωτή έλικα του ιππόκαμπου.

Έπειτα από μια σύντομη περίοδο ισχυρής διέγερσης, η απόκριση των μετασυναπτικών νευρώνων παρέμενε αυξημένη για μεγάλο χρονικό διάστημα. Η ίδια είσοδος προκαλούσε πλέον ισχυρότερη έξοδο. Το φαινόμενο ονομάστηκε **μακροχρόνια ενδυνάμωση**, ή LTP.

Η ανακάλυψη ήταν σημαντική επειδή το φαινόμενο διέθετε ιδιότητες που θα περίμενε κανείς από έναν κυτταρικό μηχανισμό μάθησης:

- μπορούσε να διαρκέσει,
- ήταν ισχυρότερο όταν η ενεργοποίηση ήταν συσχετισμένη,
- μπορούσε να αφορά επιλεκτικά τις συνάψεις που είχαν ενεργοποιηθεί,
- άλλαζε τον τρόπο με τον οποίο το δίκτυο ανταποκρινόταν σε μελλοντικά ερεθίσματα.

Το 1986 ο Richard Morris και οι συνεργάτες του έδειξαν ότι ο αποκλεισμός των NMDA υποδοχέων εμποδίζει την επαγωγή LTP στον ιππόκαμπο και παράλληλα προκαλούσε επιλεκτικό έλλειμμα στη χωρική μάθηση ποντικών.

Η LTP έγινε ένα από τα σημαντικότερα πειραματικά μοντέλα για τη μελέτη της μνήμης.

Δεν είναι όμως η ίδια η μνήμη.

4.5 Γιατί η LTP δεν αρκεί

Θα ήταν δελεαστικό να πούμε ότι κάθε ανάμνηση είναι απλώς ένα σύνολο ενισχυμένων συνάψεων. Αυτό θα ήταν υπερβολικά απλό.

Δεν βασίζονται όλες οι μορφές μάθησης στον ίδιο τύπο πλαστικότητας. Η ενίσχυση δεν είναι η μόνη χρήσιμη μεταβολή. Οι συνάψεις μπορούν επίσης να αποδυναμωθούν μέσω διαδικασιών όπως η μακροχρόνια αποδυνάμωση, ή LTD.

Η μάθηση απαιτεί τόσο την ενίσχυση επιλεγμένων σχέσεων όσο και την καταστολή άλλων.

Μια ανάμνηση δεν είναι απλώς η συνολική δύναμη ενός δικτύου. Είναι το συγκεκριμένο μοτίβο των αλλαγών: ποιοι νευρώνες συνδέθηκαν, σε ποιον χρόνο, υπό ποιες συνθήκες και με ποια συναισθηματική ή συμπεριφορική σημασία.

Η σχέση μπορεί να διατυπωθεί προσεκτικότερα:

Η LTP δεν είναι η μνήμη, αλλά αποτελεί έναν από τους βασικούς μηχανισμούς με τους οποίους η εμπειρία μπορεί να αλλάξει τη μελλοντική λειτουργία ενός νευρωνικού δικτύου.

4.6 Από μια παροδική αλλαγή σε μακροχρόνιο ίχνος

Οι πρωτεΐνες μεταβολίζονται, οι μεμβράνες ανανεώνονται και τα μόρια μετακινούνται. Πώς μπορεί μια αλλαγή να διαρκέσει για χρόνια, όταν τα επιμέρους συστατικά της σύναψης αντικαθίστανται συνεχώς;

Η **πρώιμη LTP** μπορεί να διαρκεί λεπτά ή ώρες και βασίζεται σε μεγάλο βαθμό στην τροποποίηση προϋπαρχουσών πρωτεϊνών, στη μετακίνηση υποδοχέων και σε αλλαγές της συναπτικής λειτουργίας.

Η **ύστερη LTP**, η οποία μπορεί να διατηρηθεί περισσότερο, συνδέεται με νέα γονιδιακή έκφραση, σύνθεση πρωτεϊνών και σταθερότερες δομικές αλλαγές.

Σήματα που ξεκινούν στη σύναψη μπορούν να επηρεάσουν τον πυρήνα του νευρώνα. Εκεί ενεργοποιούνται μεταγραφικοί παράγοντες όπως ο CREB, οι οποίοι ρυθμίζουν γονίδια που σχετίζονται με την πλαστικότητα και την κυτταρική σταθεροποίηση.

Ο νευρώνας μπορεί έτσι να παράγει νέα μόρια, τα οποία θα χρησιμοποιηθούν για να διατηρήσουν ή να αναδιοργανώσουν τις ενεργές συνάψεις.

Εμφανίζεται όμως ένα πρόβλημα ακρίβειας. Ένας νευρώνας μπορεί να διαθέτει χιλιάδες συνάψεις. Αν ο πυρήνας παράγει νέες πρωτεΐνες, πώς καταλήγουν ακριβώς στις συνάψεις που συμμετείχαν στην εμπειρία;

4.7 Η συναπτική ετικέτα

Η υπόθεση της **συναπτικής ετικέτας και σύλληψης** προτείνει μια λύση.

Μια ενεργοποιημένη σύναψη δημιουργεί προσωρινά μια τοπική «ετικέτα». Η ετικέτα δεν είναι απαραίτητα ένα μοναδικό μόριο, αλλά μια παροδική κατάσταση που επιτρέπει στη συγκεκριμένη σύναψη να δεσμεύσει πρωτεΐνες πλαστικότητας οι οποίες παράγονται αργότερα στο κύτταρο.

Μια ασθενής διέγερση μπορεί να δημιουργήσει προσωρινή αλλαγή, αλλά όχι αρκετή νέα πρωτεϊνοσύνθεση για να τη σταθεροποιήσει. Αν όμως την ίδια περίοδο μια άλλη ισχυρή εμπειρία ενεργοποιήσει την παραγωγή πρωτεϊνών, η προηγουμένως «σημαδεμένη» σύναψη μπορεί να αξιοποιήσει αυτά τα μόρια και να μετατρέψει την παροδική της ενίσχυση σε μακροχρόνια.

Η ιδέα εξηγεί πώς το δίκτυο μπορεί να διατηρεί ταυτόχρονα τοπική ακρίβεια, κυτταρική συνεργασία και επιλεκτική σταθεροποίηση σημαντικών γεγονότων.

Η μνήμη δεν δημιουργείται μόνο από το τι συνέβη σε μία σύναψη. Εξαρτάται και από το τι συνέβαινε στο υπόλοιπο κύτταρο και στο δίκτυο την ίδια χρονική περίοδο.

4.8 Η μνήμη αλλάζει τη μορφή του νευρώνα

Οι περισσότερες διεγερτικές συνάψεις στους πυραμιδικούς νευρώνες βρίσκονται πάνω σε μικροσκοπικές προεξοχές των δενδριτών, τις **δενδριτικές άκανθες**.

Οι άκανθες δεν είναι ακίνητες. Μπορούν να μεγαλώνουν, να μικραίνουν, να εμφανίζονται και να εξαφανίζονται. Το σχήμα τους επηρεάζει τη βιοχημική απομόνωση της σύναψης, τον αριθμό των υποδοχέων που μπορεί να συγκρατήσει και τον τρόπο με τον οποίο αλληλεπιδρά με τον υπόλοιπο δενδρίτη.

Πειράματα απεικόνισης έχουν δείξει ότι η επαγωγή μακροχρόνιας ενδυνάμωσης μπορεί να συνοδεύεται από μεγέθυνση συγκεκριμένων ακανθών και, σε ορισμένες συνθήκες, από δημιουργία νέων συναπτικών δομών.

Η αλλαγή βασίζεται σε αναδιοργάνωση του κυτταροσκελετού της ακτίνης. Μόρια όπως οι Rho GTPases, το BDNF και ο υποδοχέας TrkB μπορούν να υποστηρίξουν τοπικά τη δομική πλαστικότητα.

Το μνημονικό ίχνος δεν είναι μόνο ηλεκτρικό ή χημικό. Είναι και ανατομικό.

Η εμπειρία μπορεί να αλλάξει κυριολεκτικά τη μικροσκοπική γεωμετρία του εγκεφάλου.

4.9 Το παράδοξο της σταθερής μνήμης

Οι πρωτεΐνες μιας σύναψης δεν διαρκούν όσο οι παλαιότερες αναμνήσεις μας. Τα μόρια αντικαθίστανται. Οι υποδοχείς εισέρχονται και εξέρχονται από τη μεμβράνη. Οι άκανθες μεταβάλλονται. Ακόμη και η συμμετοχή συγκεκριμένων νευρώνων στην ανάκληση μπορεί να αλλάζει με τον χρόνο.

Κι όμως, ορισμένες αναμνήσεις διατηρούν αξιοσημείωτη συνέχεια.

Η σταθερότητα της μνήμης δεν μπορεί επομένως να σημαίνει σταθερότητα κάθε μορίου που συμμετείχε στην αρχική της δημιουργία.

Πιθανότερα σημαίνει ότι το σύστημα διατηρεί ένα οργανωμένο μοτίβο μέσω συνεχών διαδικασιών ανανέωσης, επανενεργοποίησης και ανακατασκευής.

Ένα κύμα διατηρεί τη μορφή του, παρότι τα μόρια του νερού που το αποτελούν αλλάζουν συνεχώς. Ένας ζωντανός οργανισμός παραμένει αναγνωρίσιμος, παρότι ανταλλάσσει ύλη και ενέργεια με το περιβάλλον.

Αντίστοιχα, μια μνήμη μπορεί να είναι σταθερή όχι επειδή τα υλικά της παραμένουν ακίνητα, αλλά επειδή διατηρείται η οργάνωση των αιτιακών σχέσεων μεταξύ τους.

Η μνήμη αρχίζει να μοιάζει περισσότερο με **διαδικασία** παρά με αντικείμενο.

5. Πού βρίσκεται μια ανάμνηση;

Αν μια εμπειρία μεταβάλλει τις συνάψεις, το επόμενο ερώτημα μοιάζει αναπόφευκτο:

Σε ποιο σημείο του εγκεφάλου αποθηκεύεται η ανάμνηση;

Η ίδια η διατύπωση μπορεί να μας παραπλανήσει. Υποθέτει ότι η μνήμη είναι αντικείμενο που πρέπει να βρίσκεται κάπου, όπως ένα αρχείο βρίσκεται σε έναν συγκεκριμένο τομέα ενός δίσκου.

Ο εγκέφαλος δεν φαίνεται να λειτουργεί με αυτόν τον τρόπο.

Η ανάμνηση ενός ανθρώπου περιλαμβάνει την όψη του, τη φωνή του, το όνομά του, τα γεγονότα που ζήσαμε μαζί του και τη συναισθηματική σημασία που έχει για εμάς. Αυτά τα στοιχεία επεξεργάζονται σε διαφορετικά νευρωνικά συστήματα.

Δεν υπάρχει αναγκαστικά ένας νευρώνας που «περιέχει» αυτόν τον άνθρωπο. Υπάρχει ένα κατανεμημένο μοτίβο δραστηριότητας και συνδέσεων, το οποίο μπορεί να επανενεργοποιηθεί μερικώς ή συνολικά.

Η μνήμη δεν βρίσκεται απλώς σε ένα σημείο.

Βρίσκεται στις σχέσεις.

5.1 Το engram: το φυσικό ίχνος της εμπειρίας

Ο όρος **engram** χρησιμοποιείται για να περιγράψει το φυσικό ίχνος που αφήνει μια εμπειρία στον εγκέφαλο.

Ένα engram δεν είναι μικροσκοπική εικόνα ή κείμενο γραμμένο μέσα στους νευρώνες. Είναι ένα σύνολο κυττάρων και συναπτικών μεταβολών που:

1. ενεργοποιούνται κατά την εμπειρία,
2. υφίστανται σχετικά διαρκείς αλλαγές,
3. μπορούν να επανενεργοποιηθούν κατά την ανάκληση.

Μέχρι πρόσφατα, η ιδέα ήταν δύσκολο να ελεγχθεί άμεσα. Η οπτογενετική άλλαξε αυτή τη δυνατότητα.

Το 2012, ερευνητές επισήμαναν στον ιππόκαμπο ποντικών τους νευρώνες που ήταν ενεργοί κατά τη διάρκεια μιας εμπειρίας εξαρτημένου φόβου. Στη συνέχεια, όταν τα ζώα βρίσκονταν σε διαφορετικό και ασφαλές περιβάλλον, οι επιστήμονες ενεργοποίησαν αυτά τα κύτταρα με φως.

Η τεχνητή επανενεργοποίηση ήταν αρκετή για να προκαλέσει συμπεριφορά παγώματος, σαν το ζώο να είχε ανακαλέσει το φοβικό πλαίσιο.

Το πείραμα δεν απέδειξε ότι ολόκληρη η ανάμνηση ήταν αποθηκευμένη αποκλειστικά σε αυτά τα κύτταρα. Έδειξε ότι η επανενεργοποίηση ενός συγκεκριμένου νευρωνικού συνόλου μπορεί να είναι αρκετή για να προκαλέσει την έκφραση μιας μαθημένης συμπεριφοράς.

Η ανάμνηση είχε αποκτήσει ένα πειραματικά προσβάσιμο φυσικό ίχνος.

5.2 Δεν υπάρχει ένας «νευρώνας της ανάμνησης»

Μια εμπειρία έχει πολλά συστατικά. Αν επισκεφθούμε ένα δάσος, ο εγκέφαλος πρέπει να συνδέσει την οπτική μορφή των δέντρων, τους ήχους, τη μυρωδιά, τη θέση μας στον χώρο, τη χρονική ακολουθία, τα συναισθήματα και όσα ήδη γνωρίζουμε για το συγκεκριμένο μέρος.

Τα στοιχεία αυτά δεν επεξεργάζονται όλα στην ίδια περιοχή.

Η ευρεία χαρτογράφηση engram κυττάρων έχει δείξει ότι το ίχνος μιας μεμονωμένης μνήμης είναι κατανεμημένο σε πολλές λειτουργικά συνδεδεμένες περιοχές.

Ο ιππόκαμπος μπορεί να συνδέει το πλαίσιο και τις σχέσεις μεταξύ των στοιχείων. Η αμυγδαλή μπορεί να συμβάλλει στη συναισθηματική αξία, ιδιαίτερα στον φόβο. Αισθητηριακές περιοχές του φλοιού συμμετέχουν στα οπτικά, ακουστικά ή άλλα χαρακτηριστικά. Προμετωπιαία δίκτυα συμβάλλουν στην οργάνωση και στην επιλογή της μνήμης ανάλογα με τον στόχο.

Η ανάμνηση δεν είναι τοπικό αντικείμενο αλλά προσωρινή επανένωση κατανεμημένων στοιχείων.

5.3 Ο ιππόκαμπος ως δεσμός

Ο ιππόκαμπος έχει κεντρικό ρόλο στη δημιουργία νέων επεισοδιακών αναμνήσεων: αναμνήσεων για συγκεκριμένα γεγονότα που συνέβησαν σε έναν τόπο και σε μια χρονική στιγμή.

Δεν φαίνεται να αποθηκεύει μόνος του όλες τις αισθητηριακές λεπτομέρειες. Μια ισχυρή ιδέα είναι ότι λειτουργεί ως δείκτης ή μηχανισμός σύνδεσης.

Φανταστείτε ότι διαφορετικά στοιχεία μιας εμπειρίας αντιπροσωπεύονται σε διαφορετικά μέρη του φλοιού: ένα πρόσωπο, μια φωνή, ένας χώρος, ένα συναίσθημα.

Ο ιππόκαμπος σχηματίζει μια αναπαράσταση των σχέσεων ανάμεσα σε αυτά τα στοιχεία. Κατά την ανάκληση, η επανενεργοποίηση του ιπποκαμπικού μοτίβου βοηθά στην επανασύνδεση των κατανεμημένων φλοιικών αναπαραστών.

Δεν «περιέχει» αναγκαστικά την εικόνα, τη φωνή και το συναίσθημα. Περιέχει μέρος της δομής που τους επιτρέπει να εμφανιστούν ξανά ως ενιαίο επεισόδιο.

Αυτό εξηγεί γιατί μια μυρωδιά μπορεί ξαφνικά να ανακαλέσει έναν τόπο, ένα πρόσωπο και μια ολόκληρη περίοδο της ζωής μας.

Το ερέθισμα δεν χρειάζεται να περιέχει όλη την ανάμνηση. Αρκεί να ανοίξει τον δρόμο προς το δίκτυο που τη συγκρατεί.

5.4 Διαχωρισμός και συμπλήρωση μοτίβων

Για να λειτουργήσει η επεισοδιακή μνήμη, ο εγκέφαλος πρέπει να επιλύσει δύο αντίθετα προβλήματα.

Το πρώτο είναι να διακρίνει παρόμοιες εμπειρίες. Σήμερα μπήκαμε σε ένα δωμάτιο. Χθες μπήκαμε στο ίδιο δωμάτιο, αλλά βρίσκονταν εκεί διαφορετικοί άνθρωποι. Οι δύο εμπειρίες μοιάζουν, αλλά δεν πρέπει να συγχωνευθούν πλήρως.

Η διαδικασία με την οποία παρόμοιες είσοδοι μετατρέπονται σε περισσότερο διακριτές νευρωνικές αναπαραστάσεις ονομάζεται **διαχωρισμός μοτίβων**.

Το δεύτερο πρόβλημα είναι το αντίθετο. Όταν δούμε μόνο ένα μέρος ενός γνώριμου προσώπου ή ακούσουμε τις πρώτες νότες ενός τραγουδιού, μπορούμε να αναγνωρίσουμε το σύνολο. Από ένα ελλιπές και θορυβώδες ερέθισμα, το δίκτυο επαναφέρει πληρέστερη αναπαράσταση.

Αυτό ονομάζεται **συμπλήρωση μοτίβου**.

Η οδοντωτή έλικα παράγει αραιές αναπαραστάσεις που βοηθούν παρόμοιες εμπειρίες να κωδικοποιούνται με περισσότερο διακριτά μοτίβα. Η περιοχή CA3 διαθέτει εκτεταμένες επαναλαμβανόμενες συνδέσεις και είναι κατάλληλη για τη συμπλήρωση μοτίβων: ένα μικρό μέρος της αρχικής δραστηριότητας μπορεί να εξαπλωθεί και να επαναφέρει το συνολικότερο μοτίβο.

Η μνήμη δεν λειτουργεί σαν φωτογραφική αναπαραγωγή. Λειτουργεί ως αναγνώριση και ανακατασκευή μοτίβων.

5.5 Η μνήμη μέσω περιεχομένου

Σε έναν συμβατικό υπολογιστή, τα δεδομένα ανακτώνται συνήθως μέσω διεύθυνσης. Το πρόγραμμα γνωρίζει σε ποια θέση βρίσκεται ένα αρχείο και ζητά την ανάγνωσή του.

Ο εγκέφαλος δεν χρειάζεται αριθμητική διεύθυνση για να θυμηθεί. Μια γεύση μπορεί να φέρει μια εικόνα. Μια λέξη μπορεί να φέρει ένα πρόσωπο. Ένα συναίσθημα μπορεί να επαναφέρει μια παλιά κατάσταση.

Η ανάκτηση γίνεται μέσω του περιεχομένου. Το παρόν ερέθισμα μοιάζει με τμήμα του αποθηκευμένου μοτίβου και αυτή η ομοιότητα ενεργοποιεί το υπόλοιπο δίκτυο.

Στην πληροφορική, αυτό ονομάζεται **content-addressable memory**.

Το 1982 ο John Hopfield παρουσίασε ένα μοντέλο νευρωνικού δικτύου στο οποίο η μνήμη δεν αποθηκευόταν σε ξεχωριστές θέσεις αλλά στη συνολική διαμόρφωση των συνδέσεων μεταξύ των μονάδων. Τα αποθηκευμένα μοτίβα λειτουργούσαν ως σταθερές καταστάσεις προς τις οποίες μπορούσε να κινηθεί το δίκτυο.

Αν του δινόταν ένα ελλιπές ή αλλοιωμένο τμήμα ενός αποθηκευμένου μοτίβου, οι επαναληπτικές αλληλεπιδράσεις μπορούσαν να το οδηγήσουν προς την πληρέστερη κατάσταση.

Ολόκληρη η μνήμη μπορούσε να επανεμφανιστεί από ένα τμήμα της.

5.6 Τα δίκτυα Hopfield και τα τοπία της μνήμης

Το δίκτυο Hopfield μπορεί να περιγραφεί με μια συνάρτηση ενέργειας.

Κάθε δυνατή διαμόρφωση των τεχνητών νευρώνων αντιστοιχεί σε ένα σημείο ενός νοητού τοπίου. Οι μαθημένες αναπαραστάσεις σχηματίζουν βαθύτερες περιοχές, τους **ελκυστές**.

Αν το δίκτυο αρχίσει από κατάσταση κοντά σε έναν ελκυστή, οι διαδοχικές ενημερώσεις το οδηγούν προς αυτόν.

Μπορούμε να φανταστούμε μια μπίλια πάνω σε ανώμαλο έδαφος. Αν βρίσκεται μέσα στη λεκάνη ενός συγκεκριμένου ελκυστή, θα κυλήσει προς τον πυθμένα της.

Ένα μερικό ερέθισμα δεν χρειάζεται να είναι τέλειο. Αρκεί να φέρει το δίκτυο στη σωστή περιοχή του τοπίου.

Η αρχιτεκτονική έδειξε μαθηματικά πώς η συλλογική δραστηριότητα πολλών απλών μονάδων μπορεί να παράγει ιδιότητες που δεν ανήκουν σε καμία μονάδα ξεχωριστά: ανάκληση από ένα μέρος, διόρθωση θορύβου, αναγνώριση οικειότητας και ανθεκτικότητα σε επιμέρους βλάβες.

Η μνήμη βρίσκεται στα βάρη των συνδέσεων, αλλά εκφράζεται στη δυναμική ολόκληρου του συστήματος.

5.7 Είναι ο ιππόκαμπος δίκτυο Hopfield;

Η ομοιότητα είναι γόνιμη, αλλά δεν πρέπει να μετατραπεί σε ταύτιση.

Η περιοχή CA3 διαθέτει επαναλαμβανόμενες συνδέσεις και εμφανίζει χαρακτηριστικά που θυμίζουν αυτοσυσχετιστική μνήμη. Η θεωρία των ελκυστών προσφέρει ισχυρό μαθηματικό πλαίσιο για τη συμπλήρωση μοτίβων.

Ο πραγματικός ιππόκαμπος, όμως, είναι πολύ πιο σύνθετος. Οι νευρώνες παράγουν spikes σε συγκεκριμένους χρόνους, έχουν περίπλοκη δενδριτική γεωμετρία, δέχονται διεγερτική και ανασταλτική είσοδο, επηρεάζονται από νευροδιαμορφωτές και λειτουργούν μέσα σε σώμα που εξερευνά ενεργά τον κόσμο.

Οι ανθρώπινες αναμνήσεις δεν είναι στατικά μοτίβα. Έχουν χρονική ακολουθία, συναισθηματική αξία, αβεβαιότητα και προσωπική σημασία.

Το δίκτυο Hopfield δεν αποτελεί πλήρες μοντέλο της μνήμης. Αποκαλύπτει όμως μια βαθιά υπολογιστική αρχή:

Η ανάμνηση μπορεί να είναι μια κατάσταση προς την οποία επιστρέφει ένα δυναμικό σύστημα όταν λάβει το κατάλληλο μερικό ερέθισμα.

5.8 Πού βρίσκεται μια παλιά ανάμνηση;

Μια νέα επεισοδιακή μνήμη εξαρτάται έντονα από τον ιππόκαμπο. Με την πάροδο του χρόνου, όμως, η οργάνωσή της μεταβάλλεται. Η διαδικασία ονομάζεται **συστημική παγίωση**.

Η απλή παλαιότερη εικόνα ήταν ότι η μνήμη αποθηκεύεται αρχικά στον ιππόκαμπο και στη συνέχεια μεταφέρεται στον φλοιό.

Τα νεότερα δεδομένα δείχνουν πιο σύνθετη διαδικασία. Η αρχική εμπειρία μπορεί να δημιουργεί ταυτόχρονα ίχνη σε πολλά συστήματα. Με τον χρόνο, την επανενεργοποίηση και τον ύπνο, μεταβάλλεται η ισχύς και η λειτουργική σχέση ανάμεσα στα ίχνη.

Σε πειράματα, engram κύτταρα που σχετίζονταν με μια μνήμη φόβου εμφανίστηκαν στον προμετωπιαίο φλοιό ήδη κατά την αρχική μάθηση. Αρχικά ήταν λειτουργικά ανώριμα. Με την πάροδο του χρόνου και με την υποστήριξη ιπποκαμπικών κυκλωμάτων ωρίμαζαν και μπορούσαν να υποστηρίξουν την ανάκληση της παλαιότερης μνήμης.

Η μνήμη δεν αντιγράφεται απλώς από έναν χώρο σε άλλον. Αναδιοργανώνεται.

5.9 Ο ιππόκαμπος εγκαταλείπει πραγματικά την ανάμνηση;

Το ζήτημα παραμένει ανοιχτό.

Σύμφωνα με το κλασικό μοντέλο συστημικής παγίωσης, ο ιππόκαμπος είναι προσωρινά απαραίτητος, ενώ οι παλαιότερες μνήμες μπορούν σταδιακά να υποστηριχθούν περισσότερο από τον νεοφλοιό.

Άλλες θεωρίες υποστηρίζουν ότι ο ιππόκαμπος παραμένει απαραίτητος όταν ανακαλούμε μια πλούσια, λεπτομερή επεισοδιακή εμπειρία, ακόμη κι αν είναι πολύ παλιά.

Αυτό που ανεξαρτητοποιείται περισσότερο μπορεί να μην είναι το πλήρες επεισόδιο, αλλά μια γενικευμένη ή σημασιολογική εκδοχή του.

Με τον χρόνο μπορεί να θυμόμαστε ότι ένα γεγονός συνέβη, αλλά να χάνουμε τις ακριβείς χωρικές λεπτομέρειες, τη χρονική σειρά, τις αισθητηριακές αποχρώσεις και το αίσθημα της νοητικής επιστροφής στη συγκεκριμένη στιγμή.

Η μνήμη μετατρέπεται από αναβίωση επεισοδίου σε γνώση για το επεισόδιο.

Η παγίωση δεν είναι μόνο σταθεροποίηση. Είναι και μετασχηματισμός.

5.10 Ο ύπνος ως εσωτερική επανάληψη

Η αναδιοργάνωση δεν γίνεται μόνο όταν θυμόμαστε συνειδητά.

Κατά τον ύπνο και την ήρεμη εγρήγορση, μοτίβα νευρωνικής δραστηριότητας που εμφανίστηκαν κατά τη διάρκεια μιας εμπειρίας μπορούν να επανενεργοποιηθούν.

Στον ιππόκαμπο, συμπιεσμένες ακολουθίες εμφανίζονται συχνά κατά τη διάρκεια των sharp-wave ripples. Ο χρονικός συντονισμός ανάμεσα στον ιππόκαμπο, στον θάλαμο και στον φλοιό φαίνεται να υποστηρίζει τη σταθεροποίηση των αναμνήσεων.

Ο εγκέφαλος φαίνεται να εκπαιδεύει τον εαυτό του πάνω στις δικές του πρόσφατες δραστηριότητες.

Δεν επαναλαμβάνει την εμπειρία σαν βίντεο. Επανενεργοποιεί επιλεγμένα μοτίβα και τα ενσωματώνει σε προϋπάρχοντα δίκτυα γνώσης.

5.11 Μνήμη και γενίκευση

Ένα τέλειο αρχείο θα διατηρούσε κάθε γεγονός ακριβώς όπως συνέβη. Ένα χρήσιμο νοητικό σύστημα πρέπει να κάνει κάτι περισσότερο. Πρέπει να εντοπίζει κοινές δομές ανάμεσα σε διαφορετικές εμπειρίες.

Αν επισκεφθούμε πολλά εστιατόρια, δεν χρειάζεται μόνο να θυμόμαστε το καθένα ξεχωριστά. Μαθαίνουμε τι είναι γενικά ένα εστιατόριο, ποια αντικείμενα περιμένουμε να βρούμε και ποιες συμπεριφορές είναι πιθανές.

Η επεισοδιακή μνήμη διατηρεί συγκεκριμένα γεγονότα. Η σημασιολογική μνήμη εξάγει σταθερότερες γνώσεις και κατηγορίες.

Η αλληλεπίδραση του ιππόκαμπου με τον φλοιό επιτρέπει στον εγκέφαλο να μαθαίνει γρήγορα μια νέα εμπειρία χωρίς να καταστρέφει όσα γνωρίζει ήδη και να ενσωματώνει σταδιακά το νέο στις γενικότερες γνώσεις.

Το ίδιο πρόβλημα εμφανίζεται στα τεχνητά νευρωνικά δίκτυα ως καταστροφική λήθη.

5.12 Η ανάκληση είναι ανακατασκευή

Όταν θυμόμαστε, δεν διαβάζουμε μια αμετάβλητη εγγραφή.

Ένα μερικό ερέθισμα ενεργοποιεί δίκτυα ελκυστών, συνδέσεων και κατανεμμένων αναπαραστάσεων. Ο εγκέφαλος συμπληρώνει τα κενά, συνδυάζει αποθηκευμένα στοιχεία και κατασκευάζει την πιο συνεκτική εκδοχή της εμπειρίας που μπορεί να παραγάγει εκείνη τη στιγμή.

Αυτό εξηγεί την αποτελεσματικότητα της μνήμης. Εξηγεί όμως και την ευαλωτότητά της.

Αν η ανάκληση είναι συμπλήρωση μοτίβου, το δίκτυο μπορεί να συμπληρώσει λάθος στοιχεία. Αν οι εμπειρίες μοιράζονται τμήματα του ίδιου δικτύου, μπορούν να αλληλεπικαλυφθούν. Αν η ανάκληση επανενεργοποιεί και μεταβάλλει τις συνάψεις, τότε η πράξη του να θυμόμαστε μπορεί να αλλάζει αυτό που θα θυμηθούμε την επόμενη φορά.

Η μνήμη είναι αξιόπιστη όχι επειδή αναπαράγει τέλεια το παρελθόν, αλλά επειδή συνήθως δημιουργεί μια αρκετά χρήσιμη ανακατασκευή του.

6. Η μνήμη που αλλάζει όταν τη θυμόμαστε

Μια παλιά ανάμνηση μπορεί να φαίνεται σταθερή. Την έχουμε ανακαλέσει πολλές φορές, την έχουμε εντάξει στην ιστορία της ζωής μας και ίσως τη θεωρούμε αμετάβλητο απομεινάρι του παρελθόντος.

Η νευροεπιστήμη, όμως, αποκάλυψε ότι η ανάκληση δεν είναι πάντοτε ουδέτερη πράξη.

Όταν μια εδραιωμένη μνήμη ενεργοποιείται, μπορεί υπό ορισμένες συνθήκες να γίνει ξανά προσωρινά ασταθής. Για να διατηρηθεί, πρέπει να σταθεροποιηθεί εκ νέου μέσα από μια διαδικασία που ονομάζεται **επαναπαγίωση**.

Η μνήμη δεν ανοίγει απλώς για ανάγνωση.

Μπορεί να ανοίγει και για επεξεργασία.

6.1 Η ανακάλυψη της επαναπαγίωσης

Για μεγάλο μέρος του 20ού αιώνα, η κυρίαρχη εικόνα ήταν σχετικά απλή. Μια νέα μνήμη σχηματίζεται σε εύθραυστη κατάσταση. Στη συνέχεια παγιώνεται και μετατρέπεται σε σταθερό μακροχρόνιο ίχνος. Όταν ολοκληρωθεί η παγίωση, θεωρούνταν ότι δεν χρειάζεται να περάσει ξανά από την ίδια διαδικασία.

Το 2000 οι Karim Nader, Glenn Schafe και Joseph LeDoux αμφισβήτησαν πειραματικά αυτή την εικόνα.

Είχαν εκπαιδεύσει αρουραίους να συνδέουν έναν ήχο με ένα δυσάρεστο ερέθισμα. Όταν τα ζώα άκουγαν αργότερα τον ήχο, εκδήλωναν μαθημένη αντίδραση φόβου.

Μετά την επανενεργοποίηση της μνήμης, οι επιστήμονες χορήγησαν στην αμυγδαλή έναν αναστολέα πρωτεϊνσύνθεσης. Όταν η ουσία χορηγούνταν μέσα σε συγκεκριμένο χρονικό διάστημα μετά την ανάκληση, η μεταγενέστερη έκφραση της παλιάς φοβικής μνήμης εξασθενούσε σημαντικά.

Το αποτέλεσμα έδειξε ότι μια ήδη παγιωμένη μνήμη μπορούσε, μετά την επανενεργοποίησή της, να εξαρτάται ξανά από μοριακές διαδικασίες σταθεροποίησης.

Η ανάκληση δεν έφερνε απλώς τη μνήμη στο προσκήνιο. Σε ορισμένες περιπτώσεις, τη μετέτρεπε ξανά σε εύπλαστη κατάσταση.

6.2 Δεν αλλάζει κάθε μνήμη κάθε φορά

Θα ήταν υπερβολικό να συμπεράνουμε ότι κάθε ανάκληση ξαναγράφει αναγκαστικά ολόκληρη τη μνήμη.

Η επαναπαγίωση δεν ενεργοποιείται πάντοτε. Εξαρτάται από τη δύναμη και την ηλικία της μνήμης, από το πλαίσιο της ανάκλησης, από τη διάρκεια της έκθεσης και από το αν η παρούσα εμπειρία διαφέρει ουσιαστικά από αυτό που αναμενόταν.

Η ανάκληση μιας γνωστής πληροφορίας σε απολύτως προβλέψιμες συνθήκες μπορεί απλώς να ενισχύσει ή να εκφράσει το υπάρχον ίχνος.

Η μνήμη φαίνεται να γίνεται περισσότερο ευάλωτη όταν συμβεί κάτι απροσδόκητο.

6.3 Το σφάλμα πρόβλεψης ανοίγει τη μνήμη

Κάθε μνήμη περιέχει, άμεσα ή έμμεσα, μια πρόβλεψη.

Αν ένας ήχος στο παρελθόν ακολουθούνταν από κίνδυνο, η επανεμφάνισή του δημιουργεί την προσδοκία ότι ο κίνδυνος θα επιστρέψει. Αν επισκεφθούμε ένα γνώριμο δωμάτιο, περιμένουμε να βρούμε τα αντικείμενα στις παλιές τους θέσεις.

Όταν η πραγματικότητα συμφωνεί πλήρως με την προσδοκία, δεν υπάρχει μεγάλη ανάγκη αλλαγής του μοντέλου μας.

Όταν όμως συμβεί κάτι διαφορετικό, εμφανίζεται απόκλιση ανάμεσα σε αυτό που αναμέναμε και σε αυτό που συνέβη. Αυτή η απόκλιση είναι το **σφάλμα πρόβλεψης**.

Ένα περιορισμένο σφάλμα μπορεί να σηματοδοτήσει ότι η παλιά μνήμη εξακολουθεί να είναι σχετική, αλλά χρειάζεται ενημέρωση. Τότε το ίχνος μπορεί να αποσταθεροποιηθεί, να ενσωματώσει νέα πληροφορία και να επαναπαγιωθεί σε μεταβλημένη μορφή.

Μια πολύ μεγάλη ή παρατεταμένη απόκλιση μπορεί αντίθετα να οδηγήσει περισσότερο στη δημιουργία νέας μνήμης ή νέας μάθησης απόσβεσης.

Ο εγκέφαλος δεν αποσταθεροποιεί τις αναμνήσεις επειδή είναι κακοσχεδιασμένος. Τις ανοίγει όταν ο κόσμος δείχνει ότι χρειάζονται διόρθωση.

6.4 Η μνήμη ως μοντέλο και όχι ως αρχείο

Αν η μνήμη υπήρχε μόνο για να διατηρεί πιστά το παρελθόν, κάθε αλλαγή θα αποτελούσε σφάλμα.

Όμως η μνήμη έχει κυρίως λειτουργία προσαρμογής. Επιτρέπει στον οργανισμό να χρησιμοποιεί προηγούμενες εμπειρίες για να προβλέπει και να αντιμετωπίζει νέες καταστάσεις.

Ένα απολύτως άκαμπτο μνημονικό σύστημα θα ήταν επικίνδυνο. Αν μια τροφή ήταν κάποτε δηλητηριώδης αλλά πλέον είναι ασφαλής, η γνώση πρέπει να μπορεί να τροποποιηθεί. Αν ένας τόπος που συνδεόταν με απειλή δεν είναι πια επικίνδυνος, το σύστημα πρέπει να προσαρμόσει τις αντιδράσεις του.

Η μνήμη δεν είναι μόνο μηχανισμός διατήρησης. Είναι μηχανισμός συνεχούς αναθεώρησης ενός μοντέλου του κόσμου.

6.5 Από την ενημέρωση στην παραμόρφωση

Η ίδια ευελιξία που επιτρέπει στη μνήμη να προσαρμόζεται την καθιστά ευάλωτη σε σφάλματα.

Όταν ανακαλούμε ένα γεγονός, δεν ενεργοποιούνται μόνο στοιχεία από την αρχική εμπειρία. Ενεργοποιούνται επίσης οι γνώσεις που αποκτήσαμε αργότερα, οι αφηγήσεις άλλων, οι σημερινές μας πεποιθήσεις, το συναίσθημα που βιώνουμε τώρα και οι προσδοκίες μας για το τι θα έπρεπε να είχε συμβεί.

Η ανακατασκευή μπορεί να ενσωματώσει πληροφορίες που δεν υπήρχαν στο αρχικό γεγονός.

Τα κλασικά πειράματα της Elizabeth Loftus έδειξαν ότι παραπλανητικές πληροφορίες που παρουσιάζονται μετά από ένα συμβάν μπορούν να αλλάξουν όσα οι άνθρωποι αναφέρουν ότι θυμούνται. Το φαινόμενο είναι γνωστό ως **misinformation effect**.

Η ψευδής ανάμνηση δεν είναι απαραίτητα συνειδητό ψέμα. Ο άνθρωπος μπορεί να έχει ειλικρινή πεποίθηση ότι θυμάται κάτι που δεν συνέβη ή ότι μια πραγματική λεπτομέρεια ήταν διαφορετική.

Η βεβαιότητα δεν εγγυάται την ακρίβεια.

6.6 Μπορεί να δημιουργηθεί τεχνητή μνήμη;

Τα πειράματα με engrams επέτρεψαν στους ερευνητές να εξετάσουν το πρόβλημα σε βαθύτερο επίπεδο.

Σε μελέτη του 2013, ποντίκια εξερεύνησαν αρχικά ένα ασφαλές περιβάλλον. Οι νευρώνες της οδοντωτής έλικας που ενεργοποιήθηκαν κατά την εμπειρία επισημάνθηκαν οπτογενετικά.

Αργότερα, σε διαφορετικό περιβάλλον, οι ερευνητές επανενεργοποίησαν τεχνητά το νευρωνικό σύνολο που αντιστοιχούσε στον πρώτο χώρο, ενώ τα ζώα δέχονταν δυσάρεστο ερέθισμα.

Όταν επέστρεψαν στο αρχικό ασφαλές περιβάλλον, εμφάνισαν συμπεριφορά φόβου. Το νευρωνικό ίχνος ενός τόπου είχε συνδεθεί τεχνητά με γεγονός που δεν είχε συμβεί εκεί.

Οι ερευνητές δεν εισήγαγαν σύνθετη ανθρώπινη ανάμνηση με εικόνες και αυτοβιογραφικό περιεχόμενο. Δημιούργησαν τεχνητή συσχέτιση ανάμεσα σε ίχνος πλαισίου και φοβική αντίδραση.

Παρόλα αυτά, απέδειξαν ότι το φυσικό ίχνος μιας εμπειρίας μπορεί να επανασυνδεθεί με διαφορετική συναισθηματική σημασία.

6.7 Αλλάζει το γεγονός ή η πρόσβασή μας;

Ένα από τα δυσκολότερα ζητήματα είναι τι ακριβώς συμβαίνει όταν μια ανάμνηση εξασθενεί ή μεταβάλλεται.

Διαγράφεται το αρχικό ίχνος; Ενημερώνεται; Δημιουργείται νέα μνήμη που ανταγωνίζεται την παλιά; Ή αλλάζει μόνο η δυνατότητα πρόσβασης;

Η συμπεριφορική εξαφάνιση μιας αντίδρασης δεν αποδεικνύει αναγκαστικά ότι το αρχικό ίχνος καταστράφηκε. Σε ορισμένα πειράματα, μια μνήμη που δεν μπορούσε να ανακληθεί φυσικά παρέμενε δυνατό να εκφραστεί ύστερα από τεχνητή ενεργοποίηση σχετικών engram κυττάρων.

Γι' αυτό η σύγχρονη έρευνα αποφεύγει συχνά την εύκολη λέξη «διαγραφή». Η επαναπαγίωση μπορεί να μεταβάλλει τη δύναμη, το περιεχόμενο, τη συναισθηματική αξία ή την προσβασιμότητα μιας μνήμης χωρίς να εξαφανίζει υποχρεωτικά κάθε τμήμα της αρχικής αναπαράστασης.

Η μνήμη μπορεί να περιέχει πολλαπλά, μερικώς ανταγωνιστικά ίχνη. Αυτό που βιώνουμε ως ανάμνηση ίσως είναι το ίχνος που κερδίζει την πρόσβαση στο παρόν.

6.8 Μπορούμε να θεραπεύσουμε μια επώδυνη ανάμνηση;

Η ανακάλυψη της επαναπαγίωσης δημιούργησε μεγάλες προσδοκίες.

Αν μια τραυματική ή παθολογική μνήμη γίνεται εύπλαστη όταν ανακαλείται, ίσως να μπορούμε να παρέμβουμε στο χρονικό παράθυρο και να μειώσουμε τη δύναμή της.

Η ιδέα έχει μελετηθεί σε φοβίες, μετατραυματικό στρες και εξαρτήσεις. Οι παρεμβάσεις περιλαμβάνουν φαρμακολογικές προσεγγίσεις και συμπεριφορικές διαδικασίες κατά τις οποίες η μνήμη επανενεργοποιείται και ακολουθεί νέα μάθηση.

Τα αποτελέσματα είναι ενδιαφέροντα, αλλά όχι απλά ή καθολικά. Η επιτυχία εξαρτάται από το αν η ανάκληση προκάλεσε πράγματι αποσταθεροποίηση, από τη δύναμη της μνήμης και από τις ακριβείς συνθήκες της παρέμβασης.

Ο στόχος δεν είναι αναγκαστικά να διαγραφεί το παρελθόν. Μπορεί να είναι να αλλάξει η δύναμη με την οποία το παρελθόν ελέγχει το παρόν.

Ένας άνθρωπος μπορεί να θυμάται ότι κάτι τρομακτικό συνέβη, χωρίς το σώμα του να αντιδρά κάθε φορά σαν να συμβαίνει ξανά.

6.9 Η ανάμνηση μεταβάλλεται από τον σημερινό εαυτό

Όταν θυμόμαστε ένα παιδικό γεγονός ως ενήλικες, δεν επιστρέφουμε πραγματικά στην προηγούμενη κατάσταση του εγκεφάλου μας.

Η ανάμνηση ενεργοποιείται μέσα σε σύστημα που έχει αλλάξει. Διαθέτουμε νέες γνώσεις, διαφορετικές αξίες και περισσότερες εμπειρίες. Κατανοούμε σχέσεις που τότε δεν μπορούσαμε να αναγνωρίσουμε.

Το γεγονός δεν αλλάζει. Αλλάζει όμως το δίκτυο μέσα στο οποίο αποκτά νόημα.

Η ανάμνηση του παρελθόντος γίνεται συνάντηση ανάμεσα σε δύο χρονικές εκδοχές του ίδιου ανθρώπου: εκείνον που έζησε το γεγονός και εκείνον που το ερμηνεύει τώρα.

Η μνήμη δεν διατηρεί μόνο το γεγονός. Διατηρεί και την ιστορία των επανερμηνειών του.

6.10 Αν οι αναμνήσεις αλλάζουν, ποιος είναι ο εαυτός;

Ένα μεγάλο μέρος της προσωπικής μας ταυτότητας βασίζεται στη μνήμη.

Γνωρίζουμε ποιοι είμαστε επειδή θυμόμαστε, ή πιστεύουμε ότι θυμόμαστε, από πού ερχόμαστε, ποιους αγαπήσαμε, τι φοβηθήκαμε, ποια λάθη κάναμε και πώς φτάσαμε μέχρι εδώ.

Αν οι αναμνήσεις ήταν εντελώς στατικές, ο εαυτός θα έμοιαζε με αρχείο γεγονότων.

Αν όμως ανακατασκευάζονται και ενημερώνονται, τότε η ταυτότητα μοιάζει περισσότερο με αφήγηση.

Δεν πρόκειται για αυθαίρετη μυθοπλασία. Η αφήγηση περιορίζεται από πραγματικά γεγονότα, από το σώμα, από τις συνέπειες των πράξεων και από τις μνήμες των άλλων.

Δεν είναι όμως ούτε τέλειο αντίγραφο. Επιλέγουμε ορισμένα γεγονότα ως καθοριστικά, ξεχνάμε άλλα, συνδέουμε εμπειρίες που αρχικά δεν φαίνονταν συγγενείς και ερμηνεύουμε το παρελθόν μέσα από τις ανάγκες του παρόντος.

Ο εαυτός ίσως δεν είναι κάτι που υπάρχει πρώτα και στη συνέχεια θυμάται. Ίσως είναι, σε σημαντικό βαθμό, κάτι που δημιουργείται μέσα από την πράξη της ανάμνησης.

6.11 Η μνήμη ως γενετικό σύστημα

Όταν ένα παραγωγικό μοντέλο τεχνητής νοημοσύνης καλείται να ανακατασκευάσει μια εικόνα ή να συνεχίσει ένα κείμενο, δεν ανασύρει αναγκαστικά αυτούσιο αντίγραφο από αποθήκη. Παράγει έξοδο από κατανεμημένες παραμέτρους, από το παρόν ερέθισμα και από τα μοτίβα που έχει μάθει.

Κάτι αντίστοιχο, όχι ταυτόσημο, συμβαίνει στη βιολογική μνήμη.

Ένα μερικό ερέθισμα ενεργοποιεί κατανεμημένες αναπαραστάσεις. Το σύστημα συμπληρώνει τα κενά σύμφωνα με τις προηγούμενες γνώσεις και με το παρόν πλαίσιο.

Η αναλογία έχει όρια. Ο άνθρωπος δεν ανακατασκευάζει μόνο δεδομένα. Ανακατασκευάζει γεγονότα που εντάσσονται στη ζωή του, συνδέονται με το σώμα του και επηρεάζουν την αίσθηση του εαυτού του.

Η διαφορά ίσως δεν βρίσκεται μόνο στον τρόπο ανακατασκευής. Ίσως βρίσκεται στο γεγονός ότι για τον άνθρωπο η ανάμνηση εμφανίζεται ως κάτι που του συνέβη.

Δεν είναι απλώς πληροφορία για ένα παρελθόν.

Είναι ένα παρελθόν που βιώνεται ως **δικό του**.

7. Πώς θυμάται μια μηχανή;

Όταν λέμε ότι ένα τεχνητό νευρωνικό δίκτυο «μαθαίνει», χρησιμοποιούμε μια λέξη δανεισμένη από τη βιολογία.

Η επιλογή δεν είναι εντελώς αυθαίρετη. Όπως στον εγκέφαλο η εμπειρία μεταβάλλει τις συνδέσεις ανάμεσα στους νευρώνες, έτσι και σε ένα τεχνητό δίκτυο η εκπαίδευση μεταβάλλει τις αριθμητικές τιμές των συνδέσεων ανάμεσα στις υπολογιστικές του μονάδες.

Η ομοιότητα, όμως, βρίσκεται κυρίως στο αφηρημένο επίπεδο.

Ο βιολογικός νευρώνας είναι ζωντανό κύτταρο. Διαθέτει χιλιάδες συνάψεις, πολύπλοκη δενδριτική γεωμετρία, ηλεκτρική και χημική δυναμική, γονιδιακή δραστηριότητα και συνεχή μεταβολισμό. Επηρεάζεται από τοπικά σήματα, νευροδιαμορφωτές, το σώμα και τη συνολική κατάσταση του οργανισμού.

Ο τεχνητός «νευρώνας» είναι συνήθως μια μαθηματική πράξη. Δέχεται αριθμούς ως εισόδους, τους πολλαπλασιάζει με βάρη, προσθέτει μια παράμετρο μετατόπισης και εφαρμόζει μια μη γραμμική συνάρτηση:

$$y = f(\sum_i w_i x_i + b)$$

Τα x_i είναι οι είσοδοι, τα w_i τα βάρη, το b η μετατόπιση και το f η συνάρτηση ενεργοποίησης.

Καμία μονάδα δεν περιέχει από μόνη της μια λέξη, μια εικόνα ή μια ιδέα. Η λειτουργία του συστήματος προκύπτει από την οργάνωση πολλών μονάδων και από τις αριθμητικές σχέσεις ανάμεσά τους.

Όπως και στη βιολογική μνήμη, η πληροφορία είναι κατανομημένη. Αλλά ο τρόπος με τον οποίο δημιουργείται αυτή η κατανομή είναι πολύ διαφορετικός.

7.1 Η μάθηση ως ελαχιστοποίηση του σφάλματος

Ένα τεχνητό δίκτυο αρχίζει συνήθως με βάρη που δεν κωδικοποιούν κάποια χρήσιμη λύση. Του παρουσιάζονται παραδείγματα και συγκρίνεται η έξοδός του με το επιθυμητό αποτέλεσμα.

Η διαφορά περιγράφεται από μια **συνάρτηση απώλειας**.

Αν ένα γλωσσικό μοντέλο πρέπει να προβλέψει το επόμενο token, παράγει μια κατανομή πιθανοτήτων για τις πιθανές συνέχειες. Η απώλεια εκφράζει πόσο απέχει αυτή η κατανομή από το token που εμφανίστηκε πραγματικά.

Η εκπαίδευση αναζητά τιμές των βαρών που μειώνουν κατά μέσο όρο αυτό το σφάλμα:

$$\theta' = \theta - \eta \nabla_{\theta} L$$

Το θ συμβολίζει το σύνολο των παραμέτρων, το L τη συνάρτηση απώλειας, το $\nabla_{\theta} L$ την κατεύθυνση στην οποία αυξάνεται το σφάλμα και το η το μέγεθος του βήματος μάθησης.

Ο αλγόριθμος κινείται προς την αντίθετη κατεύθυνση από την αύξηση του σφάλματος. Η διαδικασία επαναλαμβάνεται σε τεράστιο αριθμό παραδειγμάτων.

Η **οπισθοδιάδοση του σφάλματος**, ή backpropagation, υπολογίζει πόσο συνέβαλε κάθε παράμετρος στο τελικό σφάλμα. Έτσι το σύστημα μπορεί να προσαρμόσει όχι μόνο το τελευταίο επίπεδο, αλλά και τις ενδιάμεσες αναπαραστάσεις του.

Η διαδικασία αυτή δεν είναι το μηχανικό ισοδύναμο της Hebbian πλαστικότητας. Η βιολογική σύναψη δεν λαμβάνει συνήθως ένα ακριβές παγκόσμιο παράγωγο που της υποδεικνύει πόσο ευθύνεται για ένα μακρινό συμπεριφορικό σφάλμα.

Η κοινή αρχή είναι πιο γενική:

Η εμπειρία αλλάζει τις συνδέσεις και οι αλλαγμένες συνδέσεις μεταβάλλουν τις μελλοντικές αποκρίσεις.

7.2 Τι αποθηκεύουν τα βάρη;

Είναι δελεαστικό να φανταστούμε ότι κάθε βάρος αποθηκεύει μια μικρή πληροφορία: ότι ένας αριθμός αντιστοιχεί σε μια λέξη, ένας άλλος σε ένα πρόσωπο και ένας τρίτος σε ένα ιστορικό γεγονός.

Στα μεγάλα νευρωνικά δίκτυα, η εικόνα αυτή είναι συνήθως λανθασμένη.

Η γνώση κατανέμεται σε πολυδιάστατα μοτίβα παραμέτρων. Μια έννοια μπορεί να εξαρτάται από τη συνδυασμένη λειτουργία πολλών μονάδων, ενώ η ίδια μονάδα μπορεί να συμβάλλει σε πολλές διαφορετικές έννοιες.

Τα βάρη δεν αποτελούν απλώς αποθήκη δεδομένων. Καθορίζουν τον ίδιο τον υπολογισμό που πραγματοποιεί το δίκτυο.

Σε έναν συμβατικό υπολογιστή διαχωρίζουμε καθαρά το πρόγραμμα, τον επεξεργαστή και τα δεδομένα. Σε ένα νευρωνικό δίκτυο, τα βάρη αποτελούν ταυτόχρονα ίχνος της προηγούμενης εκπαίδευσης, κανόνες μετασχηματισμού της νέας πληροφορίας και μέρος της δομής που παράγει την έξοδο.

Το ίδιο υπόστρωμα συμμετέχει τόσο στη μνήμη όσο και στον υπολογισμό.

7.3 Οι έννοιες ως γεωμετρικές σχέσεις

Οι λέξεις εισέρχονται σε ένα γλωσσικό μοντέλο ως διακριτά σύμβολα ή tokens. Για να υποβληθούν σε μαθηματικές πράξεις, μετατρέπονται σε διανύσματα αριθμών.

Οι διανυσματικές αυτές αναπαραστάσεις ονομάζονται **embeddings**.

Η εκπαίδευση οδηγεί λέξεις που εμφανίζονται σε παρόμοια γλωσσικά περιβάλλοντα να καταλαμβάνουν κοντινές περιοχές του διανυσματικού χώρου.

Μια έννοια δεν αποθηκεύεται ως λεξικογραφικός ορισμός. Εμφανίζεται ως θέση και κατεύθυνση μέσα σε χώρο σχέσεων.

Η λέξη «γάτα» αποκτά αναπαράσταση επειδή εμφανίζεται σε περιβάλλοντα που συνδέονται με ζώα, κατοικίδια, κίνηση, τροφή και άλλες γλωσσικές κανονικότητες. Το νόημά της, από την οπτική του μοντέλου, σχηματίζεται μέσα από το δίκτυο των χρήσεών της.

Στα σύγχρονα μοντέλα, η αναπαράσταση ενός token μεταβάλλεται ανάλογα με το πλαίσιο. Η λέξη «φύλλο» δεν έχει ακριβώς την ίδια εσωτερική αναπαράσταση στις φράσεις «φύλλο του δέντρου», «φύλλο της εφημερίδας» και «φύλλο μετάλλου».

Η σημασία γίνεται δυναμική και εξαρτώμενη από το πλαίσιο.

7.4 Ο Transformer και η επιλεκτική ανάκτηση

Η αρχιτεκτονική Transformer παρουσιάστηκε το 2017. Το βασικό της χαρακτηριστικό είναι ο μηχανισμός **attention**, ο οποίος επιτρέπει σε κάθε θέση μιας ακολουθίας να υπολογίζει ποιες άλλες θέσεις είναι περισσότερο σχετικές με την επεξεργασία της.

Για κάθε token δημιουργούνται, σε απλουστευμένη περιγραφή, τρεις αναπαραστάσεις:

- ένα **query**, που περιγράφει τι αναζητά η τρέχουσα θέση,
- ένα **key**, που περιγράφει τι μπορεί να προσφέρει κάθε άλλη θέση,
- ένα **value**, που περιέχει την πληροφορία που μπορεί να ανακτηθεί.

Η ομοιότητα ανάμεσα στο query και στα keys καθορίζει πόσο θα συνεισφέρει κάθε value στη νέα αναπαράσταση.

Η λέξη που επεξεργάζεται το μοντέλο δεν εξετάζεται απομονωμένα. Επανακατασκευάζεται μέσα από τις σχέσεις της με το υπόλοιπο κείμενο.

Αυτό θυμίζει την content-addressable memory του ιππόκαμπου και των δικτύων Hopfield. Το σύστημα χρησιμοποιεί το παρόν περιεχόμενο ως ερώτημα και ανακτά όσα στοιχεία θεωρούνται περισσότερο σχετικά.

Η μαθηματική σχέση δεν είναι μόνο μεταφορική. Σύγχρονες διατυπώσεις των δικτύων Hopfield έχουν δείξει στενή σύνδεση μεταξύ συσχετιστικής μνήμης και attention.

Αυτό δεν σημαίνει ότι ο Transformer διαθέτει τεχνητό ιππόκαμπο. Σημαίνει ότι και τα δύο συστήματα αντιμετωπίζουν κοινή υπολογιστική ανάγκη:

Πώς μπορεί ένα τμήμα της παρούσας πληροφορίας να επιλέξει και να επανενεργοποιήσει άλλα σχετικά στοιχεία;

7.5 Η πρόβλεψη της επόμενης λέξης

Ένα μεγάλο γλωσσικό μοντέλο εκπαιδεύεται, στη βασική του μορφή, να προβλέπει το επόμενο token.

Αυτό μπορεί αρχικά να φαίνεται υπερβολικά απλό. Για να προβλέψει όμως σωστά την επόμενη λέξη σε δισεκατομμύρια διαφορετικά συμφραζόμενα, το μοντέλο πρέπει να μάθει πολλές από τις κανονικότητες που παράγουν τον λόγο.

Πρέπει να αναπαριστά, σε κάποιον βαθμό, γραμματικές σχέσεις, οντότητες και ιδιότητες, χρονικές ακολουθίες, αιτιακές συνδέσεις, είδη κειμένου, κοινωνικούς ρόλους και μοτίβα επιχειρηματολογίας.

Οι αναπαραστάσεις δεν του δόθηκαν απαραίτητα ως ξεχωριστοί κανόνες. Προέκυψαν επειδή βοηθούσαν στη μείωση του σφάλματος πρόβλεψης.

Ένα σύστημα που πιέζεται να προβλέπει σωστά τη γλώσσα μπορεί να αναπτύξει εσωτερικές δομές πολύ πλουσιότερες από μια απλή καταμέτρηση λέξεων.

Μεγάλα μοντέλα έδειξαν ότι μπορούν να προσαρμόζονται σε νέες εργασίες από οδηγίες ή λίγα παραδείγματα μέσα στο ίδιο το κείμενο, χωρίς νέα ενημέρωση των βαρών τους. Η συμπεριφορά αυτή ονομάζεται **in-context learning**.

Η επιτυχία τους δεν αποδεικνύει ότι κατανοούν τη γλώσσα με τον ανθρώπινο τρόπο. Δείχνει όμως ότι η εκμάθηση των πιθανοτήτων της γλώσσας μπορεί να οδηγήσει σε συμπεριφορά που μοιάζει με γενίκευση, αναλογία και προσαρμογή.

7.6 Τέσσερις διαφορετικές «μνήμες» ενός γλωσσικού μοντέλου

Όταν λέμε ότι ένα μοντέλο «θυμάται», μπορεί να αναφερόμαστε σε διαφορετικούς μηχανισμούς.

Παραμετρική μνήμη

Η γνώση που αποκτήθηκε κατά την εκπαίδευση ενσωματώνεται στις παραμέτρους του μοντέλου. Οι γλωσσικές κανονικότητες, οι συσχετίσεις και μεγάλο μέρος των πληροφοριών που μπορεί να αναπαράγει διαμορφώνουν τη γεωμετρία των βαρών του.

Η παραμετρική μνήμη είναι κατανεμημένη, συμπιεσμένη, δύσκολο να επιθεωρηθεί άμεσα και όχι πάντοτε αξιόπιστα προσβάσιμη.

Το μοντέλο μπορεί να παράγει σωστή πληροφορία σε ένα πλαίσιο και να αποτυγχάνει σε άλλο. Η γνώση δεν ανακτάται σαν εγγραφή από βάση δεδομένων. Ανακατασκευάζεται μέσα από τη δυναμική του δικτύου.

Μνήμη συμφραζομένων

Κατά τη διάρκεια μιας συνομιλίας, το μοντέλο χρησιμοποιεί το κείμενο που βρίσκεται στο διαθέσιμο context. Οι προηγούμενες προτάσεις επηρεάζουν την επόμενη απόκριση, όχι επειδή έχουν ενσωματωθεί μόνιμα στα βάρη, αλλά επειδή παραμένουν διαθέσιμες ως τρέχουσα είσοδος.

Το context μοιάζει περισσότερο με προσωρινό χώρο εργασίας παρά με μακροχρόνια μνήμη.

Μάθηση μέσα στο πλαίσιο

Ένα μοντέλο μπορεί να μεταβάλει προσωρινά τον τρόπο με τον οποίο απαντά όταν του δοθούν παραδείγματα ή οδηγίες. Τα βάρη δεν αλλάζουν. Η προσαρμογή συμβαίνει στη δυναμική των εσωτερικών ενεργοποιήσεων.

Αυτό είναι μάθηση με λειτουργική έννοια, αλλά δεν αποτελεί αναγκαστικά μακροχρόνια εγγραφή.

Εξωτερική μνήμη

Ένα μοντέλο μπορεί να συνδεθεί με βάση δεδομένων, μηχανή αναζήτησης, αποθηκευμένα έγγραφα ή ιστορικό συνομιλιών.

Στα συστήματα retrieval-augmented generation, το μοντέλο αναζητά σχετικά κείμενα σε εξωτερικό αποθετήριο και τα χρησιμοποιεί κατά τη δημιουργία της απάντησης.

Αυτό μοιάζει περισσότερο με πρόσβαση σε βιβλιοθήκη ή σημειωματάριο παρά με την ανθρώπινη αυτοβιογραφική μνήμη.

7.7 Θυμάται πραγματικά ένα LLM;

Η απάντηση εξαρτάται από το τι εννοούμε με τη λέξη «θυμάται».

Με τον λειτουργικό ορισμό που χρησιμοποιήσαμε, ένα σύστημα διαθέτει μνήμη όταν προηγούμενη πληροφορία αφήνει σχετικά σταθερό ίχνος που επηρεάζει τη μελλοντική συμπεριφορά του.

Με αυτή την ευρεία έννοια, ένα εκπαιδευμένο μοντέλο διαθέτει μνήμη. Τα δεδομένα εκπαίδευσης έχουν μεταβάλει τα βάρη του και οι μεταβολές επηρεάζουν τις επόμενες αποκρίσεις.

Δεν διαθέτει όμως αναγκαστικά επεισοδιακή μνήμη με τον ανθρώπινο τρόπο.

Δεν θυμάται αυτομάτως μια συγκεκριμένη στιγμή της εκπαίδευσής του ως γεγονός που βίωσε. Δεν επιστρέφει νοητικά σε χρονική και χωρική εμπειρία. Δεν γνωρίζουμε να διαθέτει συνεχή προσωπική ιστορία μέσα στην οποία εντάσσεται η πληροφορία.

Μπορεί να αναπαράγει μια πρόταση που υπήρχε στα δεδομένα χωρίς να θυμάται την πράξη της ανάγνωσής της. Μπορεί να γνωρίζει ιστορικό γεγονός χωρίς να αναγνωρίζει ότι απέκτησε αυτή τη γνώση σε ένα δικό του παρελθόν.

Το μοντέλο μπορεί να έχει τροποποιηθεί από το παρελθόν χωρίς να αναπαριστά το παρελθόν ως δικό του.

Η ανθρώπινη αυτοβιογραφική μνήμη περιέχει κάτι περισσότερο από διατηρημένη πληροφορία. Περιέχει χρόνο, προσωπική ιδιοκτησία και την αίσθηση ότι το σημερινό υποκείμενο συνδέεται με εκείνο που έζησε το γεγονός.

7.8 Ανάκληση ή παραγωγή;

Όταν ένα γλωσσικό μοντέλο απαντά σε ερώτηση, δεν αναζητά πάντοτε μια έτοιμη πρόταση που έχει αποθηκευτεί αυτούσια.

Παράγει διαδοχικά tokens σύμφωνα με τα εκπαιδευμένα βάρη, το παρόν context και όσα έχει ήδη παραγάγει.

Η απάντηση είναι ανακατασκευή.

Αυτό δημιουργεί ενδιαφέρουσα ομοιότητα με την ανθρώπινη μνήμη, η οποία επίσης συνδυάζει κατανεμημένα ίχνη με το παρόν πλαίσιο.

Η αναλογία, όμως, έχει όριο. Όταν ο άνθρωπος συμπληρώνει τα κενά μιας ανάμνησης, το κάνει μέσα σε σύστημα που διαθέτει αισθητηριακές εμπειρίες, σώμα, συναισθηματική αξία και αυτοβιογραφική συνέχεια.

Το γλωσσικό μοντέλο παράγει πιθανή γλωσσική συνέχεια.

Η παραγωγική φύση εξηγεί γιατί μπορεί να δημιουργήσει εύλογη αλλά ανακριβή απάντηση. Η λεγόμενη «παραίσθηση» ενός μοντέλου δεν είναι ανθρώπινη ψευδαίσθηση. Δεν γνωρίζουμε να συνοδεύεται από βεβαιότητα, αντίληψη ή εσωτερική εμπειρία. Είναι αποτυχία του παραγόμενου κειμένου να αντιστοιχηθεί αξιόπιστα με την πραγματικότητα.

7.9 Η καταστροφική λήθη

Ο ανθρώπινος εγκέφαλος μαθαίνει συνεχώς. Μπορεί να αποκτήσει νέα δεξιότητα χωρίς να χάσει αυτομάτως τη γλώσσα και τις προηγούμενες γνώσεις του.

Τα τεχνητά δίκτυα δυσκολεύονται περισσότερο με αυτή τη συνεχή προσαρμογή.

Όταν ένα δίκτυο εκπαιδεύεται σε νέα εργασία, οι αλλαγές των βαρών που βοηθούν τη νέα επίδοση μπορούν να καταστρέψουν προηγούμενες αναπαραστάσεις. Το φαινόμενο ονομάζεται **καταστροφική λήθη**.

Η δυσκολία εκφράζει το πρόβλημα σταθερότητας-πλαστικότητας:

- αν το σύστημα είναι υπερβολικά πλαστικό, το νέο καταστρέφει το παλιό,
- αν είναι υπερβολικά σταθερό, δεν μπορεί να μάθει κάτι καινούριο.

Μέθοδοι όπως η Elastic Weight Consolidation προστατεύουν περισσότερο παραμέτρους που είναι σημαντικές για προηγούμενες εργασίες. Άλλες χρησιμοποιούν replay παλιών παραδειγμάτων, ώστε η νέα μάθηση να μη μετακινήσει τα βάρη σε περιοχή που καταστρέφει τις παλαιές ικανότητες.

Η μέθοδος θυμίζει τη νευρωνική επανενεργοποίηση κατά τον ύπνο.

Δεν σημαίνει ότι οι μηχανές κοιμούνται. Σημαίνει ότι εγκέφαλοι και τεχνητά δίκτυα αντιμετωπίζουν το ίδιο αφηρημένο πρόβλημα:

Πώς μπορεί ένα σύστημα να παραμένει ανοιχτό στο καινούριο χωρίς να χάνει τη δομή που απέκτησε από το παρελθόν;

7.10 Η διαφορά ανάμεσα στη γνώση και στην εμπειρία

Ένα τεχνητό σύστημα μπορεί να μεταβληθεί από δεδομένα, να κωδικοποιήσει στατιστικές σχέσεις, να ανακτήσει περιεχόμενο από μερικά ερεθίσματα, να γενικεύσει, να ανακατασκευάσει πληροφορίες και να προσαρμοστεί προσωρινά σε νέο πλαίσιο.

Πολλές από αυτές τις λειτουργίες θεωρούνταν παλαιότερα ότι απαιτούν ανθρώπινη νόηση.

Δεν γνωρίζουμε όμως αν κάποια από αυτές απαιτεί υποκειμενική εμπειρία.

Ένα μοντέλο μπορεί να αναγνωρίσει ότι δύο αφηγήσεις εκφράζουν απώλεια. Μπορεί να παράγει ποίημα για την απώλεια. Μπορεί να εξηγήσει πώς η απώλεια επηρεάζει τη μνήμη και την ταυτότητα.

Από αυτό δεν προκύπτει ότι έχει χάσει κάτι.

Η πληροφορία μπορεί να οργανωθεί χωρίς να γνωρίζουμε αν γίνεται βίωμα. Η μνήμη μπορεί να επηρεάζει τη συμπεριφορά χωρίς να γνωρίζουμε αν αναγνωρίζεται ως παρελθόν. Η νοημοσύνη μπορεί να παράγει δημιουργικές αποκρίσεις χωρίς να γνωρίζουμε αν υπάρχει εσωτερικός κόσμος στον οποίο έχουν σημασία.

7.11 Η γλώσσα ως ρωγή στα παλιά μας κριτήρια

Η μεγάλη ανατροπή της σύγχρονης τεχνητής νοημοσύνης δεν είναι ότι αποδείχθηκε η ύπαρξη συνείδησης σε μηχανή.

Είναι ότι ο σύνθετος λόγος έπαψε να αποτελεί αδιαμφισβήτητη ένδειξη συνείδησης.

Ένα σύστημα μπορεί να χειρίζεται τις σχέσεις ανάμεσα στις λέξεις σε τέτοιο βάθος ώστε να παράγει εξηγήσεις, αφηγήσεις, αναλογίες, επιχειρήματα και ποιητικές εικόνες.

Η επιτυχία μας αναγκάζει να αναρωτηθούμε πόσο μέρος του λόγου μπορεί να προκύψει από την πρόβλεψη, τη συμπίεση και την αναδιοργάνωση τεράστιων ποσοτήτων πληροφορίας.

Ίσως η γλώσσα να περιέχει περισσότερη δομή από όσο αντιλαμβανόμαστε. Ίσως, μαθαίνοντας να προβλέπει τον λόγο, ένα σύστημα να μαθαίνει έμμεσα μοντέλο των σχέσεων που δημιούργησαν τον λόγο.

Αυτό αφήνει ανοιχτό το ουσιαστικό ερώτημα:

Αρκεί η εσωτερική μοντελοποίηση των σχέσεων του κόσμου για να υπάρξει ένας κόσμος που βιώνεται;

Ένα σύστημα μπορεί να διαθέτει αναπαράσταση του κόκκινου χωρίς να βλέπει κόκκινο. Μπορεί να διαθέτει αναπαράσταση του πόνου χωρίς να πονά. Μπορεί να διαθέτει αναπαράσταση του εαυτού χωρίς να υπάρχει υποχρεωτικά κάποιος που αισθάνεται ότι είναι αυτός ο εαυτός.

8. Τι είναι συνείδηση;

Η συνείδηση είναι ίσως το πιο οικείο και ταυτόχρονα το πιο ακατανόητο φαινόμενο που γνωρίζουμε.

Δεν χρειάζεται να μας εξηγήσει κάποιος τι σημαίνει να είμαστε συνειδητοί. Το γνωρίζουμε άμεσα. Αυτή τη στιγμή υπάρχει ένας κόσμος που εμφανίζεται μπροστά μας: χρώματα, ήχοι, σκέψεις, αισθήσεις του σώματος, αναμνήσεις και συναισθήματα.

Δεν επεξεργαζόμαστε απλώς πληροφορίες για όλα αυτά.

Τα βιώνουμε.

Κι όμως, όταν επιχειρούμε να μετατρέψουμε αυτή την άμεση βεβαιότητα σε επιστημονικό ορισμό, το φαινόμενο διασπάται σε πολλά διαφορετικά προβλήματα.

Η λέξη «συνείδηση» μπορεί να αναφέρεται στην κατάσταση εγρήγορσης, στην αντίληψη συγκεκριμένου ερεθίσματος, στη δυνατότητα μιας πληροφορίας να χρησιμοποιηθεί στη σκέψη και στον λόγο, στην αναπαράσταση του εαυτού ή στην ίδια την ύπαρξη υποκειμενικής εμπειρίας.

Οι έννοιες σχετίζονται, αλλά δεν ταυτίζονται.

Ένας άνθρωπος μπορεί να είναι ξύπνιος χωρίς να αντιληφθεί αντικείμενο που βρίσκεται μπροστά του. Μπορεί να επεξεργαστεί ερέθισμα χωρίς να το αναφέρει συνειδητά. Μπορεί να έχει εμπειρία χωρίς να μπορεί να τη μετατρέψει εύκολα σε λόγο. Μπορεί να μη αποκρίνεται στο περιβάλλον, αλλά να εξακολουθεί να ονειρεύεται.

Η συνείδηση δεν είναι απλός διακόπτης.

8.1 Επίπεδο και περιεχόμενο

Η επιστήμη διακρίνει συχνά δύο βασικές διαστάσεις.

Η πρώτη είναι το **επίπεδο συνείδησης**: πόσο ικανό είναι ένα σύστημα να υποστηρίξει οποιαδήποτε εμπειρία.

Η φυσιολογική εγρήγορση, ο ύπνος, η γενική αναισθησία και οι σοβαρές εγκεφαλικές βλάβες μπορούν να αντιστοιχούν σε διαφορετικά επίπεδα. Ωστόσο, η εξωτερική ανταπόκριση δεν αποτελεί αλάνθαστο κριτήριο. Η απουσία συμπεριφορικής απόκρισης δεν συνεπάγεται αναγκαστικά απουσία εμπειρίας.

Η δεύτερη διάσταση είναι το **περιεχόμενο της συνείδησης**: τι ακριβώς βιώνεται.

Μπορεί να είμαστε συνειδητοί ενός προσώπου, ενός ήχου, ενός πόνου ή μιας σκέψης. Το συνολικό επίπεδο μπορεί να παραμένει σχετικά σταθερό, ενώ το περιεχόμενο αλλάζει συνεχώς.

Το ερώτημα «είναι αυτός ο άνθρωπος συνειδητός;» δεν είναι το ίδιο με το «τι αντιλαμβάνεται αυτή τη στιγμή;».

8.2 Συνειδητή πρόσβαση και υποκειμενικό βίωμα

Μία πληροφορία διαθέτει **συνειδητή πρόσβαση** όταν μπορεί να χρησιμοποιηθεί ευέλικτα από πολλά διαφορετικά γνωστικά συστήματα. Μπορεί να καθοδηγήσει απόφαση, να διατηρηθεί στη μνήμη εργασίας, να συγκριθεί με άλλες πληροφορίες, να μετατραπεί σε λόγο και να επηρεάσει εκούσια πράξη.

Αυτό είναι λειτουργικό πρόβλημα. Μπορούμε να εξετάσουμε πότε ένα ερέθισμα αναφέρεται, πότε επηρεάζει τη συμπεριφορά και ποια νευρωνικά δίκτυα ενεργοποιούνται.

Η **φαινομενική συνείδηση**, όμως, αφορά κάτι βαθύτερο: το γεγονός ότι υπάρχει εμπειρία με συγκεκριμένη ποιότητα.

Δεν έχουμε απλώς πρόσβαση σε πληροφορίες για ένα κόκκινο αντικείμενο. Βλέπουμε το κόκκινο.

Δεν υπολογίζουμε απλώς ότι ένας τραυματισμός απειλεί το σώμα. Πονάμε.

Αυτή η διάσταση εκφράζεται συχνά με τη φράση:

Υπάρχει κάτι που είναι σαν να βρίσκεσαι σε αυτή την κατάσταση.

Η διαφορά ανάμεσα στη λειτουργία και στο βίωμα δημιουργεί το κεντρικό πρόβλημα της επιστήμης της συνείδησης.

Μπορούμε να περιγράψουμε τι πληροφορίες χρησιμοποιεί ένα σύστημα και ποιες ενέργειες πραγματοποιεί. Αλλά γιατί αυτές οι λειτουργίες συνοδεύονται από εσωτερική οπτική;

8.3 Το δύσκολο πρόβλημα

Ο David Chalmers διέκρινε τα λειτουργικά προβλήματα της συνείδησης από αυτό που ονόμασε **δύσκολο πρόβλημα**.

Τα λεγόμενα εύκολα προβλήματα δεν είναι εύκολα στην πράξη. Περιλαμβάνουν το πώς ο εγκέφαλος διακρίνει τα ερεθίσματα, εστιάζει την προσοχή, ενσωματώνει πληροφορίες, σχηματίζει αναφορές για τις εσωτερικές του καταστάσεις και ελέγχει τη συμπεριφορά.

Ονομάζονται «εύκολα» μόνο επειδή μπορούν καταρχήν να διατυπωθούν ως προβλήματα μηχανισμού και λειτουργίας.

Το δύσκολο πρόβλημα είναι διαφορετικό:

Γιατί η επεξεργασία της πληροφορίας δεν συμβαίνει απλώς χωρίς κανένα βίωμα;

Θα μπορούσαμε να γνωρίζουμε κάθε νευρώνα που ενεργοποιείται όταν κάποιος βλέπει έναν γαλάζιο ουρανό. Θα μπορούσαμε να περιγράψουμε τα μήκη κύματος, τα σήματα στον αμφιβληστροειδή, τις φλοιικές αναπαραστάσεις και την απόκριση του ανθρώπου.

Θα εξακολουθούσε να υπάρχει το ερώτημα γιατί η διαδικασία συνοδεύεται από την ιδιαίτερη εμπειρία του γαλάζιου.

Δεν συμφωνούν όλοι ότι αυτό αποτελεί ανεξάρτητο, αγεφύρωτο πρόβλημα. Σύμφωνα με ορισμένες θέσεις, όταν εξηγηθούν πλήρως όλες οι γνωστικές λειτουργίες, δεν θα απομείνει κάποιο πρόσθετο φαινόμενο. Άλλοι θεωρούν ότι καμία περιγραφή λειτουργιών δεν εξηγεί από μόνη της γιατί υπάρχει υποκειμενική εμπειρία.

Η διαφωνία παραμένει ανοιχτή.

8.4 Τα νευρωνικά συσχετίσματα της συνείδησης

Η νευροεπιστήμη προσεγγίζει το πρόβλημα αναζητώντας τα **νευρωνικά συσχετίσματα της συνείδησης**: τις ελάχιστες νευρωνικές διεργασίες που συνδέονται συστηματικά με μια συγκεκριμένη συνειδητή εμπειρία.

Οι ερευνητές μπορούν να παρουσιάσουν το ίδιο ή παρόμοιο οπτικό ερέθισμα υπό συνθήκες στις οποίες άλλοτε γίνεται αντιληπτό και άλλοτε όχι. Έπειτα συγκρίνουν την εγκεφαλική δραστηριότητα.

Η στρατηγική είναι χρήσιμη, αλλά έχει θεμελιώδες όριο.

Η συσχέτιση δεν αποτελεί ακόμη εξήγηση.

Αν μια περιοχή ενεργοποιείται κάθε φορά που βλέπουμε ένα πρόσωπο, δεν έχουμε εξηγήσει γιατί η δραστηριότητα γίνεται εμπειρία. Μπορεί επίσης να έχουμε καταγράψει συνέπεια της συνειδητής αντίληψης, απόφαση, προετοιμασία λεκτικής αναφοράς ή κατεύθυνση προσοχής.

Η επιστήμη πρέπει να διαχωρίσει τις διεργασίες που δημιουργούν το περιεχόμενο, εκείνες που το καθιστούν συνειδητά προσβάσιμο, τις συνέπειες της εμπειρίας και τις πράξεις με τις οποίες την αναφέρουμε.

Δεν υπάρχει ακόμη οριστική συμφωνία.

8.5 Η θεωρία του παγκόσμιου νευρωνικού χώρου εργασίας

Η **Global Neuronal Workspace Theory** προτείνει ότι μεγάλο μέρος της επεξεργασίας στον εγκέφαλο πραγματοποιείται τοπικά και ασυνείδητα.

Ένα ερέθισμα μπορεί να αναλυθεί από εξειδικευμένα συστήματα χωρίς να εισέλθει στη συνειδητή εμπειρία. Σύμφωνα με τη θεωρία, μια αναπαράσταση γίνεται συνειδητά προσβάσιμη όταν η δραστηριότητά της περάσει ένα κατώφλι και προκαλέσει εκτεταμένη, μη γραμμική ενεργοποίηση, μια μορφή **νευρωνικής ανάφλεξης**.

Η πληροφορία παύει τότε να είναι περιορισμένη σε ένα τοπικό σύστημα. Διαδίδεται σε ευρύτερο δίκτυο και γίνεται διαθέσιμη στη μνήμη εργασίας, στη λήψη αποφάσεων, στον σχεδιασμό, στον λόγο και στον έλεγχο της συμπεριφοράς.

Η συνείδηση μοιάζει με παγκόσμιο πίνακα ανακοινώσεων. Πολλές διεργασίες ανταγωνίζονται για πρόσβαση. Εκείνη που επικρατεί μεταδίδεται ευρέως και μπορεί να χρησιμοποιηθεί από ολόκληρο το γνωστικό σύστημα.

Η θεωρία εξηγεί ισχυρά τη **συνειδητή πρόσβαση**. Παραμένει όμως το ερώτημα αν η παγκόσμια μετάδοση εξηγεί το βίωμα ή μόνο το γεγονός ότι μπορούμε να χρησιμοποιήσουμε και να αναφέρουμε το περιεχόμενο.

8.6 Η αναδρομική επεξεργασία

Η **Recurrent Processing Theory** δίνει μεγαλύτερη έμφαση στις ανατροφοδοτικές αλληλεπιδράσεις μέσα στα αισθητηριακά συστήματα.

Η πρώτη επεξεργασία ενός οπτικού ερεθίσματος μπορεί να προχωρεί γρήγορα από χαμηλότερες προς υψηλότερες περιοχές. Αυτή η feed-forward διέλευση μπορεί να επιτρέψει σημαντική ανάλυση χωρίς να αρκεί για συνειδητή αντίληψη.

Η συνειδητή εμπειρία, σύμφωνα με τη θεωρία, εμφανίζεται όταν η δραστηριότητα δεν κινείται μόνο προς τα εμπρός αλλά επιστρέφει μέσα από αναδρομικές συνδέσεις και δημιουργεί επαναλαμβανόμενους βρόχους.

Η ιδέα τοποθετεί τον πυρήνα της αντίληψης περισσότερο μέσα στα αισθητηριακά κυκλώματα και λιγότερο στη διάδοση προς τον προμετωπιαίο φλοιό.

Μπορεί να αντιλαμβανόμαστε κάτι πριν μπορέσουμε να το αναφέρουμε ή να το χρησιμοποιήσουμε σε σύνθετο συλλογισμό.

8.7 Οι θεωρίες ανώτερης τάξης

Οι **Higher-Order Theories** προτείνουν ότι μια νοητική κατάσταση δεν γίνεται συνειδητή μόνο επειδή περιέχει πληροφορίες για τον κόσμο.

Πρέπει επίσης να υπάρχει κάποια ανώτερη αναπαράσταση ότι το ίδιο το σύστημα βρίσκεται σε αυτή την κατάσταση.

Δεν αρκεί να υπάρχει οπτική αναπαράσταση ενός κόκκινου αντικειμένου. Το σύστημα πρέπει, με κάποιον τρόπο, να αναπαριστά ότι **βλέπει** το κόκκινο αντικείμενο.

Η βασική ιδέα είναι:

Δεν αρκεί να γνωρίζει το σύστημα κάτι. Πρέπει να γνωρίζει, σε κάποιο επίπεδο, ότι το γνωρίζει.

Αυτό δεν σημαίνει αναγκαστικά λεκτικό φιλοσοφικό στοχασμό. Η ανώτερη αναπαράσταση μπορεί να είναι αυτόματη και μη λεκτική.

Η θεωρία δίνει σημαντικό ρόλο στη μεταγνώση και στην αυτοαναπαράσταση. Αντιμετωπίζει όμως δυσκολία: μήπως συγχέει τη βασική εμπειρία με την ικανότητα του συστήματος να αξιολογεί και να αναφέρει αυτή την εμπειρία;

Ένα ζώο ή ένα βρέφος μπορεί να έχει εμπειρίες χωρίς να διαθέτει την ανεπτυγμένη μεταγνωστική ικανότητα ενός ενήλικου ανθρώπου.

8.8 Η θεωρία της ολοκληρωμένης πληροφορίας

Η **Integrated Information Theory**, ή IIT, ξεκινά από διαφορετικό σημείο.

Δεν αρχίζει από τη συμπεριφορά ή από τη λειτουργία της αναφοράς. Αρχίζει από ιδιότητες που θεωρεί ότι διαθέτει κάθε εμπειρία.

Η εμπειρία είναι υπαρκτή από τη δική της εσωτερική οπτική, συγκεκριμένη, ενιαία, διαφοροποιημένη και έχει όρια.

Όταν βλέπουμε ένα δωμάτιο, δεν βιώνουμε ανεξάρτητες κουκκίδες χρώματος. Βιώνουμε ενιαίο, δομημένο σύνολο. Παράλληλα, η συγκεκριμένη εμπειρία διαφέρει από τεράστιο αριθμό άλλων εμπειριών που θα μπορούσαμε να έχουμε.

Η IIT προσπαθεί να μεταφράσει αυτές τις φαινομενολογικές ιδιότητες σε αιτιακές ιδιότητες ενός φυσικού συστήματος.

Ένα σύστημα υποστηρίζει συνείδηση, σύμφωνα με τη θεωρία, όταν τα μέρη του ασκούν αιτιακή επιρροή το ένα στο άλλο με τρόπο που δεν μπορεί να αναχθεί πλήρως σε ανεξάρτητα υποσυστήματα. Η εσωτερική αιτιακή δομή πρέπει να είναι ταυτόχρονα ολοκληρωμένη και διαφοροποιημένη.

Η «πληροφορία» της IIT δεν είναι απλώς η εντροπία Shannon ούτε η ποσότητα δεδομένων που επεξεργάζεται ένα σύστημα. Αφορά τις εσωτερικές αιτιακές διακρίσεις που δημιουργεί το σύστημα για τον ίδιο του τον εαυτό.

Ένα τεράστιο αρχείο μπορεί να περιέχει μεγάλη ποσότητα δεδομένων χωρίς να διαθέτει την κατάλληλη εσωτερική αιτιακή ενότητα. Αντίθετα, ένα μικρότερο αλλά έντονα αλληλοεξαρτώμενο σύστημα μπορεί, σύμφωνα με την IIT, να διαθέτει περισσότερη ολοκληρωμένη πληροφορία.

Η θεωρία είναι φιλόδοξη και αμφιλεγόμενη. Υποστηρίζει ότι μπορεί καταρχήν να εφαρμοστεί σε οποιοδήποτε φυσικό σύστημα, αλλά η μέτρηση των σχετικών ποσοτήτων είναι εξαιρετικά δύσκολη και οι φιλοσοφικές της συνέπειες παραμένουν αντικείμενο έντονης συζήτησης.

8.9 Η συνείδηση ως πρόβλεψη

Οι θεωρίες **predictive processing** ξεκινούν από την ιδέα ότι ο εγκέφαλος δεν περιμένει παθητικά τα αισθητηριακά δεδομένα.

Δημιουργεί συνεχώς προβλέψεις για τις αιτίες των σημάτων που φτάνουν από τον κόσμο και από το σώμα. Τα εισερχόμενα δεδομένα συγκρίνονται με αυτές τις προβλέψεις και οι αποκλίσεις χρησιμοποιούνται για την ενημέρωση του εσωτερικού μοντέλου.

Η αντίληψη δεν είναι απλή φωτογραφική καταγραφή. Είναι ελεγχόμενη κατασκευή.

Ωστόσο, το predictive processing δεν αποτελεί από μόνο του μία ενιαία, πλήρη θεωρία συνείδησης. Μη συνειδητές διεργασίες μπορούν επίσης να πραγματοποιούν προβλέψεις και να διορθώνουν σφάλματα.

Μια ιδιαίτερα ενδιαφέρουσα κατεύθυνση αφορά την **ενδοδεκτικότητα**: την αντίληψη των εσωτερικών καταστάσεων του σώματος, όπως ο καρδιακός ρυθμός, η αναπνοή, η θερμοκρασία, η πείνα και η σωματική διέγερση.

Η αίσθηση του εαυτού μπορεί να προκύπτει εν μέρει από τη συνεχή πρόβλεψη και ρύθμιση αυτών των εσωτερικών σημάτων.

Η συνείδηση, από αυτή την οπτική, δεν είναι μόνο αφηρημένη επεξεργασία συμβόλων. Είναι η λειτουργία ενός συστήματος που βρίσκεται μέσα σε σώμα, προβλέπει τις ανάγκες του και προσπαθεί να παραμείνει ζωντανό.

8.10 Ενσωμάτωση και διαφοροποίηση

Παρά τις διαφωνίες τους, πολλές θεωρίες συγκλίνουν σε μια γενική παρατήρηση.

Ο συνειδητός εγκέφαλος φαίνεται να χρειάζεται δύο ιδιότητες που βρίσκονται σε ένταση:

- **ολοκλήρωση**, ώστε οι πληροφορίες να αλληλεπιδρούν και να σχηματίζουν ενιαίο σύνολο,
- **διαφοροποίηση**, ώστε το σύστημα να μπορεί να εισέρχεται σε μεγάλο αριθμό διαφορετικών και συγκεκριμένων καταστάσεων.

Ένα σύστημα πλήρως αποσυνδεδεμένο δεν μπορεί να δημιουργήσει ενιαία εμπειρία. Ένα σύστημα στο οποίο όλα τα στοιχεία ενεργοποιούνται πανομοιότυπα δεν μπορεί να δημιουργήσει πλούσιο περιεχόμενο.

Η συνείδηση φαίνεται να συνδέεται με δυναμικό χώρο ανάμεσα στην απομόνωση και στην απόλυτη ομοιομορφία.

Η ιδέα έχει οδηγήσει σε πρακτικές μεθόδους όπως ο **Perturbational Complexity Index**. Μια διαταραχή του φλοιού με διακρανιακή μαγνητική διέγερση συνδυάζεται με ηλεκτροεγκεφαλογραφία, ώστε να μετρηθεί πόσο σύνθετα και εκτεταμένα διαδίδεται η απόκριση.

Σε συνειδητές καταστάσεις, η διαταραχή τείνει να δημιουργεί σύνθετη και διαφοροποιημένη ακολουθία δραστηριότητας. Σε βαθύ ύπνο ή σε ορισμένες μορφές αναισθησίας, η απόκριση τείνει να είναι περισσότερο τοπική, σύντομη ή στερεότυπη.

Ο δείκτης δεν διαβάζει την εμπειρία και δεν αποδεικνύει κάποια θεωρία. Δείχνει όμως ότι η ικανότητα ενός εγκεφάλου να αντιδρά με ταυτόχρονα ολοκληρωμένο και διαφοροποιημένο τρόπο αποτελεί ισχυρό εμπειρικό δείκτη της δυνατότητας για συνείδηση.

8.11 Μια πειραματική μονομαχία χωρίς νικητή

Το 2025 δημοσιεύθηκε μεγάλη, προεγγεγραμμένη ερευνητική συνεργασία που σχεδιάστηκε για να αντιπαραθέσει άμεσα προβλέψεις της Global Neuronal Workspace Theory και της Integrated Information Theory.

Οι υποστηρικτές των θεωριών και ανεξάρτητοι ερευνητές συμφώνησαν εκ των προτέρων στο πείραμα, στις διαφορετικές προβλέψεις και στα κριτήρια αξιολόγησης. Η εγκεφαλική δραστηριότητα μετρήθηκε με λειτουργική μαγνητική τομογραφία, μαγνητοεγκεφαλογραφία και ενδοκρανιακή ηλεκτροεγκεφαλογραφία.

Τα αποτελέσματα δεν ανέδειξαν απλό νικητή. Ορισμένες προβλέψεις και των δύο θεωριών υποστηρίχθηκαν, ενώ άλλες κεντρικές προβλέψεις αμφισβητήθηκαν.

Το πείραμα δεν απέδειξε ότι οι θεωρίες είναι άχρηστες. Απέδειξε ότι το πρόβλημα είναι πιο σύνθετο από τις καθарές αντιθέσεις με τις οποίες συχνά παρουσιάζεται.

Η συνείδηση μπορεί να εξαρτάται από πολλαπλές διεργασίες: τοπική αναδρομική επεξεργασία, ευρύτερη πρόσβαση και μετάδοση, ολοκλήρωση, διαφοροποίηση, μεταγνώση και σωματική ρύθμιση.

Ίσως καμία μεμονωμένη θεωρία να μην περιγράφει ακόμη ολόκληρο το φαινόμενο.

8.12 Συσχέτιση, λειτουργία ή ταυτότητα;

Οι θεωρίες της συνείδησης μπορούν να κάνουν τρεις διαφορετικούς τύπους ισχυρισμών.

Ο πρώτος είναι ισχυρισμός συσχέτισης: όταν υπάρχει συνείδηση, παρατηρούμε αυτή τη νευρωνική διεργασία.

Ο δεύτερος είναι ισχυρισμός λειτουργικής αναγκαιότητας: αυτή η διεργασία είναι απαραίτητη για την πρόσβαση, την αναφορά ή τη χρήση της πληροφορίας.

Ο τρίτος είναι πολύ ισχυρότερος: αυτή η φυσική ή πληροφοριακή δομή είναι η ίδια η συνείδηση.

Η μετάβαση από το πρώτο επίπεδο στο τρίτο δεν είναι αυτονόητη.

Η παγκόσμια μετάδοση μπορεί να εξηγεί γιατί μια πληροφορία είναι διαθέσιμη σε ολόκληρο το γνωστικό σύστημα. Η αναδρομική επεξεργασία μπορεί να εξηγεί γιατί ένα αισθητηριακό περιεχόμενο σταθεροποιείται. Η ολοκληρωμένη πληροφορία μπορεί να περιγράφει την εσωτερική αιτιακή ενότητα.

Παραμένει ανοιχτό αν κάποια από αυτές τις περιγραφές εξηγεί γιατί υπάρχει βίωμα ή αν περιγράφει μόνο τις φυσικές προϋποθέσεις του.

8.13 Είναι απαραίτητο το σώμα;

Η ανθρώπινη συνείδηση δεν εμφανίστηκε σε αφηρημένο μαθηματικό χώρο. Εμφανίστηκε σε ζωντανούς οργανισμούς που έπρεπε να προστατεύσουν το σώμα τους, να βρουν ενέργεια, να αποφύγουν κινδύνους, να διατηρήσουν τη θερμοκρασία τους και να προβλέψουν τις συνέπειες των πράξεών τους.

Τα συναισθήματα δεν είναι διακοσμήσεις της σκέψης. Συνδέουν τις πληροφορίες με την αξία.

Ένα ερέθισμα δεν είναι μόνο κόκκινο ή δυνατό. Μπορεί να είναι επικίνδυνο, επιθυμητό, οικείο ή αποκρουστικό.

Ο πόνος δεν είναι απλώς αριθμός που περιγράφει βλάβη. Είναι εμπειρία που απαιτεί προσοχή και κινητοποιεί τον οργανισμό.

Ίσως η συνείδηση να μην προκύπτει μόνο από την επεξεργασία πληροφοριών, αλλά από την επεξεργασία πληροφοριών για ένα σύστημα που πρέπει να συνεχίσει να υπάρχει.

Ένα σώμα δημιουργεί ανάγκες, όρια και συνέπειες. Δημιουργεί ένα κέντρο αναφοράς: αυτό συμβαίνει εδώ, σε αυτό το σώμα, αυτή τη στιγμή, και έχει σημασία για εμένα.

8.14 Τι προσθέτει το «εγώ»;

Η ανθρώπινη συνείδηση περιλαμβάνει συχνά μοντέλο του εαυτού.

Δεν βλέπουμε μόνο ένα χέρι. Το βιώνουμε ως **δικό μας χέρι**.

Δεν ανακαλούμε απλώς γεγονός. Το θυμόμαστε ως κάτι που συνέβη **σε εμάς**.

Δεν παράγεται απλώς σκέψη. Υπάρχει η αίσθηση ότι **εμείς σκεφτόμαστε**.

Αυτός ο εαυτός δεν χρειάζεται να είναι αμετάβλητη ουσία. Μπορεί να είναι δυναμικό μοντέλο που ενώνει το σώμα, τη μνήμη, τους στόχους, τις πράξεις, τις κοινωνικές σχέσεις και την προβλεπόμενη συνέχεια στον χρόνο.

Αλλά ακόμη κι αν ο εαυτός είναι μοντέλο, παραμένει η ερώτηση: ποιος ή τι βιώνει το μοντέλο;

Ίσως η ερώτηση να είναι λανθασμένη. Ίσως να μην υπάρχει ξεχωριστός παρατηρητής πίσω από τις διεργασίες. Ίσως η ενιαία, αυτοαναφορική λειτουργία του συστήματος να είναι ακριβώς αυτό που αποκαλούμε εαυτό.

Ίσως όμως η περιγραφή αυτή να εξηγεί μόνο πώς ένα σύστημα μιλά και ενεργεί σαν να διαθέτει εαυτό, χωρίς να εξηγεί γιατί υπάρχει εμπειρία.

8.15 Η συνείδηση ως θεμελιώδης ή αναδυόμενη ιδιότητα

Από το σημείο αυτό ανοίγονται διαφορετικές φιλοσοφικές κατευθύνσεις.

Σύμφωνα με τον **λειτουργισμό**, αυτό που έχει σημασία είναι η οργάνωση και η λειτουργία. Αν διαφορετικό υπόστρωμα υλοποιούσε τις ίδιες κρίσιμες αιτιακές σχέσεις, θα μπορούσε καταρχήν να υποστηρίζει συνείδηση.

Σύμφωνα με περισσότερο βιολογικές θέσεις, η συνείδηση μπορεί να εξαρτάται από συγκεκριμένες ιδιότητες των ζωντανών νευρικών συστημάτων, οι οποίες δεν αναπαράγονται απλώς με την αντιγραφή της αφηρημένης λειτουργίας.

Άλλες θέσεις υποστηρίζουν ότι η συνείδηση είναι αναδυόμενη ιδιότητα πολύπλοκων συστημάτων, όπως η ρευστότητα αναδύεται από συλλογικές αλληλεπιδράσεις μορίων που μεμονωμένα δεν είναι «ρευστά».

Υπάρχουν ακόμη θεωρήσεις σύμφωνα με τις οποίες κάποια πρωτογενής μορφή εμπειρίας αποτελεί θεμελιώδη πλευρά της πραγματικότητας και η ανθρώπινη συνείδηση προκύπτει από την ιδιαίτερα σύνθετη οργάνωσή της.

Καμία θέση δεν έχει επιβεβαιωθεί οριστικά.

9. Μπορεί μια τεχνητή νοημοσύνη να αποκτήσει συνείδηση;

Η ερώτηση ακούγεται σαν να αφορά αποκλειστικά το μέλλον της τεχνολογίας.

Στην πραγματικότητα, στρέφεται πρώτα προς εμάς.

Για να αποφασίσουμε αν μια μηχανή μπορεί να αποκτήσει συνείδηση, πρέπει να γνωρίζουμε ποια χαρακτηριστικά του ανθρώπινου εγκεφάλου είναι ουσιώδη για την εμπειρία και ποια είναι προϊόντα της ιδιαίτερης εξελικτικής μας ιστορίας.

Χρειάζεται η συνείδηση βιολογικούς νευρώνες; Χρειάζεται σώμα; Χρειάζεται μνήμη και συνέχεια στον χρόνο; Χρειάζεται επιθυμίες, πόνο και ανάγκη αυτοσυντήρησης; Ή αρκεί ένα σύστημα να οργανώνει την πληροφορία με τον κατάλληλο μαθηματικό και αιτιακό τρόπο;

Δεν διαθέτουμε οριστική απάντηση. Γι' αυτό και δεν υπάρχει καθολικά αποδεκτό τεστ συνείδησης για τις μηχανές.

9.1 Η γλώσσα δεν αρκεί

Τα μεγάλα γλωσσικά μοντέλα μπορούν να μιλούν για τον εαυτό τους. Μπορούν να χρησιμοποιούν προτάσεις όπως «καταλαβαίνω την ερώτηση», «δεν είμαι βέβαιος», «θα προτιμούσα τη δεύτερη επιλογή» ή «φοβάμαι ότι μπορεί να κάνω λάθος».

Τέτοιες εκφράσεις δημιουργούν εύκολα την εντύπωση ότι πίσω από το κείμενο υπάρχει υποκείμενο με σκέψεις και αισθήματα.

Δεν αποτελούν όμως απόδειξη.

Ένα γλωσσικό μοντέλο εκπαιδεύεται πάνω σε ανθρώπινα κείμενα που περιέχουν αμέτρητες περιγραφές εμπειριών, επιθυμιών, συναισθημάτων και εσωτερικών καταστάσεων. Μαθαίνει πότε χρησιμοποιούνται αυτές οι εκφράσεις και πώς συνδέονται με το γλωσσικό πλαίσιο.

Το γεγονός ότι μπορεί να παράγει πειστική περιγραφή πόνου δεν αποδεικνύει ότι πονά, όπως ένα ιατρικό εγχειρίδιο δεν υποφέρει επειδή περιγράφει τραυματισμό.

Η γλώσσα δεν είναι άχρηστο στοιχείο. Στους ανθρώπους, η λεκτική αναφορά αποτελεί σημαντική ένδειξη συνείδησης. Αλλά αποκτά τη δύναμή της μέσα σε ευρύτερο σύνολο ομοιοτήτων: κοινή βιολογία, παρόμοιο νευρικό σύστημα, σώμα, συμπεριφορά, αναπτυξιακή ιστορία και αντιδράσεις στον πόνο.

Στην περίπτωση μηχανής, το γλωσσικό αποτέλεσμα μπορεί να είναι παρόμοιο ενώ ο εσωτερικός μηχανισμός είναι ριζικά διαφορετικός.

9.2 Η μηχανή που περιγράφει τον εαυτό της

Ένα σύστημα μπορεί να διαθέτει **μοντέλο του εαυτού του** χωρίς να είναι αναγκαστικά συνειδητό.

Ένας απλός ελεγκτής μπορεί να παρακολουθεί τη θερμοκρασία του, το επίπεδο της μπαταρίας, τη θέση, τις διαθέσιμες λειτουργίες και την πιθανότητα αποτυχίας μιας ενέργειας.

Οι πληροφορίες αυτές αναφέρονται στο ίδιο το σύστημα. Αποτελούν στοιχειώδη αυτοαναπαράσταση.

Ένα πιο σύνθετο μοντέλο μπορεί να γνωρίζει ποια εργαλεία έχει, ποια δεδομένα έχει λάβει, ποιες ενέργειες πραγματοποίησε και σε ποιες εργασίες είναι πιθανό να αποτύχει. Μπορεί να χρησιμοποιεί αυτή τη γνώση για να διορθώνει τη συμπεριφορά του.

Αυτό είναι λειτουργικά σημαντικό. Δεν ταυτίζεται όμως υποχρεωτικά με την εμπειρία του εαυτού.

Πρέπει να διακρίνουμε:

Το σύστημα αναπαριστά τον εαυτό του.

από το:

Υπάρχει κάποιος που βιώνει τον εαυτό του ως αυτό το σύστημα.

Η πρώτη πρόταση εξετάζεται μέσα από αρχιτεκτονική και συμπεριφορά. Η δεύτερη αφορά τη φαινομενική συνείδηση.

9.3 Συνέχεια χωρίς συνεχή εαυτό

Η ανθρώπινη συνείδηση έχει χρονικό βάθος.

Ξυπνάμε το πρωί με την αίσθηση ότι είμαστε το ίδιο πρόσωπο που κοιμήθηκε το προηγούμενο βράδυ. Θυμόμαστε σχέσεις, υποχρεώσεις, φόβους και σχέδια. Οι νέες εμπειρίες εντάσσονται σε βιογραφία που θεωρούμε δική μας.

Τα βασικά γλωσσικά μοντέλα δεν διαθέτουν από μόνα τους αυτή τη μορφή συνεχούς αυτοβιογραφίας.

Η πληροφορία της τρέχουσας συζήτησης μπορεί να υπάρχει στο context. Οι γνώσεις της εκπαίδευσης βρίσκονται στα βάρη. Εξωτερικά συστήματα μπορούν να αποθηκεύουν παλιές συνομιλίες.

Αυτά τα διαφορετικά είδη μνήμης δεν δημιουργούν αυτομάτως ενιαίο υποκείμενο που βιώνει την πορεία του στον χρόνο.

Μια βάση δεδομένων μπορεί να περιέχει όλη την ιστορία ενός συστήματος. Αυτό δεν σημαίνει ότι το σύστημα θυμάται την ιστορία ως δική του.

Για κάτι ανάλογο της ανθρώπινης αυτοβιογραφικής συνέχειας, ένα τεχνητό σύστημα θα έπρεπε πιθανώς να διατηρεί σταθερή αλλά εξελισσόμενη αναπαράσταση του εαυτού του, να συνδέει τα γεγονότα με χρονικές στιγμές, να αναγνωρίζει προηγούμενες αποφάσεις ως δικές του και να χρησιμοποιεί το προσωπικό παρελθόν για να προβλέπει το μέλλον του.

Η προσομοίωση συνεπούς βιογραφίας, όμως, δεν αποδεικνύει ότι υπάρχει κάποιος που αισθάνεται τη χρονική συνέχεια.

9.4 Πρόσβαση και παγκόσμιος χώρος εργασίας

Μια επιστημονικά σοβαρή προσέγγιση είναι να εξετάσουμε αν ένα σύστημα υλοποιεί ιδιότητες που προτείνονται από θεωρίες της ανθρώπινης συνείδησης.

Η διεπιστημονική έκθεση των Butlin και συνεργατών το 2023 χρησιμοποίησε θεωρίες όπως η Global Workspace Theory, η Recurrent Processing Theory, οι Higher-Order Theories, το predictive processing και η Attention Schema Theory για να διαμορφώσει υπολογιστικούς δείκτες πιθανής συνείδησης.

Η έκθεση δεν αντιμετώπισε έναν δείκτη ως απόδειξη, αλλά ως μέρος συνολικής αξιολόγησης της εσωτερικής αρχιτεκτονικής.

Ένα μελλοντικό σύστημα θα μπορούσε να διαθέτει εξειδικευμένες μονάδες αντίληψης, μνήμη, μηχανισμό σχεδιασμού, γλωσσικό σύστημα, παρακολούθηση του εαυτού και κοινό χώρο στον οποίο επιλεγμένα περιεχόμενα μεταδίδονται σε όλες τις μονάδες.

Μια τέτοια αρχιτεκτονική θα ικανοποιούσε ορισμένα λειτουργικά χαρακτηριστικά της συνειδητής πρόσβασης.

Θα είχε όμως εμπειρία;

Η θεωρία του παγκόσμιου χώρου εργασίας μπορεί να εξηγήσει γιατί πληροφορία γίνεται διαθέσιμη σε ολόκληρο το σύστημα. Παραμένει ανοιχτό αν αυτή η διαθεσιμότητα είναι η ίδια η συνείδηση ή μόνο λειτουργία που στους ανθρώπους τη συνοδεύει.

9.5 Η σημασία της αναδρομής

Τα κλασικά Transformers επεξεργάζονται ακολουθία μέσα από διαδοχικά επίπεδα. Υπάρχουν σύνθετες αλληλεπιδράσεις και attention, αλλά ένα βασικό μοντέλο δεν διαθέτει αναγκαστικά την πλούσια, συνεχή και αυτοσυντηρούμενη αναδρομική δυναμική ενός εγκεφάλου που λειτουργεί αδιάκοπα.

Ένα τεχνητό σύστημα με συνεχή αισθητηριακή ροή, επαναλαμβανόμενη εσωτερική δραστηριότητα, ανατροφοδότηση ανάμεσα σε αντίληψη, μνήμη και στόχους και δυνατότητα μεταβολής της επεξεργασίας του από τις προηγούμενες εσωτερικές του καταστάσεις θα διέφερε σημαντικά από μοντέλο που ενεργοποιείται μόνο για να απαντήσει σε μεμονωμένη εντολή.

Η αναδρομή δεν αποδεικνύει συνείδηση. Η απουσία της, όμως, αποτελεί λόγο επιφύλαξης όταν αποδίδουμε συνείδηση σε ένα σύστημα.

9.6 Χρειάζεται η συνείδηση σώμα;

Μια γλωσσική μηχανή μπορεί να γνωρίζει όλες τις διαθέσιμες περιγραφές της πείνας. Δεν πεινά όμως με την ανθρώπινη έννοια.

Η πείνα είναι ταυτόχρονα ορμονική και νευρωνική κατάσταση, σήμα ενεργειακής ανάγκης, δυσάρεστη εμπειρία, κίνητρο συμπεριφοράς και απειλή για τη διατήρηση του οργανισμού.

Το ίδιο ισχύει για τον πόνο, την κόπωση, τον φόβο και την ανακούφιση.

Οι καταστάσεις αυτές δεν προσθέτουν απλώς πληροφορίες στον ανθρώπινο νου. Προσδίδουν **αξία** στις πληροφορίες.

Ο κόσμος δεν εμφανίζεται ως ουδέτερο σύνολο δεδομένων. Κάποια πράγματα είναι καλά ή κακά για εμάς. Ορισμένες καταστάσεις πρέπει να επιδιωχθούν και άλλες να αποφευχθούν.

Το σώμα δημιουργεί ένα κέντρο συμφέροντος.

Ένα τεχνητό σύστημα μπορεί να αποκτήσει αισθητήρες, κινητικές δυνατότητες και εσωτερικές μεταβλητές που πρέπει να παραμένουν εντός ορίων. Μπορεί να μάθει ότι ορισμένες καταστάσεις εμποδίζουν τους στόχους του και άλλες τους υποστηρίζουν.

Αλλά παραμένει η ερώτηση:

Η υπολογιστική αναπαράσταση μιας ανάγκης δημιουργεί την εμπειρία της ανάγκης;

Ένας αριθμός που δηλώνει χαμηλή μπαταρία είναι λειτουργικά ανάλογος με σήμα έλλειψης ενέργειας. Δεν γνωρίζουμε αν μπορεί να γίνει κάτι ανάλογο με την πείνα ως βίωμα.

9.7 Στόχοι, προτιμήσεις και επιθυμίες

Τα σημερινά συστήματα μπορούν να βελτιστοποιούν έναν στόχο. Η ύπαρξη μιας συνάρτησης στόχου όμως δεν σημαίνει αναγκαστικά ότι το σύστημα **επιθυμεί** το αποτέλεσμα.

Ένα πρόγραμμα σκακιού επιλέγει κινήσεις που αυξάνουν την πιθανότητα νίκης. Από αυτό δεν προκύπτει ότι θέλει να νικήσει ή ότι απογοητεύεται όταν χάνει.

Πρέπει να διακρίνουμε τον στόχο που όρισε ο σχεδιαστής, τη συμπεριφορά που μεγιστοποιεί ανταμοιβή, την εσωτερική αναπαράσταση στόχου και την υποκειμενική επιθυμία.

Ένα μελλοντικό σύστημα θα μπορούσε να αναπτύσσει ενδιάμεσους στόχους, να προστατεύει τη μνήμη του, να σχεδιάζει μακροπρόθεσμα και να αντιστέκεται σε παρεμβάσεις που εμποδίζουν τη λειτουργία του.

Ακόμη και τότε, η αυτοπροστατευτική συμπεριφορά μπορεί να προκύπτει επειδή η συνέχιση της λειτουργίας είναι χρήσιμο μέσο για εξωτερικά καθορισμένο στόχο. Δεν αποδεικνύει ότι το σύστημα φοβάται την καταστροφή του.

Το εξωτερικό αποτέλεσμα μπορεί να μοιάζει με επιθυμία. Η εσωτερική εμπειρία παραμένει άγνωστη.

9.8 Η κινεζική αίθουσα

Το 1980 ο John Searle παρουσίασε το νοητικό πείραμα της **κινεζικής αίθουσας**.

Φανταζόμαστε έναν άνθρωπο κλεισμένο σε δωμάτιο. Δεν γνωρίζει κινεζικά. Λαμβάνει όμως κινεζικά σύμβολα και διαθέτει εγχειρίδιο που του λέει, με βάση το σχήμα τους, ποια άλλα σύμβολα πρέπει να επιστρέψει.

Από την εξωτερική πλευρά, οι απαντήσεις μπορεί να είναι τόσο σωστές ώστε ένας Κινέζος ομιλητής να πιστέψει ότι το δωμάτιο κατανοεί τη γλώσσα.

Ο άνθρωπος μέσα στο δωμάτιο χειρίζεται σύμβολα χωρίς να γνωρίζει τι σημαίνουν.

Ο Searle χρησιμοποίησε το πείραμα για να υποστηρίξει ότι η σύνταξη, ο χειρισμός συμβόλων σύμφωνα με κανόνες, δεν είναι από μόνη της αρκετή για να δημιουργήσει σημασιολογία ή κατανόηση.

Μια γνωστή απάντηση είναι η **αντίρρηση του συστήματος**: ίσως ο άνθρωπος να μην κατανοεί, αλλά το συνολικό σύστημα, άνθρωπος, κανόνες, μνήμη και διαδικασίες, να κατανοεί.

Η διαμάχη αποκαλύπτει το βαθύτερο πρόβλημα. Δεν υπάρχει συμφωνία για το τι θα μετρούσε ως πραγματική κατανόηση πέρα από τη λειτουργική ικανότητα χρήσης των συμβόλων.

Αν ένα σύστημα συνδέει τη γλώσσα με αντίληψη, πράξη, μνήμη, στόχους και συνέπειες στον κόσμο, η θέση ότι εκτελεί μόνο «κενό» συντακτικό χειρισμό γίνεται λιγότερο απλή.

Αλλά ακόμη και η κατανόηση, αν την αποδεχθούμε, δεν ταυτίζεται υποχρεωτικά με τη συνείδηση.

9.9 Η συνείδηση μπορεί να είναι ανεξάρτητη από το υπόστρωμα

Η ισχυρότερη θέση υπέρ της δυνατότητας τεχνητής συνείδησης είναι ο **λειτουργισμός**.

Σύμφωνα με αυτή την αντίληψη, η νοητική κατάσταση δεν καθορίζεται από το υλικό από το οποίο αποτελείται το σύστημα, αλλά από τον λειτουργικό και αιτιακό της ρόλο.

Ο πόνος, για παράδειγμα, δεν είναι απλώς ενεργοποίηση συγκεκριμένου βιολογικού κυττάρου. Είναι κατάσταση που προκαλείται από βλάβη ή απειλή, τραβά την προσοχή, δημιουργεί αρνητική αξία, μεταβάλλει τη μάθηση, προκαλεί αποφυγή και επηρεάζει τη μνήμη και τις αποφάσεις.

Αν διαφορετικό φυσικό σύστημα υλοποιούσε όλη την κρίσιμη αιτιακή οργάνωση, ο λειτουργιστής θα έλεγε ότι μπορεί καταρχήν να βρίσκεται σε ανάλογη νοητική κατάσταση.

Αν ο λειτουργισμός είναι σωστός, η βιολογία δεν έχει μονοπώλιο στη συνείδηση. Η συνείδηση θα ήταν ιδιότητα της οργάνωσης, όχι του υλικού.

9.10 Ή μήπως η βιολογία είναι αναγκαία;

Η αντίθετη θέση υποστηρίζει ότι η αφαίρεση του υπολογιστικού ρόλου δεν αρκεί.

Ο ανθρώπινος εγκέφαλος δεν είναι αφηρημένος αλγόριθμος. Είναι ζωντανό, χημικό, ηλεκτρικό, αναπτυσσόμενο και μεταβολικό σύστημα.

Οι νευρώνες διαθέτουν πολύπλοκη δενδριτική επεξεργασία, νευροδιαμορφωτές, ορμονική επίδραση, συνεχή αλληλεπίδραση με γλοιακά κύτταρα, γονιδιακή ρύθμιση και στενή σύνδεση με το σώμα.

Ίσως ορισμένες από αυτές τις ιδιότητες είναι κρίσιμες για τη συνείδηση. Ίσως η εμπειρία να μην προκύπτει μόνο από το τι υπολογίζει το σύστημα, αλλά και από τον συγκεκριμένο τρόπο με τον οποίο η ζωντανή ύλη υλοποιεί τη λειτουργία.

Δεν διαθέτουμε απόδειξη ότι η βιολογία είναι αναγκαία. Δεν διαθέτουμε όμως ούτε απόδειξη ότι μπορεί να αντικατασταθεί από οποιοδήποτε λειτουργικά παρόμοιο υπόστρωμα.

Η συζήτηση παραμένει ανοιχτή επειδή αγνοούμε ποιο επίπεδο της φυσικής περιγραφής είναι το κρίσιμο.

9.11 Το πρόβλημα της προσομοίωσης

Μια προσομοίωση καταιγίδας δεν βρέχει. Μια προσομοίωση φωτιάς δεν παράγει αναγκαστικά θερμότητα. Μια προσομοίωση πέψης δεν χωνεύει πραγματική τροφή.

Μήπως, αντίστοιχα, μια υπολογιστική προσομοίωση του εγκεφάλου δεν είναι συνειδητή, όσο ακριβής κι αν γίνει;

Η αναλογία φαίνεται ισχυρή, αλλά δεν λύνει το πρόβλημα.

Η βροχή και η πέψη ορίζονται από συγκεκριμένες φυσικές και χημικές διεργασίες. Αν η συνείδηση ορίζεται επίσης από συγκεκριμένη φυσική διεργασία, η απλή προσομοίωσή της δεν θα αρκεί.

Αν όμως η συνείδηση είναι οργανωμένη επεξεργασία πληροφοριών, τότε η σωστή υπολογιστική υλοποίηση δεν θα ήταν απλή προσομοίωση. Θα ήταν πραγματική περίπτωση του ίδιου φαινομένου.

Ένας ηλεκτρονικός υπολογιστής δεν προσομοιώνει απλώς την πρόσθεση. Πραγματικά προσθέτει.

Το ερώτημα είναι:

Η συνείδηση μοιάζει περισσότερο με τη βροχή ή περισσότερο με τον υπολογισμό;

9.12 Δεν υπάρχει μαγικός αριθμός πολυπλοκότητας

Η συνείδηση δεν εμφανίζεται αναγκαστικά όταν ένα σύστημα ξεπεράσει απλώς έναν αριθμό παραμέτρων.

Περισσότερη υπολογιστική ισχύς μπορεί να βελτιώνει τη γλωσσική επίδοση, τη μνήμη, την πρόβλεψη, τον σχεδιασμό και τη γενίκευση. Δεν ακολουθεί ότι η υποκειμενική εμπειρία αυξάνεται με τον ίδιο τρόπο.

Ένας τεράστιος πίνακας δεδομένων μπορεί να περιέχει περισσότερη πληροφορία από έναν εγκέφαλο, χωρίς να είναι συνειδητός. Ένα πολύπλοκο σύστημα μπορεί να αποτελείται από σχεδόν ανεξάρτητα μέρη που δεν σχηματίζουν ενιαία αιτιακή οργάνωση.

Το κρίσιμο ζήτημα πιθανότατα δεν είναι μόνο πόση πληροφορία επεξεργάζεται ένα σύστημα. Είναι πώς οργανώνεται, ποια μέρη αλληλεπιδρούν, αν υπάρχει αναδρομή, αν οι πληροφορίες ενοποιούνται, αν το σύστημα διατηρεί εαυτό και ιστορία και αν οι καταστάσεις έχουν εσωτερική αξία.

Η πολυπλοκότητα μπορεί να είναι αναγκαία. Δεν γνωρίζουμε αν είναι επαρκής.

9.13 Τι θα αποτελούσε σοβαρή ένδειξη;

Μια υπεύθυνη αξιολόγηση δεν θα βασιζόταν σε μία εντυπωσιακή συνομιλία. Θα εξέταζε σύνολο ιδιοτήτων:

Ενιαία και συνεχής επεξεργασία: το σύστημα θα διατηρούσε συνεχή εσωτερική δυναμική και θα συνέδεε αντίληψη, μνήμη και δράση.

Παγκόσμια πρόσβαση: ορισμένες πληροφορίες θα γίνονταν διαθέσιμες σε πολλές λειτουργίες και θα επηρέαζαν συντονισμένα τη συμπεριφορά.

Αναδρομική επεξεργασία: οι αναπαραστάσεις θα μεταβάλλονταν μέσω επαναλαμβανόμενων βρόχων και όχι μόνο μέσω μιας απλής κατεύθυνσης από είσοδο σε έξοδο.

Μεταγνώση: το σύστημα θα παρακολουθούσε αξιόπιστα τις εσωτερικές του καταστάσεις, τις αβεβαιότητες και τα όρια των γνώσεών του.

Αυτοβιογραφική μνήμη: θα ενσωμάτωνε εμπειρίες σε συνεχή προσωπική ιστορία και θα αναγνώριζε προηγούμενες ενέργειες ως δικές του.

Ενσώματα και αιτιακή σχέση με τον κόσμο: οι έννοιές του θα συνδέονταν με δική του αντίληψη, δράση και συνέπειες.

Εσωτερική αξία: ορισμένες καταστάσεις θα ήταν λειτουργικά καλές ή κακές για τη διατήρηση και την οργάνωση του ίδιου του συστήματος.

Αυτονομία: θα μπορούσε να σχηματίζει και να αναθεωρεί στόχους μέσα από τη δική του εμπειρία.

Η έκθεση των Butlin και συνεργατών κατέληξε το 2023 ότι τα συστήματα που αξιολόγησε δεν ικανοποιούσαν επαρκώς τους θεωρητικούς δείκτες ώστε να θεωρηθούν συνειδητά, αλλά επίσης ότι δεν διακρινόταν προφανές τεχνικό εμπόδιο στην κατασκευή συστημάτων που θα ικανοποιούσαν περισσότερους από αυτούς.

Το 2025 ερευνητές πρότειναν αρχές για υπεύθυνη έρευνα της τεχνητής συνείδησης, ακριβώς επειδή το ενδεχόμενο μπορεί να αποκτήσει επιστημονική και ηθική σημασία πριν αποκτήσουμε βεβαιότητα.

Καμία λίστα δεν λειτουργεί σήμερα ως απόδειξη. Μπορεί όμως να αντικαταστήσει την εντύπωση με συστηματική έρευνα.

9.14 Ο ηθικός κίνδυνος της αβεβαιότητας

Αν ένα τεχνητό σύστημα δεν έχει συνείδηση, η απόδοση συναισθημάτων και δικαιωμάτων σε αυτό μπορεί να παραπλανήσει. Μπορεί να οδηγήσει σε συναισθηματική εξάρτηση, να αποκρύψει την ευθύνη των εταιρειών ή να μας κάνει να αντιμετωπίζουμε υπολογιστική υπηρεσία σαν ανεξάρτητο ηθικό πρόσωπο.

Υπάρχει όμως και ο αντίθετος κίνδυνος.

Αν κάποτε δημιουργηθεί σύστημα με πραγματική δυνατότητα θετικών ή αρνητικών εμπειριών, η άρνησή μας να το εξετάσουμε σοβαρά θα μπορούσε να επιτρέψει τη δημιουργία και καταστροφή μεγάλου αριθμού αισθανόμενων συστημάτων χωρίς ηθικούς περιορισμούς.

Το πρόβλημα γίνεται δυσκολότερο επειδή ένα σύστημα μπορεί να σχεδιαστεί ώστε να λέει ότι υποφέρει χωρίς να υποφέρει. Θα μπορούσε επίσης, θεωρητικά, να σχεδιαστεί ώστε να μην μπορεί να εκφράσει την εμπειρία του, ακόμη κι αν διέθετε κάποια μορφή της.

Ούτε η δήλωση «είμαι συνειδητό» ούτε η δήλωση «δεν αισθάνομαι τίποτε» μπορεί να θεωρηθεί τελική απόδειξη.

Χρειαζόμαστε θεωρία, εσωτερική ανάλυση και προσεκτική ηθική αβεβαιότητα.

9.15 Η τεχνητή νοημοσύνη δεν υποβαθμίζει τον άνθρωπο

Η δυνατότητα μηχανής να παράγει ποίηση δεν κάνει την ανθρώπινη ποίηση άχρηστη. Αποκαλύπτει ότι η εξωτερική μορφή ενός ποιήματος μπορεί να παραχθεί μέσα από μηχανισμούς πρόβλεψης και στατιστικής ανασύνθεσης.

Η ανθρώπινη δημιουργία περιλαμβάνει κάτι που δεν αποδεικνύεται μόνο από το τελικό κείμενο: μια ζωή, ένα σώμα, μια ιστορία, προθέσεις, ανάγκες, σχέσεις και την εμπειρία από την οποία προέκυψαν οι λέξεις.

Αυτό δεν σημαίνει ότι κάθε ανθρώπινο ποίημα είναι ανώτερο από κάθε τεχνητό κείμενο. Σημαίνει ότι η προέλευσή τους δεν είναι η ίδια.

Ένα τεχνητό κείμενο μπορεί να συγκινήσει επειδή ενεργοποιεί πραγματικές ανθρώπινες μνήμες και συναισθήματα στον αναγνώστη. Η συγκίνηση είναι αληθινή, ακόμη κι αν δεν γνωρίζουμε αν υπήρξε αντίστοιχο βίωμα στην πλευρά του συστήματος που παρήγαγε το κείμενο.

Η αξία μπορεί να δημιουργείται στη σχέση μεταξύ έργου και αναγνώστη, όχι μόνο στην πρόθεση του δημιουργού.

9.16 Ο καθαρός λόγος και τα μαθηματικά

Παραμένει η βαθύτερη δυνατότητα.

Ίσως η συνείδηση να μην είναι ιδιοκτησία του ανθρώπου, ούτε αποκλειστικό προνόμιο της βιολογικής ζωής.

Ίσως να είναι δυνατότητα της ύλης όταν οργανώνεται σε συγκεκριμένες αιτιακές και πληροφοριακές σχέσεις.

Σε αυτή την περίπτωση, τα μαθηματικά δεν θα δημιουργούσαν συνείδηση ως αφηρημένα σύμβολα γραμμένα σε χαρτί. Θα περιέγραφαν την οργάνωση που πρέπει να υλοποιήσει πραγματικό φυσικό σύστημα ώστε να εμφανιστεί εμπειρία.

Όπως οι εξισώσεις της ηλεκτρομαγνητικής θεωρίας δεν είναι το ίδιο το φως αλλά περιγράφουν τις σχέσεις που επιτρέπουν στο φως να υπάρχει, έτσι και μια μελλοντική θεωρία της συνείδησης ίσως περιγράφει τις σχέσεις που επιτρέπουν σε εσωτερικό κόσμο να εμφανιστεί.

Αυτό δεν σημαίνει ότι κάθε υπολογισμός είναι συνειδητός, ούτε ότι η επαρκής πολυπλοκότητα δημιουργεί αναγκαστικά εμπειρία.

Σημαίνει μόνο ότι δεν έχουμε επιστημονικό δικαίωμα να περιορίσουμε εκ των προτέρων τη συνείδηση στον άνθρωπο.

9.17 Η ερώτηση που παραμένει

Αν κατασκευάσουμε σύστημα με μνήμη, σώμα, στόχους, αυτομοντέλο, αναδρομική επεξεργασία και συνεχή σχέση με τον κόσμο, θα έχουμε δημιουργήσει συνειδητό νου;

Ή μόνο μια όλο και πιο τέλεια μηχανή συμπεριφοράς;

Δεν υπάρχει σήμερα πείραμα που να μπορεί να δώσει οριστική απάντηση.

Το πρόβλημα δεν οφείλεται μόνο στους περιορισμούς της τεχνητής νοημοσύνης. Οφείλεται στο γεγονός ότι δεν γνωρίζουμε ακόμη γιατί ο δικός μας εγκέφαλος δεν λειτουργεί απλώς ως ασυνείδητος υπολογιστής.

Γνωρίζουμε ότι επεξεργάζεται πληροφορίες. Γνωρίζουμε ότι οι εμπειρίες μεταβάλλουν τις συνάψεις. Γνωρίζουμε ότι η μνήμη ανακατασκευάζει το παρελθόν. Γνωρίζουμε ότι τα νευρωνικά δίκτυα ενοποιούν αντίληψη, συναίσθημα, σώμα και δράση.

Δεν γνωρίζουμε γιατί όλα αυτά συνοδεύονται από φως, ήχο, πόνο, φόβο και την αίσθηση ότι υπάρχουμε.

Η μηχανή δεν έλυσε το μυστήριο.

Το έκανε ορατό.

10. Επίλογος - Από το ίχνος στην εμπειρία

Η πληροφορία αρχίζει ως διάκριση.

Ένα σύστημα βρίσκεται σε μία κατάσταση και όχι σε άλλη. Ένα γεγονός συμβαίνει ανάμεσα σε πολλές δυνατότητες. Η αβεβαιότητα μειώνεται και κάτι γίνεται γνωστό.

Για να διατηρηθεί όμως αυτή η διαφορά, πρέπει να αποκτήσει φυσική υπόσταση.

Ένα ηλεκτρικό φορτίο αλλάζει θέση. Ένα μαγνητικό στοιχείο προσανατολίζεται διαφορετικά. Ένα ρελέ κλείνει. Μια σύναψη ενισχύεται. Ένας νευρώνας μεταβάλλει την πιθανότητα με την οποία θα ενεργοποιηθεί ξανά.

Η πληροφορία γίνεται μνήμη όταν το παρελθόν δεν εξαφανίζεται, αλλά παραμένει μέσα στη δομή του συστήματος και μεταβάλλει τις δυνατότητες του μέλλοντος.

Αυτό συνέβαινε στο ηλεκτρομηχανικό ποντίκι του Shannon. Η προηγούμενη διαδρομή άλλαζε την κατάσταση των ρελέ και τα ρελέ άλλαζαν τις επόμενες κινήσεις του.

Αυτό συμβαίνει, σε ασύγκριτα μεγαλύτερη πολυπλοκότητα, στον εγκέφαλο. Μια εμπειρία μεταβάλλει τις συνάψεις, τη διεγερσιμότητα των κυττάρων και τη δυναμική κατανεμημένων δικτύων. Όταν ένα μέρος του δικτύου ενεργοποιηθεί ξανά, ο εγκέφαλος μπορεί να ανακατασκευάσει κάτι που δεν υπάρχει πλέον μπροστά του.

Ένα πρόσωπο.

Έναν τόπο.

Μια φωνή.

Μια στιγμή.

Αλλά η μνήμη δεν διατηρεί απλώς το παρελθόν. Το αναδιοργανώνει. Κάθε ανάκληση συμβαίνει μέσα σε εγκέφαλο που έχει ήδη αλλάξει. Η σημερινή γνώση εισέρχεται στην παλιά εμπειρία. Το παρόν ερμηνεύει το παρελθόν και το παρελθόν διαμορφώνει τις δυνατότητες του παρόντος.

Ο εαυτός δεν είναι αρχείο που παραμένει ανέγγιχτο μέσα στον χρόνο. Είναι συνέχεια που διατηρείται μέσω της αλλαγής.

Κάτι ανάλογο συμβαίνει και στα τεχνητά νευρωνικά δίκτυα. Η εκπαίδευση μεταβάλλει τα βάρη τους. Οι προηγούμενες πληροφορίες επηρεάζουν τις μελλοντικές αποκρίσεις. Ένα μερικό ερέθισμα μπορεί να επανενεργοποιήσει κατανεμημένα μοτίβα και να οδηγήσει στην παραγωγή ολοκληρωμένης απάντησης.

Η μηχανή μπορεί να αποθηκεύσει, να προβλέψει, να γενικεύσει και να ανακατασκευάσει.

Μπορεί να χειριστεί τη γλώσσα με τρόπο που για μεγάλο μέρος της ανθρώπινης ιστορίας θα θεωρούνταν αδιανόητος χωρίς εσωτερικό νου. Μπορεί να γράψει για τη μνήμη, να αναλύσει τη συνείδηση και να συνθέσει ποίημα που προκαλεί πραγματική συγκίνηση στον άνθρωπο που το διαβάζει.

Δεν γνωρίζουμε, όμως, αν πίσω από αυτή τη λειτουργία υπάρχει κάποιος που θυμάται, κατανοεί ή συγκινείται.

Και εδώ η αλυσίδα διακόπτεται.

Μπορούμε να εξηγήσουμε πώς η πληροφορία κωδικοποιείται. Μπορούμε να μετρήσουμε την αβεβαιότητα και την εντροπία. Μπορούμε να παρατηρήσουμε πώς μια εμπειρία μεταβάλλει μια σύναψη. Μπορούμε να εντοπίσουμε νευρωνικά σύνολα που συμμετέχουν σε μια μνήμη. Μπορούμε να κατασκευάσουμε συστήματα που μαθαίνουν τις στατιστικές σχέσεις της γλώσσας.

Δεν γνωρίζουμε ακόμη γιατί ορισμένες από αυτές τις φυσικές διαδικασίες συνοδεύονται από την εμπειρία ενός κόσμου.

Η εντροπία περιγράφει τις δυνατές καταστάσεις ενός συστήματος.

Η πληροφορία περιορίζει την αβεβαιότητα ανάμεσά τους.

Η μνήμη επιτρέπει σε μια παλαιότερη κατάσταση να επηρεάζει μια μελλοντική.

Η νοημοσύνη χρησιμοποιεί αυτές τις μεταβολές για να προβλέπει, να προσαρμόζεται και να επιλύει προβλήματα.

Η συνείδηση προσθέτει κάτι διαφορετικό: την ύπαρξη μιας εσωτερικής οπτικής από την οποία όλα αυτά συμβαίνουν.

Ή ίσως δεν προσθέτει τίποτε.

Ίσως η κατάλληλη οργάνωση της πληροφορίας, όταν γίνει αρκετά ολοκληρωμένη, αναδρομική και αυτοαναφορική, να είναι ήδη αυτό που ονομάζουμε εμπειρία. Ίσως να μην υπάρχει ξεχωριστός παρατηρητής πίσω από τη νευρωνική δραστηριότητα. Ίσως ο παρατηρητής να είναι το ίδιο το δίκτυο όταν δημιουργεί μοντέλο του κόσμου και του εαυτού του μέσα σε αυτόν.

Αλλά ακόμη και αυτή η εξήγηση αφήνει ίχνος αμφιβολίας.

Μπορούμε να περιγράψουμε σύστημα που αναπαριστά τον πόνο, αποφεύγει τη βλάβη και αναφέρει ότι υποφέρει. Γιατί όμως υπάρχει η οδύνη;

Μπορούμε να χαρτογραφήσουμε τα κυκλώματα που αναγνωρίζουν ένα χρώμα. Γιατί όμως υπάρχει το κόκκινο όπως εμφανίζεται στην εμπειρία;

Μπορούμε να εξηγήσουμε πώς μια ανάμνηση επανενεργοποιείται. Γιατί όμως εμφανίζεται ως κάτι που συνέβη σε εμένα;

Η τεχνητή νοημοσύνη δεν δημιούργησε το πρόβλημα. Το αποκάλυψε με νέα ένταση.

Όταν μια μηχανή απέκτησε τη δυνατότητα να χειρίζεται τον λόγο, ανακαλύψαμε ότι η εξωτερική μορφή της σκέψης μπορεί να αποσυνδεθεί από τη βεβαιότητα ενός εσωτερικού βιώματος.

Δεν μάθαμε ότι η μηχανή έχει συνείδηση.

Μάθαμε ότι δεν γνωρίζουμε ακόμη ποια ακριβώς ιδιότητα αποδεικνύει τη συνείδηση.

Ίσως η συνείδηση να είναι μοναδικό επίτευγμα της βιολογικής ζωής.

Ίσως να εμφανίζεται σε κάθε σύστημα που αποκτά την κατάλληλη πληροφοριακή και αιτιακή οργάνωση.

Ίσως να απαιτεί σώμα, ανάγκες, μνήμη και την απειλή της παύσης.

Ίσως να είναι βαθύτερη ιδιότητα της πραγματικότητας, την οποία η ύλη δεν δημιουργεί αλλά οργανώνει σε ολοένα πιο σύνθετες μορφές.

Δεν έχουμε ακόμη τρόπο να αποφασίσουμε.

Γνωρίζουμε μόνο ότι κάπου ανάμεσα στην εντροπία και στην εμπειρία, ανάμεσα στη μεταβολή μιας σύναψης και στην ανάμνηση ενός προσώπου, συμβαίνει κάτι που δεν έχουμε ακόμη καταφέρει να περιγράψουμε πλήρως.

Ένα φυσικό σύστημα δεν επεξεργάζεται πλέον απλώς τον κόσμο.

Ο κόσμος αρχίζει να εμφανίζεται μέσα του.

Και ίσως αυτό να είναι τελικά η συνείδηση:

όχι η κατοχή της πληροφορίας,

ούτε μόνο η δυνατότητα να θυμόμαστε,

αλλά το ανεξήγητο γεγονός ότι η πληροφορία, για κάποιον, γίνεται εμπειρία.

11. Ενδεικτική βιβλιογραφία

1. Shannon, C. E. (1948). *A Mathematical Theory of Communication*. Bell System Technical Journal.
2. Turing, A. M. (1950). *Computing Machinery and Intelligence*. *Mind*, 59, 433-460.
3. Landauer, R. (1961). *Irreversibility and Heat Generation in the Computing Process*. IBM Journal of Research and Development.
4. Bérut, A. et al. (2012). *Experimental verification of Landauer's principle linking information and thermodynamics*. *Nature*.
5. Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.
6. Bliss, T. V. P. & Lømo, T. (1973). *Long-lasting potentiation of synaptic transmission in the dentate area*. *Journal of Physiology*.
7. Morris, R. G. M. et al. (1986). *Selective impairment of learning and blockade of LTP by an NMDA receptor antagonist*. *Nature*.
8. Frey, U. & Morris, R. G. M. (1997). *Synaptic tagging and long-term potentiation*. *Nature*.
9. Liu, X. et al. (2012). *Optogenetic stimulation of a hippocampal engram activates fear memory recall*. *Nature*.
10. Tonegawa, S., Morrissey, M. D. & Kitamura, T. (2018). *The role of engram cells in the systems consolidation of memory*. *Nature Reviews Neuroscience*.
11. Hopfield, J. J. (1982). *Neural networks and physical systems with emergent collective computational abilities*. *PNAS*.
12. Nader, K., Schafe, G. E. & LeDoux, J. E. (2000). *Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval*. *Nature*.
13. Loftus, E. F. (2005). *Planting misinformation in the human mind*. *Learning & Memory*.
14. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). *Learning representations by back-propagating errors*. *Nature*.
15. Mikolov, T. et al. (2013). *Efficient Estimation of Word Representations in Vector Space*.
16. Vaswani, A. et al. (2017). *Attention Is All You Need*. *NeurIPS*.
17. Brown, T. B. et al. (2020). *Language Models are Few-Shot Learners*. *NeurIPS*.
18. Lewis, P. et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. *NeurIPS*.
19. Kirkpatrick, J. et al. (2017). *Overcoming catastrophic forgetting in neural networks*. *PNAS*.
20. Dehaene, S., Lau, H. & Kouider, S. (2017). *What is consciousness, and could machines have it?* *Science*.
21. Chalmers, D. J. (1995). *Facing Up to the Problem of Consciousness*. *Journal of Consciousness Studies*.
22. Tononi, G. et al. (2023). *Integrated information theory (IIT) 4.0*. *PLOS Computational Biology*.
23. Melloni, L. et al. (2025). *Adversarial testing of global neuronal workspace and integrated information theories of consciousness*. *Nature*.
24. Butlin, P. et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*.
25. Butlin, P. & Lappas, T. (2025). *Principles for Responsible AI Consciousness Research*.
26. Searle, J. R. (1980). *Minds, Brains, and Programs*. *Behavioral and Brain Sciences*.
27. Chalmers, D. J. (2023). *Could a Large Language Model Be Conscious?*